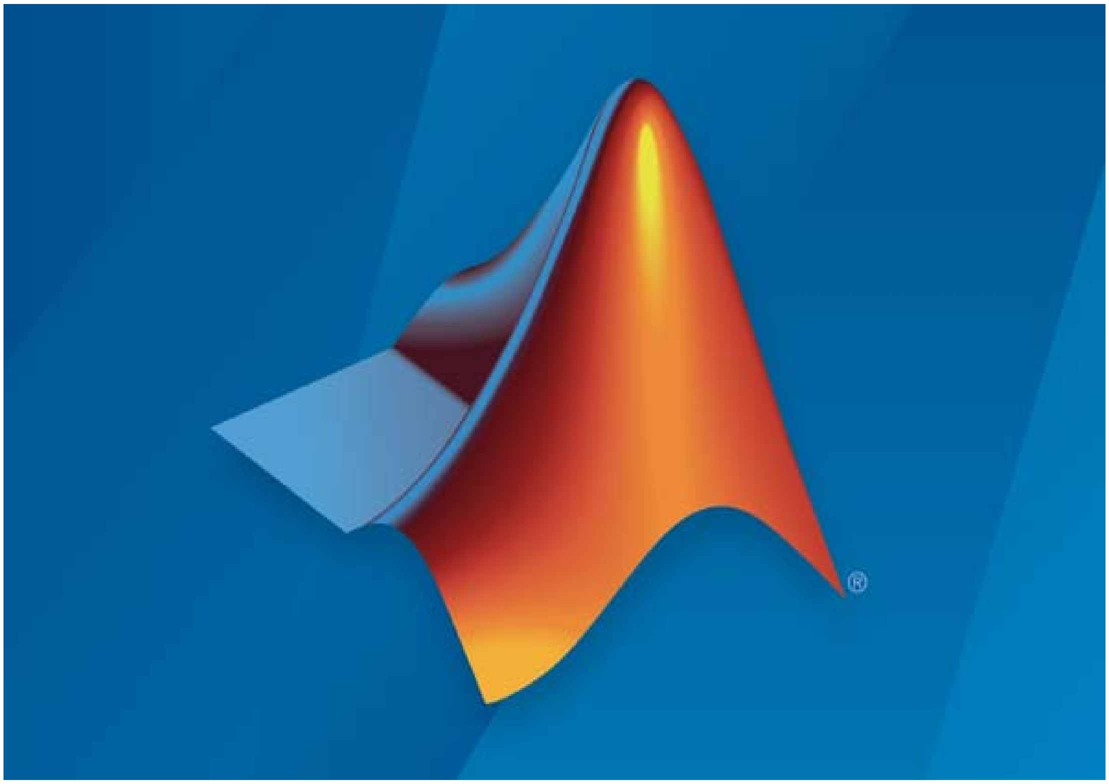
Linear Models With MATLAB



Prof. Dr. Taha Hussein Ali Alzubaydi

Department of Statistics, College of Administration and Economics Salahaddin University, Erbil, Kurdistan Region, Iraq

2021

Linear Models
With
MATLAB

Prof. Dr. Taha Hussein Ali Alzubaydi

Department of Statisticse and Informative, College of Administration and Economics, Salahaddin University, Erbil, Kurdistan Region, Iraq, 2020

Introduction

This book has been prepared for the beginners to help them understand basic to advanced functionality of MATLAB. After completing this chapter 1 (Which included an explanation of the Matlab language) you will find yourself at a moderate level of expertise in using MATLAB from where you can take yourself to next levels.

On other side, in spite of the availability of highly innovative tools in statistics, the main tool of the applied statistician remains the linear model. The linear model involves the simplest and seemingly most restrictive statistical properties: independence, normality, constancy of variance, and linearity. However, the model and the statistical methods associated with it are surprisingly versatile and robust. More importantly, mastery of the linear model is a prerequisite to work with advanced statistical tools because most advanced tools are generalizations of the linear model. The linear model is thus central to the training of any statistician, applied or theoretical.

This book develops the basic theory of linear models for regression, analysis-of variance, and analysis–of–covariance. Applications are illustrated by examples and problems using real data. This combination of theory and applications will prepare the reader to further explore the literature and to more correctly interpret the output from a linear models computer package and MATLAB.

This introductory linear models book is designed primarily for a one-semester course for advanced undergraduates or MS students. It includes more material than can be covered in one semester so as to give an instructor a choice of topics and to serve as a reference book for researchers who wish to gain a better understanding of regression and analysis-of-variance. The book would also serve well as a text for PhD classes in which the instructor is looking for a one-semester introduction, and it would be a good supplementary text or reference for a more advanced PhD class for which the students need to review the basics on their own.

Our overriding objective in the preparation of this book has been clarity of exposition. We hope that students, instructors, researchers, and practitioners will find this linear models text more comfortable than most. In the final stages of development, we asked students for written comments as they read each day's assignment. They made many suggestions that led to improvements in readability of the book. We are grateful to readers who have notified us of errors and other suggestions for improvements of the text.

Another objective of the book is to tie up loose ends. There are many approaches to teaching regression, for example. Some books present estimation of regression coefficients for fixed *x*'s only, other books use random *x*'s, some use centered models, and others define estimated regression coefficients in terms of variances and covariances or in terms of correlations. Theory for linear models has been presented using both an algebraic and a geometric approach. Many books present classical (frequents) inference for linear models, while increasingly the Bayesian approach is presented. We have tried to cover all these approaches carefully and to show how they relate to each other. We have attempted to do something similar for various approaches to analysis-of-variance. We believe that this will make the book useful as a reference as well as a textbook. An instructor can choose the approach he or she prefers, and a student or researcher has access to other methods as well.

The book includes a large number of theoretical problems and a smaller number of applied problems using real datasets. The problems, along with the extensive set of answers in Appendix A, extend the book in two significant ways: (1) the theoretical problems and answers fill in nearly all gaps in derivations and proofs and also extend the coverage of material in the text, and (2) the applied problems and answers become additional examples illustrating the theory. As instructors, we find that having answers available for the students saves a great deal of class time and enables us to cover more material and cover it better. The answers would be especially useful to a reader who is engaging this material outside the formal classroom setting.

Following a brief introduction in Chapters 2, 3, and 4 cover simple and multiple linear regression, including estimation and testing hypotheses and consequences of misspecification of the model. Chapter 5 provides diagnostics for model validation and detection of influential observations. Chapter 6 treats multiple regression with random x's.

CONTENTS

References

Agresti, A. (1984). Analysis of Ordinal Categorical Data. New York: Wiley.

Agresti, A. (1990). Categorical Data Analysis. New York: Wiley.

Anderson, E. B. (1991). The Statistical Analysis of Categorical Data (2nd ed.). New York: Springer-Verlag.

Anderson, T. W. (1984). Introduction to Multivariate Statistical Analysis (2nd ed.). New York: Wiley.

Andrews, D. F. (1974). A robust method for multiple linear regression. Technometrics 16, 523–531.

Andrews, D. F. and A. M. Herzberg (1985). Data. New York: Springer-Verlag.

Bailey, B. J. R. (1977). Tables of the Bonferroni t-statistic. Journal of the American Statistical Association 72, 469–479.

Bates, D. M. and D. G. Watts (1988). Nonlinear Regression and Its Applications. New York: Wiley.

Beckman, R. J. and R. D. Cook (1983). Outliers (with comments). Technometrics 25, 119–163.

Belsley, D. A., E. Kuh, and R. E.Welsch (1980). Regression Diagnostics: Identifying Data and Sources of Collinearity. New York: Wiley.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B: Methodological 57, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29(4), 1165–1188.

Benjamini, Y. and D. Yekutieli (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association 100(469), 71–81.

Bingham, C. and S. E. Feinberg (1982). Textbook analysis of covariance—is it correct. Biometrics 38, 747–753.

Birch, J. B. (1980). Some convergence properties of iterated reweighted least squares in the location model. Communications in Statistics B9(4), 359–369.

Bishop, Y., S. Fienberg, and P. Holland (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: Massachusetts Institute of Technology Press.

Bloomfield, P. (2000). Fourier Analysis of Time Series: An Introduction. New York: Wiley.

Bonferroni, C. E. (1936). Il calcolo delle assicurazioni su gruppi di teste. Studii in Onore del Profesor S. O. Carboni. Roma.

Box, G. E. P. and P. V. Youle (1955). The exploration and exploitation of response surfaces: An example of the link between the fitted surface and the basic mechanism of the system. Biometrics 11, 287–323.

Broadbent, K. L. (1993). A Comparison of Six Bonferroni Procedures. Master's thesis, Department of Statistics, Brigham Young University.

Brown, H. and R. Prescott (1999). Applied Mixed Models in Medicine. New York: Wiley & Sons.

Brown, H. and R. Prescott (2006). Applied Mixed Models in Medicine (2nd ed.). Hoboken, NJ: Wiley.

Bryce, G. R. (1975). The one-way model. The American Statistician 29, 69–70.

Bryce, G. R. (1998). Personal communication.

Bryce, G. R., M. W. Carter, and M. W. Reader (1976). Nonsingular and singular transformations in the fixed model. Annual Meeting of the American Statistical Association, Boston, Aug. 1976.

Bryce, G. R., M. W. Carter, and D. T. Scott (1980a). Recovery of Estimability in Fixed Models with Missing Cells. Technical Report SD-022-R, Department of Statistics, Brigham Young University.

Bryce, G. R., D. T. Scott, and M.W. Carter (1980b). Estimation and hypothesis testing in linear models—a reparameterization approach. Communications in Statistics—Series A, Theory and Methods 9, 131–150.

Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. The American Statistician 46, 167–174.

Chatterjee, S. and A. S. Hadi (1988). Sensitivity Analysis in Linear Regression. New York: Wiley.

Christensen, R. (1996). Plane Answers to Complex Questions: The Theory of Linear Models (2nd ed.). New York: Springer-Verlag.

Christensen, R. (1997). Log-Linear Models and Logistic Regression (2nd ed.). New York: Springer-Verlag.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system with applications to the analysis of variance. Proceedings, Cambridge Philosophical Society 30, 178–191.

Cochran, W. G. (1977). Sampling Techniques. New York: Wiley.

Cook, R. D. (1977). Detection of influential observations in linear regression. Technometrics 19, 15–18.

Cook, R. D. and S. Weisberg (1982). Residuals and Influence in Regression. New York: Chapman & Hall.

Crampton, E. W. and J. W. Hopkins (1934). The use of the method of partial regression in the analysis of comparative feeding trial data. Part II. J. Nutrition 8, 329–339.

Daniel, W. W. (1974). Biostatistics: A Foundation for Analysis in the Health Sciences. New York: Wiley.

Devlin, S. J., R. Gnanadesikan, and J. R. Kettenring (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika 62, 531–546.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). Analysis of Longitudinal Data. Oxford University Press.

Dobson, A. J. (1990). An Introduction to Generalized Linear Models. New York: Chapman & Hall.

Draper, N. R. and H. Smith (1981). Applied Regression Analysis (2nd ed.). New York: Wiley.

Draper, N. R. and H. Smith (1998). Applied Regression Analysis. New York: Wiley.

Driscoll, M. F. (1999). An improved result relating quadratic forms and chi-square distributions. The American Statistician 53, 273–275.

Eubank, R. L. and R. L. Eubank (1999). Nonparametric Regression and Spline Smoothing. New York: Marcel Dekker.

Evans, M. and T. Swartz (2000). Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press.

Ezekiel, M. (1930). Methods of Correlation Analysis. New York: Wiley.

Fai, A. H.-T. and P. L. Cornelius (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. Journal of Statistical Computation and Simulation 54, 363–378.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. Metron 1, 1–32.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2004). Applied Longitudinal Analysis. Hoboken, NJ: Wiley.

Flury, B. W. (1989). Understanding partial statistics and redundancy of variables in regression and discriminant analysis. The American Statistician 43(1), 27–31.

Fox, J. (1997). Applied Regresion Analysis, Linear Models, and Related Methods. Thousand Oaks, CA: SAGE Publications.

Freund, R. J. and P. D. Minton (1979). Regression Methods: A Tool for Data Analysis. New York: Marcel Dekker.

Fuller, W. A. and G. E. Battese (1973). Transformations for estimation of linear models with nested-error structure. Journal of the American Statistical Association 68, 626–632.

Gallant, A. R. (1975). Nonlinear regression. The American Statistician 29, 73–81.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). Bayesian Data Analysis (2nd ed.). Chapman & Hall/CRC.

Ghosh, B. K. (1973). Some monotonicity theorems for chi-square. F and t distributions with applications. Journal of the Royal Statistical Society 35, 480–492.

Giesbrecht, F. G. and J. C. Burns (1985). Two-stage analysis based on a mixed model: Largesample asymptotic theory and small-sample simulation results. Biometrics 41, 477–486.

Gilks, W. R. E., S. E. Richardson, and D. J. E. Spiegelhalter (1998). Markov Chain Monte Carlo in Practice. London: Chapman & Hall.

Gomez, E., G. Schaalje, and G. Fellingham (2005). Performance of the kenward-roger method when the covariance structure is selected using aic and bic. Communications in Statistics: Simulation and Computation 34(2), 377–392.

Graybill, F. A. (1954). On quadratic estimates of variance components. Annals of Mathematical Statistics 25(2), 367–372.

Graybill, F. A. (1969). Introduction to Matrices with Applications in Statistics. Belmont, CA: Wadsworth Publishing Company.

Graybill, F. A. (1976). Theory and Application of the Linear Model. North Scituate, MA: Duxbury Press.

Graybill, F. A. and H. K. Iyer (1994). Regression Analysis: Concepts and Applications. North Scituate, MA: Duxbury Press.

Graybill, F. A. and A. W. Wortham (1956). A note on uniformly best, unbiased estimators for variance components. Journal of the American Statistical Association 51, 266–268.

Gutsell, J. S. (1951). The effect of sulfamerazine on the erythrocyte and hemoglobin content of trout blood. Biometrics 7(2), 171–179.

Guttman, I. (1982). Linear Models: An Introduction. New York: Wiley.

Hald, A. (1952). Statistical Theory with Engineering Applications. New York: Wiley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association 69, 383–393.

Hartley, H. O. (1956). Programming analysis of variance for general purpose computers. Biometrics 12, 110–122.

Harville, D. A. (1997). Matrix Algebra from a Statistician's Perspective. New York: Springer-Verlag.

Healy, M. J. R. and M. Westmacott (1969). Missing values in experiments analysed on automatic computers. Applied Statistics 5, 203–206.

Helland, I. S. (1987). On the interpretation and use of R2 in regression analysis. Biometrics 43, 61–69.

Henderson, C. R. (1950). Estimation of genetic parameters. Annals of Mathetmatical Statistics 21, 309–310.

Henderson, C. R. and A. J. McAllister (1978). The missing subclass problem in two-way fixed models. Journal of Animal Science 46, 1125–1137.

Hendrix, L. J. (1967, Aug.). Auditory Discrimination Differences between Culturally Deprived and Nondeprived Preschool Children. PhD thesis, Brigham Young University.

Hendrix, L. J., M.W. Carter, and J. Hintze (1978). A comparison of five statistical methods for analyzing pretest-post designs. Journal of Experimental Education 47, 96–102.

Hendrix, L. J., M.W. Carter, and D. T. Scott (1982). Covariance analysis with heterogeneity of slopes in fixed models. Biometrics 38, 641–650.

Hilbe, J. M. (1994). Generalized linear models. The American Statistician 48, 255–265.

Hoaglin, D. C. and R. E. Welsch (1978). The hat matrix in regression and ANOVA. The American Statistician 32, 17–22.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1–51.

Hocking, R. R. (1985). The Analysis of Linear Models. Monterey, CA: Brooks/Cole.

Hocking, R. R. (1996). Methods and Applications of Linear Models. New York: Wiley.

Hocking, R. R. (2003). Methods and Applications of Linear Models (2nd ed.). New York: Wiley.

Hocking, R. R. and F. M. Speed (1975). A full rank analysis of some linear model problems. Journal of the American Statistical Association 70, 706–712.

Hoerl, A. E. and R.W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Hogg, R. V. and A. T. Craig (1995). Introduction to Mathematical Statistics (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Holland, B. (1991). On the application of three modified Bonferroni procedures to pairwise multiple comparisons in balanced repeated measures designs. Computational Statistics Quarterly 3, 219–231.

Holland, B. and M. D. Copenhaver (1987). An improved sequentially rejective Bonferroni test procedure. Biometrics 43, 417–423.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75, 383–386.

Hosmer, D., B. Jovanovic, and S. Lemeshow (1989). Best subsets logistic regression. Biometrics 45, 1265–1270.

Hosmer, D. W. and S. Lemeshow (1989). Applied Logistic Regression. New York: Wiley.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. Annals of Statistics 1, 799–821.

Hummel, T. J. and J. Sligo (1971). Empirical comparison of univariate and multivariate analysis of variance procedures. Psychological Bulletin 76, 49–57.

Jammalamadaka, S. R. and D. Sengupta (2003). Linear Models an Integrated Approach. Singapore: World Scientific Publications.

Jeske, D. R. and D. A. Harville (1988). Prediction-interval procedures and (fixed-effects) confidence-interval procedures for mixed linear models. Communications in Statistics: Theory and Methods 17, 1053–1087.

Jørgensen, B. (1993). The Theory of Linear Models. New York: Chapman & Hall.

Kackar, R. N. and D. A. Harville (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. Journal of the American Statistical Association 79, 853–862.

Kendall, M. G. and A. Stuart (1969). The Advanced Theory of Statistics (3rd ed.), Vol. 1. New York: Hafner.

Kenward, M. G. and J. H. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53, 983–997.

Keselman, H. J., R. K. Kowalchuk, J. Algina, and R. D. Wolfinger (1999). The analysis of repeated measurements: A comparison of mixed-model Satterthwaite *F* tests and a non-pooled adjusted degrees of freedom multivariate test. Communications in Statistics: Theory and Methods 28, 2967–2999.

Kleinbaum, D. G. (1994). Logistic Regression. New York: Springer-Verlag.

Krasker, W. S. and R. Welsch (1982). Efficient bounded-influence regression estimation. Journal of the American Statistical Association 77, 595–604.

Kshirsagar, A. M. (1983). A Course in Linear Models. New York: Marcel Dekker.

Ku, H. H. and S. Kullback (1974). Loglinear models in contingency table analysis. The American Statistician 28, 115–122.

Kutner, M. H., C. J. Nachtsheim, J. Neter, andW. Li (2005). Applied Linear Statistical Models (5th ed.). New York: McGraw-Hill/Irwin.

Lehmann, E. L. (1999). Elements of Large-Sample Theory. New York: Springer-Verlag.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimates for the linear model (with discussion). Journal of the Royal Statistical Society, Series B: Methodological 34, 1–41.

Lindsey, J. K. (1997). Applying Generalized Linear Models. New York: Springer-Verlag.

Little, R. J. A. and D. B. Rubin (2002). Statistical Analysis with Missing Data. Hoboken, NJ: Wiley.

Mahalanobis,P. C.(1936). On the generalized distance in statistics. Proceedings of the National Institute of Science of India 12, 49–55.

Mahalanobis, P. C. (1964). Professor Ronald Aylmer Fisher. Biometrics 20, 238–250.

Marcuse, S. (1949). Optimum allocation and variance components in nested sampling with an application to chemical analysis. Biometrics 5(3), 189–206.

McCullagh, P. and J. A. Nelder (1989). Generalized Linear Models (2nd ed.). New York: Chapman & Hall.

McCulloch, C. E. and S. R. Searle (2001). Generalized, Linear, and Mixed Models. New York: Wiley.

Mclean, R. A., W. L. Sanders, and W. W. Stroup (1991). A unifed approach to mixed linear models. American Statistician 45, 54–64.

Mendenhall, W. and T. Sincich (1996). A Second Course in Statistics: Regression Analysis. Englewood Cliffs, NJ: Prentice-Hall.

Milliken, G. A. and D. E. Johnson (1984). Analysis of Messy Data, Vol. 1: Designed Experiments. New York: Van NostrandReinhold.

Montgomery, D. C. and E. A. Peck (1992). Introduction to Linear Regresion Analysis (2nded.). New York: Wiley.

Morrison, D. F. (1983). Applied Linear Statistical Methods. Englewood Cliffs, NJ: Prentice-Hall.

Mosteller, F. and J. W. Tukey (1977). Data Analysis and Regression. Reading, MA: Addison-Wesley.

Muller, K. E. and M. C. Mok (1997). The distribution of Cook's D statistic. Communications in Statistics: Theory and Methods 26, 525–546.

Myers, R. H. (1990). Classical and Modern Regression with Applications (2nd ed.). Boston: Duxbury Press.

Myers, R. H. and J. S. Milton (1991). A First Course in the Theory of Linear Statistical Models. Boston: PWS-Kent.

Nelder, J. A. (1974). Letter to the editor. Journal of the Royal Statistical Society, Series C 23, 232.

Nelder, J. A. and P. W. Lane (1995). The computer analysis of factorial experiments: In memoriam-Frank Yates. The American Statistician 49, 382–385.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. Journal of the Royal Statistical Society, Series A 135, 370–384.

Ogden, R. T. (1997). Essential Wavelets for Statistical Applications and Data Analysis. Birkhauser.

Ostle, B. and L. C. Malone (1988). Statistics in Research: Basic Concepts and Techniques for Research Workers (4th ed.). Ames: Iowa State University Press.

Ostle, B. and R. W. Mensing (1975). Statistics in Research (3rd ed.). Ames: Iowa State University Press.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545-554.

Pawitan, Y. (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford University Press.

Pearson, E. S., R. L. Plackett, and G. A. Barnard (1990). Student: A Statistical Biography of William Sealy Gossett. New York: Oxford University Press.

Plackett, R. L. (1981). The Analysis of Categorical Data (2nd ed.). London: Griffin.

Rao, C. R. (1965). Linear Statistical Inference and Its Applications. New York: Wiley.

Rao, P. S. R. S. (1997). Variance Components Estimation. London: Chapman & Hall.

Ratkowsky, D. A. (1983). Nonlinear Regression Modelling: A Unified Approach. New York: Marcel Dekker.

Ratkowsky, D. A. (1990). Handbook of Nonlinear Regression Models. New York: Marcel Dekker.

Read, T. R. C. and N. A. C. Cressie (1988). Goodness-of-Fit Statistics for Discrete Multivariate Data. New York: Springer-Verlag.

Reader, M. W. (1973). The Analysis of Covariance with a Single Linear Covariate Having

Heterogeneous Slopes. Master's thesis, Department of Statistics, Brigham Young University.

Rencher, A. C. (1993). The contribution of individual variables to Hotelling's T2,Wilks' L and R2. Biometrics 49, 217–225.

Rencher, A. C. (1995). Methods of Multivariate Analysis. New York: Wiley.

Rencher, A. C. (1998). Multivariate Statistical Inference and Applications. New York: Wiley.

Rencher, A. C. (2002). Multivariate Statistical Inference and Applications. Hoboken, NJ: Wiley.

Rencher, A. C. and D. T. Scott (1990). Assessing the contribution of individual variables following rejection of a multivariate hypothesis. Communications in Statistics-Series B, Simulation and Computation 19, 535–553.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika 77, 663–665.

Ross, S. M. (2006). Introduction to Probability Models (9th ed.). San Diego, CA: Academic Press.

Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. Applied Statistics 32, 121–133.

Ryan, T. P. (1997). Modern Regression Methods. New York: Wiley.

Santner, T. J. and D. E. Duffy (1989). The Statistical Analysis of Discrete Data. New York: Springer-Verlag.

Satterthwaite, F. E. (1941). Synthesis of variances. Psychometrika 6, 309–316.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution (C/R: 91V45 p165–168). The American Statistician 44, 174–180.

Schaalje, G. B., J. B. McBride, and G. W. Fellingham (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. Journal of Agricultural, Biological, and Environmental Statistics 7(4), 512–524.

Scheffe´, H. (1953). A method of judging all contrasts in the analysis of variance. Biometrika 40, 87–104.

Scheffe´, H. (1959). The Analysis of Variance. New York: Wiley.

Schott, J. R. (1997). Matrix Analysis for Statistics. New York: Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Searle, S. R. (1971). Linear Models. New York: Wiley.

Searle, S. R. (1977). Analysis of Variance of Unbalanced Data from 3-Way and Higher-Order Classifications. Technical Report BU-606-M, Cornell University, Biometrics Units.

Searle, S. R. (1982). Matrix Algebra Useful for Statistics. New York: Wiley.

Searle, S. R., G. Casella, and C. E. McCulloch (1992). Variance Components. New York: Wiley.

Searle, S. R., F. M. Speed, and H. V. Henderson (1981). Some computational and model equivalencies in analysis of variance of unequal-subclass-numbers data. The American Statistician 35, 16-33.

Seber, G. A. F. (1977). Linear Regression Analysis. New York: Wiley.

Seber, G. A. F. and A. J. Lee (2003). Linear Regression Analysis (2nd ed.). Hoboken, NJ: Wiley.

Seber, G. A. F. and C. J. Wild (1989). Nonlinear Regression. New York: Wiley.

Sen, A. and M. Srivastava (1990). Regression Analaysis: Theory, Methods, and Applications. New York: Springer-Verlag.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association 81, 826–831.

Silverman, B. W. (1999). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.

Snedecor, G. W. (1948). Answer to query. Biometrics 4(2), 132–134.

Snedecor, G. W. and W. G. Cochran (1967). Statistical Methods (6th ed.). Ames: Iowa State University Press.

Snee, R. D. (1977). Validation of regression models: Methods and examples. Technometrics 19, 415–428.

Speed, F. M. (1969). A New Approach to the Analysis of Linear Models. Technical report, National Aeronautics and Space Administration, Houston, TX; a NASA Technical memo, NASA TM X-58030.

Speed, F. M., R. R. Hocking, and O. P. Hackney (1978). Methods of analysis of linear models with unbalanced data. Journal of the American Statistical Association 73, 105–112.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (pkg: P583-639). Journal of the Royal Statistical Society, Series B: Statistical Methodology 64(4), 583–616.

Stapleton, J. H. (1995). Linear Statistical Models. New York: Wiley.

Stigler, S. M. (2000). The problematic unity of biometrics. Biometrics 56(3), 653–658.

Stokes, M. E., C. S. Davis, and G. G. Koch (1995). Categorical Data Analysis Using the SAS System. Cary, NC: SAS Institute.

Theil, H. and C. Chung (1988). Information-theoretic measures of fit for univariate and multivariate linear regressions. The American Statistician 42, 249–252.

Tiku, M. L. (1967). Tables of the power of the F-test. Journal of the American Statistical Association 62, 525–539.

Turner, D. L. (1990). An easy way to tell what you are testing in analysis of variance. Communications in Statistics-Series A, Theory and Methods 19, 4807–4832.

Urquhart, N. S., D. L. Weeks, and C. R. Henderson (1973). Estimation associated with linear models: A revisitation. Communications in Statistics 1, 303–330.

Verbeke, G. and G. Molenberghs (2000). Linear Mixed Models for Longitudinal Data. Springer-Verlag.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54, 426–483.

Wang, S. G. and S. C. Chow (1994). Advanced Linear Models: Theory and Applications. New York: Marcel Dekker.

Weisberg, S. (1985). Applied Linear Regression. New York: Wiley.

Welsch, R. E. (1975). Confidence regions for robust regression. Paper presented at Annual

Meeting of the American Statistical Association, Washington, DC.

Winer, B. J. (1971). Statistical Principles in Experimental Design (2nd ed.). New York: McGraw-Hill.

Working, H. and H. Hotelling (1929). Application of the theory of error to the interpretation of trends. Journal of the American Statistical Association, Suppl. (Proceedings) 24, 73–85.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. Journal of the American Statistical Association 29, 52–66.

Chapter one

Introduction to MATLAB

1.1: Introduction

This chapter gives you aggressively a gentle introduction of MATLAB programming language. It is designed to give students fluency in MATLAB programming language. Problem-based MATLAB examples have been given in simple and easy way to make your learning fast and effective.

MATLAB is a programming language developed by MathWorks. It started out as a matrix programming language where linear algebra programming was simple. It can be run both under interactive sessions and as a batch job.

We assume you have a little knowledge of any computer programming and understand concepts like variables, constants, expressions, statements, etc. If you have done programming in any other high-level language like C, C++ or Java, then it will be very much beneficial and learning MATLAB will be like a fun for you.

MATLAB (**MAT**rix **LAB**oratory) is a fourth-generation high-level programming language and interactive environment for numerical computation, visualization and programming.

It allows matrix manipulations; plotting of functions and data; implementation of algorithms; creation of user interfaces; interfacing with programs written in other languages, including C, C++, Java, and FORTRAN; analyze data; develop algorithms; and create models and applications.

It has numerous built-in commands and math functions that help you in mathematical calculations, generating plots, and performing numerical methods.

1.2: MATLAB's Power of Computational Mathematics

MATLAB is used in every facet of computational mathematics. Following are some commonly used mathematical calculations where it is used most commonly:

- Dealing with Matrices and Arrays
- 2-D and 3-D Plotting and graphics
- Linear Algebra

- Algebraic Equations

- Non-linear Functions

- Statistics

- Data Analysis

- Calculus and Differential Equations

- Numerical Calculations

- Integration

- Transforms

- Curve Fitting

- Various other special functions

1.3: Features of MATLAB

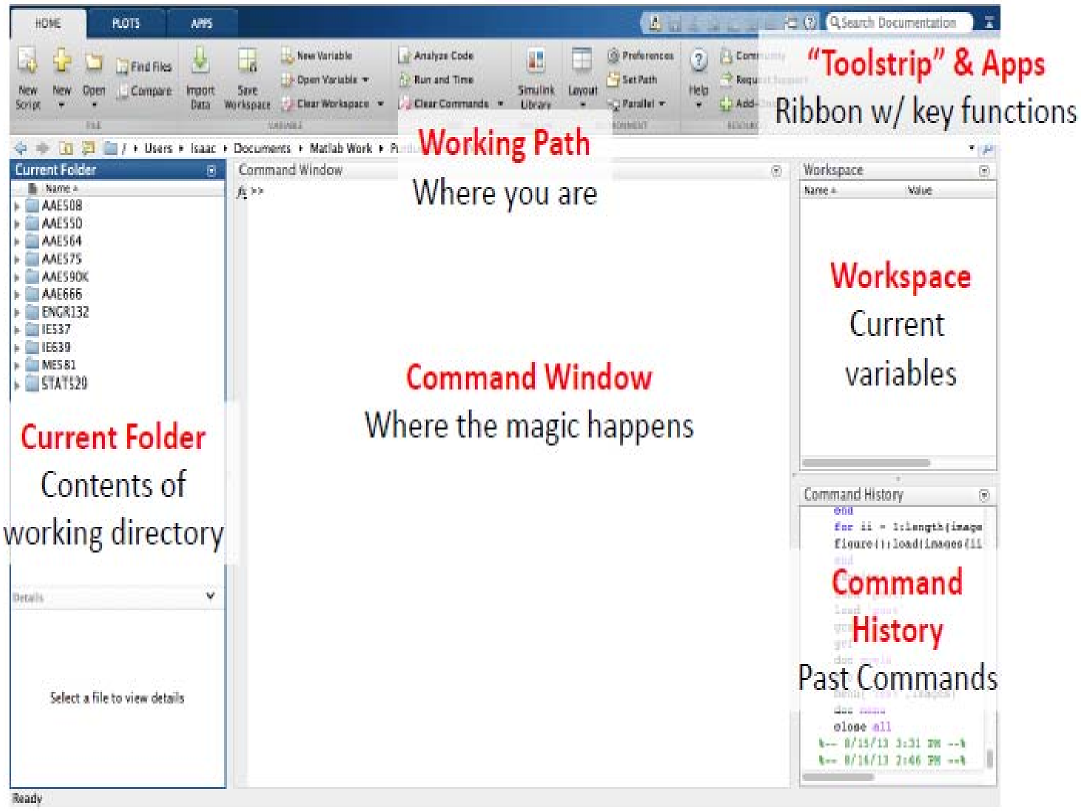
Following are the basic features of MATLAB:

- High-level language for numerical computation, visualization, and application development.
- Interactive environment for iterative exploration, design, and problem solving.
- Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration, and solving ordinary differential equations.
- Built-in graphics for visualizing data and tools for creating custom plots.
- Development tools for improving code quality and maintainability and maximizing performance.
- Tools for building applications with custom graphical interfaces.
- Functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET, and Microsoft Excel.

1.4: Desktop Basics

MATLAB development IDE can be launched from the icon created on the desktop. The main working window in MATLAB is called the desktop. When MATLAB is started, the desktop appears in its default layout:

MATLAB (R2013a) Environment



```
end
for ii = 1:length(image
figure(i);load(images{ii
close all
%-- 8/15/13 3:31 PM --%
%-- 8/16/13 2:46 PM --%
```

The desktop has the following panels:

- **Current Folder** — Access your files.
- **Command Window** — Enter commands at the command line, indicated by the prompt (>>).
- **Workspace** — Explore data that you create or import from files.

As you work in MATLAB, you issue commands that create variables and call functions.

For example, create a variable named x by typing this statement at the command line:

```
>> x = 3
```

MATLAB adds variable x to the workspace and displays the result in the Command Window.

```
x =
     3
```

Create a few more variables.

```
>> y = 5
y =
        5
>> z = x + y
z =
        8
>> d = cos(x)
d =
        -0.989995
```

When you do not specify an output variable, MATLAB uses the variable ans, short for *answer*, to store the results of your calculation.

```
>> sin(x)
ans =
        0.14112
```

If you end a statement with a semicolon, MATLAB performs the computation, but suppresses the display of output in the Command Window.

```
>> e = x * y;
```

You can recall previous commands by pressing the up- and down-arrow keys, ↑ and ↓. Press the arrow keys either at an empty command line or after you type the first few characters of a command. For example, to recall the command y = 5, type b, and then press the up-arrow key.

1.5: Matrices and Vectors

MATLAB is an abbreviation for "matrix laboratory." While other programming languages mostly work with numbers one at a time, MATLAB is designed to operate primarily on whole matrices and Vectors.

All MATLAB variables are multidimensional *Vectors*, no matter what type of data. A *matrix* is a two-dimensional *Vectors* often used for linear algebra.

Vector Creation

To create a vector with four elements in a single row, separate the elements with either a comma (,) or a space

```
>>  a = [1 2 3 4]

a =

     1     2     3     4
```

This type of array is a *row vector*.

To create a matrix that has multiple rows, separate the rows with semicolons.

```
>>  a = [1 2 3; 4 5 6; 7 8 10]

a =

         1    2    3
         4    5    6
         7    8    10
```

Another way to create a matrix is to use a function, such as ones, zeros, or rand. For example, create a 5-by-1 column vector of zeros.

```
>> z = zeros(5,1)

z  =

         0
         0
         0
         0
         0
```

And we have:

```
>> y = ones(1,5)

y =

       1    1    1    1     1
```

1.5.1: Assignment and Operators

Assignment (assign b to a)	=	a = b
Addition	+	a + b
Subtraction	-	a - b
Multiplication: Matrix	*	a * b
Multiplication: Element-by-Element	**.***	**a .* b**
Division: Matrix	/	a / b
Division: Element-by-Element	**./**	**a ./ b**
Power: Matrix	^	a ^ b
Power: Element-by-Element	**.^**	**a .^ b**

1.5.2: Extracting a Sub-Matrix

A portion of a matrix can be extracted and stored in a smaller matrix by specifying the names of both the rows and columns to extract

```
sub_matrix = matrix(r1:r2 , c1:c2)
sub_matrix = matrix(rows, columns)
```

Where r1and r2 specify the beginning and ending rows, and c1and r2 specify the beginning and ending columns to extract

Colon Operator

The colon operator helps to specify ranges

a : b Goes from **a** to **b** in increments of 1. If **a** > **b**, results in null vector

a : n : b Goes from **a** to **b** in increments of **n**. If **n** < 0 then **a** > **b**

A(: , b) The b^{th} column of A

A(a , :) The a^{th} row of A

A(: , :) All of the rows and columns of **A** (i.e., the **A** matrix)

A(a : b) Elements **a** to **b** (in increments of 1) of A. **NOTE**: Elements are counted down the columns and then across the rows!

A(: , a : b) All rows and columns **a** to **b** (in increments of 1)

A(:) All elements of **A** in a single column vector

Matrices

- Accessing single elements of a matrix:

A(a , b) → Element in row **a** and column **b**

- Accessing multiple elements of a matrix:

A(1,4) + A(2,4) + A(3,4) + A(4,4)
sum(A(1:4,4)) or **sum(A(:,end))**

– In locations, the keyword **end** refers to the *last* row or column

- Deleting rows and columns:

A(: , 2) = [ ] → Deletes the second column of A

- Concatenating matrices A and B:

C = [A ; B] for vertical concatenation
C = [A , B] for horizontal concatenation

1.5.3: Matrix Functions in Matlab

A = ones(m , n)	Creates an m×n matrix of 1's
A = zeros(n,m)	Creates an m×n matrix of 0's
A = eye(n)	Creates an n×n identity matrix
A = NaN(m,n)	Creates an m×n matrix of NaN's
A = inf(m,n)	Creates an m×n matrix of inf's
A = diag(x)	Creates a diagonal matrix A of x
x = diag(A)	Extracts diagonal elements from A
[m,n] = size(A)	Returns the dimensions of A
n = length(A)	Returns the largest dimension of A
n = numel(A)	Returns number of elements of A
x= sum(A)	Vector with sum of columns

x = prod(A)	Vector with product of columns
B = A'	Transposed matrix
d = det(A)	Determinant
[x,y] = eig(A)	Eigenvalues and eigenvectors
B = inv(A)	Inverse of square matrix
B = pinv(A)	Moore-Penrose pseudoinverse
B = chol(A)	Cholesky decomposition
[Q,R] = qr(A)	QR decomposition
[U,D,V] = svd(A)	Singular value decomposition

1.5.4: Logic in Matrices

B = any(A)	Determine if any elements in each column of A are nonzero
B = all(A)	Determine if all elements in each column of A are nonzero
B = find(A)	Find indices of all non-zero elements of A Can also use logic!
B = find(A>4 &A<5)	Elements > 4 **and**< 5
B = all(A~=9)	Elements **not equal** to 9
B = any(A==3 |A==5)	Elements **equal** to 3 **or** 5

1.6: Pre-Defined Variables

MATLAB has several pre-defined / reserved variables, **(Beware):** These variables can be overwritten with custom values!

ans	Default variable name for results
pi	Value of π
eps	Smallest incremental number (2.2204e-16)
Inf/ inf	Infinity
NaN/ nan	Not a number (e.g., 0/0)

realmin	Smallest usable positive real number (2.2251e-308)
realmax	Largest usable positive real number (1.7977e+308)
i / j	Square root of (-1)

1.7: Plotting in Matlab

- Matlab has extensive plotting capabilities
- Basic function is **plot** to plot one vector vs. another vector (vectors must have same length)

plot(x, y)

- Can also simply plot one vector vs. its index

plot(x)

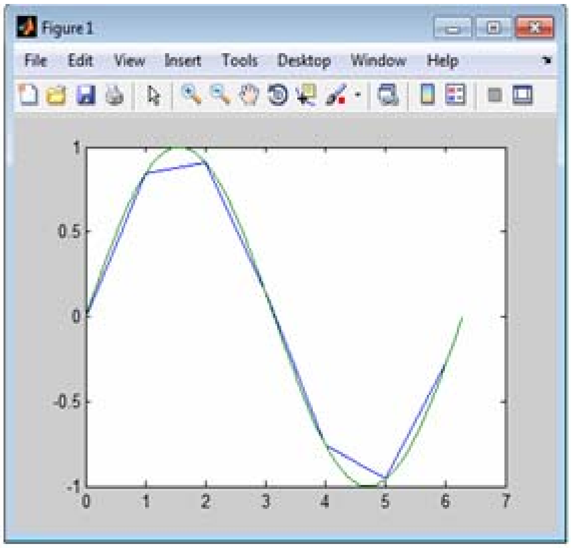
- Repeat three arguments to plot multiple vectors, different pairs of x and y data can have different sizes!

plot(x1, y1, x2, y2, x3, y3)

Example 1.1:

```
>> x1 = 0:1:2*pi;
>> y1 = sin(x1);
>> x2 = 0:0.01:2*pi;
>> y2 = sin(x2);
>> plot(x1,y1,x2,y2)
```

Matlab will automatically change the colors of the lines if plotted with one plot command!

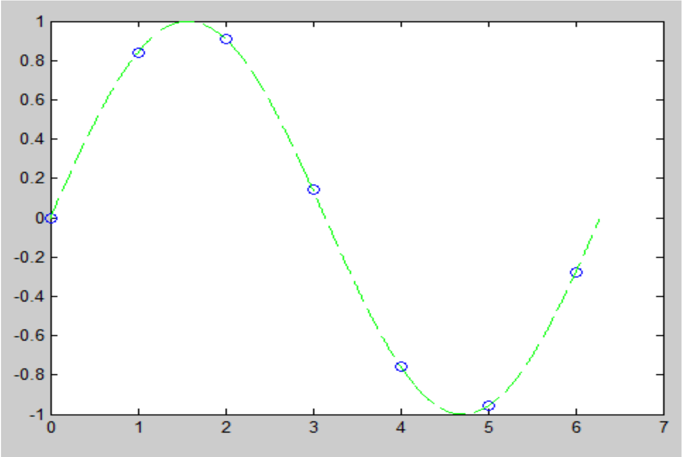


- The line style, marker symbol, and color of the plot are specified by the **Line Spec**.
- **Line Spec** is specified for each line after the y data and is optional.
- To see all options in Matlab: **doc Line Spec**
- Common formatting:

Lines	Markers	Colors
'-' solid	**'+'** plus	**'r'** red
'- -' dashed	**'o'** circle	**'g'** green
':' dotted	**'*'** star	**'b'** blue
'.-' dash-dot	**'.'** point	**'k'** black
	's' square	**'y'** yellow
	'd' diamond	**'c'** cyan
	'x' cross	**'m'** magenta

Example 1.2:

```
>> x1 = 0:1:2*pi; y1 = sin(x1);
>> x2 = 0:0.01:2*pi; y2 = sin(x2);
>> plot(x1,y1,'bo',x2,y2,'g--')
```



- Other commands allow you to modify the plot

–Annotation: title, x label, y label, z label

–Grid: grid **on**, grid **off**, grid **minor**

–Axes: **axis([xmin xmax ymin ymax])**, axis **keyword**(**doc axis** for full keyword list)

–Legend: **legend('Line 1','Line 2','Location','Position')**

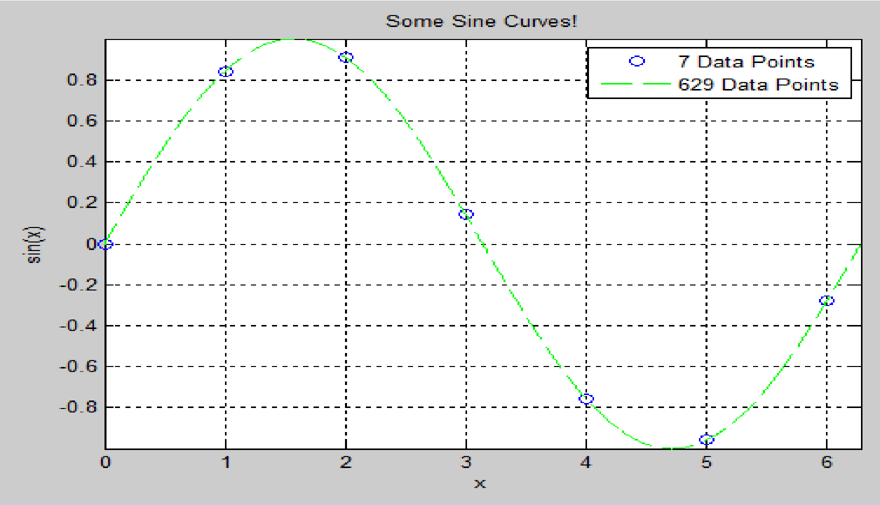
- Another way to plot multiple lines is with the **hold** command

```
hold on
plot(x1,y1)
plot(x2,y2)
hold off
```

- Unless a new figure is created using **figure()**, any plotting function will overwrite the current plot

Example 1.3:

```
x1 = 0:1:2*pi; y1 = sin(x1);
x2 = 0:0.01:2*pi; y2 = sin(x2);
plot(x1,y1,'bo',x2,y2,'g--')
legend('7 Data Points','629 Data
Points','Location','NorthEast')
title('Some Sine Curves!')
xlabel('x')
ylabel('sin(x)')
grid on
axis tight
```

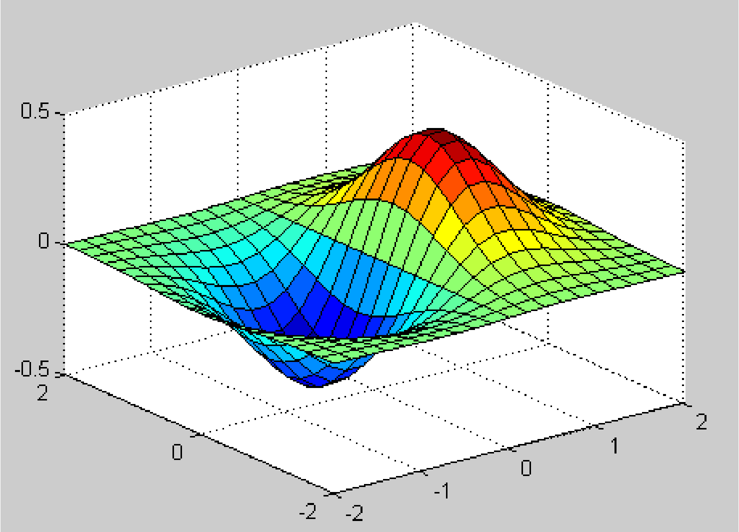


- 3-D Plots: Three-dimensional plots typically display a surface defined by a function in two variables, *z* = *f(x, y)*.

To evaluate *z*, first create a set of (*x*,*y*) points over the domain of the function using *meshgrid*.

Example 1.4:

```
>> [X,Y] = meshgrid(-2: .2: 2);
>> Z = X .* exp(-X.^2 - Y.^2);
>> surf(X,Y,Z)
```



Both the surf function and its companion mesh display surfaces in three dimensions. surf displays both the connecting lines and the faces of the surface in color. Mesh produces wireframe surfaces that color only the lines connecting the defining points.

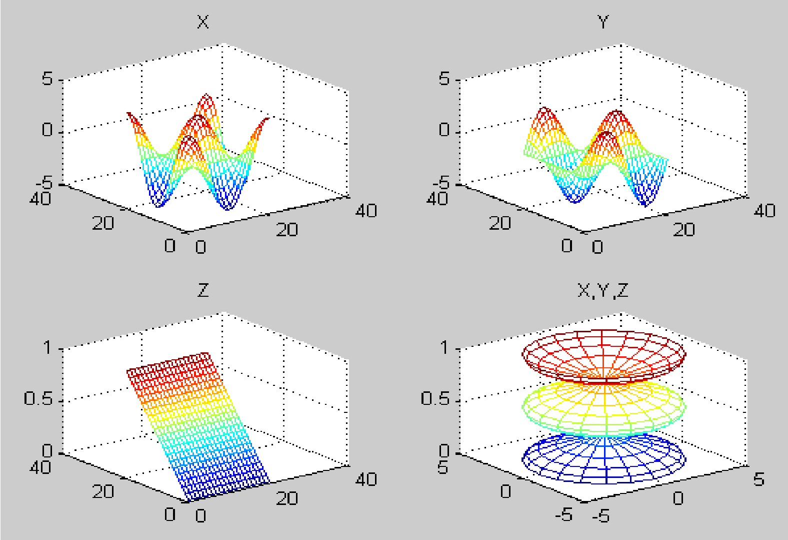
- Subplots: You can display multiple plots in different subregions of the same window using the subplot function.

The first two inputs to subplot indicate the number of plots in each row and column. The third input specifies which plot is active. As the following example shows:

Example 1.5: create four plots in a 2-by-2 grid within a figure window.

```
t = 0:pi/10:2*pi;
[X,Y,Z] = cylinder(4*cos(t));
```

```
subplot(2,2,1); mesh(X); title('X');
subplot(2,2,2); mesh(Y); title('Y');
subplot(2,2,3); mesh(Z); title('Z');
subplot(2,2,4); mesh(X,Y,Z); title('X,Y,Z');
```



• Other plotting functions in Matlab

– **Log scales:** semilogx, semilogy, loglog

– **Two y-axes scales:** plotyy

– 3D line plots: plot3

– **Surface and mesh plots:** surf, surfc, mesh, meshc, waterfall, ribbon, trisurf, trimesh

– **Histograms:** hist, histc, area, pareto

– **Bar plots:** bar, bar3, barh, bar3h

– **Pie charts:** pie, pie3, rose

– **Discrete data:** stem, stem3, stairs, scatter, scatter3, spy, plotmatrix

– **Polar plots:** polar, rose, compass

– **Contour plots:** contour, contourf, contourc, contour3, contourslice

– **Vector fields:** feather, quiver, quiver3, compass, streamslice, streamline

1.8: Logical Subscripting

The logical vectors created from logical and relational operations can be used to reference subarrays. Suppose X is an ordinary matrix and L is a matrix of the same size that is the result of some logical operation. Then X(L) specifies the elements of X where the elements of L are nonzero.

This kind of subscripting can be done in one step by specifying the logical operation as the subscripting expression. Suppose you have the following set of data:

```
x = [2.1 1.7 1.6 1.5 NaN 1.9 1.8 1.5 5.1 1.8 1.4 2.2 1.6 1.8];
```

The NaN is a marker for a missing observation, such as a failure to respond to an item on a questionnaire. To remove the missing data with logical indexing, use **isfinite(x)**, which is true for all finite numerical values and false for NaN and Inf:

```
x = x(isfinite(x))

x =

    2.1  1.7  1.6  1.5  1.9  1.8  1.5  5.1  1.8  1.4  2.2  1.6  1.8
```

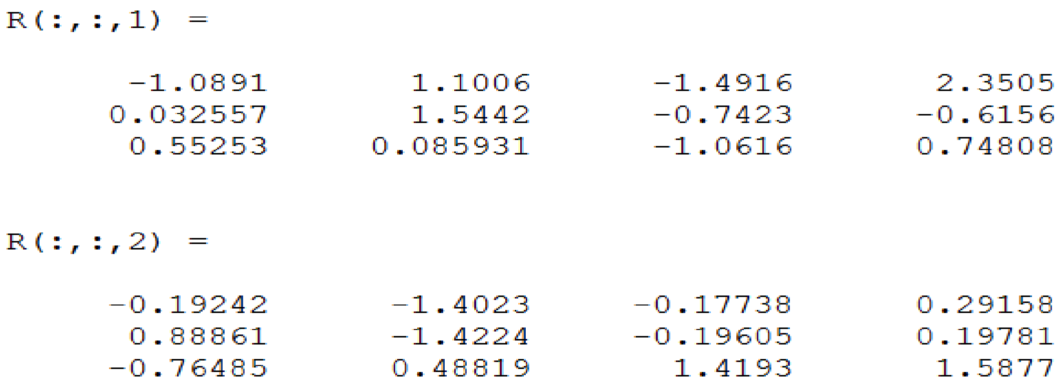
Now there is one observation, 5.1, which seems to be very different from the others. It is an *outlier*. The following statement removes outliers, in this case those elements more than three standard deviations from the mean:

```
x = x(abs(x-mean(x)) <= 3*std(x))
x =
       2.1 1.7 1.6 1.5 1.9 1.8 1.5 1.8 1.4 2.2 1.6 1.8
```

1.9: Multidimensional Arrays

Multidimensional arrays in the MATLAB environment are arrays with more than two subscripts. One way of creating a multidimensional array is by calling zeros, ones, rand, or randn with more than two arguments. For example,

```
R = randn(3,4,2)
```



```
R(:,:,1) =

      -1.0891        1.1006       -1.4916        2.3505
     0.032557        1.5442       -0.7423       -0.6156
      0.55253      0.085931       -1.0616       0.74808


R(:,:,2) =

     -0.19242       -1.4023      -0.17738       0.29158
      0.88861       -1.4224      -0.19605       0.19781
     -0.76485       0.48819        1.4193        1.5877
```

Creates a 3-by-4-by-2 array, with a total of (3*4*2 = 24) normally distributed random elements.

A three-dimensional array might represent three-dimensional physical data; say the temperature in a room, sampled on a rectangular grid. Or it might represent a sequence of matrices, $A^{(k)}$, or samples of a time-dependent matrix, $A(t)$. In these latter cases, the $(i, j)^{th}$ element of the k^{th} matrix, or the t^{kth} matrix, is denoted by A(i, j, k).

MATLAB and Dürer's versions of the magic square of order 4 differ by an interchange of two columns. Many different magic squares can be generated by interchanging columns. The statement

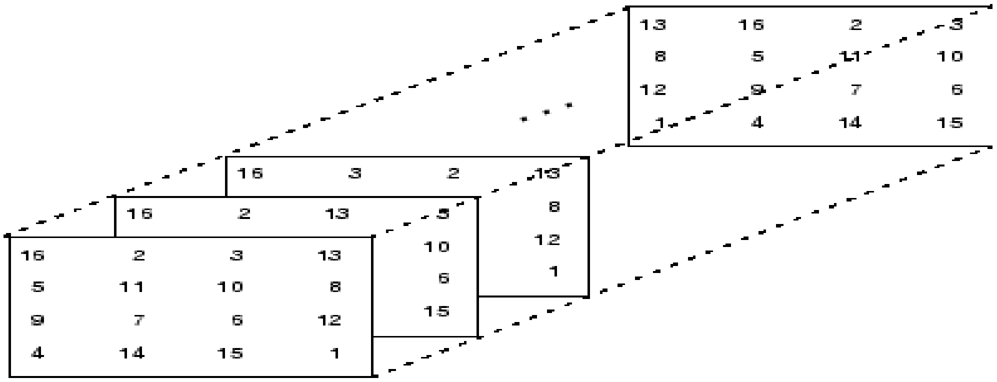
p = perms(1:4);

Generates the 4! = 24 permutations of 1:4. The k^{th} permutation is the row vector p(k,:). Then stores the sequence of (24) magic squares in a three-dimensional array, M. The size of M is

size(M)

ans =

    4   4 24



Note: The order of the matrices shown in this illustration might differ from your results. The perms function always returns all permutations of the input vector, but the order of the permutations might be different for different MATLAB versions.

The statement

```
sum(M,d)
```

Computes sums by varying the d^{th} subscript. So

```
sum(M,1)
```

Is a 1-by-4-by-24 array containing 24 copies of the row vector:

```
34  34  34  34
```

And

```
sum(M,2)
```

Is a 4-by-1-by-24 array containing 24 copies of the column vector

```
34
34
34
34
```

Finally,

```
S = sum(M,3)
```

Adds the (24) matrices in the sequence. The result has size 4-by-4-by-1, so it looks like a 4-by-4 array:

```
S =
        204  204  204  204
        204  204  204  204
        204  204  204  204
        204  204  204  204
```

1.10: Programming in Matlab

• Elements of Matlabas a programming language:

– Expressions

– Flow Control Blocks

  • Conditional

- Iterations (Loops)

– Scripts

– Functions

– Objects and classes (not covered here)

- Be mindful of existing variables and function names!

– Creating a variable or function that is already used by Matlab will cause troubles and errors!

– Example: Saving a variable as **sin = 10** will prevent you from using the sine function! Use something more descriptive such as **sin_x= 10**

1.10.1: Relational Operators

- Matlab has six relational Operators

– Less Than	<
– Less Than or Equal	<=
– Greater Than	>
– Greater Than or Equal	>=
– Equal to	==
– Not Equal to	~=

- Relational operators can be used to compare scalars to scalars, scalars to matrices/vectors, or matrices/vectors to matrices/vectors of the same size

- Relational operators to precedence after addition / subtraction

1.10.2: Logical Operators

- Matlab supports four logical operators

– Not	~
– And	**&** or **&&**
– Or	| or ||
– Exclusive Or (xor)	**xor()**

- Not has the highest precedence and is evaluated after parentheses and exponents

- And, or, xor have lowest precedence and are evaluated last

1.10.3: Conditional Structures

- If / Then Structure

```
if expression
   commands
end
```

- Example

```
if (x > 4)  &&  (y < 10)
    z = x + y;
end
```

- If / Else Structure

```
if expression
   commands
else
   commands
end
```

- Example

```
if (x > 4) && (y < 10)
    z = x + y;
else
    z = x * y;
end
```

- If / Elseif/ Else Structure

```
if expression
           commands
elseif expression
          commands
else
          commands
end
```

- Example

```
if  (x > 4) && (y < 10)
     z = x + y;
elseif (x < 3)
          z = 10 * x;
elseif (y > 12)
          z = 5 / y;
else
          z = x * y;
end
```

- Conditional Structures can be nested inside each other

```
if (x > 3)
if (y > 5)
    z = x + y;
elseif (y < 5)
    z = x -y;
end
elseif (y < 10)
          z = x * y;
else
          z = x / y;
end
```

- Matlab will auto-indent for you, but indentation is not required
- Switch / Case / Otherwise function used if known cases of a variable will exist

– Used in place of If / Elseif/ Else structure

- Syntax

```
switch switch_expression
case case_expression
    statements
case case_expression
    statements
otherwise
    statements
end
```

if–elseif-else	switch –case -otherwise
if x == 1	**switch** x
z = 5;	**case** 1
elseif x == 2	z = 5;
z = 4;	**case** 2
elseif x == 3	z = 4;
z = 3;	**case** 3
elseif (x == 4) || (x == 5)	z = 3;
z = 2;	**case**{4 , 5}
else	z = 2;
z = 1;	**otherwise**
end	**z = 1;**
	end

1.11: Matlab Iteration Structures

- Definite looping structures (**for**)

```
for variable = expression
    commands
end
```

- Example

```
for i = 1:1:25
    A(i) = i^2;
end
```

- Can also nest loops!

– Can mix for / while loops

- Nested For Loop Example

```
for i = 1:1:25
    for j = 1:1:4
        A(i,j) = i*j;
    end
end
```

- Indefinite looping structures (**while**)

```
while expression
    commands
end
```

- Example

```
x = 0; y = 0;
while x < 10
    y = y + x;
    x = x + 1;
end
```

- You need to make sure the variable in the while loop expression is changed during the loop!

– May lead to an infinite loop!

•Example for infinite Loop

```
x = 0;
while x < 10
    y = x;
end
```

1.12: M-Files

- Text files containing Matlab programs

– Can be called from the command line or from other M-Files

- Contain "**.m**" file extension

- Two main types of M-Files

– Scripts

– Functions

- Comment character is %

– % will comment out rest of line

1.12.1: M-Files –Scripts

• Scripts are simply M-Files with a set of commands to run

– Do not require input values or have output values

– Execute commands similarly to how they would be done if typed into the command window

– Ctrl + N

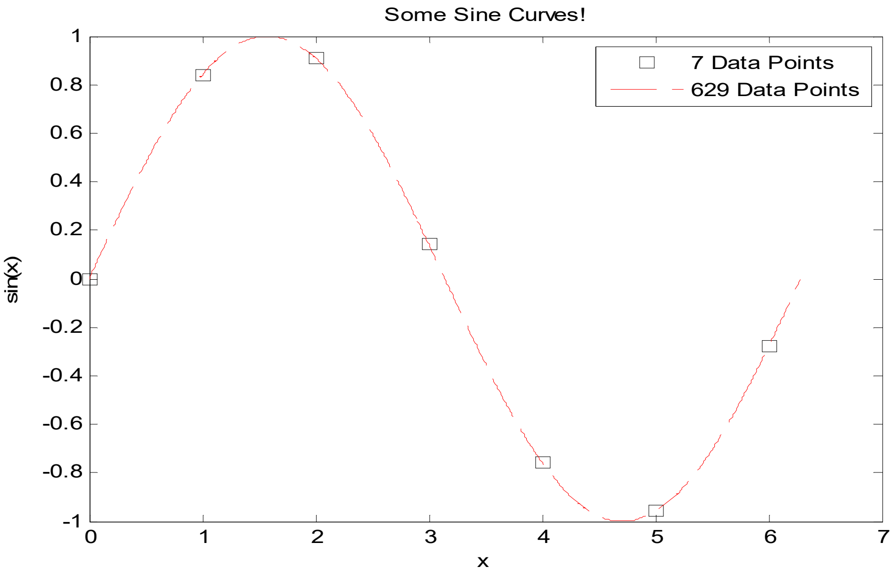
– Select New → Script from Menu

• To run M-File:

–>> F5 or Run Run

Example 1.6:

```
figure() % New Figure
x1 = 0:1:2*pi; y1 = sin(x1); % First Data Set
x2 = 0:0.01:2*pi; y2 = sin(x2); % Second Data
Set
plot(x1,y1,'sk',x2,y2,'r--') % Make Plot
title('Some Sine Curves!') % Add Title, Labels,
Legend, etc.
xlabel('x')
ylabel('sin(x)')
legend('7 Data Points','629 Data
Points','Location','NorthEast')
```



1.12.2: M-Files –Functions

- Functions typically require input or output values

- "What happens in the function, stays in the function"
  - Only variables visible *after* function executes are those variables defined as output

- Usually one file for each function defined

- Structure:

```
function [outputs] = funcName (inputs)
commands;
end
```

- Function Definition Line Components

1. Function keyword → Identifies M-File as a function

2. Output Variables → Separated by commas, contained in **square brackets**

   - Output variables must match the name of variables inside the function!

3. Function Name → must match the name of the .m file!

4. Input Variables → Separated by commas, contained in **parentheses**
   - Input variables must match the name of variables inside the function!

- When calling a function, you can use any name for the variable as input or output

  - The names **do not** have to match the names of the .m file

Example 1.7: Explain function to calculate the area and perimeter of a rectangle

```
function [area, perimeter] = dF(base, height)
% "df" Demo func. to calculate the area and perimeter of a rectangle
% Function can handle scalar and vector inputs
% Isaac Tetzloff -Aug 2013
area = base .* height; % Calculate the area
perimeter = 2 * (base + height); % Calculate the perimeter
end
```

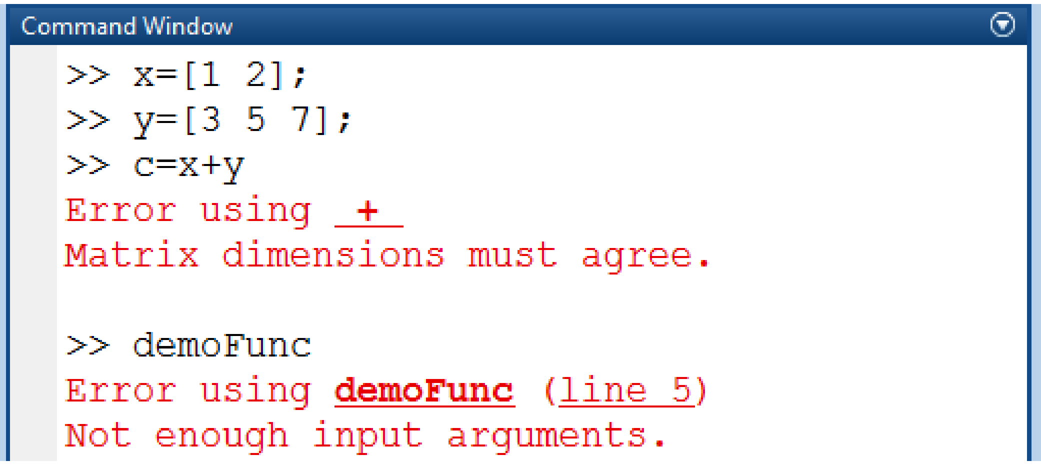
```
>> [a, p] = dF(10, 15); % Returns both values as a & p
>> area = dF(10, 5);% Returns area and saves as area
>> perim= dF(5, 15);% Returns area and saves as perim!
>> [perim, area] = dF(5, 15);% Saves area as perim, and vice versa!
>> x = [1 2 3]; y = [5 4 3];
>> [x, y] = dF(x, y);% Returns both and overwrites input!
```

- In modified function below, only variables output are **area** and **perimeter**
  - Matlab and other functions will not have access to **depth**, **mult**, **add**, or **volume**!
  - **REMEMBER:** *What happens in the function stays in the function!*

```
function [area, perimeter] = dF(base, height)
depth = 10;                 % Assume 3D prism has depth of 10
mult= base .* height;       % Multiply base by height
add = base + height;        % Add base and height
area = mult;                % Calculate the area
perimeter = 2 * add;        % Calculate the perimeter
volume = mult* depth;       % Calculate the volume
end
```

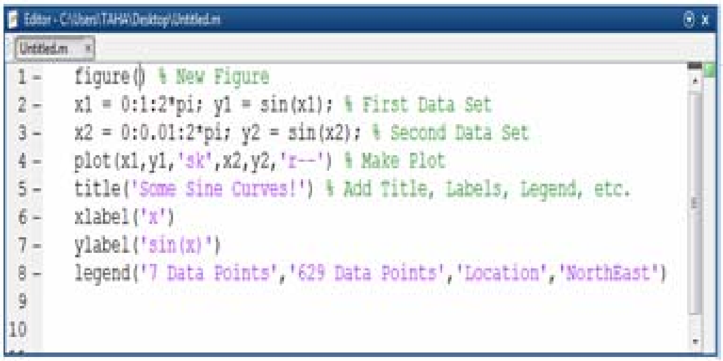
1.13: Debugging in Matlab

- Matlab errors are very descriptive and provide specifics about error
  - If a function or script causes an error, Matlab will give the line of code and file with the error



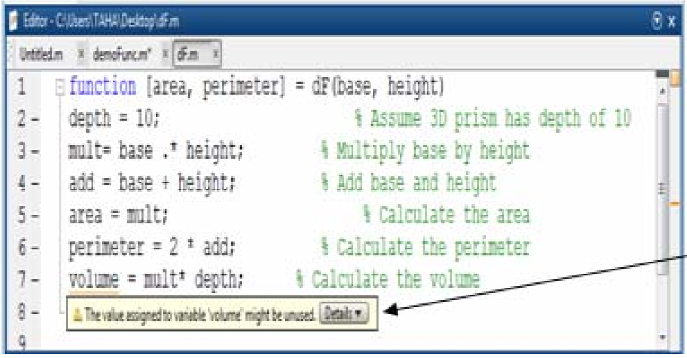
```
>> x=[1 2];
>> y=[3 5 7];
>> c=x+y
Error using  + 
Matrix dimensions must agree.

>> demoFunc
Error using demoFunc (line 5)
Not enough input arguments.
```

- The Matlab Editor provides on-the-fly debugging help!



```
figure() % New Figure
x1 = 0:1:2*pi; y1 = sin(x1); % First Data Set
x2 = 0:0.01:2*pi; y2 = sin(x2); % Second Data Set
plot(x1,y1,'sk',x2,y2,'r--') % Make Plot
title('Some Sine Curves!') % Add Title, Labels, Legend, etc.
xlabel('x')
ylabel('sin(x)')
legend('7 Data Points','629 Data Points','Location','NorthEast')
```

Green square
No errors or warning



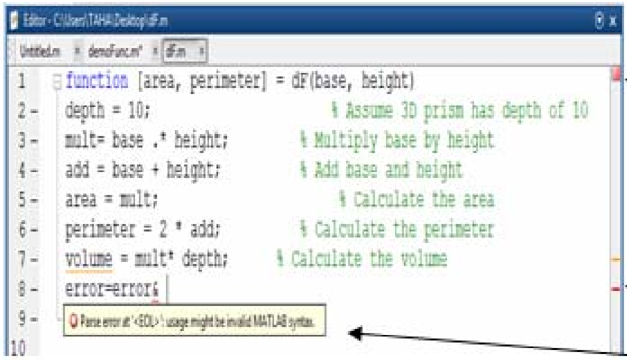
```
function [area, perimeter] = dF(base, height)
depth = 10;                        % Assume 3D prism has depth of 10
mult= base .* height;          % Multiply base by height
add = base + height;           % Add base and height
area = mult;                        % Calculate the area
perimeter = 2 * add;          % Calculate the perimeter
volume = mult* depth;      % Calculate the volume
```

⚠ The value assigned to variable 'volume' might be unused.

Orange Square
Warning present, but code will still run
Indicated by orange bar

Mouse over for warning message

- The Matlab Editor provides on-the-fly debugging help!



```
function [area, perimeter] = dF(base, height)
depth = 10;                        % Assume 3D prism has depth of 10
mult= base .* height;          % Multiply base by height
add = base + height;           % Add base and height
area = mult;                        % Calculate the area
perimeter = 2 * add;          % Calculate the perimeter
volume = mult* depth;      % Calculate the volume
error=error&
```

Parse error at '<EOL>': usage might be invalid MATLAB syntax.

Red square
Error present and
Code will not run!
Indicated by red bar

Mouse over for error message

1.14: Advanced Features to Explore

Symbolic Math

- Allows for symbolic manipulation of equations, including solving, simplifying, differentiating, etc.

Inline Functions

- Creates a workspace variable that is a simple equation

```
>> f =  x^2 + 2*x + 1
>> y = f(3) →y = 16
```

Optimization

- Solve constrained problems with **fmincon**, unconstrained with **fminunc**, bounded problems with **fminbnd**, etc.

Many Others!

- Matlab is extremely powerful and has a lot of advanced features, too many to go through here!
- Within Matlab:
  - –Type **help function** to provide information about the function in the command window
  - – Type **doc function** to open the documentation about the function
  - – Type **doc** to pull up the documentation within Matlab to explore
- Online
  - – Documentation: http://www.mathworks.com/help/matlab/
  - – Tutorials: http://www.mathworks.com/academia/student_center/tutorials/
  - – Matlab Primer / Getting Started with Matlab(pdf): http://www.mathworks.com/help/pdf_doc/matlab/getstart.pdf

1.15: Descriptive statistics with the Statistics Toolbox of MATLAB

Some of the functions to compute the most frequent statistics are the following:

```
mean(x)              % Mean value of the elements in x.
median(x)            % Median value of the elements in x.
std(x),var(x)          % Standard deviation and variance of x normalized
                         by n − 1.
std(x,1),var(x,1)    % Standard deviation and variance of x normalized
                       by n.
range(x)             % Range of x.
iqr(x)               % Interquartile range of x.
```

```
mad(x)                % Mean absolute deviation of x.
max(x),min(x)         % Maximum and minimum element of x.
skewness(x), kurtosis(x)       % Skewness and kurtosis of x.
moment(x, order)          % Central moment of x specified by order.
prctile(x,p)        % pth percentile of x (if p=50, returns the median of x)
```

Observe that if x is a matrix, then the result of these functions is a row vector containing the statistic for each column of x.

Other two interesting functions are **cov** and **corrcoef**. For vectors, the command **cov** returns the variance:

```
>> x=rand(100,1); cov(x)
```

For matrices, where each row is an observation, and each column a variable, returns the covariance matrix:

```
>> x=rand(100,5); cov(x)
```

For two vectors, z and w, of equal length, cov(z , t) returns a matrix with the variances of z and w in the diagonal and the covariance of z and w in the two off-diagonal entries.

```
>> z=rand(100,1); t=rand(100,1); cov(z , t)
```

Observe that cov(z , t) = cov([z t]). For two matrices,

cov(X,Y)=cov(X(:),Y(:)). Finally, cov(x) or cov(x,y) normalizes by (n − 1) and cov(x,1) or cov(x,y,1) normalizes by n, where n is the number of observations.

The corrcoef(X) command calculates a matrix of correlation coefficients for an array X, in which each row is an observation and each column is a variable. Observe that corrcoef(X,Y), where X and Y are column vectors, is the same as corrcoef([X Y]).

```
>> corrcoef(x)
```

The Statistics Toolbox and some built-in functions of MATLAB allows to plot a number of useful graphics in descriptive statistics.

```
hist(x)            % Histogram.
boxplot(x)         % Boxplots of a data matrix (one per column).
cdfplot(x)         % Plot of empirical cumulative distribution function.
normplot(x)        % Normal probability plot (one per column).
qqplot(x,y)        % Quantile-Quantile plot.
```

You can change the way any toolbox function works by copying and renaming the M-file, then modifying your copy. You can also extend the toolbox by adding your own M-files.

For example, imagine we are interested in plotting a variant of the histogram where the counts are replaced by the normalized counts, that is, the relative histogram. By normalized count, we mean the count in a class divided by the total number of observation times the class width. For this normalization, the area (or integral) under the histogram is equal to one. Now, we can look for the file hist.m and modify it. This file is usually in the following path (or something similar):

```
c:\MATLAB6p5\toolbox\matlab\datafun
```

Open it and let's try to change it. Observe that the hist command produces a histogram bar plot if there are no output arguments, that is, we look for the sentences:

```
if nargout == 0

bar(x,nn,'hist');

...
```

The sentence bar(x,nn,'hist') draws the values of the vector nn (frequency) as a group of vertical bars whose midpoints are the values of x, see help bar. For example, we can change the previous sentences by the following ones to obtain a white normalized histogram:

```
if nargout == 0

bar(x,nn/(length(y)*(x(2)-x(1))),'hist','w');
...
```

You can also change the help section including for example a sentence like this:

```
    % HIST(...) without output arguments produces a normalized
    histogram bar
    % plot of the results.
```

And now, save the changed file as histn.m, for example. If you want histn to be a global function, you can save it in the same folder hist.m was. Otherwise, you can save it in a different folder and then histn will only work if you are in this directory or if you add it to the MATLAB's search path, (see path).

1.16: Simulation of linear models

The reporting of a simulation experiment should receive the same care and consideration that would be accorded the reporting of other scientific experiments. Hoaglin and Andrews (1975) outline the items that should be included in a report of a simulation study. In addition to a careful general description of the experiment, the report should include mention of the random number generator used, any variance-reducing methods employed, and a justification of the simulation sample size. The *Journal of the American Statistical Association* includes these reporting standards in its style guide for authors.

Closely related to the choice of the sample size is the standard deviation of the estimates that result from the study. The sample standard deviations actually achieved should be included as part of the report. Standard deviations are often reported in parentheses beside the estimates with which they are associated. A formal analysis, of course, would use the sample variance of each estimate to assess the significance of the differences observed between points in the design space; that is, a formal analysis of the simulation experiment would be a standard analysis of variance.

1.16.1: Simulation of simple linear model

Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + E$$

Where a response or "dependent variable", y, is modeled as a linear function of a single regressor or "independent variable", *x*, plus a random variable, *E*, called the "error". Because *E* is a random variable, y is also a random variable. The statistical problem is to make inferences about the unknown, constant parameters β_0 and β_1 and about distributional parameters of the random variable, *E*.

We also generally assume that the realizations of the random error are independent and are unrelated to the value of *x*.

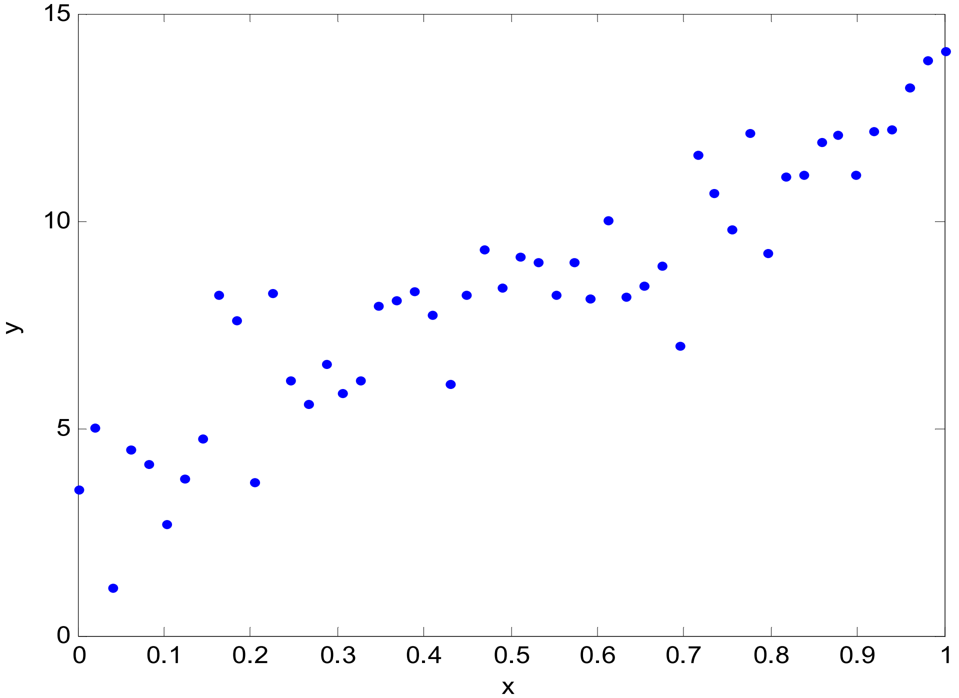
A bivariate scatter plot is a simple plot of *x* versus *y* between two variables. A bivariate scatter plot is a convenient first step to visualize the relationship between the two variables.

Assume that we have two variables that are linearly related, except some Gaussian noise term with mean 0 and standard deviation 1:

$y = 3 + 10x + \text{noise}$

Assuming that the variable x is a linearly spaced row vector of length 50, between 0 and 1, generate the y vector:

```
n=50; % number of observations
x=linspace(0,1,n); % linearly spaced vector a length n
beta0=3;
beta1=10;
E=randn(1,n);
y= beta0+beta1*x +E;
plot(x,y,'.')
xlabel('x')
ylabel('y')
```



Each time the command is used, a different number will be generated. The “random” numbers generated by Matlab (and others) are actually pseudorandom numbers as they are computed using a deterministic algorithm. The algorithm, however, is very complicated, and the output does not appear to follow a predictable pattern. For this reason the output can be treated as random for most practical purposes. The same sequence of numbers will not be generated unless the same starting point is used. This starting point is called the “seed”. Each time you start Matlab, the random number generator is initialized to the same seed value. The current seed value can be seen using:

```
randn('seed',1)  % specify a seed (optional)
```

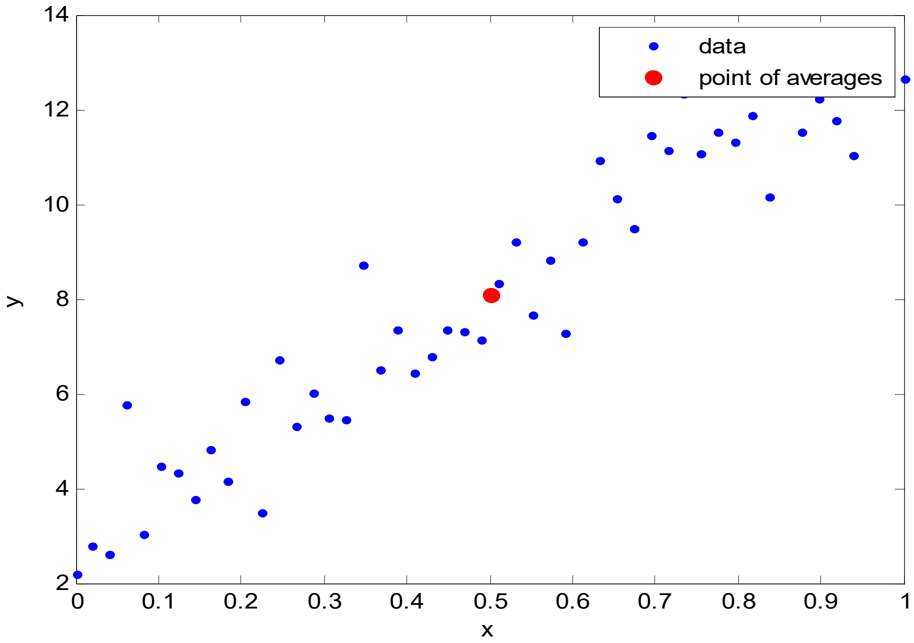
By setting a seed value, we ensure that the same results will be produced each time the script is executed. The seed can be set to a value (say, 1234) as follows:

```
randn('seed',1234)
```

The purpose here is to make sure that the program starts from the same seed. The value of the seed is not important.

In a bivariate scatter plot (*x*,*y*), the point with coordinates (mean(*x*), mean(*y*)) , is known as the point of averages.

```
mx=mean(x);
my=mean(y);
hold on;
plot(mx,my, 'ro', 'markerfacecolor','r')
legend('data', 'point of averages')
```



Covariance:

Covariance between vectors *x* and *y* can be computed in “unbiased” and “biased” versions as:

```
c= mean((x-mx).*(y-my))    % covariance (biased)
n=length(x);
cs= c*n/(n-1)              % sample covariance(unbiased)
```

Ans:

```
c = 0.85307   cs =0.87048
```

Correlation coefficient:

The correlation coefficient between two variables is a measure of the linear relationship between them. The correlation coefficient between two vectors can be found using the average of the product of the z-scores of x and y. The “biased” version is:

```
zx=zscore(x,1);
zy=zscore(y,1) ;
r=mean(zx.*zy)
```

Ans:

```
r =
     0.94845
```

Correlation coefficient can also be computed from the covariance, as follows:

```
sx=std(x,1);
sy=std(y,1);
r=c/(sx*sy)
```

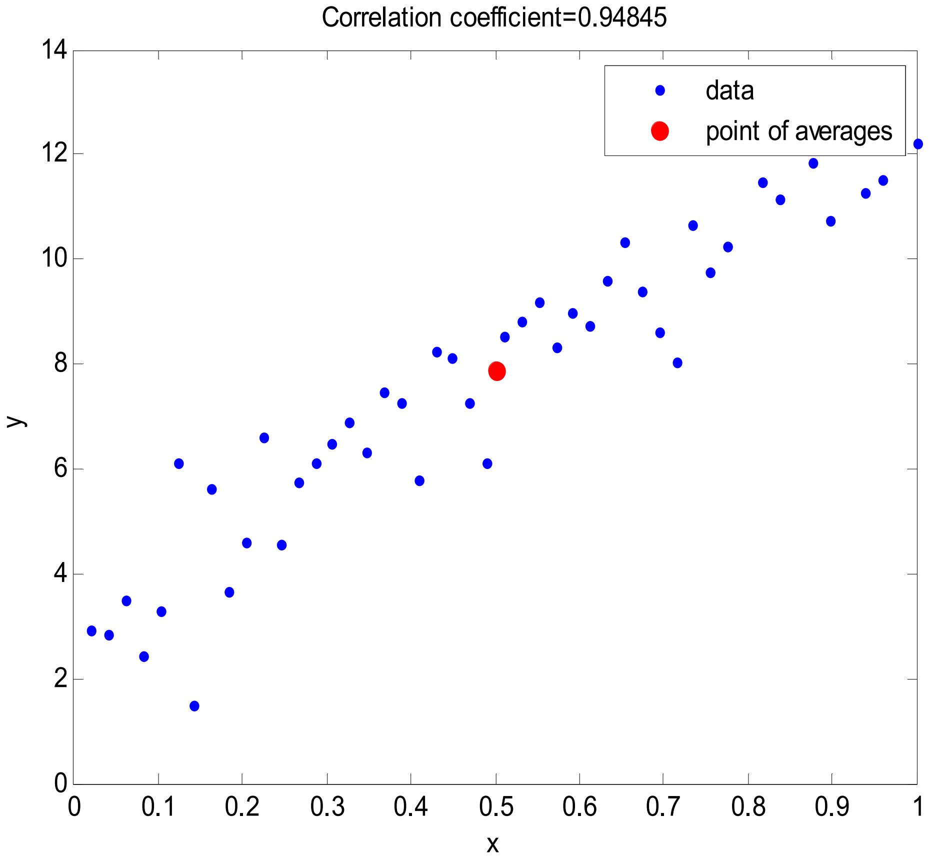
Ans:

```
r =
     0.94845
```

The “unbiased” version (sample correlation coefficient) is computed the same way, except that the flag “1” is replaced by “0”.

Add a title that shows the correlation coefficient to the previous plot. For this, we need to convert the numerical value to a string, using the num2str command:

```
title(['Correlation coefficient=',num2str(r)])
```



The correlation coefficient is sensitive to outliers. To see this, change the first element of y to 40 and recomputed the correlation coefficient:

```
y(1)=40;

zx=zscore(x,1)

zy=zscore(y,1)

r=mean(zx.*zy)
```

Ans:

```
r =
     0.31003
```

Notice that a single outlier has significantly reduced the correlation coefficient.

1.16.2: Ordinary Least Squares Regression

Regression is a way to understand the mathematical relationship between variables. This relationship can then be used to

- Describe the linear dependence of one variable on another.
- Predict values of one variable from values of another.
- Correct for the linear dependence of one variable on another, in order to clarify other features of its variability.

Unlike the correlation coefficient, which measures the strength of a linear relationship, regression focuses on the mathematical form of the relationship.

In simple linear regression, the mathematical problem is as follows: Given a set of *k* points (x_i, y_i), *i* =1,2,…,*k*, which are related through the equation $y_i = b_0 + b_1 x_i + n_i$, where b_0 and b_1 are constant (unknown) coefficients and n_i is a realization of zero-mean Gaussian noise with variance σ^2. That is, $n_i \sim \mathrm{N}(0,\sigma^2)$. As the noise term n_i is a realization of a random variable, so is y_i. Because of the random noise, the coefficients b_0 and b_1 cannot be determined with certainty. Our goal is to find the best fit line $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$ minimizing the sum of squared errors:

$$S = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2$$

The $\hat{b}_1$ and $\hat{b}_0$ values minimizing *S* are found by setting $\frac{\partial S}{\partial b_1} = 0$, $\frac{\partial S}{\partial b_0} = 0$.

The result is:

$$\hat{b}_1 = \frac{Covariance \;\; between \;\; x \;\; and \;\; y}{Variance \;\; of \;\; x}$$

$$\hat{b}_0 = (mean \; of \;\; y) - \hat{b}_1 (mean \; of \;\; x)$$

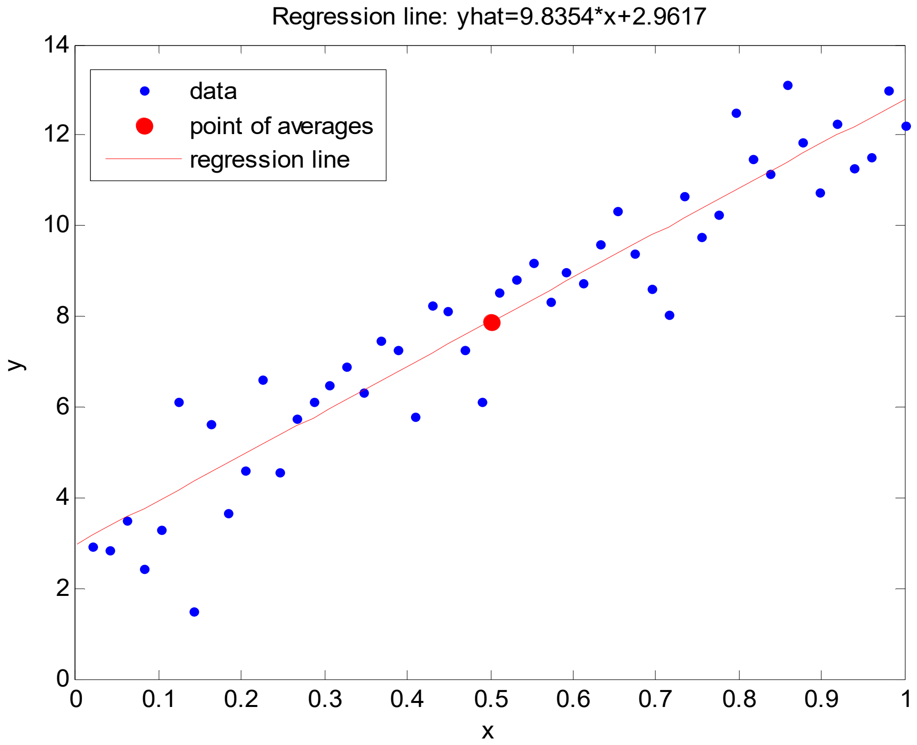
These $\hat{b}_1$ and $\hat{b}_0$ values are the Ordinary Least Square (OLS) estimates of b_1 and b_0, respectively. The equation of the regression line (also known as the “best fit line”) is then $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$

```
bh1=c/sx^2;             % covariance divided by variance of x
bh0=my-bh1*mx;
yhat=bh0+bh1*x;         % regression line
```

```
Ans:

bh1 =

    9.8354

bh0 =

    2.9617
```

Plot the regression line in red, and update the legend and the title:

```
plot(x,yhat,'r')
legend('data', 'point of averages','regression line')
title(['Regression line: yhat=',num2str(bh1),'*x+',num2str(bh0)])
```



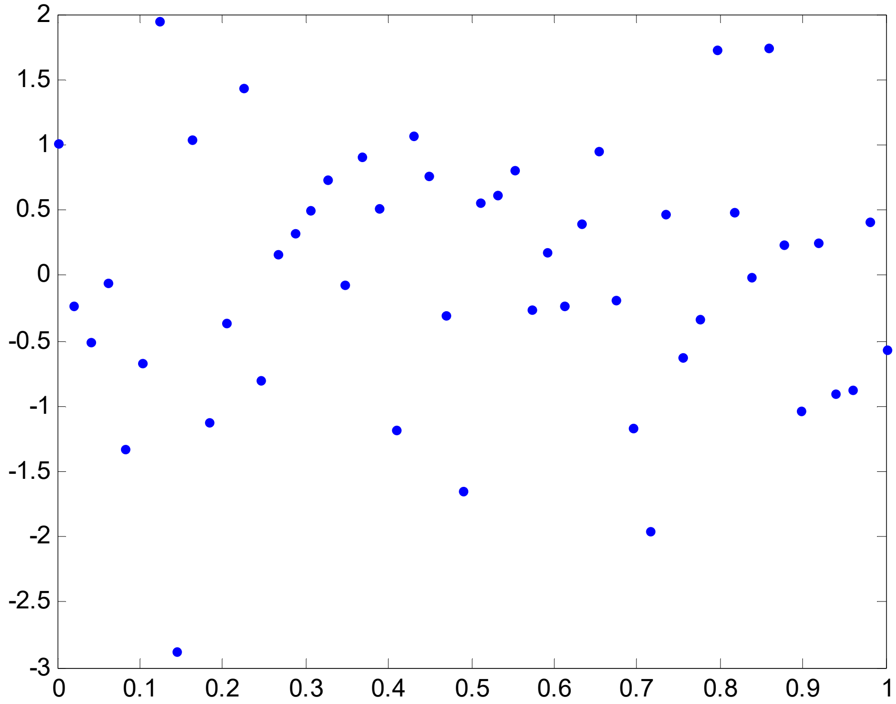
Note that the regression line passes through the point of averages. The equation of the regression line shown in the title should be close to the original equation from which the data was generated:

$y = 3 + 10x + \text{noise}$

Because of the noise, the predictions will not exactly coincide with the observations. The residuals e_i are defined as the deviations of each observation from its estimate:

$$e_i = y_i - \hat{y}_i$$

```
e=y-yhat; %residuals
figure;
plot(x,e,'.')
```



Ideally, the residuals should be more or less symmetrically distributed around zero (have mean $\cong 0$):

```
M = mean(e) % average residual
```

Ans:

```
M =

        -2.1583e-15
```

In addition, the amount of scatter should not show a systematic increase or decrease with increasing values of *x*. In other words, the scatter plot should be homoscedastic, not heteroscedastic. The variance of the noise can be estimated from the residuals (MSE) as follows:

$$\mathrm{MSE} = \hat{\sigma}^2 = \frac{\sum_{i=1}^{\mathrm{n}} e_i^2}{\mathrm{n}-2}$$

```
MSE = sum(e.^2)/(n-2)      % OLS estimator of noise variance
```

```
Ans:
MSE =
      0.97588
```

The n-2 in the denominator is known as the “degrees of freedom”, and is computed by subtracting the number of parameters estimated (b_0 and b_1) from the number of observations.

The estimated noise variance for this particular problem should be close to 1, which is the variance of the noise used in generating the data.

The coefficient of determination (R^2) is a measure of how well the regression line represents the data. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{\mathrm{n}} e_i^2}{\sum_{i=1}^{\mathrm{n}} (y_i - \bar{y})^2}, \quad where \quad \bar{y} = \frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}} y_i$$

In simple linear regression, R^2 is equal to the square of the correlation coefficient (r^2) between *x* and *y*. If *r* = 0.9, then $R^2 = r^2$ = 0.81 which means that 81% of the total variation in *y* can be explained by the linear relationship between *x* and *y*. The other 19% of the total variation in *y* remains unexplained.

```
R2=1-sum(e.^2)/sum((y-my).^2)  % coefficient of determination
r2=r^2                         % correlation coefficient squared
```

```
Ans:

R2 =
    0.89956
r2 =
    0.89956
```

Save the code as chapter1simsimple.m. This file will be used in future chapters.

1.16.3: Simple linear regression in matrix form

Consider the simple linear regression equation $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$.

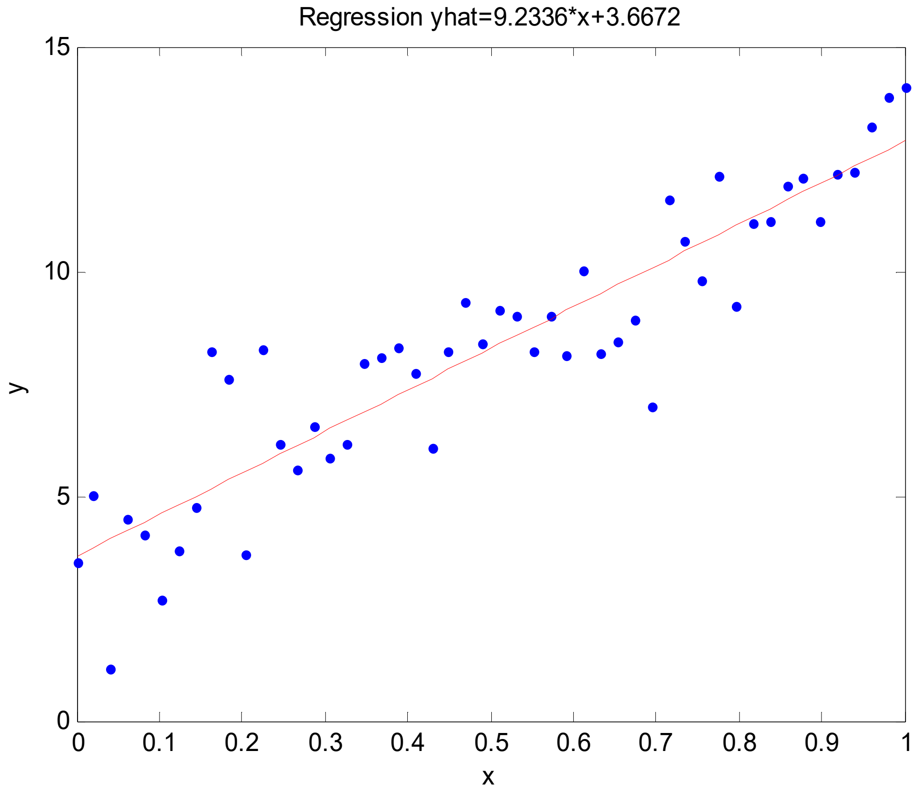
Note that same equation can be written as $\hat{y}_i = [1 \quad x_i]\begin{bmatrix}\hat{b}_0 \\ \hat{b}_1\end{bmatrix}$.

This means that if the two coefficients are combined into a single column vector $\hat{\mathrm{b}} = \begin{bmatrix}\hat{b}_0 \\ \hat{b}_1\end{bmatrix}$, and the independent variable is augmented by adding a "1" to the front $\breve{x}_i = [1 \quad x_i]$, the i^{th} predicted value can be computed as $\hat{y}_i = \breve{x}_i\,\hat{\mathrm{b}}$. For the entire set of observations, we can write $\hat{\mathrm{Y}} = \mathrm{X}\hat{\mathrm{b}}$ where $\hat{\mathrm{Y}}$ is a column of predicted values, X is the design matrix, where the first column consists of ones, the second column is the values of the independent variables, and $\hat{\mathrm{b}} = \begin{bmatrix}\hat{b}_0 \\ \hat{b}_1\end{bmatrix}$.

The OLS (ordinary least squares) estimate of the regression coefficients is given by $\hat{\mathrm{b}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$. Recall the simple linear regression data generated from

$y = 3 + 10x + \text{noise}$

```
n=50;
x=linspace(0,1,n);                % linearly spaced vector a length n
y= 10*x + 3 + randn(1,n);
mx=mean(x), my=mean(y), sx=std(x,1);
c= mean((x-mx).*(y-my))           %covariance
bh1=c/sx^2
bh0=my-bh1*mx
yhat=bh0+bh1*x;                   %regression line
figure;
plot(x,y,'.')
hold on
plot(x,yhat,'r')
xlabel('x'), ylabel('y')
title(['Regression yhat=',num2str(bh1),'*x+',num2str(bh0)])
```



The same estimates of the regression coefficients can be obtained using the matrix form:

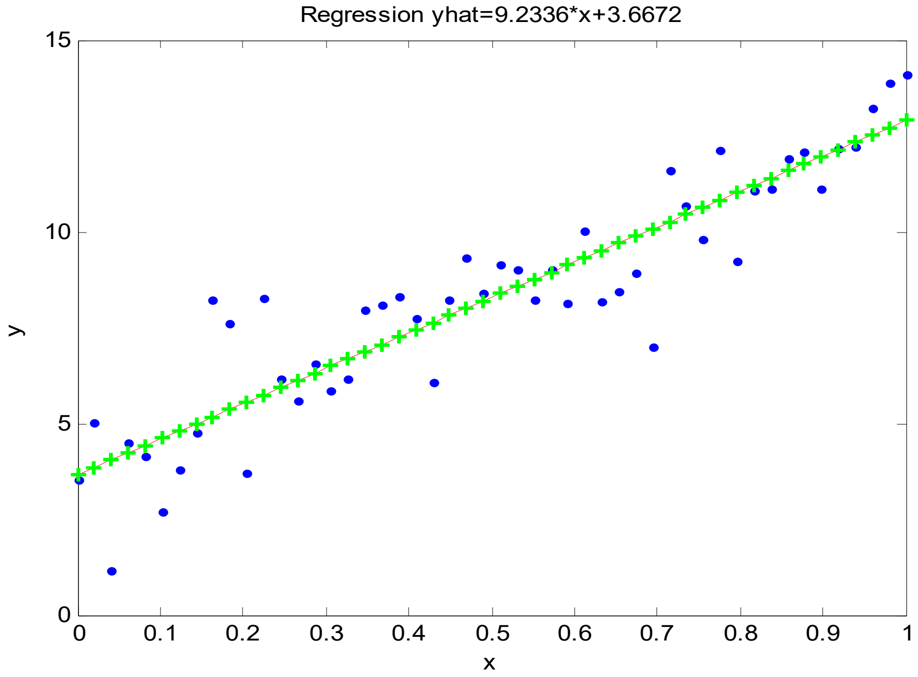
```
x=x(:);                        % make x a column
y=y(:);                        % make y a column
XX=[ones(n,1),x];              % create the design matrix
bh=(XX'*XX)^-1*XX'*y           % OLS estimate of b
```

Ans.

```
bh =

    3.6672

    9.2336
```

The $\hat{b}$ vector should contain the previously computed b_0 and b_1 values. The new regression line should also coincide with the previous line.

```
yhat=XX*bh;
hold on
plot(x,yhat,'g+','linewidth',2)
```



The residuals and the estimated noise variance are computed as

```
e=y-yhat;                 % residuals
dof= n-rank(XX);          % degrees of freedom
MSE=sum(e.^2)/dof         % estimated noise variance
Ans.

MSE=

    1.5741
```

Save the code as SIMSIMPLEMATRIX.m. This file will be used in future chapters.

1.16.4: Multiple Linear Regression

In multiple linear regression, the regression equation is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$$

And each observation is equal to the predicted value and a residual term e_i: $y_i = \hat{y}_i + e_i$

The matrix-based analysis presented in the previous section is equally applicable to multiple independent variables. For each additional independent variable, another column is added to the design matrix, X.

With k independent variables, the design matrix contains k+1 columns, the first column containing 1's. One difficulty with multiple independent variables is that the entire analysis cannot be summarized in a single figure, and the residuals need to be plotted with respect to each independent variable separately.

By using matrices, the multiple linear regression model, $\mathrm{Y} = \mathrm{X}\beta + \varepsilon$

Where $\varepsilon \sim \mathrm{N}\left(0, \sigma^{2}\, \mathrm{I}_{n}\right)$ and Y is an n×1 vector of observations, X is an n×k matrix of regressors, β is a n×1 vector of parameters and ε is an n×1 vector of random disturbances. The least squares estimator of β is given by,

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\, \mathrm{X}'\mathrm{Y}$$

Whose variance is,

$$\mathrm{Var}\left(\hat{\beta}\right) = \sigma^{2}\left(\mathrm{X}'\mathrm{X}\right)^{-1}$$

The predicted values are given by,

$$\hat{\mathrm{Y}} = \mathrm{X}\hat{\beta}$$

The residuals are,

$$\mathrm{e} = \mathrm{Y} - \hat{\mathrm{Y}}$$

And the residual variance is,

$$\mathrm{MSE} = \hat{\sigma}^{2} = \frac{\sum_{i=1}^{\mathrm{n}} e_i^2}{\mathrm{n} - k - 1}$$

We can now define the following function to solve the regression problem:

The coefficient of determination (R^2) is computed the same way as in the simple linear case:

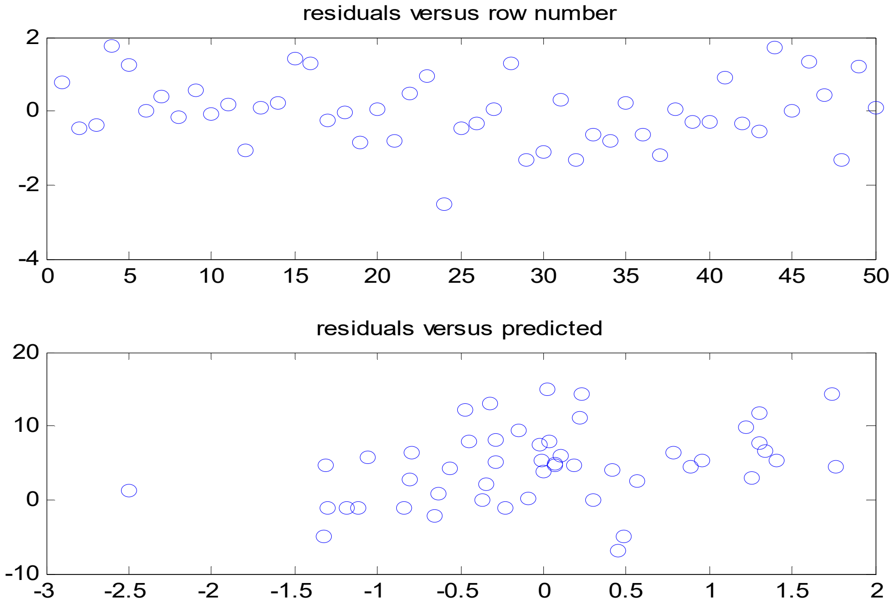
$$R^{2} = 1 - \frac{\sum_{i=1}^{\mathrm{n}} e_i^2}{\sum_{i=1}^{\mathrm{n}} \left(y_i - \bar{y}\right)^2}, \quad where \quad \bar{y} = \frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{n}} y_i$$

The R^2 value in multiple linear regression is often called the "coefficient of multiple determination."

```
randn('seed',1234) % specify a seed (optional)
n = 50; k = 4;
X = [ones(n,1) randn(n,k)];
b = [5;1;2;3;4];
y = X*b + randn(n,1);
[beta,Var_beta,resid,sR2] = regress(y,X)
MSE = sum(resid.^2)/(n - k - 1)
Var_Cov_beta=inv(X'*X)*MSE
R2=1-sum(resid.^2)/sum((y-mean(y)).^2)
subplot(2,1,1),plot(resid,'o'),title('residuals versus row number')
subplot(2,1,2),plot(resid,ypred,'o'),title('residuals versus predicted')
```

Ans.

```
              beta          R2            MSE
             5.1611       0.96567       0.87179
            0.78908
             2.1569
             2.9181
             4.0902

Var_Cov_beta =

     0.018533   -0.002176   -0.0023977    0.0011898    0.0028394
    -0.002176    0.022436    0.0048211    0.0030259   -0.0016523
   -0.0023977   0.0048211     0.020029    0.0066967     0.001665
    0.0011898   0.0030259    0.0066967     0.016782   -0.0006353
    0.0028394  -0.0016523     0.001665   -0.0006353     0.024338
```



Save the code as SIMMULTIPLEMATRIX.m. This file will be used in future chapters.

1.16.5: Multiple linear regression with the Statistics Toolbox of MATLAB

The Statistics Toolbox provides the regress function to address the multiple linear regression problems. regress uses QR decomposition of X followed by the backslash operator to compute $\hat{\beta}$. The QR decomposition is not necessary for computing $\hat{\beta}$, but the matrix R is useful for computing confidence intervals.

b = regress(y,X) returns the least squares estimator $\hat{\beta}$.

[b, bint, r, rint, stats] = regress(y, X) returns an estimate of β

Interval for β in the k×2 array bint. The residuals are returned in r and a 95% confidence interval for each residual is returned in the n × 2 array rint. The vector stats contain the R^2 statistic along with the F and p values for the regression.

[b,bint,r,rint,stats] = regress(y,X,alpha) gives 100(1 - alpha)% confidence intervals for bint and rint. For example, alpha = 0.2 gives 80% confidence intervals. Let’s see an example. Suppose the true model is, $\mathrm{Y} = \mathrm{X}\begin{pmatrix}10\\1\end{pmatrix} + \varepsilon \quad , \quad \varepsilon \sim \mathrm{N}(0, 0.01\,\mathrm{I}_n)$

Where I is the identity matrix. Suppose we have the following data:

```
randn('seed',1234);n=10; X = [ones(n,1) (1:n)']
y = X * [5;2] + normrnd(0,0.1,n,1)
[b,bint] = regress(y,X,0.05)
```

X		y
1	1	6.92063102736475
1	2	8.95834974723594
1	3	10.9217439183399
1	4	13.2145703970460
1	5	14.9213956160792
1	6	17.0448135509485
1	7	19.0098435509121
1	8	20.9326093816663
1	9	23.0200396628726
1	10	24.9311656046701

```
b =
     4.9845
     2.0005
bint =
     4.8304      5.1386
     1.9757      2.0254
```

Compare b to [10 1]’. Note that bint includes the true model values.

Another example comes from Chatterjee and Hadi (1986) in a paper on regression diagnostics. The data set (originally from Moore (1975)) has five predictor variables and one response.

```
load moore
```

```
X = [ones(size(moore,1),1) moore(:,1:5)];
```

Matrix X has a column of ones, and then one column of values for each of the five predictor variables. The column of ones is necessary for estimating the y-intercept of the linear model.

```
y = moore(:,6);
[beta, beta_interval, resid, resid_interval, STATS] = regress(y,X)
```

Where **regress** Multiple linear regression using least squares.

beta = regress(Y,X) returns the vector beta of regression coefficients in the linear model Y = X* beta. X is an n-by-p design matrix, with rows corresponding to observations and columns to predictor variables. Y is an n-by-1 vector of response observations.

[beta, beta_interval] = regress(Y,X) returns a matrix beta_interval of 95% confidence intervals for beta.

[beta, beta_interval, resid] = regress(Y,X) returns a vector resid of residuals.

[beta, beta_interval, resid, resid_interval] = regress(Y,X) returns a matrix resid_interval of intervals that can be used to diagnose outliers. If RINT(i,:) does not contain zero, then the i-th residual is larger than would be expected, at the 5% significance level. This is evidence that the I-th observation is an outlier.

[beta, beta_interval, resid, resid_interval, STATS] = regress(Y,X) returns a vector STATS containing, in the following order, the R-square statistic, the F statistic and p value for the full model, and an estimate of the error variance.

Ans.

beta	beta_interval	
-2.1561	-4.11538	-0.19691
-9.0116e-06	-0.00112	0.001103
0.0013159	-0.00139	0.004026
0.0001278	-3.71e-05	0.000293
0.0078989	-0.02213	0.037926
0.00014165	-1.65e-05	0.0003

resid	resid_interval	
0.562317	0.225802	0.898832
-0.14555	-0.54763	0.256525
0.088524	-0.32617	0.50322
-0.04788	-0.55146	0.455704
-0.2307	-0.70433	0.242926
0.170682	-0.28023	0.621592
-0.34134	-0.83769	0.155007
-0.07079	-0.62602	0.484439
-0.01029	-0.47488	0.454305
-0.10945	-0.63998	0.421089
0.171722	-0.3311	0.674541
0.050437	-0.49066	0.591533
-0.03991	-0.59383	0.514003
0.022723	-0.49909	0.544541
-0.39447	-0.87015	0.081217
0.081334	-0.41688	0.579544
0.072986	-0.08787	0.233845
0.011354	-0.4987	0.521405
-0.22227	-0.66763	0.223093
0.380568	-0.00711	0.768246

STATS =

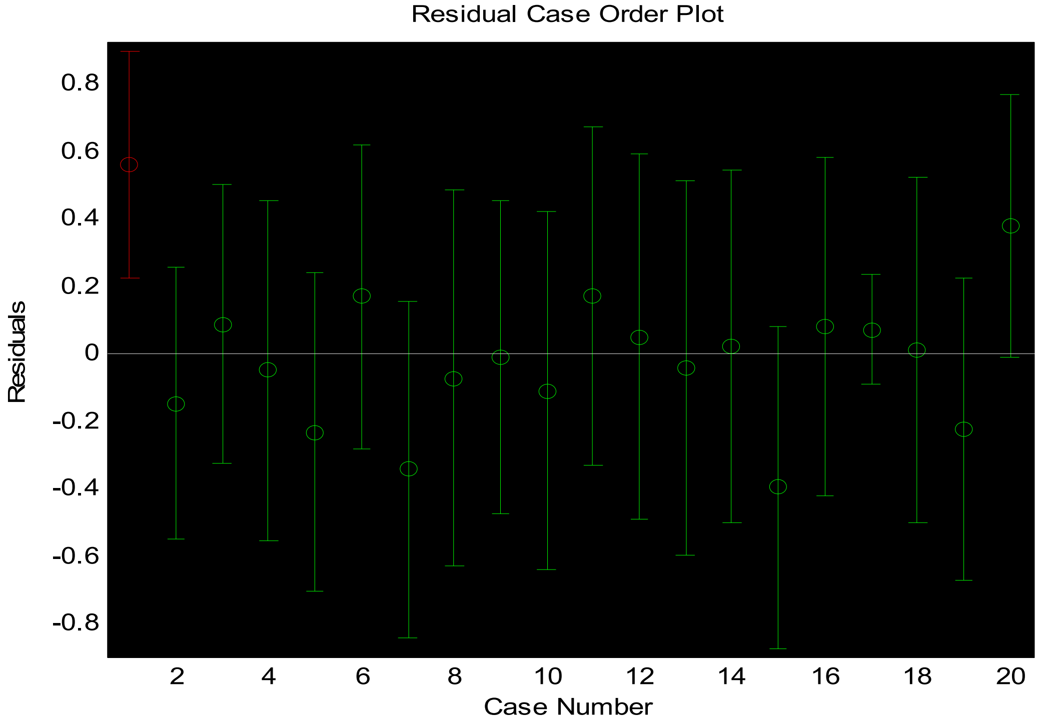
R^2	F	p-value	error variance
0.810665	11.98861	0.000118	0.068538

The y-intercept is b_0, which corresponds to the column index of the column of ones.

The elements of the vector stats are the regression R^2 statistic, the F statistic (for the hypothesis test that all the regression coefficients are zero), the p-value associated with this F statistic, and error variance

R^2 is 0.8107 indicating the model accounts for over 80% of the variability in the observations.

The F statistic of about 12 and its p-value of 0.0001 indicate that it is highly unlikely that all of the regression coefficients are zero.



The plot shows the residuals plotted in case order (by row). The 95% confidence intervals about these residuals are plotted as error bars. The first observation is an outlier since its error bar does not cross the zero reference line. [The program name: CONFIDENC]

1.17: Simulation of Stochastic processes

In this section, we will simulate and represent graphically various simple stochastic processes.

1.17.1: Simulation of Bernoulli process

A Bernoulli process is a discrete-time stochastic process consisting of finite or infinite sequence of independent random variables $x_1, x_2, x_3, \cdots$ such that,

$$x_i = \begin{cases} 1, & with \quad prop = p \\ -1, & with \quad prop = 1 - p \end{cases}$$

Random variables associated with the Bernoulli process include:

- The number of successes in the first n trials; this has a binomial distribution;
- The number of trials needed to get r successes; this has a negative binomial distribution.
- The number of trials needed to get one success; this has a geometric distribution, which is a special case of the negative binomial distribution.

We can simulate a realization of size 100 of a Bernoulli process with p = 0.5 as follows.

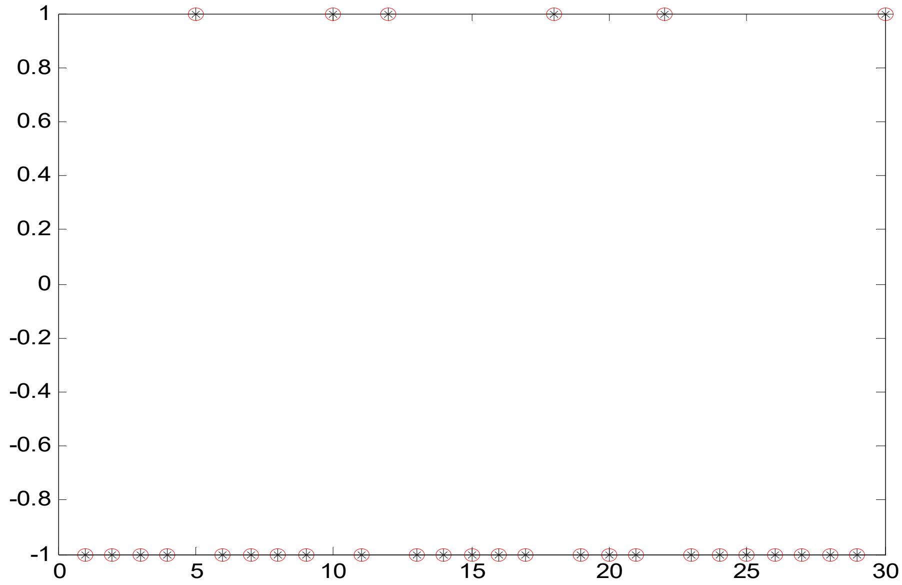
```
u=rand(10,1);
```

```
X=1-2*floor(u*2)
```

Where (floor) Round towards minus infinity,
floor(X) rounds the elements of X to the nearest integers towards minus infinity.

We can simulate another realization of a Bernoulli process with p = 0.25 and observe the differences.[The program name BERNOULLI.m]

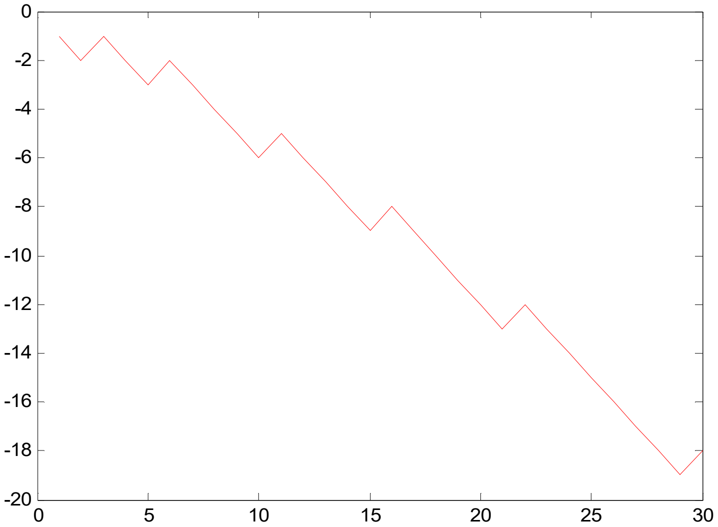
```
u=rand(30,1);
Y(u<0.25)=1;Y(u>0.25)=-1;
plot(1:30,Y,'ro',1:30,Y,'k*')
```



1.17.2: Simulation of Random walk

By using the **cumsum** command, we can simulate random walks from the Bernoulli processes simulated previously. [The program name RANDOMWALK.m].

```
u=rand(30,1);

Y(u<0.25)=1;Y(u>0.25)=-1;
plot(1:30,cumsum(Y),'r')
```



1.17.3: Simulation of Poisson process

Firstly, observe that continuous time processes are only possible to simulate by discretization of the unit time.

A Poisson process, x_t, with rate λ verifies the following property:

$$x_t = \text{Number of occurrences in } [0, t) \sim Po(\lambda t).$$

If we want simulate a realization with 10 occurrences from a Poisson process of rate λ = 2, we can first simulate 10 exponential times of mean 1/λ = 0.5 between occurrences. [The program name POISSONPROCES.m].

```
x=exprnd(0.5,1,10);
```

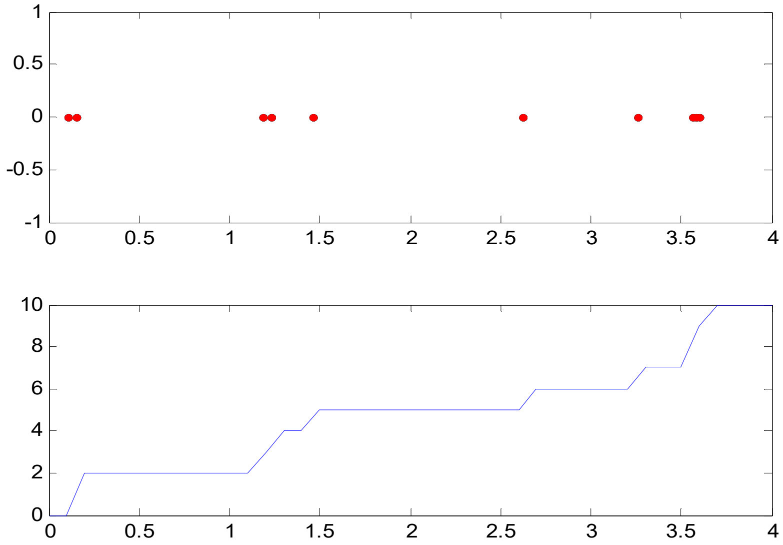
Then, we can obtain the occurrence times as follows.

```
x=cumsum(x);
subplot(2,1,1),plot(x,zeros(length(x)),'.')
```

Suppose we want to know the value of the process x_t at the following instant times:

Then, we can compute:

```
for i=1:length(t);X(i)=sum(x<t(i));end
subplot(2,1,2),plot(t,X)
```



1.17.4: Simulation of Autoregressive process

Suppose we want to simulate T = 100 values from an autoregressive model AR(1),

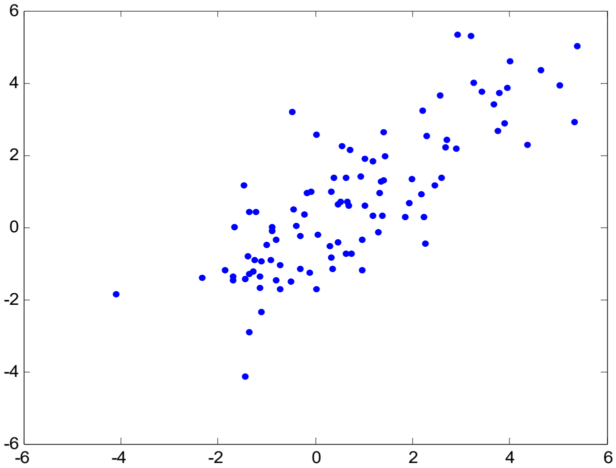
$$x_t = \alpha x_t + e_t$$

where e_t are i.i.d. N (0, 1) and assume three values for α □ {0.8, 0.5,−0.8}. One possibility is to assume x1 = e1 and then obtain recursively the remaining values. [The program name AR1.m].

```
e=randn(100,1);
x=zeros(100,1);
x(1)=e(1);
alpha=0.8;
for i=2:100, x(i)=alpha*x(i-1)+e(i); end
```

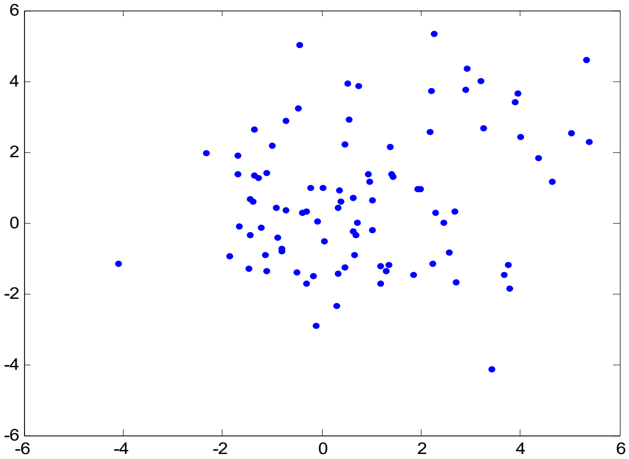
We can calculate the sample coefficient of the autocorrelation function. For example, the first coefficient is the sample correlation coefficient of x_{t-1} and x_t:

```
corrcoef(x(1:99),x(2:100));
plot(x(1:99),x(2:100),'.')
```



Observe that after 10 lags, there is almost no relation between of x_{t-1} and x_t:

```
plot(x(1:90),x(11:100),'.')
```



1.17.5: Simulation of Moving average process

Suppose now that we want to simulate T = 100 values from a moving average model MA(1),

$$x_t = \theta e_{t-1} + e_t$$

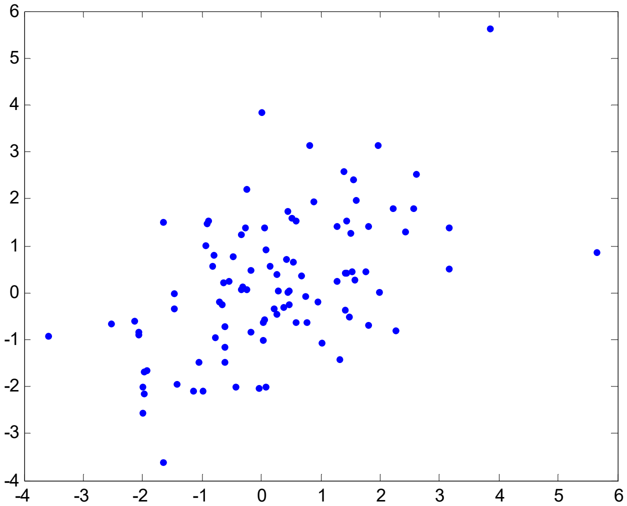
Where e_t are i.i.d. N (0, 1) and assume three values for θ □ {0.8, 0.5,−0.8}. [The program name MA1.m].

This process is easier to initialize because we just have to simulate e_0.

```
e=randn(101,1);
theta=0.8;
x=theta*e(1:100,1)+e(2:101,1);
```

Compute the first two coefficients of the autocorrelation function and observe the following plots:

```
plot(x(1:99),x(2:100),'.')
plot(x(1:98),x(3:100),'.');
```



1.18: Nonlinear Regression

When the relationship between the independent variable(s) and the dependent variable cannot be approximated as a line (or a hyperplane), approaches beyond linear regression are needed. There are many different methods for dealing with nonlinear relationships, but we will focus on two approaches: (a) Using a nonlinear transformation which makes the data approximately linear; (b) Polynomial fitting.

1.18.1: Nonlinear Transformations

Sometimes a non-linear relationship can be transformed into a linear one by a mathematical transformation. Examples include the exponential growth equation:

$$y = \mathrm{A}\, e^{bx} u \Leftrightarrow \log(y) = \log(\mathrm{A}) + bx + \log(u)$$

And the constant-elasticity equation

$$y = \mathrm{A}\,x^b u \Leftrightarrow \log(y) = \log(\mathrm{A}) + b.\log(x) + \log(u)$$

Linear regression can now be performed using the transformed variables.

Example 1.8: The table below shows data to test the relationship between porosity and sandstone strength.

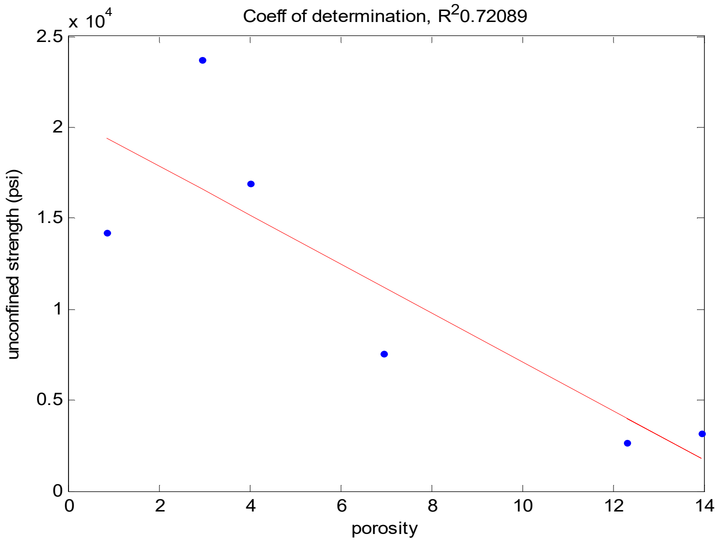
x=porosity	y=unconfined strength (psi)	Source: Hale, P. A. & Shakoor, A., 2003, A laboratory investigation of the Effects of Cyclic Heating and Cooling, Wetting and Drying, and Freezing and Thawing on the Compressive Strength of Selected Sandstones: Environmental and Engineering geoscience, vol IX, p. 117-130.
12.32	2636	
13.94	3162	
6.94	7580	
4.0	16899	
2.94	23739	
0.86	14224	

Plot the data and the regression line, and compute the coefficient of determination. [The program name example118.m].

```
x=[12.32,13.94,6.94,4,2.94,0.86];
y=[ 2636, 3162, 7580, 16899, 23739, 14224];
x=x(:); y=y(:);
n=length(x);
XX=[ones(n,1),x];
b=(XX'*XX)^-1*XX'*y
yhat=XX*b;
e=y-yhat;
my=mean(y);
R2=1-sum(e.^2)/sum((y-my).^2)
figure;
plot(x,y,'.')
hold on , plot(x,yhat,'r')
title(['Coeff of determination, R^2' ,num2str(R2)])
xlabel('porosity'), ylabel('unconfined strength (psi)')
MSE=sum(e.^2)/(n-2)
```

Ans.

```
b =                  R2 =                MSE =
     20560               0.72089               2.4403e+07
   -1344.4
```

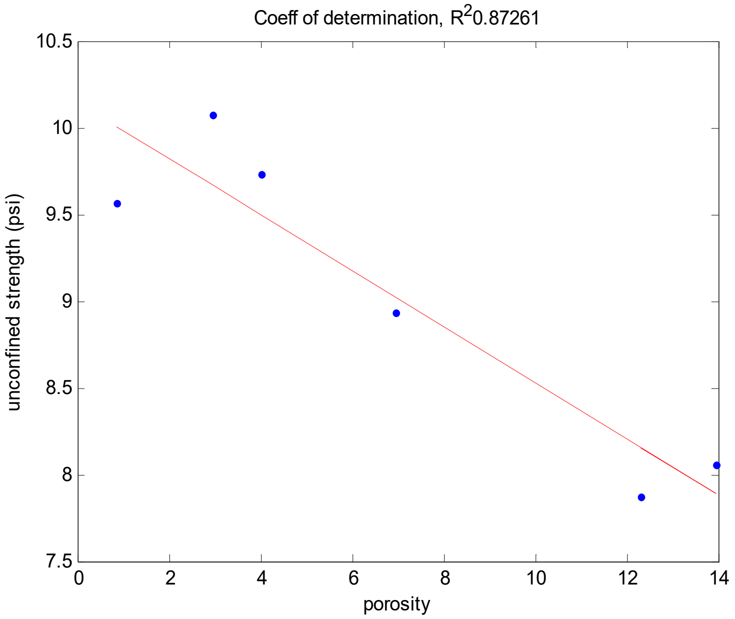


The coefficient of determination is $R^2 = 0.72$, indicating that the regression equation can explain 72% of the variation in unconfined strength. And MSE equals 2.4403e+07

Repeat the same analysis, using a nonlinear transformation: [The program name example118.m].

y=log(y)

```
b =                 R2 =                      MSE =
      10.142              0.87261                        0.13228
     -0.1612
```



The coefficient of determination has increased to $R^2 = 0.87$ and MSE has decreased to 0.13228

There are a few points to keep in mind when using this method. First, we are assuming that the errors in the transformed equation follow a zero-mean Gaussian distribution, which may not be a reasonable assumption. Second, once we get the estimates from the transformed equation, going back to the original equation can be tricky. Some parameter estimates are biased, and the confidence intervals are no longer symmetrical around the predicted values. We need to get the confidence interval from the transformed equation and then transform the bounds back.

1.18.2: Polynomial fitting

The commands **polyfit** and **polyval** can be used whenever the data can be approximated by a polynomial.

1- **polyfit** Fit polynomial to data.

P = polyfit(X,Y,N) finds the coefficients of a polynomial P(X) of degree N that fits the data Y best in a least-squares sense. P is a row vector of length N+1 containing the polynomial coefficients in descending powers,

P(1)*X^N + P(2)*X^(N-1) +...+ P(N)*X + P(N+1).

[P,S] = polyfit(X,Y,N) returns the polynomial coefficients P and a structure S for use with POLYVAL to obtain error estimates for predictions. S contains fields for the triangular factor (R) from a QR decomposition of the Vandermonde matrix of X, the degrees of freedom (df), and the norm of the residuals (normr). If the data Y are random, an estimate of the covariance matrix of P is (Rinv*Rinv')*normr^2/df, where Rinv is the inverse of R.

[P,S,MU] = polyfit(X,Y,N) finds the coefficients of a polynomial in

XHAT = (X-MU(1))/MU(2) where MU(1) = MEAN(X) and MU(2) = STD(X). This centering and scaling transformation improves the numerical properties of both the polynomial and the fitting algorithm.

Warning messages result if N is >= length(X), if X has repeated, or nearly repeated, points, or if X might need centering and scaling.
Class support for inputs X,Y: float: double, single

2- **polyval** Evaluate polynomial.

Y = polyval(P,X) returns the value of a polynomial P evaluated at X. P is a vector of length N+1 whose elements are the coefficients of the polynomial in descending powers.

Y = P(1)*X^N + P(2)*X^(N-1) + ... + P(N)*X + P(N+1)

If X is a matrix or vector, the polynomial is evaluated at all points in X. See POLYVALM for evaluation in a matrix sense.

[Y,DELTA] = polyval(P,X,S) uses the optional output structure S created by POLYFIT to generate prediction error estimates DELTA. DELTA is an estimate of the standard deviation of the error in predicting a future observation at X by P(X).

If the coefficients in P are least squares estimates computed by POLYFIT, and the errors in the data input to POLYFIT are independent, normal, with constant variance, then Y +/- DELTA will contain at least 50% of future observations at X.

Y = polyval(P,X,[],MU) or [Y,DELTA] = polyval(P,X,S,MU) uses XHAT = (X-MU(1))/MU(2) in place of X. The centering and scaling parameters MU are optional output computed by POLYFIT.

Consider the following nonlinear system:

```
randn('seed', 1);
x=(1:50)';
y = sin(x/50)./ x + 0.002 * randn(50,1)
```

Fit a polynomial of order 5:

```
order=5;
poly = polyfit(x, y, order);
```

Evaluate the polynomial at the data points:

```
yhat= polyval(poly,x)
```

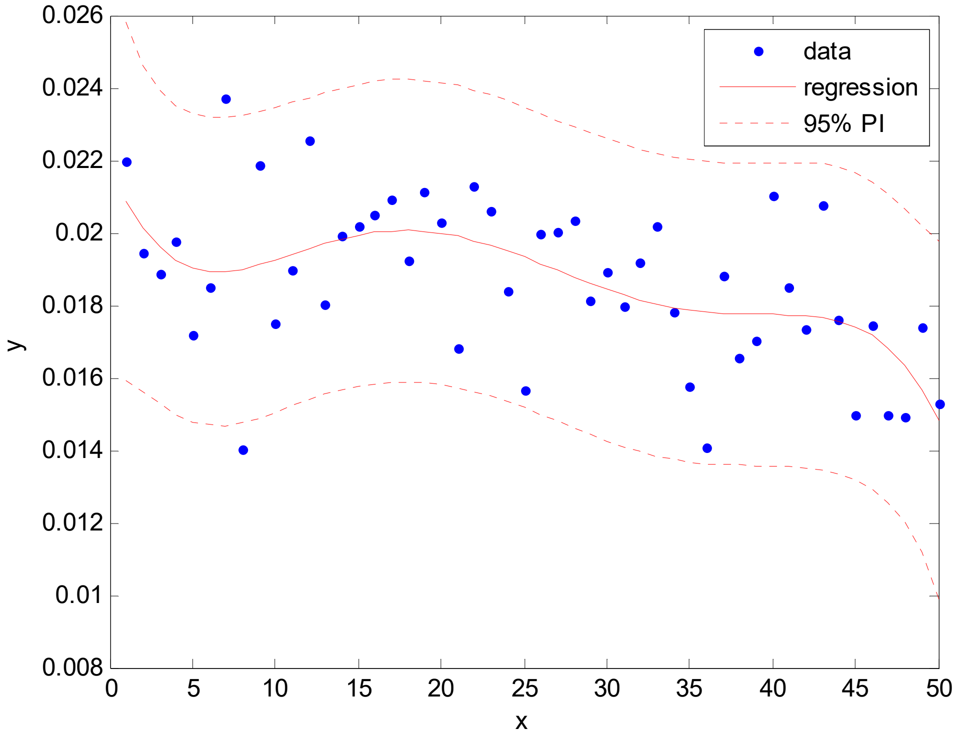
An approximate 95% prediction interval for y (including the noise) can be constructed as follows: [The program name NONLINEAR.m].

```
randn('seed', 1);
x=(1:50)'; y = sin(x/50)./ x + 0.002 * randn(50,1); n=length(x);
order=5; poly = polyfit(x, y, order); yhat= polyval(poly,x)
```

```
[poly model] = polyfit(x, y, order); % fit a polynomial
[yhat s] = polyval(poly, x, model); % evaluate the polynomial
alpha=0.05; % for 95% confidence
p=1-alpha/2; % probability to be used in CDF
df=50-(5+1); % degrees of freedom
t=tinv(p,df); % t-value, may need tinv558
PI_lower=yhat-t*s; PI_upper=yhat+t*s;
figure;
plot(x,y,'.')
hold on
plot(x,yhat, 'r')
plot(x, PI_lower, 'r:')
plot(x, PI_upper, 'r:')
legend('data','regression','95% PI')
xlabel('x'), ylabel('y')
my=mean(y); e=y-yhat;
MSE=sum(e.^2)/(n-2); R1=1-sum(e.^2)/sum((y-my).^2)
```

Ans.

MSE = 0.30322, R2 = 3.616e-06



PROBLEMS

1.1: Define MATLAB

1.2: What is interest MATLAB?

1.3: where the name came from MATLAB?

1.4: What MATLAB language characterized for other programming languages?

1.5: What magic matrix and how do we get them?

1.6: In analyzing linear equations if you know that:

$$A = \begin{bmatrix} 9 & 4 & 1 \\ 8 & 5 & 2 \\ 6 & 3 & 4 \end{bmatrix}$$

Find the following:

1- The inverse of the matrix.
2- Cholesky factorization.
3- Upper and lower trigonometric matrix.
4- Pseudoinverse matrix.

1.7: In the analysis of the Eigenvalues if you know that:

$$B = \begin{bmatrix} 3 & 4 & 1 \\ 5 & 7 & 8 \\ 1 & 2 & 1 \end{bmatrix}$$

A- Eigen values and Eigen vector.

B- Singular value decomposition.

1.8: Analysis functions of matrices if you know that:

$$C = \begin{bmatrix} 5 & 1 & 2 & 4 \\ 6 & 2 & 5 & 1 \\ 4 & 3 & 1 & 5 \\ 8 & 9 & 3 & 2 \end{bmatrix}$$

Find the following:

1- Matrix exponential
2- Matrix logarithm
3- Matrix square root

1.9: Explain the command Kronecker with a practical example?

1.10: Solving linear systems following:

$$A*X=B$$

If you know that A represents Pascal matrix (Dim.3) and

$$B=\begin{bmatrix}3 & 1 & 4\end{bmatrix}^T$$

1.11: Estimate and draw the negative exponential model using (OLS) method for the following data:

t = [0 .3 .8 1.1 1.6 2.3]' and y = [.82 .72 .63 .60 .55 .50]'

Where $y(t) = c_1 + c_2 \mathrm{e}^{-t}$

1.12: Estimate the Simple Linear Model using method (OLS) for the following data:

y	2	3	5	7	8	10	12	15
x	8	10	14	16	17	20	22	26

Where $y_i = c_1 + c_2 x_i$

Find the following:

1- Average of D.V.

2- Variance of I.V.

3- Standard Deviation of the D.V.

4- Simple Linear Correlation Coefficient.

5- Mean Square Error.

6- The Coefficient of Determination.

7- Standard Error.

8- Covariance between the I.V. and D.V.

1.13: Draw the scatter plot of the following data:

$$z=[0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7]$$

$$x=[3 \quad 5 \quad 7 \quad 9 \quad 11 \quad 13 \quad 15 \quad 17]$$

$$y=[1 \quad 6 \quad 8 \quad 12 \quad 16 \quad 20 \quad 24 \quad 29]$$

1.14: Select outlier values for the following data:

$$z=[0 \quad 1 \quad 2 \quad 3 \quad 20 \quad 5 \quad 6 \quad 7]$$

$$x=[3 \quad 5 \quad 7 \quad 9 \quad 11 \quad 13 \quad 15 \quad 17]$$

$$y=[1 \quad 6 \quad 8 \quad 12 \quad 16 \quad 20 \quad 24 \quad 29]$$

1.15: Estimate the Multiple Linear Model using method (OLS) for the following data:

Where $y_i = c_1 + c_2 x_i + c_3 z_i$

The required account the following:

1 - Average of D.V. 2 - Variance of x. 3 - Mean Square Error.
4 - Standard Error. 5 - Covariance between the variables.

1.16: Write a computer program to implement for generating a F-distribution with (8) & (11) degrees of freedom respectively, for n = 30

1.17: Write a computer program to implement for generating a Exp(6) random deviate, n = 20

1.18: Compute possible some cases Normal output matrix of random matrix generated from Uniform distribution $^{(3\times2)}$ multiplied by 10 for just the integer values.

1.19: Write a computer program to implement for generating a *t*-distribution with (20) degree of freedom, for n = 25 by using Direct Method.

1.20: Write a computer program to implement for generating a multivariate normal distribution for (k=4) variables, n=30 and:

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 & 1 \\ & 2 & 3 & 4 \\ & & 6 & 10 \\ & & & 20 \end{bmatrix}$$

For means equal to [2 15 6 12], find mean, variance and correlation matrix.

1.21: Write a computer program to implement for generating:

- Poisson(5) random deviate, n = 20
- Exp(2) random deviate, n = 10

Chapter two

Simple Linear Regression

2.1: The model

The simple linear regression model for *n* observations can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i , \quad 1, 2, \ldots, \mathrm{n} \qquad (2.1)$$

The designation *simple* indicates that there is only one *x* to predict the response *y*, and *linear* means that the model (2.1) is linear in β_0 and β_1. [Actually, it is the assumption $E(y) = \beta_0 + \beta_1 x$ that is linear; see assumption 1 below.] For example, a model such as $y_i = \beta_0 + e^{\beta_1 x_i} + \varepsilon_i$ is linear in β_0 and β_1, whereas the model is not linear.

In this chapter, we assume that y_i and ε_i are random variables and that the values of x_i are known constants, which means that the same values of $x_1, x_2, \ldots, x_n$ would be used in repeated sampling.

To complete the model in (2.1), we make the following additional assumptions:

1- $E(\varepsilon_i)$ for all *i* = 1, 2, …, *n*, or, equivalently, $E(y_i) = \beta_0 + \beta_1 x_i$

2- $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all *i* = 1, 2, … , *n*, or, equivalently, $\mathrm{Var}(y_i) = \sigma^2$

3- $\mathrm{Cov}(\varepsilon_i , \varepsilon_i) = 0$ for all $i \neq j$, or, equivalently, $\mathrm{Cov}(y_i , y_j) = 0$

Assumption 1 states that the model (2.1) is correct, implying that y*ᵢ* depends only on x_i and that all other variation in y_i is random. Assumption 2 asserts that the variance of ε or *y* does not depend on the values of x_i. (Assumption 2 is also known as the assumption of *homoscedasticity*, *homogeneous variance* or *constant variance*.) Under assumption 3, the ε variables (or the y variables) are uncorrelated with each other. In Section 2.3, we will add a normality assumption, and the *y* (or the ε) variables will thereby be independent as well as uncorrelated. Each assumption has been stated in terms of the y. For example, if $\mathrm{Var}(\varepsilon_i) = \sigma^2$ then $\mathrm{Var}(y_i) = E[y_i - E(y_i)]^2 = E(y_i - \beta_0 + \beta_1 x_i)^2 = E(\varepsilon_i^2) = \sigma^2$

Any of these assumptions may fail to hold with real data. A plot of the data will often reveal departures from assumptions 1 and 2 (and to a lesser extent assumption 3). Techniques for checking on the assumptions are discussed in Chapter 5.

2.2: Estimation of β_0, β_1 and σ^2

Using a random sample of *n* observations *y*1, *y*2, .. . , *yn* and the accompanying fixed values *x*1, *x*2, .. . , *xn*, we can estimate the parameters β_0, β_1 and σ^2. To obtain the estimates $\hat{\beta}_0$ *and* $\hat{\beta}_1$, we use the method of least squares, which does not require any distributional assumptions (for maximum likelihood estimators based on normality, see Section 3.6.2).

In the least-squares approach, we seek estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squares of the deviations $y_i - \hat{y}_i$ of the n observed y_i's from their predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$\hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_o - \hat{\beta}_1 x_i)^2 \tag{2.2}$$

Note that the predicted value $\hat{y}_i$ estimates $E(y_i)$, not y_i; that is, $\hat{\beta}_0 + \hat{\beta}_1 x_i$ estimate $\beta_0 + \beta_1 x_i$ not $\beta_0 + \beta_1 x_i + \varepsilon_i$. A better notation would be $E(\hat{y}_i)$ but $\hat{y}_i$ is commonly used.

To find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $\hat{\varepsilon}'\hat{\varepsilon}$ in (2.2), we differentiate with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set the results equal to 0:

$$\frac{\partial \hat{\varepsilon}'\hat{\varepsilon}}{\partial \hat{\beta}_o} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_o - \hat{\beta}_1 x_i)0 = 0 \tag{2.3}$$

$$\frac{\partial \hat{\varepsilon}'\hat{\varepsilon}}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_o - \hat{\beta}_1 x_i)x_i = 0 \tag{2.4}$$

The solution to (2.3) and (2.4) is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.5}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.6}$$

To verify that $\hat{\beta}_0$ and $\hat{\beta}_1$ in (2.5) and (2.6) minimize $\hat{\varepsilon}'\hat{\varepsilon}$ in (2.2), we can examine the second derivatives or simply observe that $\hat{\varepsilon}'\hat{\varepsilon}$ has no maximum and therefore the first derivatives yield a minimum. For an algebraic proof that $\hat{\beta}_0$ and $\hat{\beta}_1$minimize (2.2).

Example 2.1: Students in a statistics class (taught by one of the authors) claimed that doing the homework had not helped prepare them for the midterm exam. The exam score y and homework score x (averaged up to the time of the midterm) for the 18 students in the class were as follows:

y	x	y	x	y	x
95	96	72	89	35	0
80	77	66	47	50	30
0	0	98	90	72	59
0	0	90	93	55	77
79	78	0	18	75	74
77	64	95	86	66	67

Using (2.5) and (2.6), we obtain

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{81195 - 18(58.056)(61.389)}{80199 - 18(58.056)^2} = 0.8726$$

$$\hat{\beta}_0 = 61.389 - 0.8726(58.056) = 10.73$$

The prediction equation is thus given by

$$\hat{y}_i = 10.73 + 0.8726 x_i$$

This equation and the 18 points are plotted in Figure 2.1. It is readily apparent in the plot that the slope $\hat{\beta}_1$ is the rate of change of $\hat{y}$ as x varies and that the intercept $\hat{\beta}_0$ is the value of $\hat{y}$ at x = 0.

The apparent linear trend in Figure 2.1 does not establish cause and effect between homework and test results (for inferences that can be drawn, see Section 2.3). The assumption $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (constant variance) for all i = 1, 2, . . . , 18 appears to be reasonable.

Note that the three assumptions in Section 2.1 were not used in deriving the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in (2.5) and (2.6). It is not necessary that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be based on $E(y_i) = \beta_0 + \beta_1 x_i$; that is, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ can be fit to a set of data for which $E(y_i) \neq \beta_0 + \beta_1 x_i$. This is illustrated in Figure 2.2, where a straight line has been fitted to curve data.

Example 2.1[The program name ta1.m]	Applications using MATLAB

```
clc
y=[95 80 0 0 79,77 72 66 98 90 0 95 35 50 72 55 75 66]';
x=[96 77 0 0 78 64 89 47 90 93 18 86 0 30 59 77 74 67]';
n=length(x);
E=[ones(size(x)) x];
beta=E\y
X=(0:1:100)';
Yhat=[ones(size(X)) X]*beta;
plot(x,y,'ko',X,Yhat,'-')
xlabel('x'), ylabel('y')
legend('data','regression line')
title(['Regression line: Yhat=',num2str(beta(1)),'+',num2str(beta(2)),'*x'])
```

Ans.

```
beta =  10.727
        0.87265
```

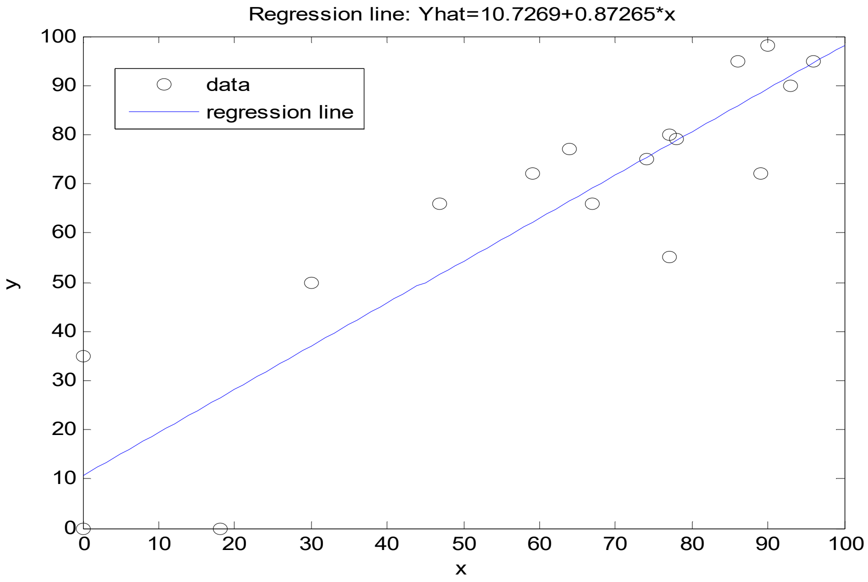


Figure 2.1: Regression line and data for homework and test scores

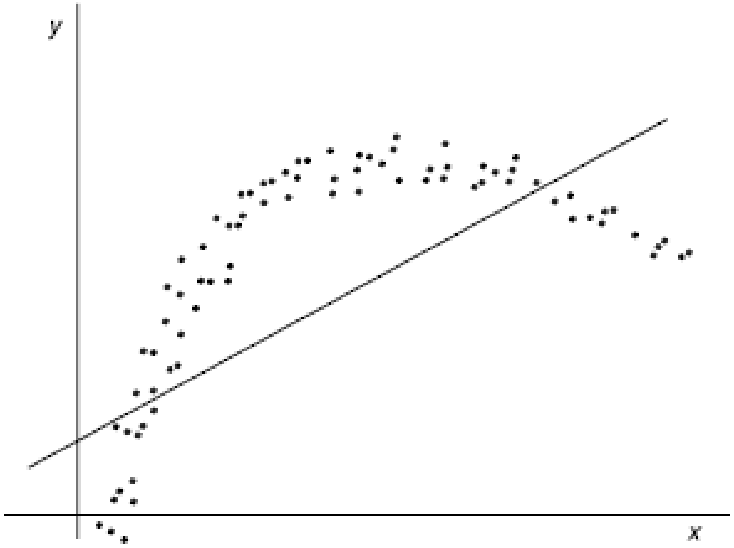


Figure 2.2: A straight line fitted to data with a curved trend

However, if the three assumptions in Section 2.1 hold, then the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all linear unbiased estimators (for the minimum variance property, see Theorem 3.3d in Section 3.3.2; note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of $y_1, y_2, \cdots, y_n$). Using the three assumptions, we obtain the following means and variances of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$E\left(\hat{\beta}_1\right) = \beta_1 \tag{2.7}$$

$$E\left(\hat{\beta}_0\right) = \beta_0 \tag{2.8}$$

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.9}$$

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \tag{2.10}$$

Note that in discussing $E\left(\hat{\beta}_1\right)$ and $\mathrm{Var}\left(\hat{\beta}_1\right)$, for example, we are considering random variation of $\hat{\beta}_1$ from sample to sample. It is assumed that the *n* values $x_1, x_2, \cdots, x_n$ would remain the same in future samples so that $\mathrm{Var}\left(\hat{\beta}_1\right)$ and $\mathrm{Var}\left(\hat{\beta}_0\right)$ are constant.

In (2.9), we see that $\mathrm{Var}\left(\hat{\beta}_1\right)$ is minimized when $\sum_{i=1}^{n} (x_i - \bar{x})^2$ is maximized. If the x_i values have the range $a \le x_i \le b$, then $\sum_{i=1}^{n} (x_i - \bar{x})^2$ is maximized if half the x's are selected equal to *a* and half equal to *b* (assuming that *n* is even; see Problem 2.4). In (2.10), it is clear that $\mathrm{Var}\left(\hat{\beta}_0\right)$ is minimized when $\bar{x} = 0$.

The method of least squares does not yield an estimator of $\mathrm{Var}(y_i) = \sigma^2$, minimization of $\hat{\varepsilon}'\hat{\varepsilon}$ yields only $\hat{\beta}_0$ and $\hat{\beta}_1$. To estimate σ^2, we use the definition $\sigma^2 = E[y_i - E(y_i)]^2$. By assumption 2 in Section 2.1, σ^2 is the same for each y_i, *i* = 1, 2, . . . , *n*. Using $\hat{y}_i$ as an estimator of $E(y_i)$, we estimate σ^2 by an average from the sample, that is

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2} = \frac{SSE}{n-2} \qquad (2.11)$$

Where $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by (2.5) and (2.6) and $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The deviation $\hat{\varepsilon}_i = (y_i - \hat{y}_i)^2$ is often called the residual of y_i, and SSE is called the residual sum of squares or error sum of squares. With $n-2$ in the denominator, S^2 is an unbiased estimator of σ^2:

$$E\left(S^2\right) = \frac{E(SSE)}{n-2} = \frac{(n-2)\sigma^2}{(n-2)} = \sigma^2 \qquad (2.12)$$

Intuitively, we divide by $(n-2)$ in (2.11) instead of $(n-1)$ as in $S^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 / (n-1)$, because $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ has two estimated parameters and should there by be a better estimator of $E(y_i)$ than $\bar{y}$. Thus we expect $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ to be less than $\sum_{i=1}^{n}(y_i - \bar{y})^2$. In fact, using (2.5) and (2.6), we can write the numerator of (2.11) in the form

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2.13)$$

Which shows that $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is indeed smaller than $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

2.3: Hypothesis Test and Confidence Interval for β_1

Typically, hypotheses about β_1 are of more interest than hypotheses about β_0, since our first priority is to determine whether there is a linear relationship between y and x. (See Problem 2.9 for a test and confidence interval for β_0) In this section, we consider the hypothesis $H_0 : \beta_1 = 0$, which states that there is no linear relationship between y and x in the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The hypothesis $H_0 : \beta_1 = \mathrm{c}$ (for c= 0) is of less interest.

In order to obtain a test for $H_0 : \beta_1 = 0$, we assume that y_i is $N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$. Then $\hat{\beta}_1$ and S^2 have the following properties (these are special cases of results established in Theorem 2.6b in Section 2.6.3):

1. $\hat{\beta}_1$ is $N[\beta_1, \sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2]$.

2. $(n-2)S^2/\sigma^2$ is $\chi^2_{(n-2)}$.

3. $\hat{\beta}_1$ and S^2 are independent.

From these three properties it follows:

$$t = \frac{\hat{\beta}_1}{S \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{2.14}$$

Is distributed as $t(n-2,\delta)$, the non-central t with non-centrality parameter δ. And δ is given by

$$\delta = E(\hat{\beta}_1) \Big/ \sqrt{\mathrm{Var}(\hat{\beta}_1)} = \beta_1 \Big/ \left[\sigma^2 \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].$$

If $\beta_1 = 0$, t is distributed as $t(n-2)$. For a two-sided alternative hypothesis $H_1 : \beta_1 \neq 0$, we reject $H_0 : \beta_1 = 0$ if $|t| \geq t_{\alpha/2,n-2}$, where $t_{\alpha/2,n-2}$, is the upper $\alpha/2$ percentage point of the central *t* distribution and α is the desired significance level of the test (probability of rejecting H_0 when it is true). Alternatively, we reject H_0 if $p \leq \alpha$, where *p* is the *p* value. For a two sided test, the *p* value is defined as twice the probability that $t(n-2)$ exceeds the absolute value of the observed *t*.

A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{2.15}$$

Confidence intervals are defined and discussed further in Section 4.6. A confidence interval for E(y) and a prediction interval for y are also given in Section 4.6.

Example 2.2: We test the hypothesis $H_0 : \beta_1 = 0$ for the grades data in Example 2.1. By (2.14), the t statistic is

$$t = \frac{\hat{\beta}_1}{S \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{0.8726}{(13.8547)/(139.753)} = 8.8025$$

Since $t = 8.8025 > t_{0.025,16} = 2.120$, we reject $H_0 : \beta_1 = 0$ at the $\alpha = 0.05$ level of significance. Alternatively, the p value is 1.571×10^{-7}, which is less than 0.05.

A 95% confidence interval for β_1 is given by (2.15) as

$$\hat{\beta}_1 \pm t_{0.025,16} \frac{S}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$0.8726 \pm 2.120(0.09914)$$

$$(0.6624 \quad , \quad 1.0828)$$

Example 2.2[The program name ta2.m]	Applications using MATLAB

```
clc
y=[95 80 0 0 79,77 72 66 98 90 0 95 35 50 72 55 75 66]';
x=[96 77 0 0 78 64 89 47 90 93 18 86 0 30 59 77 74 67]';
n=length(x);E=[ones(size(x)) x];
beta=E\y
Yhat=E*beta;e=y-Yhat;
MSE=e'*e/(n-2),S=sqrt(MSE)
Sxx=sum((x-mean(x)).^2)
t=beta(2)/(S/sqrt(Sxx))
beta1l=beta(2)-2.12*S/sqrt(Sxx);
beta1u=beta(2)+2.12*S/sqrt(Sxx);
beta1LUL=[beta1l beta1u]
```

Ans.

```
beta =          MSE =         S =           Sxx =         t =

    10.727
                  191.95        13.855        19531         8.8025
   0.87265

beta1LUL =

              0.66248       1.0828
```

2.4: Coefficient of Determination

The coefficient of determination r^2 is defined as

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (2.16)$$

Where $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ is the regression sum of squares and $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares. The total sum of squares can be partitioned into SST = SSR+ SSE, that is,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2.17)$$

Thus r^2 in (2.16) gives the proportion of variation in y that is explained by the model or, equivalently, accounted for by regression on x.

We have labeled (2.16) as r^2 because it is the same as the square of the sample correlation coefficient *r* between y and x.

$$r = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right]}} \qquad (2.18)$$

Where S_{xy} is $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ (see Problem 2.11). When x is a random variable, *r* estimates the population correlation, ($\rho_{ij} = Corr(y_i, y_j) = \sigma_{ij} / \sigma_i \sigma_j$). The coefficient of determination r^2 is discussed further in Section 3.7.

Example 2.3: For the grades data of Example 2.2, we have

$$r^2 = \frac{SSR}{SST} = \frac{14873}{17944.3} = 0.8288$$

The correlation between homework score and exam score is $r = \sqrt{0.8288} = 0.910$.

The *t* statistic in (2.14) can be expressed in terms of *r* as follows:

$$t = \frac{\hat{\beta}_1}{S \Big/ \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{2.19}$$

$$= \frac{\sqrt{n-2}.r}{\sqrt{1-r^2}} \tag{2.20}$$

If $H_0 : \beta_1 = 0$ is true, then, as noted following (2.14), the statistic in (2.19) is distributed as $t(n-2)$ under the assumption that the x_i's are fixed and the y_i's are independently distributed as $N(\beta_0 + \beta_1 x_i, \sigma^2)$. If x is a random variable such that x and y have a bivariate normal distribution, then $t = \sqrt{n-2}.r \big/ \sqrt{1-r^2}$ in (2.20) also has the $t(n-2)$ distribution provided that $H_0 : \rho = 0$ is true, where *r* is the population correlation coefficient defined as ($\rho_{ij} = Corr(y_i, y_j) = \sigma_{ij} / \sigma_i \sigma_j$). However, (2.19) and (2.20) have different distributions if $H_0 : \beta_1 = 0$ and $H_0 : \rho = 0$ are false. If $\beta_1 \neq 0$, then (2.19) has a non-central *t* distribution, but if $\rho \neq 0$, (2.20) does not have a non-central *t* distribution.

Example 2.3[The program name ta3.m]	Applications using MATLAB

```
clc
y=[95 80 0 0 79,77 72 66 98 90 0 95 35 50 72 55 75 66]';
x=[96 77 0 0 78 64 89 47 90 93 18 86 0 30 59 77 74 67]';
n=length(x);E=[ones(size(x)) x];beta=E\y;
Yhat=E*beta;e=y-Yhat;MSE=e'*e/(n-2);S=sqrt(MSE);
Sxx=sum((x-mean(x)).^2);t=beta(2)/(S/sqrt(Sxx))
p = 1-tcdf(t,n-2),beta1l=beta(2)-2.12*S/sqrt(Sxx);
r=corr(y,x);R=r^2,SSE=(n-2)*MSE,SST=(n-1)*var(y),
SSR=SST-SSE,R1=SSR/SST;
tr=sqrt(n-2)*r/sqrt(1-r^2)% test r
```

Ans.

```
t =              p =              R =              SSE =            SST =

    8.8025        7.8534e-08        0.82885          3071.2            17944
```

```
SSR =            tr =

   14873             8.8025
```

PROBLEMS

2.1: Obtain the L-S solutions (2.5) and (2.6) from (2.3) and (2.4).

2.2: (a) Show that $E(\hat{\beta}_1) = \beta_1$ as in (2.7).
(b) Show that $E(\hat{\beta}_0) = \beta_0$ as in (2.8).

2.3: (a) Show that $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \big/ \sum_{i=1}^{n} (x_i - \bar{x})^2$ as in (2.9).
(b) Show that $\mathrm{Var}(\hat{\beta}_0) = \sigma^2 [1/n + \bar{x}^2 \big/ \sum_{i=1}^{n} (x_i - \bar{x})^2]$ as in (2.10).

2.4: Suppose that *n* is even and the *n* values of x_i can be selected anywhere in the interval from *a* to *b*. Show that $\mathrm{Var}(\hat{\beta}_1)$ is a minimum if *n/2* values of x_i are equal to *a* and *n/2* values are equal to *b*.

2.5: Show that $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ in (6.11) can be expressed in the form given in (2.13).

2.6: Show that $E(S^2) = \sigma^2$ as in (2.12).

2.7: Show that $t = \hat{\beta}_1 \big/ [S \big/ \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}]$ in (2.14) is distributed as $t(n-2, \delta)$, where $\delta = \hat{\beta}_1 \big/ [\sigma \big/ \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}]$.

2.8: Obtain a test for $H_0 : \beta_1 = \mathrm{c}$ versus $H_1 : \beta_1 \neq \mathrm{c}$.

2.9: (a) Obtain a test for $H_0 : \beta_0 = \mathrm{a}$ versus $H_0 : \beta_0 \neq \mathrm{a}$.
(b) Obtain a confidence interval for β_0.

2.10: Show that $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ as in (2.17).

2.11: Show that r^2 in (2.16) is the square of the correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

As given by (2.18).

Table 2.1: Eruptions of Old Faithful Geyser, August 1–4, 1978a

y	x	y	x	y	x	y	x	y	x
80	3.5	42	1.8	88	4.7	79	3.7	75	4.0
84	4.1	91	4.1	51	1.8	60	3.8	73	3.7
50	2.3	51	1.8	80	4.6	86	3.4	67	3.7
93	4.7	79	3.2	49	1.9	76	4.5	68	4.3
55	1.7	53	1.9	82	3.5	82	3.9	86	3.6
76	4.9	82	4.6	80	4.3	84	4.3	72	3.8
58	1.7	51	2.0	56	1.7	53	2.3	75	3.8
74	4.6	78	4.4	80	3.9	86	3.8	75	3.8
75	3.4	74	3.9	69	3.7	51	1.9	66	2.5
		68	4.0	57	3.1	85	4.6	84	4.5
		76	4.0	90	4.0	45	1.8	70	4.1

Where x = duration, y = interval (both in minutes).

2.12: Show that $r = \cos\theta$, where θ is the angle between the vectors x−$\bar{x}$j and y−$\bar{x}$j where x− $\bar{x}$ j$=(x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x})'$ and y− $\bar{x}$ j$=(y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_n - \bar{y})'$.

2.13: Show that $t = \hat{\beta}_1 \big/ \left[s \big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$ in (2.19) is equal to $\sqrt{n-2}.r \big/ \sqrt{1-r^2}$ in (2.20).

2.14: Table 2.1 (Weisberg 1985, p. 231) gives the data on daytime eruptions of Old Faithful Geyser in Yellowstone National Park during August 1–4, 1978. The variables are x = duration of an eruption and y = interval to the next eruption. Can x be used to successfully predict y using a simple linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$?

(a) Find $\hat{\beta}_0$ and $\hat{\beta}_1$. (b) Test $H_0 : \beta_1 = 0$ using (2.14).

(c) Find a confidence interval for β_1. (d) Find r^2 using (2.16).

Chapter three

Multiple Regression: Estimation

3.1: Introduction

In multiple regression, we attempt to predict a dependent or response variable y on the basis of an assumed linear relationship with several independent or predictor variables $x_1, x_2, \cdots, x_k$. In addition to constructing a model for prediction, we may wish to assess the extent of the relationship between y and the x variables. For this purpose, we use the multiple correlation coefficient R.

In this chapter, y is a continuous random variable and the x variables are fixed constants (either discrete or continuous) that are controlled by the experimenter.

Useful applied expositions of multiple regression for the fixed-x case can be found in Morrison (1983), Myers (1990), Montgomery and Peck (1992), Graybill and Iyer (1994), Mendenhall and Sincich (1996), Ryan (1997), Draper and Smith (1998), and Kutner et al. (2005). Theoretical treatments are given by Searle (1971), Graybill (1976), Guttman (1982), Kshirsagar (1983), Myers and Milton (1991), Jørgensen (1993), Wang and Chow (1994), Christensen (1996), Seber and Lee (2003), and Hocking (1976, 1985, 2003).

3.2: The Model

The multiple linear regression model, can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{3.1}$$

We discuss estimation of the β parameters when the model is linear in the β's. An example of a model that is linear in the β's but not the x's is the second-order response surface model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon \tag{3.2}$$

To estimate the β's in (3.1), we will use a sample of n observations on y and the associated x variables. The model for the *i*th observation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i\,, \quad i = 1,2,\cdots,n \tag{3.3}$$

The assumptions for ε_i or y_i are essentially the same as those for simple linear regression in Section 2.1:

1. $E(\varepsilon_i) = 0$ for $i = 1,2,\cdots,n$, or, equivalently,

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

2. $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for $i = 1,2,\cdots,n$, or, equivalently, $\mathrm{Var}(y_i) = \sigma^2$.
3. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, or, equivalently, $\mathrm{Cov}(y_i, y_j) = 0$.

Assumption 1 states that the model is correct, in other words that all relevant x's are included and the model is indeed linear. Assumption 2 asserts that the variance of y is constant and therefore does not depend on the x's. Assumption 3 states that the y's are uncorrelated with each other, which usually holds in a random sample (the observations would typically be correlated in a time series or when repeated measurements are made on a single plant or animal). Later we will add a normality assumption (Section 3.6), under which the y variable will be independent as well as uncorrelated.

When all three assumptions hold, the least-squares estimators of the β's have some good properties (Section 3.3.2). If one or more assumptions do not hold, the estimators may be poor. Under the normality assumption (Section 3.6), the maximum likelihood estimators have excellent properties.

Any of the three assumptions may fail to hold with real data. Several procedures have been devised for checking the assumptions. These diagnostic techniques are discussed in Chapter 5.

Writing (3.3) for each of the n observations, we have

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{aligned}$$

These n equations can be written in matrix form as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{(n\times 1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}_{(n\times k+1)} . \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{(k+1\times 1)} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{(n\times 1)}$$

Or

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon \qquad (3.4)$$

The preceding three assumptions on ε_i or y_i can be expressed in terms of the model in (3.4):

1. $E(\mathrm{\varepsilon}) = 0$ or $E(\mathrm{y}) = \mathrm{X\beta}$.

2. $\mathrm{Cov}(\mathrm{\varepsilon}) = \sigma^2\,\mathrm{I}$ or $\mathrm{Cov}(\mathrm{y}) = \sigma^2\,\mathrm{I}$

Note that the assumption $\mathrm{Cov}(\mathrm{\varepsilon}) = \sigma^2\,\mathrm{I}$ includes both the previous assumptions $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

The matrix **X** in (3.4) is $n \times (k+1)$. In this chapter we assume that $n > (k+1)$ and rank (**X**) = *k+1*. If $n < (k+1)$ or if there is a linear relationship among the x's, for example, $x_5 = \sum_{i=1}^{n} x_j / 4$, then **X** will not have full column rank. If the values of the x_{ij}'s are planned (chosen by the researcher), then the **X** matrix essentially contains the experimental design and is sometimes called the *design matrix*.

The β parameters in (3.1) or (3.4) are called *regression coefficients*. To emphasize their collective effect, they are sometimes referred to as *partial regression coefficients*. The word *partial carries* both a mathematical and a statistical meaning. Mathematically, the partial derivative of $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ with respect to x_1, for example, is β_1. Thus β_1 indicates the change in E(*y*) with a unit increase in x_1 when $x_2, x_3, \cdots, x_k$ are held constant. Statistically, β_1 shows the effect of x_1 on E(*y*) in the presence of the other x's. This effect would typically be different from the effect of x_1 on E(*y*) if the other x's were not present in the model. Thus, for example, β_0 and β_1 in

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Will usually be different from β_0^* and β_1^* in

$$y = \beta_0^* + \beta_1^* x_1 + \varepsilon^*$$

[If x_1 and x_2 are orthogonal, that is, if $\mathrm{x}_1'\mathrm{x}_2 = 0$ or if $(\mathrm{x}_1 - \bar{\mathrm{x}}_1\,\mathrm{j})'(\mathrm{x}_2 - \bar{\mathrm{x}}_2\,\mathrm{j}) = 0$, where x_1 and x_2 are columns in the **X** matrix, then $\beta_0 = \beta_0^*$ and $\beta_1 = \beta_1^*$; see Corollary 1 to Theorem 3.9a and Theorem 3.10]. The change in parameters when an *x* is deleted from the model is illustrated (with estimates) in the following example.

Example 3.1: [See Freund and Minton (1979, pp. 36–39)]. Consider the (contrived) data in Table 3.1.

TABLE 3.1: Data for Example 3.1

Observation Number	y	x_1	x_2
1	2	0	2
2	3	2	6
3	2	2	7
4	7	2	5
5	6	4	9
6	8	4	8
7	10	4	7
8	7	6	10
9	8	6	11
10	12	6	9
11	11	8	15
12	14	8	13

Using (2.5) and (2.6) from Section 2.2 and (3.6) in Section 3.3, we obtain prediction equations for *y* regressed on x_1 alone, on x_2 alone, and on both x_1 and x_2:

$$\hat{y} = 1.86 + 1.30x_1$$

$$\hat{y} = 0.86 + 0.78x_2$$

$$\hat{y} = 5.37 + 3.01x_1 - 1.29x_2$$

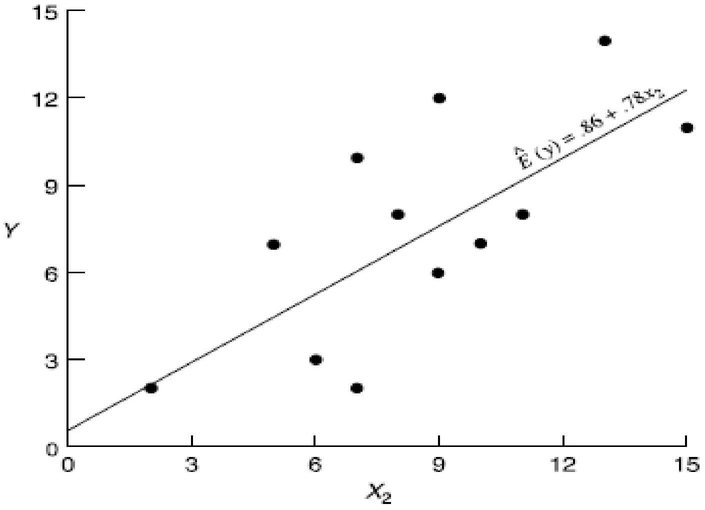


Figure: 3.1 Regression of *y* on x_2 ignoring x_1

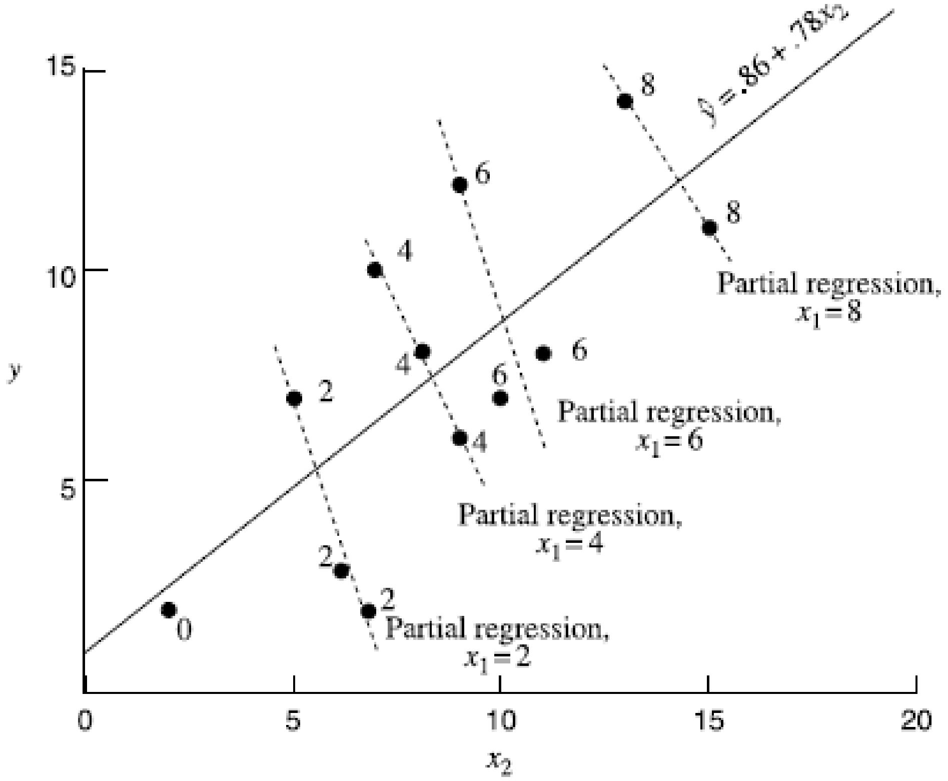


Figure 3.2: Regression of *y* on x_2 showing the value of x_1 at each point and partial regressions of *y* on x_2.

As expected, the coefficients change from either of the reduced models to the full model. Note the sign change as the coefficient of x_2 changes from .78 to -1.29.

The values of y and x_2 are plotted in Figure 3.1 along with the prediction equation $\hat{y} = 0.86 + 0.78x_2$. The linear trend is clearly evident.

In Figure 3.2 we have the same plot as in Figure 3.1, except that each point is labeled with the value of x_1. Examining values of y and x_2 for a fixed value of x_1 (2, 4, 6, or 8) shows a negative slope for the relationship. These negative relationships are shown as partial regressions of y on x_2 for each value of x_1. The partial regression coefficient $\hat{\beta}_2 = -1.29$ reflects the negative slopes of these four partial regressions.

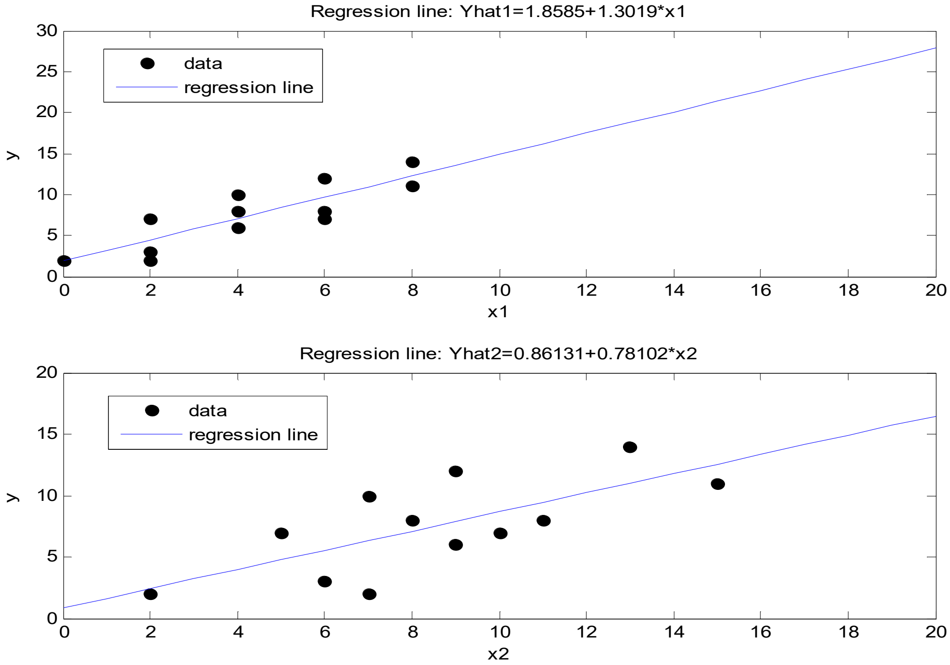
Further insight into the meaning of the partial regression coefficients is given in Section 3.10.

Example 3.1[The program name ta4.m] | Applications using MATLAB

```
clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]'; x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]';
n=length(x1);E1=[ones(size(x1)) x1];
beta1=E1\y,E2=[ones(size(x2)) x2];
beta2=E2\y,E3=[ones(size(x1)) x1 x2]; beta3=E3\y,X=(0:1:20)';
Y1=[ones(size(X)) X]*beta1; Y2=[ones(size(X)) X]*beta2;
subplot(2,1,1),plot(x1,y,'ko',X,Y1,'-')
xlabel('x1'), ylabel('y'), legend('data','regression line')
title(['Regression line: Yhat1=',num2str(beta1(1)),
'+',num2str(beta1(2)),'*x1'])
subplot(2,1,2),plot(x2,y,'ko',X,Y2,'-')
xlabel('x2'), ylabel('y'), legend('data','regression line')
title(['Regression line: Yhat2=',num2str(beta2(1)),
'+',num2str(beta2(2)),'*x2'])
```

Ans.

```
beta1 =            beta2 =              beta3 =
      1.8585             0.86131              5.3754
      1.3019             0.78102              3.0118
                                              -1.2855
```



3.3: Estimation of β and σ^2

3.3.1: Least-Squares Estimator for β

In this section, we discuss the least-squares approach to estimation of theβ’s in the fixed-x model (3.1) or (3.4). No distributional assumptions on y are required to obtain the estimators.

For the parameters $\beta_0, \beta_1, \cdots, \beta_k$, we seek estimators that minimize the sum of squares of deviations of the *n* observed y’s from their predicted values $\hat{y}$. By extension of (2.2), we seek $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ that minimize

$$\begin{aligned} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}\right)^2 \qquad (3.5) \end{aligned}$$

Note that the predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$ estimates $E(y_i)$, not y_i. A better notation would be $E(\hat{y}_i)$, but $\hat{y}_i$ is commonly used.

To obtain the least-squares estimators, it is not necessary that the prediction equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$ be based on $E(y_i)$. It is only necessary to postulate an empirical model that is linear in the $\hat{\beta}$’s, and the least-squares method will find the “best” fit to this model. This was illustrated in Figure 2.2.

To find the values of $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ that minimize (3.5), we could differentiate $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ with respect to each $\hat{\beta}_j$ and set the results equal to zero to yield *k* +1 equations that can be solved simultaneously for the $\hat{\beta}_j$’s. However, the procedure can be carried out in more compact form with matrix notation. The result is given in the following theorem.

Theorem 3.3a: If $y = X\beta + \varepsilon$, where **X** is $n \times (k+1)$ of rank *k* +1 < *n*, then the value of $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k\right)'$ that minimizes (3.5) is

$$\hat{\beta} = (X'X)^{-1} X'y \qquad (3.6)$$

Proof. We can write (3.5) as

$$\hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n} \left(y_i - x_i'\hat{\beta}\right)^2 = \left(y - X\hat{\beta}\right)'\left(y - X'\hat{\beta}\right) \qquad (3.7)$$

Where $\mathrm{x}_i' = (1, x_{i1}, \cdots, x_{ik})$ is the *i*th row of **X**. When the product $\left(\mathrm{y} - \mathrm{X}'\hat{\beta}\right)'\left(\mathrm{y} - \mathrm{X}'\hat{\beta}\right)$ in (3.7) is expanded as two of the resulting four terms can be combined to yield

$$\hat{\varepsilon}'\hat{\varepsilon} = \mathrm{y}'\mathrm{y} - 2\,\mathrm{y}'\mathrm{X}\hat{\beta} + \hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}$$

We can find the value of $\hat{\beta}$ that minimizes $\hat{\varepsilon}'\hat{\varepsilon}$ by differentiating $\hat{\varepsilon}'\hat{\varepsilon}$ with respect to $\hat{\beta}$ and setting the result equal to zero:

$$\frac{\partial\,\hat{\varepsilon}'\hat{\varepsilon}}{\partial\,\hat{\beta}} = 0 - 2\,\mathrm{X}'\mathrm{y} + 2\,\mathrm{X}'\mathrm{X}\hat{\beta} = 0$$

This gives the normal equations

$$\mathrm{X}'\mathrm{X}\hat{\beta} = \mathrm{X}'\mathrm{y} \qquad (3.8)$$

If X is full-rank, $\mathrm{X}'\mathrm{X}$ is non-singular, and the solution to (3.8) is given by (3.6).

Since $\hat{\beta}$ in (3.6) minimizes the sum of squares in (3.5), $\hat{\beta}$ is called the *least squares estimator*. Note that each $\hat{\beta}_j$ in $\hat{\beta}$ is a linear function of y; that is, $\hat{\beta}_j = \mathrm{a}_j'\,\mathrm{y}$, where a_j' is the *j*th row of $(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{X}'$. This usage of the word *linear* in *linear estimator* is different from that in *linear model*, which indicates that the model is linear in the β's.

We now show that $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{X}'\,\mathrm{y}$ minimizes $\hat{\varepsilon}'\hat{\varepsilon}$. Let *b* be an alternative estimator that may do better than $\hat{\beta}$ so that $\hat{\varepsilon}'\hat{\varepsilon}$ is

$$\hat{\varepsilon}'\hat{\varepsilon} = (\mathrm{y} - \mathrm{X}b)'(\mathrm{y} - \mathrm{X}b)$$

Now adding and subtracting $\mathrm{X}\hat{\beta}$, we obtain

$$\hat{\varepsilon}'\hat{\varepsilon} = \left(\mathrm{y} - \mathrm{X}\hat{\beta} + \mathrm{X}\hat{\beta} - \mathrm{X}b\right)'\left(\mathrm{y} - \mathrm{X}\hat{\beta} + \mathrm{X}\hat{\beta} - \mathrm{X}b\right) \qquad (3.9)$$

$$= \left(\mathrm{y} - \mathrm{X}\hat{\beta}\right)'\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right) + \left(\hat{\beta} - b\right)'\mathrm{X}'\mathrm{X}\left(\hat{\beta} - b\right) + 2\left(\hat{\beta} - b\right)'\left(\mathrm{X}'\mathrm{y} - \mathrm{X}'\mathrm{X}\hat{\beta}\right) \qquad (3.10)$$

The third term on the right side of (3.10) vanishes because of the normal equations $\mathrm{X}'\mathrm{y} = \mathrm{X}'\mathrm{X}\hat{\beta}$ in (3.8). The second term is a positive definite quadratic form (assuming that **X** is full-rank; and $\hat{\varepsilon}'\hat{\varepsilon}$ is therefore minimized when $b = \hat{\beta}$.

To examine the structure of $\mathrm{X'X}$ and $\mathrm{X'y}$, the $(k+1)\times(k+1)$ matrix $\mathrm{X'X}$ can be obtained as products of columns of **X**; similarly, $\mathrm{X'y}$ contains products of columns of **X** and y:

$$\mathrm{X'X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^{2} & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^{2} & \cdots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^{2} \end{pmatrix}_{(k+1\times k+1)}$$

$$\mathrm{X'y} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}\,y_i \end{pmatrix}_{(k+1\times 1)}$$

If $\hat{\beta} = (\mathrm{X'X})^{-1}\,\mathrm{X'y}$ as in (3.6), then

$$\hat{\varepsilon} = \mathrm{y} - \mathrm{X}\hat{\beta} = \mathrm{y} - \hat{\mathrm{y}} \qquad (3.11)$$

Is the vector of residuals, $\hat{\varepsilon}_1 = y_1 - \hat{y}_1, \hat{\varepsilon}_2 = y_2 - \hat{y}_2, \ldots, \hat{\varepsilon}_n = y_n - \hat{y}_n$. The residual vector $\hat{\varepsilon}$ estimates ε in the model $\mathrm{y} = \mathrm{X}\beta + \varepsilon$ and can be used to check the validity of the model and attendant assumptions; see Chapter 5.

Example 3.3.1a: We use the data in Table 3.1 to illustrate computation of $\hat{\beta}$ using (3.6).

$$\mathrm{X'} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 & 4 & 4 & 4 & 6 & 6 & 6 & 8 & 8 \\ 2 & 6 & 7 & 5 & 9 & 8 & 7 & 10 & 11 & 9 & 15 & 13 \end{pmatrix}$$

$$\mathrm{y'} = \begin{pmatrix} 2 & 3 & 2 & 7 & 6 & 8 & 10 & 7 & 8 & 12 & 11 & 14 \end{pmatrix}$$

$$X'X=\begin{pmatrix}12 & 52 & 102\\ 52 & 395 & 536\\ 102 & 536 & 1004\end{pmatrix},\quad X'y=\begin{pmatrix}90\\ 482\\ 872\end{pmatrix},$$

$$(X'X)^{-1}=\begin{pmatrix}0.97476 & 0.24290 & -0.22871\\ 0.24290 & 0.16207 & -0.11120\\ -0.22871 & -0.11120 & 0.08360\end{pmatrix}$$

$$\hat{\beta}=(X'X)^{-1}X'y=\begin{pmatrix}5.3754\\ 3.0118\\ -1.2855\end{pmatrix}$$

Example 3.3.1a [The program name ta5.m]	Applications using MATLAB

```
clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]';
x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]';
n=length(x1);X=[ones(size(x1)) x1 x2];
XTX=X'*X,InvXTX=inv(X'*X),XTy=X'*y
beta=X\y
```

Ans.

```
XTX =

        12        52        102

        52       296        536

       102       536       1004

InvXTX =

     0.97476      0.2429     -0.22871
      0.2429     0.16207      -0.1112
    -0.22871     -0.1112     0.083596

XTy =        beta =

    90             5.3754
   482             3.0118
   872            -1.2855
```

Example 7.3.1b: Simple linear regression from Chapter 2 can also be expressed in matrix terms:

$$\mathrm{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathrm{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \ \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}, \ \mathrm{X'X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$$(\mathrm{X'X})^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \ \mathrm{X'y} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Then $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained using (3.6), $\hat{\beta} = (\mathrm{X'X})^{-1}\mathrm{X'y}$:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right) \\ -\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) + n\sum_{i=1}^{n} x_i y_i \end{pmatrix} \qquad (3.12)$$

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in (3.11) are the same as those in (2.5) and (2.6).

3.3.2: Properties of the Least-Squares Estimator $\hat{\beta}$

The least-squares estimator $\hat{\beta} = (\mathrm{X'X})^{-1}\mathrm{X'y}$ in Theorem 3.3a was obtained without using the assumptions $E(\mathrm{y}) = \mathrm{X}\beta$ and $Cov(y) = \sigma^2 I$ given in Section 3.2. We merely postulated a model $\mathrm{y} = \mathrm{X}\beta + \varepsilon$ as in (3.4) and fitted it. If $E(\mathrm{y}) \neq \mathrm{X}\beta$, the model $\mathrm{y} = \mathrm{X}\beta + \varepsilon$ could still be fitted to the data, in which case, $\hat{\beta}$ may have poor properties. If $Cov(y) \neq \sigma^2 I$, there may be additional adverse effects on the estimator $\hat{\beta}$. However, if $E(\mathrm{y}) = \mathrm{X}\beta$ and $Cov(y) = \sigma^2 I$ hold, $\hat{\beta}$ has some good properties, as noted in the four theorems in this section. Note that $\hat{\beta}$ is a random vector (from sample to sample). We discuss its mean vector and covariance matrix in this section (with no distributional assumptions on y) and its distribution (assuming that the y variables are normal) in Section 3.6.3. In the following theorems, we assume that **X** is fixed (remains constant in repeated sampling) and full rank.

Theorem 3.3b: If $E(\mathrm{y}) = \mathrm{X}\beta$, then $\hat{\beta}$ is an unbiased estimator for β.

Proof

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}] \\ &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'E(\mathrm{y}) \\ &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}\beta \\ &= \beta \end{aligned} \tag{3.13}$$

Theorem 3.3c: If $Cov(\mathrm{y}) = \sigma^2 I$, the covariance matrix for $\hat{\beta}$ is given by $\sigma^2(\mathrm{X}'\mathrm{X})^{-1}$.

Proof

$$\begin{aligned} Cov(\hat{\beta}) &= Cov[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}] \\ &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\,Cov(\mathrm{y})[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']' \\ &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\sigma^2 I)\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} \\ &= \sigma^2(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} \\ &= \sigma^2(\mathrm{X}'\mathrm{X})^{-1} \end{aligned} \tag{3.14}$$

Example 3.3.2a: Using the matrix $(\mathrm{X}'\mathrm{X})^{-1}$ for simple linear regression given in example 3.3.1, we obtain

$$Cov(\hat{\beta}) = Cov\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$$

$$= \frac{\sigma^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \tag{3.15}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 \big/ n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \tag{3.16}$$

Thus

$$Var\left(\hat{\beta}_0\right)=\frac{\sigma^2\sum_{i=1}^{n}x_i^2\Big/n}{\sum_{i=1}^{n}(x_i-\bar{x})^2},\ Var\left(\hat{\beta}_1\right)=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},\ Cov\left(\hat{\beta}_0,\hat{\beta}_1\right)=\frac{-\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

We found $Var\left(\hat{\beta}_0\right)$ and $Var\left(\hat{\beta}_1\right)$ in Section 2.2 but did not obtain $Cov\left(\hat{\beta}_0,\hat{\beta}_1\right)$. Note that if $\bar{x}>0$, then $Cov\left(\hat{\beta}_0,\hat{\beta}_1\right)$ is negative and the estimated slope and intercept are negatively correlated. In this case, if the estimate of the slope increases from one sample to another, the estimate of the intercept tends to decrease (assuming the *x*'s stay the same).

Example 3.3.2b: For the data in Table 3.1, $(\mathrm{X}'\mathrm{X})^{-1}$ is as given in Example 3.3.1. Thus, $Cov\left(\hat{\beta}\right)$ is given by

$$Cov\left(\hat{\beta}\right)=\sigma^2(\mathrm{X}'\mathrm{X})^{-1}=\sigma^2\begin{pmatrix}0.975 & 0.243 & -0.229\\ 0.243 & 0.162 & -0.111\\ -0.229 & -0.111 & 0.084\end{pmatrix}$$

The negative value of $Cov\left(\hat{\beta}_1,\hat{\beta}_2\right)=-0.111\sigma^2$ indicates that in repeated sampling (using the same 12 values of x_1 and x_2), $\hat{\beta}_1$ and $\hat{\beta}_2$ would tend to move in opposite directions; that is, an increase in one would be accompanied by a decrease in the other.

In addition to $E\left(\hat{\beta}\right)=\beta$ and $Cov\left(\hat{\beta}\right)=\sigma^2(\mathrm{X}'\mathrm{X})^{-1}$, a third important property of $\hat{\beta}$ is that under the standard assumptions, the variance of each $\hat{\beta}_j$ is minimum (see the following theorem).

Example 3.3.2b [The program name ta6.m]	Applications using MATLAB

```
clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]';
x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]';
n=length(x1);X=[ones(size(x1)) x1 x2];
XTX=X'*X;InvXTX=inv(X'*X),XTy=X'*y;beta=X\y;
Yhat=X*beta;e=y-Yhat;
MSE=e'*e/(n-2)
Covbeta=MSE*InvXTX
```

```
Ans.
InvXTX =

        0.97476      0.2429    -0.22871
         0.2429     0.16207     -0.1112
       -0.22871     -0.1112    0.083596

MSE =

         2.5459

Covbeta =

         2.4816      0.6184    -0.58226
         0.6184      0.4126     -0.2831
       -0.58226     -0.2831     0.21283
```

Theorem 7.3d: (Gauss–Markov Theorem). If $E(\mathrm{y}) = \mathrm{X}\beta$ and $Cov(y) = \sigma^2 I$, the least-squares estimators $\hat{\beta}_j$, *j = 0, 1, . . . , k*, have minimum variance among all linear unbiased estimators.

Proof. We consider a linear estimator **Ay** of β and seek the matrix **A** for which **Ay** is a minimum variance unbiased estimator of β. In order for **Ay** to be an unbiased estimator of β, we must have E(**Ay**) $= \beta$. Using the assumption $E(\mathrm{y}) = \mathrm{X}\beta$, this can be expressed as

$$E(\mathrm{Ay}) = \mathrm{A}\,E(\mathrm{y}) = \mathrm{AX}\beta = \beta$$

which gives the unbiasedness condition

$$\mathrm{AX} = \mathrm{I}$$

Since the relationship $\mathrm{AX}\beta = \beta$ must hold for any possible value of β.

The covariance matrix for the estimator **Ay** is given by

$$Cov(\mathrm{Ay}) = \mathrm{A}(\sigma^2\,\mathrm{I})\mathrm{A}' = \sigma^2\,\mathrm{AA}'$$

The variances of the $\hat{\beta}_j$'s are on the diagonal of $\sigma^2\mathrm{AA}'$, and we therefore need to choose **A** (subject to $\mathrm{AX} = \mathrm{I}$) so that the diagonal elements of AA' are minimized. To relate **Ay** to $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$, we add and subtract $(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ to obtain

$$\mathrm{AA}' = [\mathrm{A} - (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' + (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'][\mathrm{A} - (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' + (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']'$$

Expanding this in terms of $\mathrm{A}-(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ and $(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$, we obtain four terms, two of which vanish because of the restriction $\mathrm{AX}=\mathrm{I}$. The result is

$$\mathrm{AA}'=[\mathrm{A}-(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'][\mathrm{A}-(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']'+(\mathrm{X}'\mathrm{X})^{-1} \qquad (3.17)$$

The matrix $[\mathrm{A}-(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'][\mathrm{A}-(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']'$ on the right side of (3.17) is positive semi-definite, and the diagonal elements are greater than or equal to zero. These diagonal elements can be made equal to zero by choosing $\mathrm{A}=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$. (This value of **A** also satisfies the unbiasedness condition $\mathrm{AX}=\mathrm{I}$) The resulting minimum variance estimator of β is

$$\mathrm{A}\,\mathrm{y}=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

which is equal to the least–squares estimator $\hat{\beta}$.

The Gauss–Markov theorem is sometimes stated as follows. If $E(\mathrm{y})=\mathrm{X}\beta$ and $Cov(\mathrm{y})=\sigma^2 I$, the least-squares estimators $(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k)$ are best linear unbiased estimators (BLUE). In this expression, best means minimum variance and linear indicates that the estimators are linear functions of **y**.

The remarkable feature of the Gauss–Markov theorem is its distributional generality. The result holds for any distribution of **y**; normality is not required. The only assumptions used in the proof are $E(\mathrm{y})=\mathrm{X}\beta$ and $Cov(\mathrm{y})=\sigma^2 I$. If these assumptions do not hold, $\hat{\beta}$ may be biased or each $\hat{\beta}_j$ may have a larger variance than that of some other estimator.

The Gauss–Markov theorem is easily extended to a linear combination of the $\hat{\beta}$'s, as follows.

Corollary 1: If $E(\mathrm{y})=\mathrm{X}\beta$ and $Cov(\mathrm{y})=\sigma^2 I$, the best linear unbiased estimator of $\mathrm{a}'\beta$ is $\mathrm{a}'\hat{\beta}$, where $\hat{\beta}$ is the least–squares estimator $\hat{\beta}=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$.

Proof. See Problem 3.7.

Note that Theorem 3.3d is concerned with the form of the estimator $\hat{\beta}$ for a given **X** matrix. Once **X** is chosen, the variances of the $\hat{\beta}_j$'s are minimized by $\hat{\beta}=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$. However, in Theorem 3.3c, we have

$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$ and therefore $Var(\hat{\beta}_j)$ and $Cov(\hat{\beta}_i, \hat{\beta}_j)$ depend on the values of the x_j's. Thus the configuration of $X'X$ is important in estimation of the β_j's (this was illustrated in Problem 2.4).

In both estimation and testing, there are advantages to choosing the x's (or the centred x's) to be orthogonal so that $X'X$ is diagonal. These advantages include minimizing the variances of the $\hat{\beta}_j$'s and maximizing the power of tests about the β_j's (Chapter 4). For clarification, we note that orthogonality is necessary but not sufficient for minimizing variances and maximizing power. For example, if there are two x's, with values to be selected in a rectangular space, the points could be evenly placed on a grid, which would be an orthogonal pattern. However, the optimal orthogonal pattern would be to place one-fourth of the points at each corner of the rectangle.

A fourth property of $\hat{\beta}$ is as follows. The predicted value $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k = \hat{\beta}' x$ is invariant to simple linear changes of scale on the x's, where $x = (1, x_1, x_2, \cdots, x_k)'$. Let the rescaled variables be denoted by $z_j = c_j x_j$, *j = 1, 2, . . . , k*, where the c_j terms are constants. Thus **x** is transformed to $z = (1, c_1 x_1, \cdots, c_k x_k)'$. The following theorem shows that $\hat{y}$ based on **z** is the same as $\hat{y}$ based on **x**.

Theorem 3.3e: If $x = (1, x_1, x_2, \cdots, x_k)'$ and $z = (1, c_1 x_1, \cdots, c_k x_k)'$, then $\hat{y} = \hat{\beta}' x = \hat{\beta}_z' z$, where $\hat{\beta}_z$ is the least squares estimator from the regression of **y** on **z**.

Proof: We can rewrite **z** as **z = Dx**, where **D** = $\text{diag}(1, c_1 x_1, \cdots, c_k x_k)$. Then, the **X** matrix is transformed to Z = **XD**. We substitute **Z** = **XD** in the least-squares estimator $\hat{\beta}_Z = (Z'Z)^{-1} Z' y$ to obtain

$$\begin{aligned} \hat{\beta}_Z = (Z'Z)^{-1} Z' y &= [(XD)'(XD)]^{-1} (XD)' y \\ &= D^{-1} (X'X)^{-1} X'y \\ &= D^{-1} \hat{\beta} \qquad (3.18) \end{aligned}$$

Where $\hat{\beta}$ is the usual estimator for **y** regressed on the *x*'s. Then

$$\hat{\beta}_z' z = \left(D^{-1}\hat{\beta}\right)' Dx = \hat{\beta}' x$$

In the following corollary to Theorem 3.3e, the invariance of $\hat{y}$ is extended to any full-rank linear transformation of the *x* variables.

Corollary 1: The predicted value $\hat{y}$ is invariant to a full-rank linear transformation on the *x*'s.

Proof: We can express a full-rank linear transformation of the *x*'s as

$$Z = XK = (j, X_1)\begin{pmatrix} 1 & 0' \\ 0 & K_1 \end{pmatrix} = (j + X_1\, 0, j0' + X_1K_1) = (j, X_1K_1)$$

Where K_1 is non-singular and

$$X_1 = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \qquad (3.19)$$

We partition **X** and **K** in this way so as to transform only the *x*'s in X_1, leaving the first column of **X** unaffected. Now $\hat{\beta}_Z$ becomes

$$\hat{\beta}_Z = (Z'Z)^{-1} Z' y = K^{-1}\hat{\beta} \qquad (3.20)$$

And we have

$$\hat{y} = \hat{\beta}_z' z = \hat{\beta}' x \qquad (3.21)$$

Where $z = K' x$

In addition to $\hat{y}$, the sample variance S^2 (Section 3.3.3) is also invariant to changes of scale on the *x* variable (see Problem 3.10). The following are invariant to changes of scale on **y** as well as on the *x*'s (but not to a joint linear transformation on **y** and the *x*'s): *t* statistics (Section 4.5), *F* statistics (Chapter 4), and R^2 (Sections 3.7 and 6.3).

3.3.3: An Estimator for σ^2

The method of least squares does not yield a function of the *y* and *x* values in the sample that we can minimize to obtain an estimator of σ^2. However, we can devise an unbiased estimator for σ^2 based on the least-squares estimator $\hat{\beta}$. By assumption 2 following (3.3), σ^2 is the same for each y_i, i = 1, 2, . . . , n. σ^2 is defined by $\sigma^2 = E[y_i - E(y_i)]^2$, and by assumption 1, we obtain

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \mathrm{x}_i'\beta$$

Where x_i' is the ith row of **X**. Thus σ^2 becomes

$$\sigma^2 = E[y_i - \mathrm{x}_i'\beta]^2$$

We estimate σ^2 by a corresponding average from the sample

$$S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(y_i - \mathrm{x}_i'\hat{\beta}\right)^2 \tag{3.22}$$

Where n is the sample size and k is the number of x's. Note that, by the corollary to Theorem 3.3d, $\mathrm{x}_i'\hat{\beta}$ is the *BLUE* of $\mathrm{x}_i'\beta$.

Using (3.7), we can write (3.22) as

$$S^2 = \frac{1}{n-k-1}\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right)'\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right) \tag{3.23}$$

$$S^2 = \frac{\mathrm{y}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\mathrm{y}}{n-k-1} = \frac{SSE}{n-k-1} \tag{3.24}$$

Where $SSE = \left(\mathrm{y} - \mathrm{X}\hat{\beta}\right)'\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right) = \mathrm{y}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\mathrm{y}$. With the denominator n-k-1, S^2 is an unbiased estimator of σ^2, as shown below.

Theorem 3.3f: If S^2 is defined by (3.22), (3.23), or (3.24) and if $E(\mathrm{y}) = \mathrm{X}\beta$ and $Cov(\mathrm{y}) = \sigma^2 I$, then

$$E\left(S^2\right) = \sigma^2 \tag{3.25}$$

Proof: Using (3.24) and (3.6), we write *SSE* as a quadratic form:

$$\begin{aligned} SSE &= \mathrm{y}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\mathrm{y} = \mathrm{y}'\mathrm{y} - \mathrm{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y} \\ &= \mathrm{y}'[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\mathrm{y} \end{aligned} \tag{3.26}$$

We have

$$\begin{aligned} E(SSE) &= tr\{[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\sigma^2\,\mathrm{I}\} + E(\mathrm{y}')[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']E(\mathrm{y}) \\ &= \sigma^2 tr\{\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\} + \beta'\mathrm{X}'[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\mathrm{X}\beta \\ &= \sigma^2\left\{n - tr\left[\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right]\right\} + \beta'\mathrm{X}'\mathrm{X}\beta - \beta'\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}\beta \end{aligned}$$

$$= \sigma^2\left\{n - tr\left[X(X'X)^{-1}X'\right]\right\} + \beta'X'X\beta - \beta'X'X\beta$$

Since $X'X$ is $(k+1)\times(k+1)$, this becomes

$$E(SSE) = \sigma^2\left\{n - tr\left[I_{k+1}\right]\right\} = \sigma^2(n-k-1)$$

Corollary 1: An unbiased estimator of $Cov(\hat{\beta})$ in (3.14) is given by

$$\hat{C}ov(\hat{\beta}) = S^2(X'X)^{-1} \qquad (3.27)$$

Note the correspondence between $n-(k+1)$ and $y'y-\hat{\beta}'X'y$; there are *n* terms in $y'y$ and *k*+1 terms in $\hat{\beta}'X'y = \hat{\beta}'X'X\hat{\beta}$ [see (3.8)]. A corresponding property of the sample is that each additional *x* (and $\hat{\beta}$) in the model reduces *SSE* (see Problem 3.13).

Since *SSE* is a quadratic function of **y**, it is not a best linear unbiased estimator. The optimality property of S^2 is given in the following theorem.

Theorem 3.3g: If $E(\varepsilon)=0$, $Cov(\varepsilon)=\sigma^2$, and $E(\varepsilon_i^4)=3\sigma^4$ for the linear model $y = X\beta + \varepsilon\, y$, then S^2 in (3.23) or (3.24) is the best (minimum variance) quadratic unbiased estimator of σ^2.

Proof: See Graybill (1954), Graybill and Wortham (1956), or Wang and Chow (1994, pp. 161–163).

Example 3.3.3: For the data in Table 3.1, we have

$$SSE = y'y - \hat{\beta}'X'y \text{ then } SSE = 840 - (5.3754, 3.0118, -1.2855)\begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}$$

$$SSE = 840 - 814.541 = 25.459 \text{ and } S^2 = \frac{SSE}{n-k-1} = \frac{25.459}{12-2-1} = 2.829$$

Example 3.3.3[The program name ta7.m]	Applications using MATLAB

```
clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]';
x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]';
n=length(x1);X=[ones(size(x1)) x1 x2];k=2;
XTX=X'*X;InvXTX=inv(X'*X);XTy=X'*y;beta=X\y;
SSE=y'*y-beta'*X'*y,Ssquare=SSE/(n-k-1)
```

Ans.

SSE = 25.459   Ssquare = 2.8288

3.4: Geometry of Least Squares

In Sections 3.1–3.3 we presented the multiple linear regression model as the matrix equation $y = X\beta + \varepsilon$ in (3.4). We defined the principle of least-squares estimation in terms of deviations from the model [see (3.7)], and then used matrix calculus and matrix algebra to derive the estimators of β in (3.6) and of σ^2 in (3.23) and (3.24). We now present an alternate but equivalent derivation of these estimators based completely on geometric ideas.

It is important to clarify first what the geometric approach to least squares is not. In two dimensions, we illustrated the principle of least squares by creating a two dimensional scatter plot (Fig. 2.1) of the *n* points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. We then visualized the least-squares regression line as the best-fitting straight line to the data. This approach can be generalized to present the least-squares estimate in multiple linear regression on the basis of the best-fitting hyperplane in $(k+1)$-dimensional space to the *n* points $(x_{11}, x_{12}, \cdots, x_{1k}, y_1)$, $(x_{21}, x_{22}, \cdots, x_{2k}, y_2)$, …, $(x_{n1}, x_{n2}, \cdots, x_{nk}, y_n)$. Although this approach is somewhat useful in visualizing multiple linear regression, the geometric approach to least-squares estimation in multiple linear regression does not involve this high-dimensional generalization.

The geometric approach to be discussed below is appealing because of its mathematical elegance. For example, the estimator is derived without the use of matrix calculus. Also, the geometric approach provides deeper insight into statistical inference. Several advanced statistical methods including kernel smoothing (Eubank and Eubank 1999), Fourier analysis (Bloomfield 2000), and wavelet analysis (Ogden 1997) can be understood as generalizations of this geometric approach. The geometric approach to linear models was first proposed by Fisher (Mahalanobis 1964).Christensen (1996) and Jammalamadaka and Sengupta (2003) discuss the linear statistical model almost completely from the geometric perspective.

3.4.1: Parameter Space, Data Space, and Prediction Space

The geometric approach to least squares begins with two high-dimensional spaces, a (*k*+1)-dimensional space and an *n*-dimensional space. The unknown parameter vector β can be viewed as a single point in (*k*+1)-dimensional space, with axes corresponding to the *k* +1 regression coefficients $\beta_0, \beta_1, \cdots, \beta_k$. Hence we call this space the parameter space (Fig. 3.3). Similarly, the data vector **y** can be viewed as a single point in *n*-dimensional space with axes corresponding to the *n* observations. We call this space the data space.

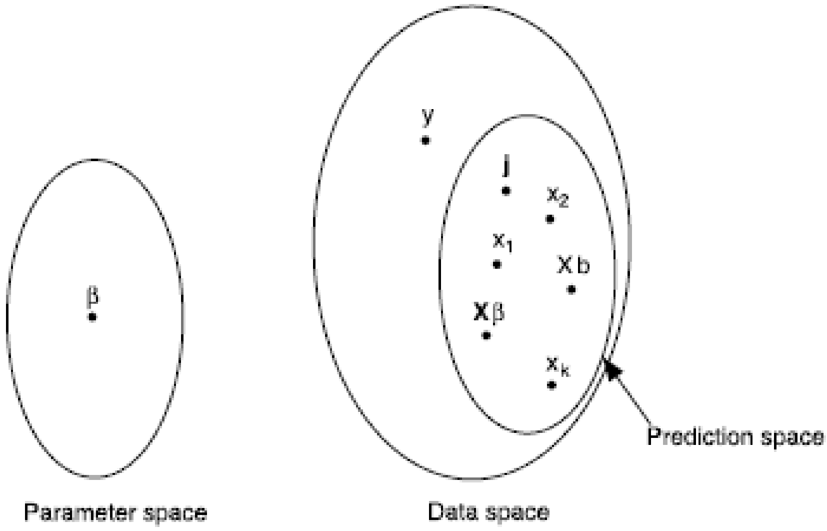


Figure 3.3: Parameter space, data space, and prediction space with representative elements.

The **X** matrix of the multiple regression model (3.4) can be written as a partitioned matrix in terms of its *k*+1 columns as

$$\mathrm{X} = (\mathrm{j}, \mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3 \cdots, \mathrm{x}_k)$$

The columns of **X**, including **j**, are all *n*-dimensional vectors and are therefore points in the data space. Note that because we assumed that **X** is of rank *k* +1, these vectors are linearly independent. The set of all possible linear combinations of the columns of **X** constitutes a subset of the data space. Elements of this subset can be written as

$$\mathrm{Xb} = b_0\,\mathrm{j} + b_1\,\mathrm{x}_1 + b_2\,\mathrm{x}_2 + \cdots + +b_k\,\mathrm{x}_k \qquad (3.28)$$

where **b** is any *k* +1 vector, that is, any vector in the parameter space. This subset actually has the status of a subspace because it is closed under addition and scalar multiplication (Harville 1997, pp. 28–29). This subset is said to be the subspace generated or spanned by the columns of **X**, and we will call this subspace the prediction space. The columns of **X** constitute a basis set for the prediction space.

3.4.2: Geometric Interpretation of the Multiple Linear Regression Model

The multiple linear regression model (3.4) states that **y** is equal to a vector in the prediction space, $E(y)=X\beta$, plus a vector of random errors, ε (Fig. 3.4). The problem is that neither β nor ε is known. However, the data vector **y**, which is not in the prediction space, is known. And it is known that E(**y**) is in the prediction space.

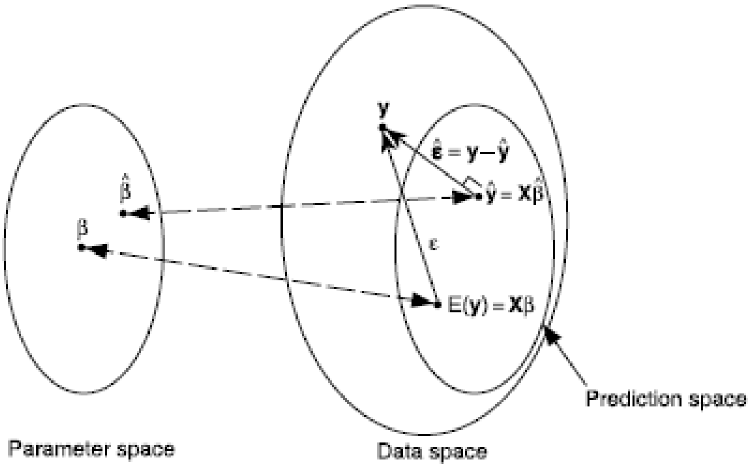


Figure 3.4: Geometric relationships of vectors associated with the multiple linear regression model.

Multiple linear regression can be understood geometrically as the process of finding a sensible estimate of E(**y**) in the prediction space and then determining the vector in the parameter space that is associated with this estimate (Fig. 3.4). The estimate of E(**y**) is denoted as $\hat{y}$, and the associated vector in the parameter space is denoted as $\hat{\beta}$.

A reasonable geometric idea is to estimate E(**y**) using the point in the prediction space that is closest to **y**. It turns out that $\hat{y}$, the closest point in the prediction space to **y**, can be found by noting that the difference vector $\hat{\varepsilon} = y - \hat{y}$ must be orthogonal (perpendicular) to the prediction space (Harville 1997, p. 170). Furthermore, because the prediction space is spanned by the columns of **X**, the point $\hat{y}$ must be such that $\hat{\varepsilon}$ is orthogonal to the columns of **X**. We therefore seek $\hat{y}$ such that

$$X'\hat{\varepsilon} = 0$$

or

$$X'(y - \hat{y}) = X'(y - X\hat{\beta}) = X'y - X'X\hat{\beta} = 0 \qquad (3.29)$$

which implies that

$$X'X\hat{\beta} = X'y$$

Thus, using purely geometric ideas, we obtain the normal equations (3.8) and consequently the usual least-squares estimator $\hat{\beta}$ in (3.6). We can then calculate $\hat{y}$ as $X\hat{\beta} = X'(X'X)^{-1}X'y = Hy$. Also, $\hat{\varepsilon} = y - X\hat{\beta} = (I - H)y$ can be taken as an estimate of ε. Since $\hat{\varepsilon}$ is a vector in (*n*-*k*-1)-dimensional space, it seems reasonable to estimate σ^2 as the squared length of $\hat{\varepsilon}$ divided by *n*-*k*-1. In other words, a sensible estimator of σ^2 is $S^2 = y'(I - H)y/(n - k - 1)$, which is equal to (3.25).

3.5: The Model in Centered Form

The model in (3.3) for each y_i can be written in terms of centered *x* variables as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$y_i = \alpha + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + \varepsilon_i \qquad (3.30)$$

$i = 1,2,\cdots,n$ where

$$\alpha = \beta_0 + \beta_1\bar{x}_1 + \beta_2\bar{x}_2 + \cdots + \beta_k\bar{x}_k \qquad (3.31)$$

And $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$, $j = 1,2,\cdots,k$. The centered form of the model is useful in expressing certain hypothesis tests (Section 4.1), in a search for influential observations (Section 5.2), and in providing other insights.

In matrix form, the centered model (3.30) for $y_1, y_2, \cdots, y_n$ becomes

$$y = (j, X_c)\begin{pmatrix}\alpha \\ \beta_1\end{pmatrix} + \varepsilon \qquad (3.32)$$

Where $\beta_1 = (\beta_1, \beta_2, \cdots, \beta_k)'$

$$\mathrm{X}_c = \left(\mathrm{I} - \frac{1}{n}\mathrm{J}\right)\mathrm{X}_1 = \begin{pmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1k}-\bar{x}_k \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2k}-\bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{nk}-\bar{x}_k \end{pmatrix} \qquad (3.33)$$

and X_1 is as given in (3.19). The matrix $\mathrm{I}-(1/n)\mathrm{J}$ is sometimes called the *centering matrix.*

As in (3.8), the normal equations for the model in (3.32) are

$$(\mathrm{j},\mathrm{X}_c)'(\mathrm{j},\mathrm{X}_c)\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}_1\end{pmatrix} = (\mathrm{j},\mathrm{X}_c)'\,\mathrm{y} \qquad (3.34)$$

The product $(\mathrm{j},\mathrm{X}_c)'(\mathrm{j},\mathrm{X}_c)$ on the left side of (3.34) becomes

$$(\mathrm{j},\mathrm{X}_c)'(\mathrm{j},\mathrm{X}_c) = \begin{pmatrix}\mathrm{j}'\\ \mathrm{X}_\mathrm{c}'\end{pmatrix}(\mathrm{j},\mathrm{X}_c) = \begin{pmatrix}\mathrm{j}'\mathrm{j} & \mathrm{j}'\mathrm{X}_c\\ \mathrm{X}_\mathrm{c}'\mathrm{j} & \mathrm{X}_\mathrm{c}'\mathrm{X}_c\end{pmatrix}$$

$$= \begin{pmatrix} n & 0' \\ 0 & \mathrm{X}_\mathrm{c}'\mathrm{X}_c\end{pmatrix} \qquad (3.35)$$

Where $\mathrm{j}'\mathrm{X}_c = 0'$ because the columns of X_c sum to zero (Problem 3.16). The right side of (3.34) can be written as

$$(\mathrm{j},\mathrm{X}_c)'\,\mathrm{y} = \begin{pmatrix}\mathrm{j}'\\ \mathrm{X}_\mathrm{c}'\end{pmatrix}\mathrm{y} = \begin{pmatrix} n\bar{y}\\ \mathrm{X}_\mathrm{c}'\mathrm{y}\end{pmatrix}$$

The least-squares estimators are then given by

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}_1\end{pmatrix} = [(\mathrm{j},\mathrm{X}_c)'(\mathrm{j},\mathrm{X}_c)]^{-1}(\mathrm{j},\mathrm{X}_c)'\,\mathrm{y} = \begin{pmatrix} n & 0' \\ 0 & \mathrm{X}_\mathrm{c}'\mathrm{X}_c\end{pmatrix}^{-1}\begin{pmatrix} n\bar{y}\\ \mathrm{X}_\mathrm{c}'\mathrm{y}\end{pmatrix}$$

$$= \begin{pmatrix} 1/n & 0' \\ 0 & (\mathrm{X}_\mathrm{c}'\mathrm{X}_c)^{-1}\end{pmatrix}\begin{pmatrix} n\bar{y}\\ \mathrm{X}_\mathrm{c}'\mathrm{y}\end{pmatrix} = \begin{pmatrix} \bar{y}\\ (\mathrm{X}_\mathrm{c}'\mathrm{X}_c)^{-1}\,\mathrm{X}_\mathrm{c}'\mathrm{y}\end{pmatrix}$$

Or

$$\hat{\alpha} = \bar{y} \qquad (3.36)$$

$$\hat{\beta}_1 = (\mathrm{X}_\mathrm{c}'\mathrm{X}_c)^{-1}\,\mathrm{X}_\mathrm{c}'\mathrm{y} \qquad (3.37)$$

These estimators are the same as the usual least-squares estimators $\hat{\beta}_1 = (\mathrm{X}_\mathrm{c}'\mathrm{X}_c)^{-1}\,\mathrm{X}_\mathrm{c}'\mathrm{y}$ in (3.6), with the adjustment

$$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 - \cdots - \hat{\beta}_k\bar{x}_k = \bar{y} - \hat{\beta}_1' \mathrm{x} \tag{3.38}$$

Obtained from an estimator of α in (3.31) (see Problem 3.17).

When we express $\hat{\mathrm{y}}$ in centered form

$$\hat{\mathrm{y}} = \hat{\alpha} + \hat{\beta}_1(x_1 - \bar{x}_1) + \hat{\beta}_2(x_2 - \bar{x}_2) + \cdots + \hat{\beta}_k(x_k - \bar{x}_k)$$

It is clear that the fitted regression plane passes through the point $(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k, \bar{y})$.

Adapting the expression for *SSE* (3.24) to the centered model with centered $\hat{\mathrm{y}}$'s, we obtain

$$SSE = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1' \mathrm{X}_c' \mathrm{y} \tag{3.39}$$

which turns out to be equal to $SSE = \mathrm{y}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\mathrm{y}$ (see Problem 3.19).

We can use (3.36)–(3.38) to express $\hat{\beta}_1$ and $\hat{\beta}_0$ in terms of sample variances and covariances, which will be useful in comparing these estimators with those for the random-*x* case in Chapter 6. We first define a sample covariance matrix for the *x* variables and a vector of sample covariances between *y* and the *x*'s.

$$S_{xx} = \begin{pmatrix} S_1^2 & S_{12} & \cdots & S_{1k} \\ S_{21} & S_2^2 & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_k^2 \end{pmatrix}, \quad S_{yx} = \begin{pmatrix} S_{y1} \\ S_{y2} \\ \vdots \\ S_{yk} \end{pmatrix} \tag{3.40}$$

where, S_i^2, S_{ij}, and S_{yi} are analogous to S^2 and S_{xy}; for example

$$S_2^2 = \frac{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}{n-1} \tag{3.41}$$

$$S_{12} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{n-1} \tag{3.42}$$

$$S_{y2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{i2} - \bar{x}_2)}{n-1} \tag{3.43}$$

with $\bar{x}_2 = \sum_{i=1}^{n} x_{i2} / n$. However, since the *x*'s are fixed, these sample variances and covariances do not estimate population variances and covariances. If the x's were random variables, as in Chapter 6, the s_i^2, s_{ij}, and s_{yi} values would estimate population parameters.

To express $\hat{\beta}_1$ and $\hat{\beta}_0$ in terms of S_{xx} and s_{yx}, we first write S_{xx} and s_{yx} in terms of the centered matrix X_c:

$$S_{xx} = \frac{\mathrm{X}_c'\mathrm{X}_c}{n-1} \tag{3.44}$$

$$s_{yx} = \frac{\mathrm{X}_c'\mathrm{y}}{n-1} \tag{3.45}$$

Note that $\mathrm{X}_c'\mathrm{y}$ in (3.45) contains terms of the form $\sum_{i=1}^n (x_{ij} - \bar{x}_j) y_i$ rather than $\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y})$ as in (3.43). It can readily be shown that $\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y}) = \sum_{i=1}^n (x_{ij} - \bar{x}_j) y_i$ (see Problem 6.2).

From (3.37), (3.44), and (3.45), we have

$$\hat{\beta}_1 = (n-1)(\mathrm{X}_c'\mathrm{X}_c)^{-1}\frac{\mathrm{X}_c'\mathrm{y}}{n-1} = \left(\frac{\mathrm{X}_c'\mathrm{X}_c}{n-1}\right)^{-1}\frac{\mathrm{X}_c'\mathrm{y}}{n-1} = S_{xx}^{-1}s_{yx} \tag{3.46}$$

and from (3.38) and (3.46), we obtain

$$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1'\bar{\mathrm{x}} = \bar{y} - s_{yx}'S_{xx}^{-1}\bar{\mathrm{x}} \tag{3.47}$$

Example 3.5: For the data in Table 3.1, we calculate $\hat{\beta}_1$ and $\hat{\beta}_0$ using (3.46) and (3.47).

$$\begin{aligned}
\hat{\beta}_1 &= S_{xx}^{-1}s_{yx} \\
&= \begin{pmatrix} 6.4242 & 8.5455 \\ 8.5455 & 12.4545 \end{pmatrix}^{-1}\begin{pmatrix} 8.3636 \\ 9.7273 \end{pmatrix} \\
&= \begin{pmatrix} 3.0118 \\ -1.2855 \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - s_{yx}'S_{xx}^{-1}\bar{\mathrm{x}} \\
&= 7.50 - (3.0118, -1.2855)\begin{pmatrix} 4.3333 \\ 8.5000 \end{pmatrix} \\
&= 7.50 - 2.1246 = 5.3754
\end{aligned}$$

These values are the same as those obtained in Example 3.3.1a.

Example 3.5[The program name ta8.m] | Applications using MATLAB

```
clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]';
x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]';
n=length(x1);
X=[x1 x2];k=2;
My=mean(y),Mx=mean(X)'
Sxx=[cov(x1,x2)],
Syx1=cov(y,x1);
Syx2=cov(y,x2);
Syx=[Syx1(1,2);Syx2(1,2)]
beta1=inv(Sxx)*Syx,
beta0=My-Syx'*inv(Sxx)*Mx
```

Ans.

```
My =           Mx =            Sxx =                    Syx =

     7.5           4.3333         6.4242     8.5455         8.3636
                   8.5            8.5455     12.455         9.7273

beta1 =                        beta0 =

      3.0118                         5.3754
     -1.2855
```

3.6: Normal Model

3.6.1 Assumptions

Thus far we have made no normality assumptions about the random variables $y_1, y_2, \cdots, y_n$. To the assumptions in Section 3.2, we now add that

$$\mathrm{y} \quad is \quad N_n\left(\mathrm{X}\beta, \sigma^2\,\mathrm{I}\right) \text{ or } \varepsilon \quad is \quad N_n\left(0, \sigma^2\,\mathrm{I}\right)$$

Under normality, $\sigma_{ij} = 0$ implies that the **y** (or ε) variables are independent, as well as uncorrelated.

3.6.2: Maximum Likelihood Estimators for β and σ^2

With the normality assumption, we can obtain maximum likelihood estimators. The likelihood function is the joint density of the *y*'s, which we denote by $L(\beta, \sigma^2)$. We seek values of the unknown β and σ^2 that maximize $L(\beta, \sigma^2)$ for the given *y* and *x* values in the sample.

In the case of the normal density function, it is possible to find maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$ by differentiation. Because the normal density involves a product and an exponential, it is simpler to work with $\text{Ln}\, L(\beta, \sigma^2)$, which achieves its maximum for the same values of β and σ^2 as does $L(\beta, \sigma^2)$.

The maximum likelihood estimators for β and σ^2 are given in the following theorem.

Theorem 3.6a: If **y** is $N_n(\mathrm{X}\beta, \sigma^2 \mathrm{I})$, where **X** is $n \times (k+1)$ of rank $k+1 < n$, the maximum likelihood estimators of β and σ^2 are

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y} \tag{3.48}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right)'\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right) \tag{3.49}$$

Proof: We sketch the proof. For the remaining steps, see Problem 3.21. The likelihood function (joint density of $y_1, y_2, \cdots, y_n$):

$$L(\beta, \sigma^2) = f(\mathrm{y}; \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\left|\sigma^2 \mathrm{I}\right|^{1/2}} e^{-\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)'\left(\sigma^2\right)^{-1}\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)/2}$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} e^{-\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)'\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)/2\sigma^2} \tag{3.50}$$

[Since the y_i's are independent, $L(\beta, \sigma^2)$ can also be obtained as $\prod_{i=1}^{n} f\left(y_i; \mathrm{x}_i'\beta, \sigma^2\right)$.] Then $\text{Ln}\, L(\beta, \sigma^2)$ becomes

$$\text{Ln}\, L(\beta, \sigma^2) = -\frac{n}{2}\text{Ln}(2\pi) - \frac{n}{2}\text{Ln}\left(\sigma^2\right) - \frac{1}{2\sigma^2}\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right)'\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right) \tag{3.51}$$

Taking the partial derivatives of $\text{Ln}\, L(\beta, \sigma^2)$ with respect to β and σ^2 and setting the results equal to zero will produce (3.48) and (3.49). To verify that $\hat{\beta}$ maximizes (3.50) or (3.51), see (3.10).

The maximum likelihood estimator $\hat{\beta}$ in (3.48) is the same as the least-squares estimator $\hat{\beta}$ in Theorem 3.3a. The estimator $\hat{\sigma}^2$ in (3.49) is biased since the denominator is *n* rather than *n*- *k* -1.We often use the unbiased estimator S^2 given in (3.23) or (3.24).

3.6.3: Properties of $\hat{\beta}$ and $\hat{\sigma}^2$

We now consider some properties of $\hat{\beta}$ and $\hat{\sigma}^2$ (or S^2) under the normal model. The distributions of $\hat{\beta}$ and $\hat{\sigma}^2$ are given in the following theorem.

Theorem 3.6b: Suppose that **y** is $N_n(X\beta, \sigma^2 I)$, where **X** is $n \times (k+1)$ of rank $k+1<n$, and $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$. Then the maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$ given in Theorem 3.6a have the following distributional properties:

(i) $\hat{\beta}$ is $N_{k+1}(\beta, \sigma^2 (X'X)^{-1})$.

(ii) $n\hat{\sigma}^2/\sigma^2$ is $\chi^2_{(n-k-1)}$, or equivalently, $(n-k-1)S^2/\sigma^2$ is $\chi^2_{(n-k-1)}$.

(iii) $\hat{\beta}$ and $\hat{\sigma}^2$ (or S^2) are independent.

Proof:

(i) Since $\hat{\beta} = (X'X)^{-1}X'y$ is a linear function of **y** of the form $\hat{\beta} = Ay$, where $A = (X'X)^{-1}X'$ is a constant matrix, $\hat{\beta}$ is $N_{k+1}(\beta, \sigma^2 (X'X)^{-1})$.

(ii) If **y** is $N_p(\mu, \sigma^2 I)$, then $y'Ay/\sigma^2$ is $\chi^2_{(r, \mu' A \mu/2\sigma^2)}$ if and only if A is idempotent of rank *r*.

(iii) If **y** is $N_p(\mu, \sigma^2 I)$, then **By** and $y'Ay$ are independent if and only if **BA = 0**

Another property of $\hat{\beta}$ and $\hat{\sigma}^2$ under normality is that they are sufficient statistics. Intuitively, a statistic is sufficient for a parameter if the statistic summarizes all the information in the sample about the parameter. Sufficiency of $\hat{\beta}$ and $\hat{\sigma}^2$ can be established by the Neyman factorization theorem [see Hogg and Craig (1995, p. 318) or Graybill (1976, pp. 69–70)], which states $\hat{\beta}$ and $\hat{\sigma}^2$ are jointly sufficient for β and σ^2 if the density $f(y;\beta,\sigma^2)$ can be factored as $f(y;\beta,\sigma^2) = g(\hat{\beta}, \hat{\sigma}^2, \beta, \sigma^2)h(y)$, where $h(\mathbf{y})$ does not depend on β and σ^2. The following theorem shows that $\hat{\beta}$ and $\hat{\sigma}^2$ satisfy this criterion.

Theorem 3.6c: If **y** is $N_n(\mathrm{X}\beta, \sigma^2 \mathrm{I})$, then $\hat{\beta}$ and $\hat{\sigma}^2$ are jointly sufficient for β and σ^2.

Proof: The density $f(\mathrm{y};\beta,\sigma^2)$ is given in (3.50). In the exponent, we add and subtract $\mathrm{X}\hat{\beta}$ to obtain

$$
\begin{aligned}
(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta) &= \left(\mathrm{y}-\mathrm{X}\hat{\beta}+\mathrm{X}\hat{\beta}-\mathrm{X}\beta\right)'\left(\mathrm{y}-\mathrm{X}\hat{\beta}+\mathrm{X}\hat{\beta}-\mathrm{X}\beta\right) \\
&= \left[\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)+\mathrm{X}\left(\hat{\beta}-\beta\right)\right]'\left[\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)+\mathrm{X}\left(\hat{\beta}-\beta\right)\right]
\end{aligned}
$$

Expanding this in terms of $\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)$ and $\mathrm{X}\left(\hat{\beta}-\beta\right)$, we obtain four terms, two of which vanish because of the normal equations $\mathrm{X}'\mathrm{X}\hat{\beta} = \mathrm{X}'\mathrm{y}$. The result is

$$
\begin{aligned}
(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta) &= \left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)'\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)+\left(\hat{\beta}-\beta\right)'\mathrm{X}'\mathrm{X}\left(\hat{\beta}-\beta\right) \qquad (3.52) \\
&= n\hat{\sigma}^2 + \left(\hat{\beta}-\beta\right)'\mathrm{X}'\mathrm{X}\left(\hat{\beta}-\beta\right)
\end{aligned}
$$

We can now write the density (3.50) as

$$
f\left(\mathrm{y};\beta,\sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} e^{-[n\hat{\sigma}^2 + \left(\hat{\beta}-\beta\right)'\mathrm{X}'\mathrm{X}\left(\hat{\beta}-\beta\right)]/2\sigma^2}
$$

which is of the form

$$
f\left(\mathrm{y};\beta,\sigma^2\right) = g\left(\hat{\beta},\hat{\sigma}^2,\beta,\sigma^2\right)h(\mathrm{y})
$$

where $h(\mathbf{y}) = 1$. Therefore, by the Neyman factorization theorem, $\hat{\beta}$ and $\hat{\sigma}^2$ are jointly sufficient for β and σ^2.

Note that $\hat{\beta}$ and $\hat{\sigma}^2$ are jointly sufficient for β and σ^2, not independently sufficient; that is, $f(\mathrm{y};\beta,\sigma^2)$ does not factor into the form $f\left(\mathrm{y};\beta,\sigma^2\right) = g_1\left(\hat{\beta},\beta\right)g_2\left(\hat{\sigma}^2,\sigma^2\right)h(\mathrm{y})$. Also note that because $s^2 = n\hat{\sigma}^2/(n-k-1)$, the proof to Theorem 3.6c can be easily modified to show that $\hat{\beta}$ and s^2 are also jointly sufficient for β and σ^2.

Since $\hat{\beta}$ and S^2 are sufficient, no other estimators can improve on the information they extract from the sample to estimate β and σ^2. Thus, it is not surprising that $\hat{\beta}$ and S^2 are minimum variance unbiased estimators (each $\hat{\beta}_j$ in $\hat{\beta}$ has minimum variance). This result is given in the following theorem.

Theorem 3.6d: If **y** is $N_n(X\beta, \sigma^2 I)$, then $\hat{\beta}$ and S^2 have minimum variance among all unbiased estimators.

Proof: See Graybill (1976, p. 176) or Christensen (1996, pp. 25–27).

In Theorem 3.3d, the elements of $\hat{\beta}$ were shown to have minimum variance among all linear unbiased estimators. With the normality assumption added in Theorem 3.6d, the elements of $\hat{\beta}$ have minimum variance among all *unbiased* estimators. Similarly, by Theorem 3.3g, S^2 has minimum variance among all *quadratic unbiased* estimators. With the added normality assumption in Theorem 3.6d, S^2 has minimum variance among all *unbiased* estimators.

The following corollary to Theorem 3.6d is analogous to Corollary 1 of Theorem 3.3d.

Corollary 1: If **y** is $N_n(X\beta, \sigma^2 I)$, then the minimum variance unbiased estimator of $a'\beta$ is $a'\hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimator given in (3.48).

3.7: R^2 In Fixed-x Regression

In (3.39), we have $SSE = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1' X_c' y$. Thus the corrected total sum of squares $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ can be partitioned as

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \hat{\beta}_1' X_c' y + SSE \qquad (3.53)$$

$$SST = SSR + SSE$$

where $SSR = \hat{\beta}_1' X_c' y$ is the regression sum of squares. From (3.37), we obtain $X_c' y = X_c' X_c \hat{\beta}_1$, and multiplying this by $\hat{\beta}_1'$ gives $\hat{\beta}_1' X_c' y = \hat{\beta}_1' X_c' X_c \hat{\beta}_1$. Then $SSR = \hat{\beta}_1' X_c' y$ can be written as

$$SSR = \hat{\beta}_1' X_c' X_c \hat{\beta}_1 = (X_c \hat{\beta}_1)'(X_c \hat{\beta}_1) \qquad (3.54)$$

In this form, it is clear that *SSR* is due to $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \cdots, \beta_k)'$.

The proportion of the total sum of squares due to regression is

$$R^2 = \frac{\hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{X}_c \hat{\boldsymbol{\beta}}_1}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{SSR}{SST} \tag{3.55}$$

which is known as the *coefficient of determination* or the *squared multiple correlation*. The ratio in (3.55) is a measure of model fit and provides an indication of how well the *x*'s predict *y*.

The partitioning in (3.53) can be rewritten as the identity

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2 = \left(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2\right) + \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}\right)$$

$$= SSR + SSE$$

which leads to an alternative expression for R^2:

$$R^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2} \tag{3.56}$$

The positive square root *R* obtained from (3.55) or (3.56) is called the *multiple correlation coefficient*. If the *x* variables were random, *R* would estimate a population multiple correlation (see Section (6.4)).

We list some properties of R^2 and *R*:

1. The range of R^2 is $0 \le R^2 \le 10$. If all the $\hat{\beta}_j$'s were zero, except for $\hat{\beta}_0$, R^2 would be 0. (This event has probability 0 for continuous data.) If all the *y* values fell on the fitted surface, that is, if $y_i = \hat{y}_i$, $i = 1, 2, \cdots, n$, then R^2 would be 1.
2. $R = r_{y\hat{y}}$; that is, the multiple correlation is equal to the simple correlation [see (2.18)] between the observed y_i's and the fitted $\hat{y}_i$'s.
3. Adding a variable *x* to the model increases (cannot decrease) the value of R^2.
4. If $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, then

$$R^2 = \frac{k}{n-1} \tag{3.57}$$

Note that the $\hat{\beta}_j$'s will not be 0 when the β_j's are 0.

5. R^2 Can not be partitioned into k components, each of which is uniquely attributable to an x_j, unless the x's are mutually orthogonal, that is, $\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{im}-\bar{x}_m)=0$ for $j \neq m$.
6. R^2 is invariant to full-rank linear transformations on the x's and to a scale change on y (but not invariant to a joint linear transformation including y and the x's).

In properties 3 and 4 we see that if k is a relatively large fraction of n, it is possible to have a large value of R^2 that is not meaningful. In this case, x's that do not contribute to predicting y may appear to do so in a particular example, and the estimated regression equation may not be a useful estimator of the population model. To correct for this tendency, an adjusted R^2, denoted by R_a^2, was proposed by Ezekiel (1930). To obtain R_a^2, we first subtract $k/(n-1)$ in (3.57) from R^2 in order to correct for the bias when $\beta_1=\beta_2=\cdots=\beta_k=0$. This correction, however, would make R_a^2 too small when the β's are large, so a further modification is made so that $R_a^2=1$ when $R^2=1$. Thus R_a^2 is defined as

$$R_a^2=\frac{\left(R^2-\frac{k}{n-1}\right)(n-1)}{n-k-1}=\frac{(n-1)R^2-k}{n-k-1} \qquad (3.58)$$

Example 3.7: For the data in Table 3.1 in Example 3.2, we obtain R^2 by (3.56) and R_a^2 by (3.58). The values of $\hat{\beta}'X'y$ and $y'y$ are given in Example 3.3.3.

$$R^2=\frac{\hat{\beta}'X'y-n\bar{y}^2}{y'y-n\bar{y}^2}$$

$$=\frac{814.5410-12(7.5)^2}{840-12(7.5)^2}$$

$$=\frac{139.5410}{165.0000}=0.8457$$

$$R_a^2=\frac{(n-1)R^2-k}{n-k-1}=\frac{(11)(0.8457)-2}{9}=0.8114$$

Example 3.7[The program name ta9.m]	Applications using MATLAB

```
clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]';
x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]';
n=length(x1);X=[ones(size(x1)) x1 x2];k=2;
My=mean(y);beta=X\y;
RS=(beta'*X'*y-n*My^2)/(y'*y-n*My^2)
RSa=((n-1)*RS'-k)/(n-k-1)
```

Ans.

```
RS =                    RSa =
      0.8457                  0.81141
```

Using (3.44) and (3.46), we can express R^2 in (3.55) in terms of sample variances and covariances:

$$R^2 = \frac{\hat{\beta}_1' \mathrm{X}_c' \mathrm{X}_c \hat{\beta}_1}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{S_{yx}' S_{xx}^{-1} (n-1) S_{xx} S_{xx}^{-1} S_{yx}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{S_{yx}' S_{xx}^{-1} S_{yx}}{S_y^2} \qquad (3.59)$$

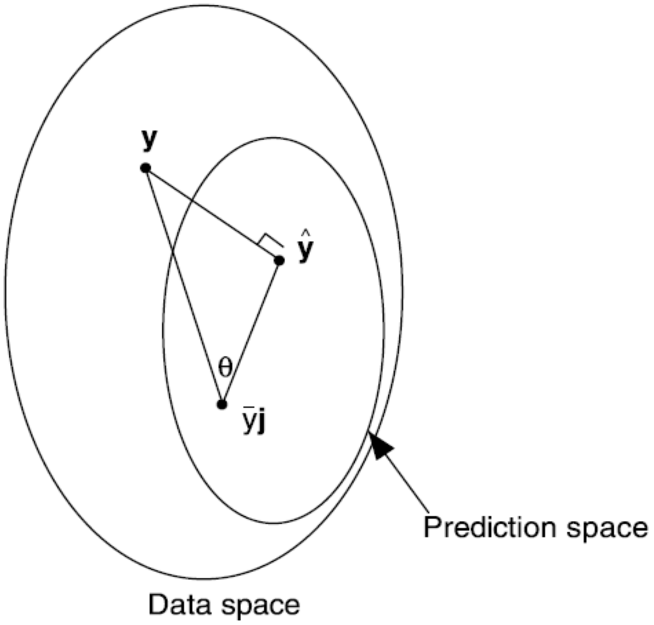


Figure 3.5: Multiple correlation R as cosine of θ, the angle between $\mathrm{y} - \bar{y}\,\mathrm{j}$ and $\hat{\mathrm{y}} - \bar{y}\,\mathrm{j}$

This form of R^2 will facilitate a comparison with R^2 for the random-x case in Section (6.4) [see (6.34)].

Geometrically, R is the cosine of the angle θ between y and $\hat{\mathrm{y}}$ corrected for their means. The mean of $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ is $\bar{y}$, the same as the mean of $y_1, y_2, \cdots, y_n$ (see Problem 3.30). Thus the centered forms of y and $\hat{\mathrm{y}}$ are $\mathrm{y} - \bar{y}\mathrm{j}$ and $\hat{\mathrm{y}} - \bar{y}\mathrm{j}$. The angle between them is illustrated in Figure 3.5. (Note that $\bar{y}\mathrm{j}$ is in the estimation space since it is a multiple of the first column of **X**.)

To show that $\cos\theta$ is equal to the square root of R^2 as given by (3.56), we use the cosine of the angle between two vectors:

$$\cos\theta = \frac{(\mathrm{y} - \bar{y}\mathrm{j})'(\hat{\mathrm{y}} - \bar{y}\mathrm{j})}{\sqrt{[(\mathrm{y} - \bar{y}\mathrm{j})'(\mathrm{y} - \bar{y}\mathrm{j})][(\hat{\mathrm{y}} - \bar{y}\mathrm{j})'(\hat{\mathrm{y}} - \bar{y}\mathrm{j})]}} \tag{3.60}$$

To simplify (3.60), we use the identity $(\mathrm{y} - \bar{y}\mathrm{j}) = (\hat{\mathrm{y}} - \bar{y}\mathrm{j}) + (\mathrm{y} - \hat{\mathrm{y}})$, which can also be seen geometrically in Figure 3.5. The vectors $\hat{\mathrm{y}} - \bar{y}\mathrm{j}$ and $\mathrm{y} - \hat{\mathrm{y}}$ on the right side of this identity are orthogonal since $\hat{\mathrm{y}} - \bar{y}\mathrm{j}$ is in the prediction space. Thus the numerator of (3.60) can be written as

$$\begin{aligned}(\mathrm{y} - \bar{y}\mathrm{j})'(\hat{\mathrm{y}} - \bar{y}\mathrm{j}) &= [(\hat{\mathrm{y}} - \bar{y}\mathrm{j}) + (\mathrm{y} - \hat{\mathrm{y}})]'(\hat{\mathrm{y}} - \bar{y}\mathrm{j}) \\ &= (\hat{\mathrm{y}} - \bar{y}\mathrm{j})'(\hat{\mathrm{y}} - \bar{y}\mathrm{j}) + (\mathrm{y} - \hat{\mathrm{y}})'(\hat{\mathrm{y}} - \bar{y}\mathrm{j}) \\ &= (\hat{\mathrm{y}} - \bar{y}\mathrm{j})'(\hat{\mathrm{y}} - \bar{y}\mathrm{j}) + 0\end{aligned}$$

Then (3.60) becomes

$$\cos\theta = \frac{\sqrt{(\hat{\mathrm{y}} - \bar{y}\mathrm{j})'(\hat{\mathrm{y}} - \bar{y}\mathrm{j})}}{\sqrt{[(\mathrm{y} - \bar{y}\mathrm{j})'(\mathrm{y} - \bar{y}\mathrm{j})]}} \tag{3.61}$$

which is easily shown to be the square root of R^2 as given by (3.56). This is equivalent to property 2 following (3.56): $R = r_{y\hat{y}}$.

We can write (3.61) in the form

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST}$$

In which $SSR=\sum_{i=1}^{n}(y_i-\bar{y})^2$ is a sum of squares for the $\hat{y}_i$'s. Then the partitioning $SST=SSR+SSE$ below (3.53) can be written as

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

which is analogous to (2.17) for simple linear regression.

3.8: Generalized Least Squares: $Cov(\mathrm{y})=\sigma^2\,\mathrm{V}$

We now consider models in which the y variables are correlated or have differing variances, so that $Cov(\mathrm{y})\neq\sigma^2\,\mathrm{I}$. In simple linear regression, larger values of x_i may lead to larger values of $Var(y_i)$. In either simple or multiple regression, if $y_1,y_2,\cdots,y_n$ occur at sequential points in time, they are typically correlated. For cases such as these, in which the assumption $Cov(\mathrm{y})=\sigma^2\,\mathrm{I}$ is no longer appropriate, we use the model

$$\mathrm{y}=\mathrm{X}\beta+\varepsilon,\quad E(\mathrm{y})=\mathrm{X}\beta\quad ,Cov(\mathrm{y})=\Sigma=\sigma^2\,\mathrm{V} \qquad (3.62)$$

Where **X** is full-rank and **V** is a known positive definite matrix. The usage $\Sigma=\sigma^2\,\mathrm{V}$ permits estimation of σ^2 in some convenient contexts (see Examples 3.8.1 and 3.8.2). The $n\times n$ matrix **V** has *n* diagonal elements and C_2^n elements above (or below) the diagonal. If **V** were unknown, these C_2^n+n distinct elements could not be estimated from a sample of *n* observations. In certain applications, a simpler structure for **V** is assumed that permits estimation. Such structures are illustrated in Examples 3.8.1 and 3.8.2.

3.8.1: Estimation of β and σ^2 when $Cov(\mathrm{y})=\sigma^2\,\mathrm{V}$

In the following theorem we give estimators of β and σ^2 for the model in (3.62).

Theorem 3.8a: Let $\mathrm{y}=\mathrm{X}\beta+\varepsilon$, let $E(\mathrm{y})=\mathrm{X}\beta$, and let $Cov(\mathrm{y})=Cov(\varepsilon)=\sigma^2\,\mathrm{V}$, where **X** is a full-rank matrix and **V** is a known positive definite matrix. For this model, we obtain the following results:

(i) The best linear unbiased estimator (BLUE) of β is

$$\hat{\beta}=\left(\mathrm{X}'\,\mathrm{V}^{-1}\mathrm{X}\right)^{-1}\mathrm{X}'\mathrm{V}^{-1}\mathrm{y} \qquad (3.63)$$

(ii) The covariance matrix for $\hat{\beta}$ is

$$Cov(\hat{\beta}) = \sigma^2 (X'V^{-1}X)^{-1} \qquad (3.64)$$

(iii) An unbiased estimator of σ^2 is

$$S^2 = \frac{(y - X\hat{\beta})' V^{-1} (y - X\hat{\beta})}{n-k-1} \qquad (3.65)$$

$$S^2 = \frac{y'[V^{-1} - V^{-1} X(X'V^{-1}X)^{-1} X'V^{-1}]y}{n-k-1} \qquad (3.66)$$

Where $\hat{\beta}$ is as given by (3.63).

Proof: We prove part (i). For parts (ii) and (iii), see Problems (3.32) and (3.33).

Since **V** is positive definite, there exists an $n \times n$ nonsingular matrix **P** such that $V = PP'$. Multiplying $y = X\beta + \varepsilon$ by P^{-1}, we obtain $P^{-1}y = P^{-1}X\beta + P^{-1}\varepsilon$, for which $E(P^{-1}\varepsilon) = P^{-1} E(\varepsilon) = P^{-1}\,0 = 0$ and

$$\begin{aligned} Cov(P^{-1}\varepsilon) &= P^{-1}\,Cov(\varepsilon)(P^{-1})' \\ &= P^{-1}\sigma^2 V(P^{-1})' = \sigma^2 P^{-1}PP'(P')^{-1} = \sigma^2 I \end{aligned}$$

Thus the assumptions for Theorem 3.3d are satisfied for the model $P^{-1}y = P^{-1}X\beta + P^{-1}\varepsilon$, and the least-squares estimator $\hat{\beta} = [(P^{-1}X)'(P^{-1}X)]^{-1}(P^{-1}X)' P^{-1}y$ is BLUE. This can be written as

$$\begin{aligned} \hat{\beta} &= [X'(P^{-1})' P^{-1}X]^{-1} X'(P^{-1})' P^{-1}y \\ &= [X'(P')^{-1} P^{-1}X]^{-1} X'(P')^{-1} P^{-1}y \\ &= [X'(PP')^{-1} X]^{-1} X'(PP')^{-1} y \\ &= (X'V^{-1}X)^{-1} X'V^{-1}y \end{aligned}$$

Note that, since **X** is full-rank, $X'V^{-1}X$ is positive definite. The estimator $\hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}y$ is usually called the *generalized least-squares* estimator. The same estimator is obtained under a normality assumption.

Theorem 3.8b: If **y** is $N_n(X\beta, \sigma^2 I)$, where **X** is full-rank and **V** is a known positive definite matrix, where **X** is $n \times (k+1)$ of rank k +1, then the maximum likelihood estimators for β and σ^2 are

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(y - X\hat{\beta}\right)' V^{-1}\left(y - X\hat{\beta}\right)$$

Proof: The likelihood function is

$$L\left(\beta, \sigma^2\right) = \frac{1}{(2\pi)^{n/2}\left|\sigma^2 V\right|^{1/2}} e^{-(y-X\beta)'\left(\sigma^2 V\right)^{-1}(y-X\beta)/2}$$

And $\left|\sigma^2 V\right| = \left(\sigma^2\right)^n |V|$. Hence

$$L\left(\beta, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}|V|^{1/2}} e^{-(y-X\beta)' V^{-1}(y-X\beta)/2\sigma^2}$$

The results can be obtained by differentiation of $\mathrm{Ln}L\left(\beta,\sigma^2\right)$ with respect to β and with respect to σ^2.

We illustrate an application of generalized least squares.

Example 3.8.1: Consider the centered model in (3.32)

$$y = \left(j, X_c\right)\begin{pmatrix}\alpha \\ \beta_1\end{pmatrix} + \varepsilon$$

With covariance pattern

$$\Sigma = \sigma^2[(1-\rho)I + \rho J] = \sigma^2 V \qquad (3.67)$$

$$= \sigma^2\begin{pmatrix}1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1\end{pmatrix}$$

In which all variables have the same variance σ^2 and all pairs of variables have the same correlation ρ. Assume for certain repeated measures and intraclass correlation designs.

By (3.63), we have

$$\hat{\beta} = \begin{pmatrix}\hat{\alpha} \\ \hat{\beta}_1\end{pmatrix} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y$$

For the centered model with $X = \left(j, X_c\right)$, the matrix $X'V^{-1}X$ becomes

$$\mathrm{X}'\mathrm{V}^{-1}\mathrm{X} = \begin{pmatrix} \mathrm{j}' \\ \mathrm{X}_\mathrm{c}' \end{pmatrix} \mathrm{V}^{-1}(\mathrm{j}, \mathrm{X}_c)$$

$$= \begin{pmatrix} \mathrm{j}'\mathrm{V}^{-1}\mathrm{j} & \mathrm{j}'\mathrm{V}^{-1}\mathrm{X}_c \\ \mathrm{X}_\mathrm{c}'\mathrm{V}^{-1}\mathrm{j} & \mathrm{X}_\mathrm{c}'\mathrm{V}^{-1}\mathrm{X}_c \end{pmatrix}$$

The inverse of the $n \times n$ matrix $\mathrm{V} = (1-\rho)\mathrm{I} + \rho\mathrm{J}$ in (3.67) is given by

$$\mathrm{V}^{-1} = a(\mathrm{I} - b\rho\mathrm{J}) \qquad (3.68)$$

Where $a = 1/(1-\rho)$ and $b = 1/[1+(n-1)\rho]$. Using V^{-1} in (3.68), $\mathrm{X}'\mathrm{V}^{-1}\mathrm{X}$ becomes

$$\mathrm{X}'\mathrm{V}^{-1}\mathrm{X} = \begin{pmatrix} bn & 0' \\ 0 & a\mathrm{X}_\mathrm{c}'\mathrm{X}_c \end{pmatrix} \qquad (3.69)$$

Similarly

$$\mathrm{X}'\mathrm{V}^{-1}\,\mathrm{y} = \begin{pmatrix} bn\bar{y} \\ a\mathrm{X}_\mathrm{c}'\,\mathrm{y} \end{pmatrix} \qquad (3.70)$$

We therefore have

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{pmatrix} = \left(\mathrm{X}'\mathrm{V}^{-1}\mathrm{X}\right)^{-1}\mathrm{X}'\mathrm{V}^{-1}\mathrm{y} = \begin{pmatrix} \bar{y} \\ (\mathrm{X}_\mathrm{c}'\mathrm{X}_c)^{-1}\mathrm{X}_\mathrm{c}'\,\mathrm{y} \end{pmatrix}$$

which is the same as (3.36) and (3.37). Thus the usual least-squares estimators are BLUE for a covariance structure with equal variances and equal correlations.

3.8.2: Misspecification of the Error Structure

Suppose that the model is $\mathrm{y} = \mathrm{X}\beta + \varepsilon$ with $Cov(\mathrm{y}) = \sigma^2\mathrm{V}$, as in (3.62), and we mistakenly (or deliberately) use the ordinary least-squares estimator $\hat{\beta}^* = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ in (3.6), which we denote here by $\hat{\beta}^*$ to distinguish it from the BLUE estimator $\hat{\beta} = \left(\mathrm{X}'\mathrm{V}^{-1}\mathrm{X}\right)^{-1}\mathrm{X}'\mathrm{V}^{-1}\mathrm{y}$ in (3.63). Then the mean vector and covariance matrix for $\hat{\beta}^*$ are

$$E\left(\hat{\beta}^*\right) = \beta \qquad (3.71)$$

$$Cov\left(\hat{\beta}^*\right) = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{V}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} \qquad (3.72)$$

Thus the ordinary least-squares estimators are unbiased, but the covariance matrix differs from (3.64). Because of Theorem 3.8a(i), the variances of the $\hat{\beta}_j^*$'s in (3.72) cannot be smaller than the variances in $Cov(\hat{\beta}) = \sigma^2 (X'V^{-1}X)^{-1}$ in (3.64). This is illustrated in the following example.

Example 3.8.2: Suppose that we have a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $Var(y_i) = \sigma^2 x_i$ and $Cov(y_i, y_j) = 0$ for $i \neq j$. Thus

$$Cov(y) = \sigma^2 V = \sigma^2 \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_n \end{pmatrix}$$

This is an example of weighted least squares, which typically refers to the case where **V** is diagonal with functions of the *x*'s on the diagonal. In this case

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

And by (3.63), we have

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X'V^{-1}X)^{-1} X'V^{-1}y$$

$$= \frac{1}{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n 1/x_i\right) - n^2} \begin{pmatrix} \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i/x_i\right) - n\left(\sum_{i=1}^n y_i\right) \\ \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n 1/x_i\right) - n\left(\sum_{i=1}^n y_i/x_i\right) \end{pmatrix} \qquad (3.73)$$

The covariance matrix for $\hat{\beta}$ is given by (3.64):

$$Cov(\hat{\beta}) = \sigma^2 (X'V^{-1}X)^{-1}$$

$$= \frac{\sigma^2}{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n 1/x_i\right) - n^2} \begin{pmatrix} \left(\sum_{i=1}^n x_i\right) & -n \\ -n & \left(\sum_{i=1}^n 1/x_i\right) \end{pmatrix} \qquad (3.74)$$

If we use the ordinary least-squares estimator $\hat{\beta}^* = (X'X)^{-1} X'y$ as given in (2.5) and (2.6) or in (3.12) in Example 3.3.1b, then $Cov(\hat{\beta}^*)$ is given by (3.72); that is,

$$Cov\left(\hat{\beta}^{*}\right)=\sigma^{2}(X'X)^{-1}X'VX(X'X)^{-1}$$

$$=\sigma^{2}\begin{pmatrix} n & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \end{pmatrix}^{-1}\begin{pmatrix} \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \\ \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i^3 \end{pmatrix}\begin{pmatrix} n & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \end{pmatrix}^{-1}$$

$$=\sigma^{2}c\begin{pmatrix} \sum_{i=1}^{n}x_i^3\left(\sum_{i=1}^{n}x_i\right)^2-\sum_{i=1}^{n}x_i\left(\sum_{i=1}^{n}x_i^2\right)^2 & n\left(\sum_{i=1}^{n}x_i^2\right)^2-n\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i^3 \\ n\left(\sum_{i=1}^{n}x_i^2\right)^2-n\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i^3 & n^2\sum_{i=1}^{n}x_i^3-2n\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i^2+\left(\sum_{i=1}^{n}x_i\right)^3 \end{pmatrix} \qquad (3.75)$$

where $c=1\Big/\left[n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right]^2$. The variance of the estimator $\hat{\beta}_1$ is given by the lower right diagonal element of (3.75):

$$Var\left(\hat{\beta}_1^{*}\right)=\sigma^{2}\frac{n^2\sum_{i=1}^{n}x_i^3-2n\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i^2+\left(\sum_{i=1}^{n}x_i\right)^3}{\left[n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right]^2} \qquad (3.76)$$

And the variance of the estimator $\hat{\beta}_1^{*}$ is given by the corresponding element of (3.74):

$$Var\left(\hat{\beta}_1\right)=\sigma^{2}\frac{\sum_{i=1}^{n}(1/x_i)}{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}(1/x_i)-n^2} \qquad (3.77)$$

Consider the following seven values of *x*: 1, 2, 3, 4, 5, 6, 7. Using (3.76), we obtain $Var\left(\hat{\beta}_1^{*}\right)=0.1429\sigma^{2}$, and from (3.77), we have $Var\left(\hat{\beta}_1\right)=0.1099\sigma^{2}$. Thus for these values of *x*, the use of ordinary least squares yields a slope estimator with a larger variance, as expected.

Further consequences of using a wrong model are discussed in the next section.

3.9: Model Misspecification

In Section 3.8.2, we discussed some consequences of misspecification of $Cov(y)$. We now consider consequences of misspecification of E(**y**). As a framework for discussion, let the model $y=X\beta+\varepsilon$ be partitioned as

$$y=X\beta+\varepsilon=\begin{pmatrix} X_1 & X_2 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}+\varepsilon$$

$$y=X_1\beta_1+X_2\beta_2+\varepsilon \qquad (3.78)$$

If we leave out $X_2\beta_2$ when it should be included (i.e., when $\beta_2 \neq 0$), we are *underfitting*. If we include $X_2\beta_2$ when it should be excluded (i.e., when $\beta_2 = 0$), we are *overfitting*. We discuss the effect of underfitting or overfitting on the bias and the variance of the $\hat{\beta}_j$, $\hat{y}$, and s^2 values.

We first consider estimation of β_1 when *underfitting*. We write the reduced model as

$$y = X_1\beta_1^* + \varepsilon^* \qquad (3.79)$$

Using β_1^* to emphasize that these parameters (and their estimates $\hat{\beta}_1^*$) will be different from β_1 (and $\hat{\beta}_1$) in the *full* model (3.78) (unless the *x*'s are orthogonal; see Corollary 1 to Theorem 3.9a and Theorem 3.10). This was illustrated in Example 3.2. In the following theorem, we discuss the bias in the estimator $\hat{\beta}_1^*$ obtained from (3.79) and give the covariance matrix for $\hat{\beta}_1^*$.

Theorem 3.9a: If we fit the model $y = X_1\beta_1^* + \varepsilon^*$ when the correct model is $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ with $Cov(y) = \sigma^2 I$, then the mean vector and covariance matrix for the least-squares estimator $\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'y$ are as follows:

$$(i) \quad E\left(\hat{\beta}_1^*\right) = \beta_1 + A\beta_2 \; Where \quad A = (X_1'X_1)^{-1}X_1'X_2 \qquad (3.80)$$

$$(ii) \quad Cov\left(\hat{\beta}_1^*\right) = \sigma^2 (X_1'X_1)^{-1} \qquad (3.81)$$

Proof:

$$(i) \quad E\left(\hat{\beta}_1^*\right) = E[(X_1'X_1)^{-1}X_1'y] = (X_1'X_1)^{-1}X_1'E(y)$$

$$= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2)$$

$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$

$$(ii) \quad Cov\left(\hat{\beta}_1^*\right) = Cov[(X_1'X_1)^{-1}X_1'y]$$

$$= (X_1'X_1)^{-1}X_1'(\sigma^2 I)X_1(X_1'X_1)^{-1}$$

$$= \sigma^2 (X_1'X_1)^{-1}$$

Thus, when underfitting, $\hat{\beta}_1^*$ is biased by an amount that depends on the values of the *x*'s in both X_1 and X_2. The matrix $A = (X_1'X_1)^{-1} X_1' X_2$ in (3.81) is called the *alias* matrix.

Corollary 1: If $X_1'X_2 = O$, that is, if the columns of X_1 are orthogonal to the columns of X_2, then $\hat{\beta}_1^*$ is unbiased: $E(\hat{\beta}_1^*) = \beta_1$.

In the next three theorems, we discuss the effect of underfitting oroverfitting on $\hat{y}$, S^2, and the variances of the $\hat{\beta}_j$'s. In some of the proofs we follow Hocking (1996, pp. 245–247).

Let $x_0 = (1, x_{01}, x_{02}, \cdots x_{0k})'$ be a particular value of **x** for which we desire to estimate $E(y_0) = x_0' \beta$. If we partition x_0' into (x_{01}', x_{02}') corresponding to the partitioning $X = (X_1, X_2)$ and $\beta' = (\beta_1', \beta_2')$, then we can use either $\hat{y}_0 = x_0'\hat{\beta}$ or $\hat{y}_{01} = x_{01}'\hat{\beta}_1^*$ to estimate $x_0'\beta$. In the following theorem, we consider the mean of $\hat{y}_{01}$.

Theorem 3.9b: Let $\hat{y}_{01} = x_{01}'\hat{\beta}_1^*$, where $\hat{\beta}_1^* = (X_1'X_1)^{-1} X_1'y$. Then, if $\beta_2 \neq 0$, we obtain

$$E(x_{01}'\hat{\beta}_1^*) = x_{01}'(\beta_1 + A\beta_2) \qquad (3.82)$$

$$= x_0'\beta - (x_{02} - A'x_{01})' \beta_2 \neq x_0'\beta \qquad (3.83)$$

Proof: See Problem 3.43.

In Theorem 3.9b, we see that, when underfitting, $x_{01}'\hat{\beta}_1^*$ is biased for estimating $x_0'\beta$. [When overfitting, $x_0'\hat{\beta}$ is unbiased since $E(x_0'\hat{\beta}) =$ $x_0'\beta =$ $x_{01}'\beta_1 + x_{02}'\beta_2$, which is equal to $x_{01}'\beta_1$ if $\beta_2 = 0$].

In the next theorem, we compare the variances of $\hat{\beta}_j^*$ and $\hat{\beta}_j$, where $\hat{\beta}_j^*$ is from $\hat{\beta}_1^*$ and $\hat{\beta}_j$ is from $\hat{\beta}_1$. We also compare the variances of $x_{01}'\hat{\beta}_1^*$ and $x_0'\hat{\beta}$.

Theorem 3.9c: Let $\hat{\beta} = (X'X)^{-1} X'y$ from the full model be partitioned as $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$, and let $\hat{\beta}_1^* = (X_1'X_1)^{-1} X_1'y$ be the estimator from the reduced model. Then

(i) $Cov(\hat{\beta}_1) - Cov(\hat{\beta}_1^*) = \sigma^2 AB^{-1}A'$, which is a positive definite matrix, where $A = (X_1'X_1)^{-1}X_1'X_2$ and $B = (X_2'X_2)^{-1} - X_2'X_1A$. Thus $Var(\hat{\beta}_j) > Var(\hat{\beta}_j^*)$

(ii) $\quad Var(x_0'\hat{\beta}) \ \geq \ Var(x_{01}'\hat{\beta}_j^*)$

Proof:

(i) Using $X'X$ partitioned to conform to $X = (X_1, X_2)$, we have

$$Cov(\hat{\beta}) = Cov\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} = \sigma^2 (X'X)^{-1} = \sigma^2 \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1}$$

$$= \sigma^2 \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{pmatrix}$$

Where $G_{ij} = X_i'X_j$ and G^{ij} is the corresponding block of the partitioned inverse matrix $(X'X)^{-1}$. Thus $Cov(\hat{\beta}_1) = \sigma^2 G^{11}$. And $G^{11} = G_{11}^{-1} + G_{11}^{-1}G_{12}B^{-1}G_{21}G_{11}^{-1}$, where $B = G_{22} - G_{21}G_{11}^{-1}G_{12}$. By (3.81), $Cov(\hat{\beta}_1^*) = \sigma^2 (X_1'X_1)^{-1} = \sigma^2 G_{11}^{-1}$.Hence

$$Cov(\hat{\beta}_1) - Cov(\hat{\beta}_1^*) = \sigma^2 (G^{11} - G_{11}^{-1})$$

$$= \sigma^2 (G_{11}^{-1} + G_{11}^{-1}G_{12}B^{-1}G_{21}G_{11}^{-1} - G_{11}^{-1})$$

$$= \sigma^2 AB^{-1}A'$$

(ii) $\quad Var(x_0'\hat{\beta}) = \sigma^2 x_0'(X'X)^{-1}x_0$

$$= \sigma^2 (x_{01}', x_{02}') \begin{pmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{pmatrix} \begin{pmatrix} x_{01} \\ x_{02} \end{pmatrix}$$

$$= \sigma^2 (x_{01}'G^{11}x_{01} + x_{01}'G^{12}x_{02} + x_{02}'G^{21}x_{01} + x_{02}'G^{22}x_{02})$$

Can be shown that

$$Var(x_0'\hat{\beta}) - Var(x_{01}'\hat{\beta}_1^*) = \sigma^2 (x_{02} - A'x_{01})' G^{22} (x_{02} - A'x_{01}) \geq 0$$

because G^{22} is positive definite.

By Theorem 3.9c(i), $Var(\hat{\beta}_j)$ in the full model is greater than $Var(\hat{\beta}_j^*)$ in the reduced model. Thus underfitting reduces the variance of the $\hat{\beta}_j$'s but introduces bias. On the other hand, overfitting increases the variance of the $\hat{\beta}_j$'s. In Theorem 3.9c (ii), $Var(\hat{y}_0)$ based on the full model is greater than $Var(\hat{y}_{01})$ based on the reduced model. Again, underfitting reduces the variance of the estimate of $E(y_0)$ but introduces bias. Overfitting increases the variance of the estimate of $E(y_0)$.

We now consider S^2 for the full model and for the reduced model. For the full model $\mathrm{y} = \mathrm{X}\beta + \varepsilon = \mathrm{X}_1\beta_1 + \mathrm{X}_2\beta_2 + \varepsilon$, the sample variance S^2 is given by (3.23) as

$$S^2 = \frac{(\mathrm{y} - \mathrm{X}\hat{\beta})'(\mathrm{y} - \mathrm{X}\hat{\beta})}{n - k - 1}$$

In Theorem 3.3f, we have $E(S^2) = \sigma^2$. The expected value of S^2 for the reduced model is given in the following theorem.

Theorem 3.9d: If $\mathrm{y} = \mathrm{X}\beta + \varepsilon$ is the correct model, then for the reduced model $\mathrm{y} = \mathrm{X}_1\beta_1^* + \varepsilon^*$ (underfitting), where X_1 is $n \times (p+1)$ with $p < k$, the variance estimator

$$S_1^2 = \frac{(\mathrm{y} - \mathrm{X}_1\hat{\beta}_1^*)'(\mathrm{y} - \mathrm{X}_1\hat{\beta}_1^*)}{n - p - 1} \tag{3.84}$$

has expected value

$$E(S_1^2) = \sigma^2 + \frac{\beta_2' \mathrm{X}_2' [\mathrm{I} - \mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1} \mathrm{X}_1'] \mathrm{X}_2 \beta_2}{n - p - 1} \tag{3.85}$$

Proof: We write the numerator of (3.84) as

$$\begin{aligned} SSE_1 &= \mathrm{y}'\mathrm{y} - \hat{\beta}_1^{*\prime} \mathrm{X}_1'\mathrm{y} = \mathrm{y}'\mathrm{y} - \mathrm{y}'\mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{y} \\ &= \mathrm{y}'[\mathrm{I} - \mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1']\mathrm{y} \end{aligned}$$

Since $E(\mathrm{y}) = \mathrm{X}\beta$ by assumption, we have, by Theorem 3.2a,

$$\begin{aligned} E(SSE_1) &= tr\{[\mathrm{I} - \mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1']\sigma^2 \mathrm{I}\} + \beta'\mathrm{X}'[\mathrm{I} - \mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1']\mathrm{X}\beta \\ &= (n - p - 1)\sigma^2 + \beta_2'\mathrm{X}_2'[\mathrm{I} - \mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1']\mathrm{X}_2\beta_2 \end{aligned}$$

(see Problem 3.45).

Since the quadratic form in (3.85) is positive semidefinite, S^2 is biased upward when underfitting (see Fig. 3.6). We can also examine (3.85) from the perspective of overfitting, in which case $\beta_2 = 0$ and S^2 is unbiased.

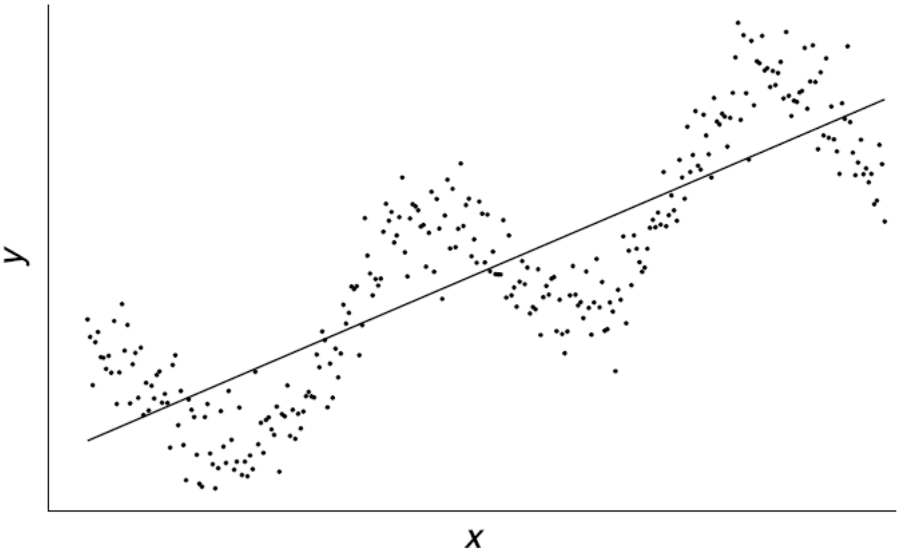


Figure 3.6: Straight-line fit to a curved pattern of points

To summarize the results in this section, underfitting leads to biased $\hat{\beta}_j$'s, biased $\hat{y}$'s, and biased s^2. Overfitting increases the variances of the $\hat{\beta}_j$'s and of the $\hat{y}$'s. We are thus compelled to seek an appropriate balance between a biased model and one with large variances. This is the task of the model builder and serves as motivation for seeking an optimum subset of *x*'s.

Example 3.9a: Suppose that the model $y_i = \beta_0^* + \beta_1^* x_i + \varepsilon_i^*$ has been fitted when the true model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$. (This situation is similar to that illustrated in Figure 3.2.) In this case, β_0^*, β_1^*, and s_1^2 would be biased by an amount dependent on the choice of the x_i's [see (3.80) and (3.86)]. The error term ε_i^* in the misspecified model $y_i = \beta_0^* + \beta_1^* x_i + \varepsilon_i^*$ does not have a mean of 0:

$$\begin{aligned} E(\varepsilon_i^*) &= E(y_i - \beta_0^* - \beta_1^* x_i) \\ &= E(y_i) - \beta_0^* - \beta_1^* x_i \\ &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 - \beta_0^* - \beta_1^* x_i \\ &= \beta_0 - \beta_0^* + (\beta_1 - \beta_1^*) x_i + \beta_2 x_i^2 \end{aligned}$$

Example 3.9b: Suppose that the true model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and we fit the model $y_i = \beta_1^* x_i + \varepsilon_i^*$, as illustrated in Figure 3.7.

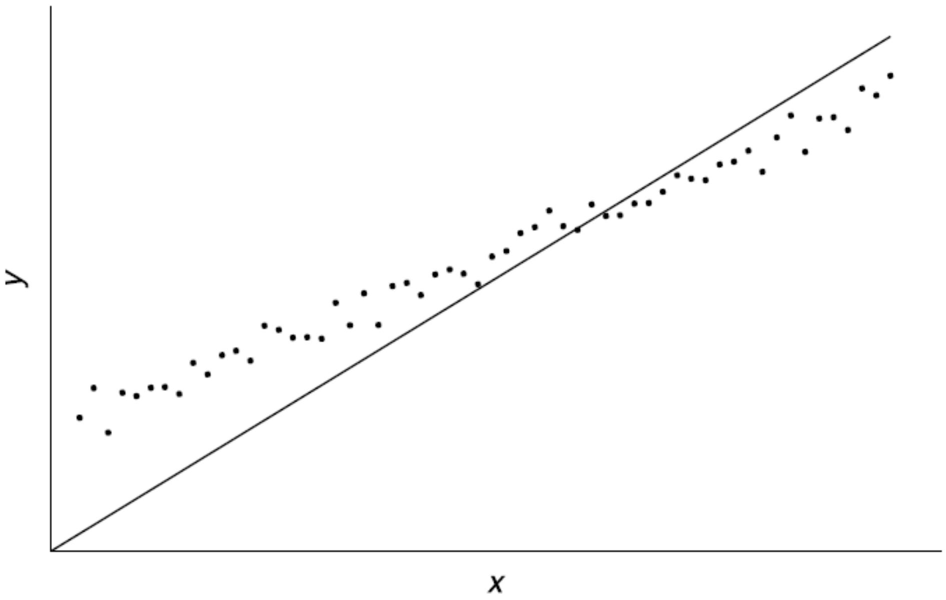


Figure 3.7: No-intercept model fit to data from an intercept model

For the model $y_i = \beta_1^* x_i + \varepsilon_i^*$, the least-squares estimator is

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \tag{3.86}$$

(see Problem 3.46). Then, under the full model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, we have

$$E\left(\hat{\beta}_1^*\right) = \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n x_i E\left(y_i\right)$$

$$= \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n x_i \left(\beta_0 + \beta_1 x_i\right)$$

$$= \frac{1}{\sum_{i=1}^n x_i^2} \left(\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2\right)$$

$$= \beta_0 \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} + \beta_1$$

Thus $\hat{\beta}_1^*$ is biased by an amount that depends on β_0 and the values of the *x*’s.

3.10: Orthogonalization

In Section 3.9, we discussed estimation of β_1^* in the model $y = X_1\beta_1^* + \varepsilon^*$ when the true model is $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$. By Theorem 3.9a, $E\left(\hat{\beta}_1^*\right) = \beta_1 + \left(X_1'X_1\right)^{-1} X_1'X_2\beta_2$, so that estimation of β_1 is affected by the presence of X_2, unless $X_1'X_2 = O$, in which case, $E\left(\hat{\beta}_1^*\right) = \beta_1$. In the following theorem, we show that if $X_1'X_2 = O$, the estimators of β_1^* and β_1 not only have the same expected value, but are exactly the same.

Theorem 3.10: If $X_1'X_2 = O$, then the estimator of β_1 in the full model $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ is the same as the estimator of β_1^* in the reduced model $y = X_1\beta_1^* + \varepsilon^*$.

Proof: The least-squares estimator of β_1^* is $\hat{\beta}_1^* = (X_1'X_1)^{-1} X_1'y$. For the estimator of β_1 in the full model, we partition $\hat{\beta} = \left(X'X\right)^{-1} X'y$ to obtain

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}$$

Using the notation in the proof of Theorem 3.9c, this becomes

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}^{-1} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}$$

$$= \begin{pmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{pmatrix} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}$$

We obtain

$$\begin{aligned} \hat{\beta}_1 &= G^{11}X_1'y + G^{12}X_2'y \\ &= \left(G_{11}^{-1} + G_{11}^{-1}G_{12}\, B^{-1}G_{21}G_{11}^{-1}\right)X_1'y - G_{11}^{-1}G_{12}B^{-1}X_2'y \end{aligned}$$

Where $B = G_{22} - G_{21}G_{11}^{-1}G_{12}$. If $G_{12} = X_1'X_2 = O$, then $\hat{\beta}_1$ reduces to

$$\hat{\beta}_1 = G_{11}^{-1}X_1'y = (X_1'X_1)^{-1} X_1'y$$

Which is the same as $\hat{\beta}_1^*$.

Note that Theorem 3.10 will also hold if X_1 and X_2 are “essentially orthogonal,” that is, if the centered columns of X_1 are orthogonal to the centered columns of X_2.

In Theorem 3.9a, we discussed estimation of β_1^* in the presence of β_2 when $X_1'X_2 \neq O$. We now consider a process of orthogonalization to give additional insights into the meaning of partial regression coefficients.

In Example 3.2, we illustrated the change in the estimate of a regression coefficient when another *x* was added to the model. We now use the same data to further examine this change. The prediction equation obtained in Example 3.2 was

$$\hat{y} = 5.3754 + 3.0118 x_1 - 1.2855 x_2 \tag{3.88}$$

And the negative partial regressions of **y** on x_2 were shown in Figure 3.2. By means of orthogonalization, we can give additional meaning to the term $-1.2855 x_2$. In order to add x_2 to the prediction equation containing only x_1, we need to determine how much variation in **y** is due to x_2 after the effect of x_1 has been accounted for, and we must also correct for the relationship between x_1 and x_2. Our approach is to consider the relationship between the residual variation after regressing **y** on x_1 and the residual variation after regressing x_2 on x_1. We follow a three-step process.

1. Regress **y** on x_1, and calculate residuals [see (3.11)]. The prediction equation is

$$\hat{y} = 1.8585 + 1.3019 x_1 \tag{3.89}$$

and the residuals $y_i - \hat{y}_i(x_1)$ are given in Table 3.2, where $\hat{y}_i(x_1)$ indicates that $\hat{y}$ is based on a regression of **y** on x_1 as in (3.89).

2. Regress x_2 on x_1 and calculate residuals. The prediction equation is

$$\hat{x}_2 = 2.7358 + 1.3302 x_1 \tag{3.90}$$

and the residuals $x_{2i} - \hat{x}_{2i}(x_1)$ are given in Table 3.2, where $\hat{x}_{2i}(x_1)$ indicates that x_2 has been regressed on x_1 as in (3.90).

3. Now regress $y - \hat{y}(x_1)$ on $x_2 - \hat{x}_2(x_1)$, which gives

$$\overline{y - \hat{y}} = -1.2855(x_2 - \hat{x}_2) \tag{3.91}$$

There is no intercept in (3.91) because both sets of residuals have a mean of 0.

TABLE 3.2: Data from Table 3.1 and Residuals

y	x_1	x_2	$y-\hat{y}(\mathrm{x}_1)$	$x_2-\hat{x}_2(\mathrm{x}_1)$
2	0	2	0.141509	-0.735849
3	2	6	-1.462264	0.603774
2	2	7	-2.462264	1.603774
7	2	5	2.537736	-0.396226
6	4	9	-1.06604	0.943396
8	4	8	0.93396	-0.056604
10	4	7	2.93396	-1.056604
7	6	10	-2.66981	-0.716981
8	6	11	-1.66981	0.2830189
12	6	9	2.330189	-1.716981
11	8	15	-1.273584	1.622642
14	8	13	1.726415	-0.377358

In (3.91), we obtain a clearer insight into the meaning of the partial regression coefficient -1.2855 in (3.88). We are using the "unexplained" portion of x_2 (after x_1 is accounted for) to predict the "unexplained" portion of **y** (after x_1 is accounted for).

Since $x_2-\hat{x}_2(\mathrm{x}_1)$ is orthogonal to x_1 [see Section 3.4.2, in particular (3.29)], fitting $y-\hat{y}(\mathrm{x}_1)$ to $x_2-\hat{x}_2(\mathrm{x}_1)$ yields the same coefficient, -1.2855, as when fitting **y** to x_1 and x_2 together. Thus -1.2855 represents the additional effect of x_2 beyond the effect of x_1 and also after taking into account the overlap between x_1 and x_2 in their effect on **y**. The orthogonality of x_1 and $x_2-\hat{x}_2(\mathrm{x}_1)$ makes this simplified breakdown of effects possible.

We can substitute $\hat{y}(x_1)$ and $\hat{x}_2(x_1)$ in (3.91) to obtain

$$\overline{\mathrm{y}-\hat{\mathrm{y}}} = \hat{y}(x_1,x_2)-\hat{y}(x_1) = -1.2855[x_2-\hat{x}_2(x_1)]$$

Or

$$\hat{y}-(1.8585+1.3019x_1) = -1.2855[x_2-(2.7358+1.3302x_1)] \qquad (3.92)$$

which reduces to

$$\hat{y} = 5.3754+3.0118x_1-1.2855x_2 \qquad (3.93)$$

the same as (3.88). If we regress *y* (rather than $y-\hat{y}$ on $x_2-\hat{x}_2(\mathrm{x}_1)$, we will still obtain $-1.2855x_2$, but we will not have $5.3754+3.0118x_1$.

The correlation between the residuals $y-\hat{y}(x_1)$ and $x_2-\hat{x}_2(x_1)$ is the same as the (sample) partial correlation of y and x_2 with x_1 held fixed:

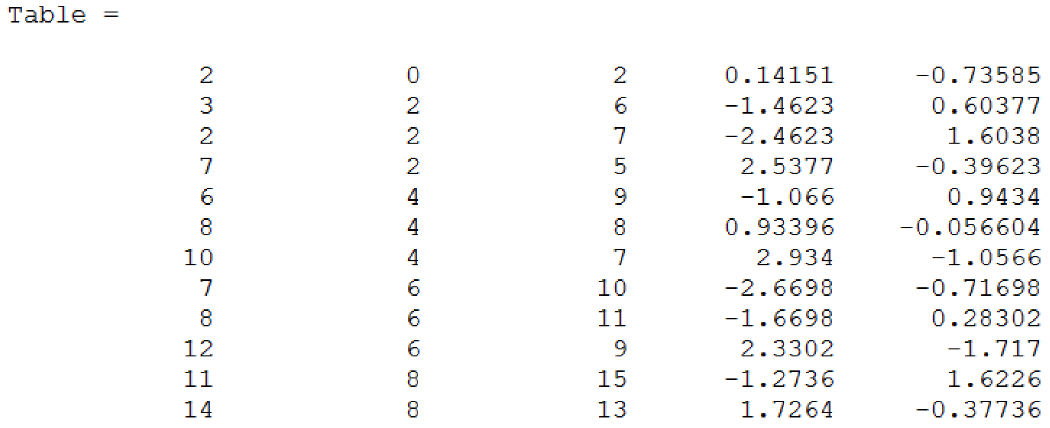
$$r_{y2.1} = r_{y-\hat{y},x_2-\hat{x}_2} \tag{3.94}$$

This is discussed further in Section 6.8.

Example 3.2a[The program name ta10.m]	Applications using MATLAB

```
clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]'; x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]'; n=length(x1);
% Regress y on x1
EX1=[ones(size(x1)) x1];
beta1=EX1\y;YX1=EX1*beta1;eX1=y-YX1;
['Reg. line 1: yhat = ',num2str(beta1(1)),' + ',num2str(beta1(2)), '*x1']
% Regress x2 on x1
EX1=[ones(size(x1)) x1];
beta2=EX1\x2;X2X1=EX1*beta2;eX2= x2-X2X1;
['Reg. line 2: xhat2 = ',num2str(beta2(1)),' + ',num2str(beta2(2)), '*x1']
% Regress eX1 on eX2
beta3=eX2\eX1;data=[y x1 x2];partialcorr(data);
['Reg. line 3: y-yhat = ',num2str(beta3),'*(x2-x2hat)']
Table=[y x1 x2 eX1 eX2],ry2and1=partialcorr(data);
ry2andfixed1=ry2and1(3,1),reX1andeX2=corr(eX1,eX2)
```

```
Ans =

Reg. line 1: yhat = 1.8585 + 1.3019*x1
Reg. line 2: xhat2 = 2.7358 + 1.3302*x1
Reg. line 3: y-yhat = -1.2855*(x2-x2hat)
```



```
Table =

     2         0         2     0.14151    -0.73585
     3         2         6     -1.4623     0.60377
     2         2         7     -2.4623      1.6038
     7         2         5      2.5377    -0.39623
     6         4         9      -1.066      0.9434
     8         4         8     0.93396   -0.056604
    10         4         7       2.934     -1.0566
     7         6        10     -2.6698    -0.71698
     8         6        11     -1.6698     0.28302
    12         6         9      2.3302      -1.717
    11         8        15     -1.2736      1.6226
    14         8        13      1.7264    -0.37736
```

```
ry2andfixed1 =
             -0.66112
reX1andeX2 =
              -0.66112
```

We now consider the general case with full model

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

and reduced model

$$y = X_1\beta_1^* + \varepsilon^*$$

We use an orthogonalization approach to obtain an estimator of β_2, following the same three steps as in the illustration with x_1 and x_2 above:

1. Regress *y* on x_1 and calculate residuals $y - \hat{y}(X_1)$, where $\hat{y} = X_1\hat{\beta}_1^*$ $= X_1(X_1'X_1)^{-1}X_1'y$ [see (3.11)].
2. Regress the columns of X_2 on X_1 and obtain residuals $X_{2.1} =$ $X_2 - \hat{X}_2(X_1)$. If X_2 is written in terms of its columns as $X_2 =$ $(x_{21}, \cdots, x_{2j}, \cdots x_{2p})$, then the regression coefficient vector for x_{2j} on X_1 is $b_j = (X_1'X_1)^{-1}X_1'x_{2j}$, and $\hat{x}_{2j} = X_1 b_j = X_1(X_1'X_1)^{-1}X_1'x_{2j}$. For all columns of X_2, this becomes $\hat{X}_2(X_1) = X_1(X_1'X_1)^{-1}X_1'X_2 = X_1A$, where $A = (X_1'X_1)^{-1}X_1'X_2$ is the alias matrix defined in (3.80). Note that $X_{2.1} =$ $X_2 - \hat{X}_2(X_1)$. is orthogonal to X_1:

$$X_1'X_{2.1} = O \tag{3.95}$$

   Using the alias matrix **A**, the residual matrix can be expressed as

$$X_{2.1} = X_2 - \hat{X}_2(X_1) \tag{3.96}$$

$$= X_2 - X_1(X_1'X_1)^{-1}X_1'X_2 = X_2 - X_1A \tag{3.97}$$

3. Regress $y - \hat{y}(X_1)$ on $X_{2.1} = X_2 - \hat{X}_2(X_1)$. Since $X_{2.1}$ is orthogonal to X_1, we obtain the same $\hat{\beta}_2$ as in the full model $\hat{y} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2$. Adapting the notation of (3.91) and (3.92), this can be expressed as

$$\hat{y}(X_1, X_2) - \hat{y}(X_1) = X_{2.1}\hat{\beta}_2 \tag{3.98}$$

If we substitute $\hat{\mathrm{y}}(\mathrm{X}_1)=\mathrm{X}_1\hat{\beta}_1^*$ and $\mathrm{X}_{2.1}=\mathrm{X}_2-\mathrm{X}_1\mathrm{A}$ into (3.98) and use $\hat{\beta}_1^*=\hat{\beta}_1+\mathrm{A}\hat{\beta}_2$ from (3.80), we obtain

$$\begin{aligned}\hat{\mathrm{y}}(\mathrm{X}_1,\mathrm{X}_2)&=\mathrm{X}_1\hat{\beta}_1^*+(\mathrm{X}_2-\mathrm{X}_1\mathrm{A})\hat{\beta}_2\\&=\mathrm{X}_1(\hat{\beta}_1+\mathrm{A}\hat{\beta}_2)+(\mathrm{X}_2-\mathrm{X}_1\mathrm{A})\hat{\beta}_2\\&=\mathrm{X}_1\hat{\beta}_1+\mathrm{X}_2\hat{\beta}_2\end{aligned}$$

which is analogous to (3.93). This confirms that the orthogonality of X_1 and $\mathrm{X}_{2.1}$ leads to the estimator $\hat{\beta}_2$ in (3.98). For a formal proof, see Problem 3.50.

PROBLEMS

3.1: Show that $\sum_{i=1}^{n}(y_i-\mathrm{x}_i'\hat{\beta})^2=(\mathrm{y}-\mathrm{X}\hat{\beta})'(\mathrm{y}-\mathrm{X}\hat{\beta})$, thus verifying (3.7).

3.2: Show that (3.10) follows from (3.9). Why is $\mathrm{X}'\mathrm{X}$ positive definite, as noted below (3.10)?

3.3: Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ in (3.12) in Example 3.3.1 are the same as in (2.5) and (2.6).

3.4: Obtain $Cov(\hat{\beta})$ in (3.16) from (3.15).

3.5: Show that $Var(\hat{\beta}_0)=\sigma^2\left(\sum_{i=1}^{n}x_i^2\Big/n\right)\Big/\sum_{i=1}^{n}(x_i-\bar{x})^2$ in (3.16) in Example 3.3.2a is the same as $Var(\hat{\beta}_0)$ in (3.10).

3.6: Show that AA' can be expressed as $\mathrm{AA}'=[\mathrm{A}-(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'][\mathrm{A}-(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']'+(\mathrm{X}'\mathrm{X})^{-1}$ as in (3.17) in Theorem 3.3d.

3.7: Prove Corollary 1 to Theorem 3.3d in the following two ways: (a) Use an approach similar to the proof of Theorem 3.3d. (b) Use the method of Lagrange multipliers.

3.8: Show that if the *x*'s are rescaled as $z_j=c_jx_j$, *j* = 1, 2, . . . , *k*, then $\hat{\beta}_z=\mathrm{D}^{-1}\hat{\beta}$, as in (3.18) in the proof of the Theorem 3.3e.

3.9: Verify (3.20) and (3.21) in the proof of Corollary 1 to Theorem 3.3e.

3.10: Show that s^2 is invariant to changes of scale on the *x*'s, as noted following Corollary 1 to Theorem 3.3e.

3.11: Show that $(\mathrm{y}-\mathrm{X}\hat{\beta})'(\mathrm{y}-\mathrm{X}\hat{\beta})=\mathrm{y}'\mathrm{y}-\hat{\beta}'\mathrm{X}'\mathrm{y}$ as in (3.24).

3.12: Show that $E(SSE)=\sigma^2(n-k-1)$, as in Theorem 3.3f, using the following approach. Show that $SSE=\mathrm{y}'\mathrm{y}-\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}$. Show that $E(\mathrm{y}'\mathrm{y})=n\sigma^2+\beta'\mathrm{X}'\mathrm{X}\beta$ and that $E(\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta})=(k+1)\sigma^2+\beta'\mathrm{X}'\mathrm{X}\beta$.

3.13: Prove that an additional *x* reduces *SSE*, as noted following Theorem 3.3f.

3.14: Show that the non-centered model preceding (3.30) can be written in the centered form in (3.30), with α defined as in (3.31).

3.15: Show that $\mathrm{X}_c=[\mathrm{I}-(1/n)\mathrm{J}]\mathrm{X}_1$ as in (3.33), where X_1 is as given in (3.19).

3.16: Show that $\mathrm{j}'\mathrm{X}_c=0'$, as in (3.35), where X_c is the centered X matrix defined in (3.33).

3.17: Show that the estimators $\hat{\alpha}=\bar{y}$ and $\hat{\beta}_1=(\mathrm{X}_c'\mathrm{X}_c)^{-1}\mathrm{X}_c'\mathrm{y}$ in (3.36) and (3.37) are the same as $\hat{\beta}=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ in (3.6). Use the following two methods:

(a) Work with the normal equations in both cases.

(b) Use the inverse of $\mathrm{X}'\mathrm{X}$ in partitioned form:

$$(\mathrm{X}'\mathrm{X})^{-1}=[(\mathrm{j},\mathrm{X}_1)'(\mathrm{j},\mathrm{X}_1)]^{-1}$$

3.18: Show that the fitted regression plane $\hat{y}=\hat{\alpha}+\hat{\beta}_1(x_1-\bar{x}_1)+\cdots+\hat{\beta}_k(x_k-\bar{x}_k)$ passes through the point $(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_k,\bar{y})$, as noted below (3.38).

3.19: Show that $SSE=\sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{\beta}_1'\mathrm{X}_c'\mathrm{y}$ in (3.39) is the same as $SSE=\mathrm{y}'\mathrm{y}-\hat{\beta}'\mathrm{X}'\mathrm{y}$ in (3.24).

3.20: (a) Show that $S_{xx}=\mathrm{X}_c'\mathrm{X}_c/(n-1)$ as in (3.44).

(b) Show that $S_{yx}=\mathrm{X}_c'\mathrm{y}/(n-1)$ as in (3.45).

3.21: (a) Show that if **y** is $N_n(\mathrm{X}\beta,\sigma^2\mathrm{I})$, the likelihood function is

$$L(\beta,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{n/2}}e^{-(y-X\hat{\beta})'(y-X\hat{\beta})/2\sigma^2}$$

as in (3.50) in the proof of Theorem 3.6a.

(b) Differentiate $\text{Ln}\,L(\beta,\sigma^2)$ in (3.51) with respect to $\hat{\beta}$ to obtain $\hat{\beta}=(X'X)^{-1}X'y$ in (3.48).

(c) Differentiate $\text{Ln}\,L(\beta,\sigma^2)$ with respect to σ^2 to obtain $\hat{\sigma}^2 = (y-X\hat{\beta})'(y-X\hat{\beta})/n$ as in (3.49).

3.22: Prove parts (ii) and (iii) of Theorem 3.6b.

3.23: Show that $(y-X\beta)'(y-X\beta)=(y-X\hat{\beta})'(y-X\hat{\beta})+(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)$ as in (3.52) in the proof of Theorem 3.6c.

3.24: Explain why $f(y;\beta,\sigma^2)$ does not factor into $g_1(\hat{\beta},\beta)g_2(\hat{\sigma}^2,\sigma^2)h(y)$, as noted following Theorem 3.6c.

3.25: Verify the equivalence of (3.55) and (3.56); that is, show that $\hat{\beta}'X'y-n\bar{y}^2=\hat{\beta}_1'X_c'X_c\hat{\beta}_1$.

3.26: Verify the comments in property 1 in Section 3.7, namely, that if $\hat{\beta}_1=\hat{\beta}_2=\cdots=\hat{\beta}_k=0$, then $R^2=0$, and if $y_i=\hat{y}_i$, $i=1,2,\cdots,n$, then $R^2=1$.

3.27: Show that adding an x to the model increases (cannot decrease) the value of R^2, as in property 3 in Section 3.7.

3.28: (a) Verify that R^2 is invariant to full-rank linear transformations on the x's as in property 6 in Section 3.7.

(b) Show that R^2 is invariant to a scale change $z=cy$ on y.

3.29: (a) Show that R^2 in (3.55) can be written in the form $R^2=1-SSE\big/\sum_{i=1}^n(y_i-\bar{y})^2$.

(b) Replace SSE and $\sum_{i=1}^n(y_i-\bar{y})^2$ in part (a) by variance estimators $SSE/(n-k-1)$ and $\sum_{i=1}^n(y_i-\bar{y})^2\big/(n-1)$ and show that the result is the same as R_a^2 in (3.56).

3.30: Show that $\sum_{i=1}^n\hat{y}_i\big/n=\sum_{i=1}^n y_i\big/n$, as noted following (3.59) in Section 3.7.

3.31: Show that $\cos\theta = R$ as in (3.61), where R^2 is as given by (3.56).

3.32: (a) Show that $E(\hat{\beta}) = \beta$, where $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$ as in (3.63).
(b) Show that $Cov(\hat{\beta}) = \sigma^2(X'V^{-1}X)^{-1}$ as in (3.64).

3.33: (a) Show that the two forms of s^2 in (3.65) and (3.66) are equal.
(b) Show that $E(s^2) = \sigma^2$, where s^2 is as given by (3.66).

3.34: Complete the steps in the proof of Theorem 3.8b.

3.35: Show that for $V = (1-\rho)I + \rho J$ in (3.67), the inverse is given by $V^{-1} = a(I - b\rho J)$ as in (3.68), where $a = 1/(1-\rho)$ and $b = 1/[1+(n-1)\rho]$.

3.36: (a) Show that $X'V^{-1}X = \begin{pmatrix} bn & 0' \\ 0 & aX_c'X_c \end{pmatrix}$ as in (3.69).

(b) Show that $X'V^{-1}y = \begin{pmatrix} bn\bar{y} \\ aX_c'y \end{pmatrix}$ as in (3.70).

3.37: Show that $Cov(\hat{\beta}^*) = \sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$ as in (3.72), where $\hat{\beta}^* = (X'X)^{-1}X'y$ and $Cov(y) = \sigma^2 V$.

3.38: (a) Show that the weighted least-squares estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ for the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $Var(y_i) = \sigma^2 x_i$ has the form given in (3.73).
(b) Verify the expression for $Cov(\hat{\beta})$ in (3.74).

3.39: Obtain the expression for $Cov(\hat{\beta}^*)$ in (3.75).

3.40: As an alternative derivation of $Var(\hat{\beta}_1^*)$ in (3.76), use the following two steps to find $Var(\hat{\beta}_1^*)$ using $\hat{\beta}_1^* = \sum_{i=1}^n (x_i - \bar{x})y_i \big/ \sum_{i=1}^n (x_i - \bar{x})^2$ from the answer to Problem 2.2:

(a) Using $Var(y_i) = \sigma^2 x_i$, show that $Var(\hat{\beta}_1^*) = \sigma^2 \sum_{i=1}^n (x_i - \bar{x})^2 x_i \big/ [\sum_{i=1}^n (x_i - \bar{x})^2]^2$.

(b) Show that this expression for $Var(\hat{\beta}_1^*)$ is equal to that in (3.76).

3.41: Using x = 2, 3, 5, 7, 8, 10, compare $Var(\hat{\beta}_1^*)$ in (3.76) with $Var(\hat{\beta}_1)$ in (3.77).

3.42: Provide an alternative proof of $Cov\left(\hat{\beta}_1^*\right) = \sigma^2\left(\mathrm{X}_1'\mathrm{X}_1\right)^{-1}$ in (3.81) , $Cov\left(\hat{\beta}_1^*\right) = E\{[\hat{\beta}_1^* - E(\hat{\beta}_1^*)][\hat{\beta}_1^* - E(\hat{\beta}_1^*)]'\}$

3.43: Prove Theorem 3.9b.

3.44: Provide the missing steps in the proof of Theorem 3.9c(ii).

3.45: Show that $\mathrm{x}_{01}\hat{\beta}_1^*$ is biased for estimating $\mathrm{x}_{01}\beta_1$ if $\beta_2 \neq 0$ and $\mathrm{X}_1'\mathrm{X}_2 \neq \mathrm{O}$.

3.46: Show that $Var\left(\mathrm{x}_{01}\hat{\beta}_1\right) \geq Var\left(\mathrm{x}_{01}\hat{\beta}_1^*\right)$.

3.47: Complete the steps in the proof of Theorem 3.9d.

3.48: Show that for the no-intercept model $y_i = \beta_1^* x_i + \varepsilon_i^*$, the least-squares estimator is $\hat{\beta}_1^* = \sum_{i=1}^n x_i y_i \Big/ \sum_{i=1}^n x_i^2$ in (3.86).

3.49: Obtain $E\left(\hat{\beta}_1^*\right) = \beta_0 \sum_{i=1}^n x_i \Big/ \sum_{i=1}^n x_i^2 + \beta_1$ in (3.87) using (3.80), $E\left(\hat{\beta}_1^*\right) = \beta_1 + \mathrm{A}\beta_2$.

3.50: Suppose that we use the model $y_i = \beta_0^* + \beta_1^* x_i + \varepsilon_i^*$ when the true model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$.

(a) Using (3.80), find $E\left(\hat{\beta}_0^*\right)$ and $E\left(\hat{\beta}_1^*\right)$ if observations are taken at x = -3, -2, -1, 0, 1, 2, 3.

(b) Using (3.85), find $E\left(s_1^2\right)$ for the same values of *x*.

3.51: Show that $\mathrm{X}_{2.1} = \mathrm{X}_2 - \hat{\mathrm{X}}_2(\mathrm{X}_1)$ is orthogonal to X_1, that is, $\mathrm{X}_1'\mathrm{X}_{2.1} = \mathrm{O}$, as in (3.95).

3.52: Show that $\hat{\beta}_2$ in (3.98) is the same as in the full fitted model $\hat{\mathrm{y}} = \mathrm{X}_1\hat{\beta}_1 + \mathrm{X}_2\hat{\beta}_2$.

3.53: When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere. An experiment was conducted to determine whether *y*, the amount of vapor, can be predicted using the following four variables based on initial conditions of the tank and the dispensed gasoline:

x_1 = tank temperature (^0F)
x_2 = gasoline temperature (^0F)

$x_3 =$ vapor pressure in tank (psi)
$x_4 =$ vapor pressure of gasoline (psi)

The data are given in Table 3.3 (Weisberg 1985, p. 138).

(a) Find $\hat{\beta}$ and s^2.

(b) Find an estimate of $Cov(\hat{\beta})$.

(c) Find $\hat{\beta}_1$ and $\hat{\beta}_0$ using S_{xx} and S_{yx} as in (3.46) and (3.47).

(d) Find R^2 and R_a^2.

TABLE 3.3: Gas Vapor Data

y	x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4
29	33	53	3.32	3.42	40	90	64	7.32	6.70
24	31	36	3.10	3.26	46	90	60	7.32	7.20
26	33	51	3.18	3.18	55	92	92	7.45	7.45
22	37	51	3.39	3.08	52	91	92	7.27	7.26
27	36	54	3.20	3.41	29	61	62	3.91	4.08
21	35	35	3.03	3.03	22	59	42	3.75	3.45
33	59	56	4.78	4.57	31	88	65	6.48	5.80
34	60	60	4.72	4.72	45	91	89	6.70	6.60
32	59	60	4.60	4.41	37	63	62	4.30	4.30
34	60	60	4.53	4.53	37	60	61	4.02	4.10
20	34	35	2.90	2.95	33	60	62	4.02	3.89
36	60	59	4.40	4.36	27	59	62	3.98	4.02
34	60	62	4.31	4.42	34	59	62	4.39	4.53
23	60	36	4.27	3.94	19	37	35	2.75	2.64
24	62	38	4.41	3.49	16	35	35	2.59	2.59
32	62	61	4.39	4.39	22	37	37	2.73	2.59

3.54: In an effort to obtain maximum yield in a chemical reaction, the values of the following variables were chosen by the experimenter:

$x_1 =$ temperature (${}^0C$)
$x_2 =$ concentration of a reagent (%)
$x_3 =$ time of reaction (hours)

Two different response variables were observed:

$y_1 =$ percent of unchanged starting material

$y_2 =$ percent converted to the desired product

The data are listed in Table 3.4 (Box and Youle 1955, Andrews and Herzberg 1985, p. 188). Carry out the following for y_1:

(a) Find $\hat{\beta}$ and s^2.

(b) Find an estimate of $Cov(\hat{\beta})$.

(c) Find R^2 and R_a^2.

(d) In order to find the maximum yield for y_1, a second-order model is of interest. Find $\hat{\beta}$ and s^2 for the model $y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3 + \varepsilon$.

(e) Find R^2 and R_a^2 for the second-order model.

TABLE 3.4: Chemical Reaction Data

y_1	y_2	x_1	x_2	x_3
41.5	45.9	162	23	3
33.8	53.3	162	23	8
27.7	57.5	162	30	5
21.7	58.8	162	30	8
19.9	60.6	172	25	5
15	58	172	25	8
12.2	58.6	172	30	5
4.30	52.4	172	30	8
19.3	56.9	167	27.5	6.5
6.40	55.4	177	27.5	6.5
37.6	46.9	157	27.5	6.5
18	57.3	167	32.5	6.5
26.3	55	167	22.5	6.5
9.90	58.9	167	27.5	9.5
25	50.3	167	27.5	3.5
14.1	61.1	177	20	6.5
15.2	62.9	177	20	6.5
15.9	60	160	34	7.5
19.6	60.6	160	34	7.5

3.55: The following variables were recorded for several counties in Minnesota in 1977:

$y =$ Average rent paid per acre of land with alfalfa

$x_1 =$ Average rent paid per acre for all land

$x_2 =$ Average number of dairy cows per square mile

$x_3 =$ Proportion of farmland in pasture

The data for 34 counties are given in Table 3.5 (Weisberg 1985, p.162). Can rent for alfalfa land be predicted from the other three variables?

(a) Find $\hat{\beta}$ and s^2.

(b) Find $\hat{\beta}_1$ and $\hat{\beta}_0$ using S_{xx} and S_{yx} as in (3.46) and (3.47).

(c) Find R^2 and R_a^2.

TABLE 3.5: Land Rent Data

y	x_1	x_2	x_3	y	x_1	x_2	x_3
18.38	15.50	17.25	0.24	8.50	9	8.89	0.08
20	22.29	18.51	0.20	36.5	20.64	23.81	0.24
11.5	12.36	11.13	0.12	60	81.40	4.54	0.05
25	31.84	5.54	0.12	16.25	18.92	29.62	0.72
52.50	83.90	5.44	0.04	50	50.32	21.36	0.19
82.50	72.25	20.37	0.05	11.50	21.33	1.53	0.10
25	27.14	31.20	0.27	35	46.85	5.42	0.08
30.67	40.41	4.29	0.10	75	65.94	22.10	0.09
12	12.42	8.69	0.41	31.56	38.68	14.55	0.17
61.2	69.42	6.63	0.04	48.50	51.19	7.59	0.13
60	48.46	27.40	0.12	77.50	59.42	49.86	0.13
57.50	69	31.23	0.08	21.67	24.64	11.46	0.21
31	26.09	28.50	0.21	19.75	26.94	2.48	0.10
60	62.83	29.98	0.17	56	46.20	31.62	0.26
72.50	77.06	13.59	0.05	25	26.86	53.73	0.43
60.33	58.83	45.46	0.16	40	20	40.18	0.56
49.75	59.48	35.90	0.32	56.67	62.52	15.89	0.05

.

Chapter Four

Multiple Regression: Tests of Hypotheses and Confidence Intervals

4: Introduction

In this chapter we consider hypothesis tests and confidence intervals for the parameters $\beta_0, \beta_1, \cdots, \beta_k$ in β in the model $y = X\beta + \varepsilon$. We also provide a confidence interval for $\sigma^2 = Var(y_i)$. We will assume throughout the chapter that **y** is $N(X\beta, \sigma^2 I)$, where **X** is $n \times (k+1)$ of rank $k+1<n$.

4.1: Test of Overall Regression

We noted in Section 3.9 that the problems associated with both overfitting and underfitting motivate us to seek an optimal model. Hypothesis testing is a formal tool for, among other things, choosing between a reduced model and an associated full model. The hypothesis H_0, expresses the reduced model in terms of values of a subset of the β_j's in β. The alternative hypothesis, H_1, is associated with the full model.

To illustrate this tool we begin with a common test, the test of the overall regression hypothesis that none of the *x* variables predict *y*. This hypothesis (leading to the reduced model) can be expressed as $H_0 : \beta_1 = 0$, where $\beta_1 = (\beta_1, \beta_2, \cdots, \beta_k)'$. Note that we wish to test $H_0 : \beta_1 = 0$, not $H_0 : \beta_0 = 0$, where

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Since β_0 is usually not zero, we would rarely be interested in including $\beta_0 = 0$ in the hypothesis. Rejection of $H_0 : \beta = 0$ might be due solely to β_0, and we would not learn whether the *x* variables predict *y*. For a test of $H_0 : \beta = 0$, see Problem 4.6.

We proceed by proposing a test statistic that is distributed as a central *F* if H_0 is true and as a non-central *F* otherwise. Our approach to obtaining a test statistic is somewhat simplified if we use the centered model (3.32)

$$y = (j, X_c) \begin{pmatrix} \alpha \\ \beta_1 \end{pmatrix} + \varepsilon$$

where $X_c = [I-(1/n)J]X_1$ is the centered matrix [see (3.33)] and X_1 contains all the columns of **X** except the first [see (3.19)]. The corrected total sum of squares $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ can be partitioned as

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \hat{\beta}_1' X_c' y + \left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1' X_c' y\right], \quad \text{[by (3.53)]}$$

$$= \hat{\beta}_1' X_c' X_c \hat{\beta}_1 + SSE = SSR + SSE \qquad \text{[by } (3.54)], \qquad (4.1)$$

where *SSE* is as given in (3.39). The regression sum of squares $SSR = \hat{\beta}_1' X_c' X_c \hat{\beta}_1$ is clearly due to β_1.

In order to construct an *F* test, we first express the sums of squares in (4.1) as quadratic forms in **y** so that *SSR* and *SSE* have chi-square distributions and are independent. Using $\sum_{i=1}^{n}(y_i - \bar{y})^2 = y'[I-(1/n)J]y$, $\hat{\beta}_1 = (X_c' X_c)^{-1} X_c' y$ in (3.37), and $SSE = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1' X_c' y$ in (3.39), we can write (4.1) as

$$y'[I-(1/n)J]y = SSR + SSE$$

$$= y'X_c(X_c'X_c)^{-1}X_c'y + y'[I-(1/n)J]y - y'X_c(X_c'X_c)^{-1}X_c'y$$

$$= y'H_c y + y'[I-(1/n)J - H_c]y \qquad (4.2)$$

Where $H_c = X_c(X_c'X_c)^{-1}X_c'$

In the following theorem we establish some properties of the three matrices of the quadratic forms in (4.2).

Theorem 4.1a: The matrices $I-(1/n)J$, $H_c = X_c(X_c'X_c)^{-1}X_c'$, and $I-(1/n)J-H_c$ have the following properties:

(i) $H_c[I-(1/n)J] = H_c \qquad (4.3)$

(ii) H_c is idempotent of rank *k*.

(iii) $I-(1/n)J-H_c$ is idempotent of rank $n-k-1$.

(iv) $H_c[I-(1/n)J-H_c] = O \qquad (4.4)$

Proof: Part (i) follows from $X_c' j = 0$, which was established in Problem 3.16. Part (ii) can be shown by direct multiplication. Parts (iii) and (iv) follow from (i) and (ii).

The distributions of SSR/σ^2 and SSE/σ^2 are given in the following theorem.

Theorem 4.1b: If **y** is $N_n(\mathrm{X}\beta,\sigma^2 \mathrm{I})$, then $SSR/\sigma^2 = \hat{\beta}_1'\mathrm{X}_c'\mathrm{X}_c\hat{\beta}_1/\sigma^2$ and $SSE/\sigma^2 = \left[\sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1'\mathrm{X}_c'\mathrm{X}_c\hat{\beta}_1\right]/\sigma^2$ have the following distributions:

(i) SSR/σ^2 is $\chi^2_{(k,\lambda_1)}$, where $\lambda_1 = \mu'\mathrm{A}\,\mu/2\sigma^2 = \beta_1'\mathrm{X}_c'\mathrm{X}_c\beta_1/2\sigma^2$
(i) SSE/σ^2 is $\chi^2_{(n-k-1)}$.

Proof: These results follow from (4.2), Theorem 4.1a(ii) and (iii). The independence of SSR and SSE is demonstrated in the following theorem.

Theorem 4.1c: If **y** is $N_n(\mathrm{X}\beta,\sigma^2 \mathrm{I})$, then *SSR* and *SSE* are independent, where *SSR* and *SSE* are defined in (4.1) and (4.2).

Proof: This follows from Theorem 4.1a(iv).

We can now establish an *F* test for $H_0 : \beta_1 = 0$ versus $H_0 : \beta_1 \neq 0$.

Theorem 4.1d: If **y** is $N_n(\mathrm{X}\beta,\sigma^2 \mathrm{I})$, the distribution of

$$F = \frac{SSR/(k\sigma^2)}{SSE/[(n-k-1)\sigma^2]} = \frac{SSR/k}{SSE/(n-k-1)} \tag{4.5}$$

is as follows:

(i) If $H_0 : \beta_1 = 0$ is false, then

F is distributed as $F_{(k,n-k-1,\lambda_1)}$, where $\lambda_1 = \beta_1'\mathrm{X}_c'\mathrm{X}_c\beta_1/2\sigma^2$

(ii) If $H_0 : \beta_1 = 0$ is true, then $\lambda_1 = 0$ and *F* is distributed as $F_{(k,n-k-1)}$,

Proof:

(i) This result follows from Theorems 4.1b and 4.1c.
(ii) This result follows from Theorems 4.1b and 4.1c.

Note that $\lambda_1 = 0$ if and only if $\beta_1 = 0$, since $\mathrm{X}_c'\mathrm{X}_c$ is positive definite.

The test for $H_0 : \beta_1 = 0$ is carried out as follows. Reject H_0 if $F \geq F_{(\alpha,k,n-k-1)}$, where $F_{(\alpha,k,n-k-1)}$ is the upper α percentage point of the (central) *F* distribution. Alternatively, a *p*-value can be used to carry out the test. A *p*-value is the tail area of the central *F* distribution beyond the calculated *F* value, that is, the probability of exceeding the calculated *F* value, assuming $H_0 : \beta_1 = 0$ to be true. A *p*-value less than α is equivalent to $F > F_{(\alpha,k,n-k-1)}$.

TABLE 4.1: ANOVA Table for the *F* Test of $H_0 : \beta_1 = 0$

Source of Variation	df	Sum of Squares	Mean Square	Expected Mean Square
Due to β_1	*k*	$SSR = \hat{\beta}_1' X_c' y = \hat{\beta}' X' y - n\bar{y}^2$	SSR/k	$\sigma^2 + \frac{1}{k}\beta_1' X_c' X_c \beta_1$
Error	*n*-*k*-1	$SSE = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1' X_c' y$ $= y'y - \hat{\beta}' X' y$	$\dfrac{SSE}{n-k-1}$	σ^2
Total	*n*-1	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$		

The analysis-of-variance (ANOVA) table (Table 4.1) summarizes the results and calculations leading to the overall *F* test. Mean squares are sums of squares divided by the degrees of freedom of the associated chi-square (χ^2) distributions.

The entries in the column for expected mean squares in Table 4.1 are simply $E(SSR/k)$ and $E[SSE/(n-k-1)]$. It was established by Theorem 3.3f.

If $H_0 : \beta_1 = 0$ is true, both of the expected mean squares in Table 4.1 are equal to σ^2, and we expect *F* to be near 1. If $\beta_1 \neq 0$, then $E(SSR/k) > \sigma^2$ since $X_c' X_c$ is positive definite, and we expect *F* to exceed 1.We therefore reject H_0 for large values of *F*.

The test of $H_0 : \beta_1 = 0$ in Table 4.1 has been developed using the centered model (3.32). We can also express *SSR* and *SSE* in terms of the non-centered model $y = X\beta + \varepsilon$ in (3.4):

$$SSR = \hat{\beta}' X' y - n\bar{y}^2, \quad SSE = y'y - \hat{\beta}' X' y \qquad (4.6)$$

These are the same as *SSR* and *SSE* in (4.1) [see (3.24), (3.39), (3.54), and Problems 3.19, 3.25].

Example 4.1: Using the data in Table 3.1, we illustrate the test of $H_0 : \beta_1 = 0$ where, in this case, $\beta_1 = (\beta_1, \beta_2)'$. In Example 3.3.1(a), we found $X' y = (90, \ 482, \ 872)'$ and $\hat{\beta} = (5.3754, \ 3.0118, \ -1.2855)'$. The quantities $y'y$, $\hat{\beta}' X' y$, and $n\bar{y}^2$ are given by

$$y'y = \sum_{i=1}^{12} y_i^2 = 2^2 + 3^2 + \cdots + 14^2 = 840$$

$$\hat{\beta}'\,X'\,y = (5.3754,\quad 3.0118,\quad -1.2855)\begin{pmatrix}90\\482\\872\end{pmatrix} = 814.5410$$

TABLE 4.2: ANOVA for Overall Reg. Test for Data in Table 3.1

Source	df	SS	MS	*F*
Due to β_1	2	139.5410	69.7705	24.665
Error	9	25.4590	2.8288	
Total	11	165.000		

$$n\bar{y}^2 = n\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2 = 12\left(\frac{90}{12}\right)^2 = 675$$

Thus, by (4.6), we obtain

$$SSR = \hat{\beta}'\,X'\,y - n\bar{y}^2 = 139.5410\ , \qquad SSE = y'y - \hat{\beta}'\,X'\,y = 25.4590$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = y'y - n\bar{y}^2 = 165$$

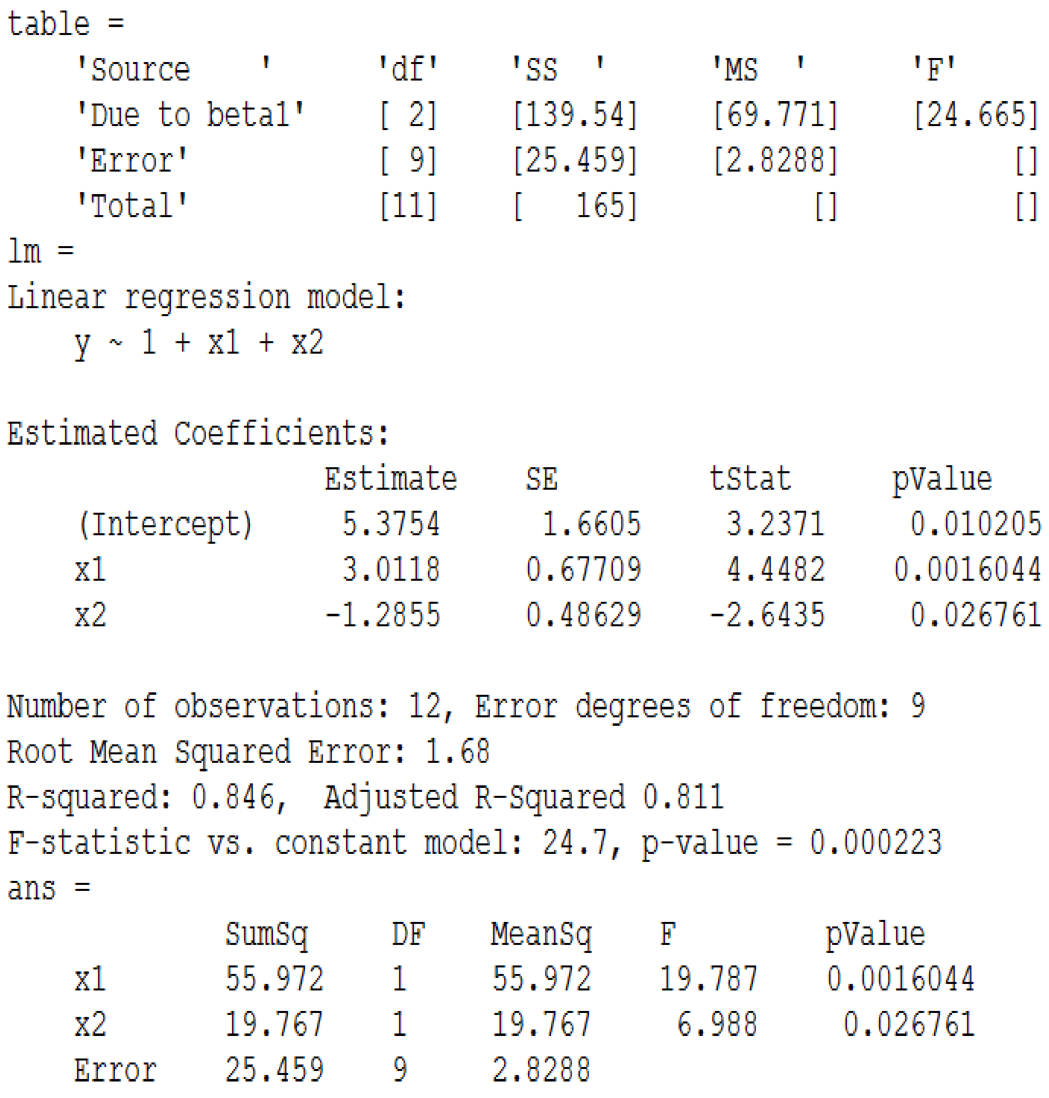
The F test is given in Table 4.2. Since $24.665 > F_{0.05,2,9} = 4.26$, we reject $H_0 : \beta_1 = 0$ and conclude that at least one of β_1 or β_2 is not zero. The *p*-value is .000223.

Example 4.1[The program name ta11.m]	Applications using MATLAB

```
Clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]'; x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]';
n=length(x1);k=2;x=[ones(size(x1)) x1 x2];
beta=x\y, Y=x*beta; e=Y-y;MSE=e'*e/(n-3);
% test of H0:beta1=0
beta1=beta(2:3),xy=x'*y,yy=y'*y,betaxy=beta'*x'*y
SSR=betaxy-n*mean(y)^2,SSE=yy-betaxy
SST=yy-n*mean(y)^2
MSR=SSR/k, MSE=SSE/(n-k-1),F=MSR/MSE
table=[{'Source  ', 'df' ,'SS ' ,'MS ' , 'F'} ; {'Due to beta1' k SSR MSR
F};{'Error' n-k-1 SSE MSE []};{'Total' n-1 SST [] []}]
% direect method
d=[x1 x2];lm=LinearModel.fit(d,y),anova(lm)
```

Ans.

```
beta =         beta1 =       xy =       yy =
       5.3754         3.0118     90         840
       3.0118        -1.2855    482
      -1.2855                   872
betaxy =       SSR =         SSE =      SST =
       814.54         139.54    25.459     165
MSR =          MSE =         F =
       69.771         2.8288    24.665
```



```
table =
    'Source    '      'df'    'SS    '    'MS    '    'F'
    'Due to beta1'    [ 2]    [139.54]    [69.771]    [24.665]
    'Error'           [ 9]    [25.459]    [2.8288]          []
    'Total'           [11]    [   165]          []          []
lm =
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate    SE         tStat      pValue
    (Intercept)     5.3754      1.6605     3.2371     0.010205
    x1              3.0118     0.67709     4.4482    0.0016044
    x2             -1.2855     0.48629    -2.6435     0.026761

Number of observations: 12, Error degrees of freedom: 9
Root Mean Squared Error: 1.68
R-squared: 0.846,  Adjusted R-Squared 0.811
F-statistic vs. constant model: 24.7, p-value = 0.000223
ans =
             SumSq     DF    MeanSq    F         pValue
    x1       55.972    1     55.972    19.787    0.0016044
    x2       19.767    1     19.767     6.988     0.026761
    Error    25.459    9     2.8288
```

4.2: Test on A Subset of The β's

In more generality, suppose that we wish to test the hypothesis that a subset of the *x*'s is not useful in predicting *y*. A simple example is $H_0 : \beta_j = 0$ for a single β_j. If H_0 is rejected, we would retain $\beta_j x_j$ in the model. As another illustration, consider the model in (3.2).

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$

For which we may wish to test the hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$. If H_0 is rejected, we would choose the full second-order model over the reduced first-order model.

Without loss of generality, we assume that the β's to be tested have been arranged last in $\mathrm{\beta}$, with a corresponding arrangement of the columns of **X**. Then $\mathrm{\beta}$ and **X** can be partitioned accordingly, and by (3.78), the model for all *n* observations becomes

$$\mathrm{y} = \mathrm{X}\mathrm{\beta} + \varepsilon = (\mathrm{X}_1, \mathrm{X}_2)\begin{pmatrix} \mathrm{\beta}_1 \\ \mathrm{\beta}_2 \end{pmatrix} + \varepsilon$$

$$= \mathrm{X}_1\mathrm{\beta}_1 + \mathrm{X}_2\mathrm{\beta}_2 + \varepsilon \qquad (4.7)$$

Where $\mathrm{\beta}_2$ contains the β's to be tested. The intercept β_0 would ordinarily be included in $\mathrm{\beta}_1$.

The hypothesis of interest is $H_0 : \mathrm{\beta}_2 = 0$. If we designate the number of parameters in $\mathrm{\beta}_2$ by *h*, then X_2 is $n \times h$, $\mathrm{\beta}_1$ is $(k-h+1)\times 1$, and X_1 is $n \times (k-h+1)$. Thus $\mathrm{\beta}_1 = (\beta_0, \beta_1, \cdots, \beta_{k-h})'$ and $\mathrm{\beta}_2 = (\beta_{k-h+1}, \beta_{k-h+2}, \cdots, \beta_k)'$. In terms of the illustration at the beginning of this section, we would have $\mathrm{\beta}_1 = (\beta_0, \beta_1, \beta_2)'$ and $\mathrm{\beta}_2 = (\beta_3, \beta_4, \beta_5)'$. Note that $\mathrm{\beta}_1$ in (4.7) is different from $\mathrm{\beta}_1$ in Section 4.1, in which $\mathrm{\beta}$ was partitioned as $\mathrm{\beta} = \begin{pmatrix} \beta_0 \\ \mathrm{\beta}_1 \end{pmatrix}$ and $\mathrm{\beta}_1$ constituted all of $\mathrm{\beta}$ except β_0.

To test $H_0 : \mathrm{\beta}_2 = 0$ versus $H_1 : \mathrm{\beta}_2 \neq 0$, we use a full–reduced-model approach. The full model is given by (4.7). Under $H_0 : \mathrm{\beta}_2 = 0$, the reduced model becomes

$$\mathrm{y} = \mathrm{X}_1\mathrm{\beta}_1^* + \varepsilon^* \qquad (4.8)$$

We use the notation β_1^* and ε^* as in Section 3.9, because in the reduced model, β_1^* and ε^* will typically be different from β_1 and ε in the full model (unless X_1 and X_2 are orthogonal; see Theorem 3.9a and its corollary). The estimator of β_1^* in the reduced model (4.8) is $\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'y$, which is, in general, not the same as the first $k-h+1$ elements of $\hat{\beta} = (X'X\ )^{-1}X'y$ from the full model (4.7) (unless X_1 and X_2 are orthogonal; see Theorem 3.10).

In order to compare the fit of the full model (4.7) to the fit of the reduced model (4.8), we add and subtract $\hat{\beta}'X'y$ and $\hat{\beta}_1^{*'}X_1'y$ to the total corrected sum of squares $\sum_{i=1}^n (y_i - \bar{y})^2 = y'y - n\bar{y}^2$ so as to obtain the partitioning

$$y'y - n\bar{y}^2 = \left(y'y - \hat{\beta}'X'y\right) + \left(\hat{\beta}'X'y - \hat{\beta}_1^{*'}X_1'y\right) + \left(\hat{\beta}_1^{*'}X_1'y - n\bar{y}^2\right) \qquad (4.9)$$

Or

$$SST = SSE + SS(\beta_2|\beta_1) + SSR(reduced) \qquad (4.10)$$

where $SS(\beta_2|\beta_1) = \hat{\beta}'X'y - \hat{\beta}_1^{*'}X_1'y$ is the “extra” regression sum of squares due to β_2 after adjusting for β_1. Note that $SS(\beta_2|\beta_1)$ can also be expressed as

$$\begin{aligned} SS(\beta_2|\beta_1) &= \hat{\beta}'X'y - n\bar{y}^2 - \left(\hat{\beta}_1^{*'}X_1'y - n\bar{y}^2\right) \\ &= SSR(full) - SSR(reduced) \end{aligned}$$

which is the difference between the overall regression sum of squares for the full model and the overall regression sum of squares for the reduced model [see (4.6)].

If $H_0:\beta_2 = 0$ is true, we would expect $SS(\beta_2|\beta_1)$ to be small so that *SST* in (4.10) is composed mostly of $SSR(reduced)$ and *SSE*. If $\beta_2 \neq 0$, we expect $SS(\beta_2|\beta_1)$ to be larger and account for more of *SST*. Thus we are testing $H_0:\beta_2 = 0$ in the full model in which there are no restrictions on β_1. We are not ignoring β_1 (assuming $\beta_1 = 0$) but are testing $H_0:\beta_2 = 0$ *in the presence* of β_1, that is, above and beyond whatever β_1 contributes to *SST*.

To develop a test statistic based on $SS(\beta_2|\beta_1)$, we first write (4.9) in terms of quadratic forms in **y**. Using $\hat{\beta} = (X'X)^{-1}X'y$ and $\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'y$, (4.9) becomes

$$y'[I-(1/n)J]y = y'y - y'X(X'X)^{-1}X'y + y'X(X'X)^{-1}X'y$$

$$- y'X_1(X_1'X_1)^{-1}X_1'y + y'X_1(X_1'X_1)^{-1}X_1'y - y'\frac{1}{n}Jy$$

$$= y'[I - X(X'X)^{-1}X']y + y'[X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X']y$$

$$+ y'\left[X_1(X_1'X_1)^{-1}X_1' - \frac{1}{n}J\right]y \qquad (4.11)$$

$$= y'(I-H)y + y'(H-H_1)y + y'\left(H_1 - \frac{1}{n}J\right)y \qquad (4.12)$$

where $H = X(X'X)^{-1}X'$ and $H_1 = X_1(X_1'X_1)^{-1}X_1'$. The matrix $I-H$ was shown to be idempotent; with rank $n-k-1$, where k +1 is the rank of **X** (k+1 is also the number of elements in β). The matrix $H-H_1$ is shown to be idempotent in the following theorem.

Theorem 8.2a: The matrix $H - H_1 = X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1'$ is idempotent with rank *h*, where *h* is the number of elements in β_2.

Proof: Premultiplying **X** by **H**, we obtain

$$HX = X(X'X)^{-1}X'X = X$$

Or

$$X = [X(X'X)^{-1}X']X \qquad (4.13)$$

Partitioning **X** on the left side of (4.13) and the last **X** on the right side, we obtain

$$(X_1, X_2) = [X(X'X)^{-1}X'](X_1, X_2)$$

$$= [X(X'X)^{-1}X'X_1, X(X'X)^{-1}X'X_2]$$

Thus

$$\left.\begin{aligned} X_1 &= X(X'X)^{-1}X'X_1 \\ X_2 &= X(X'X)^{-1}X'X_2 \end{aligned}\right\} \qquad (4.14)$$

Simplifying HH_1 and $\mathrm{H}_1\mathrm{H}$ by (4.14) and its transpose, we obtain

$$\mathrm{HH}_1 = \mathrm{H}_1 \quad \textit{and} \quad \mathrm{H}_1\mathrm{H} = \mathrm{H}_1 \tag{4.15}$$

The matrices H and H_1 are idempotent. Thus

$$\begin{aligned}(\mathrm{H}-\mathrm{H}_1)^2 &= \mathrm{H}^2 - \mathrm{HH}_1 - \mathrm{H}_1\mathrm{H} + \mathrm{H}_1^2 \\ &= \mathrm{H} - \mathrm{H}_1 - \mathrm{H}_1 + \mathrm{H}_1 \\ &= \mathrm{H} - \mathrm{H}_1\end{aligned}$$

and $\mathrm{H}-\mathrm{H}_1$ is idempotent. For the rank of $\mathrm{H}-\mathrm{H}_1$, we have

$$\begin{aligned}rank(\mathrm{H}-\mathrm{H}_1) &= tr(\mathrm{H}-\mathrm{H}_1) \\ &= tr(\mathrm{H}) - tr(\mathrm{H}_1) \\ &= tr[\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'] - tr[\mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1'] \\ &= tr[\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}] - tr[\mathrm{X}_1'\mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}] \\ &= tr(\mathrm{I}_{k+1}) - tr(\mathrm{I}_{k-h+1}) = k+1-(k-h+1) = h\end{aligned}$$

We now find the distributions of $\mathrm{y}'(\mathrm{I}-\mathrm{H})\mathrm{y}$ and $\mathrm{y}'(\mathrm{H}-\mathrm{H}_1)\mathrm{y}$ in (4.12) and show that they are independent.

Theorem 8.2b: If **y** is $N_n(\mathrm{X}\beta, \sigma^2\mathrm{I})$ and H and H_1 are as defined in (4.11) and (4.12), then

(i) $\mathrm{y}'(\mathrm{I}-\mathrm{H})\mathrm{y}/\sigma^2$ is $\chi^2_{(n-k-1)}$.

(ii) $\mathrm{y}'(\mathrm{H}-\mathrm{H}_1)\mathrm{y}/\sigma^2$ is $\chi^2_{(h,\lambda_1)}$, $\lambda_1 = \beta_2'[\mathrm{X}_2'\mathrm{X}_2 - \mathrm{X}_2'\mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{X}_2]\beta_2/2\sigma^2$.

(iii) $\mathrm{y}'(\mathrm{I}-\mathrm{H})\mathrm{y}$ and $\mathrm{y}'(\mathrm{H}-\mathrm{H}_1)\mathrm{y}$ are independent.

Proof: Adding $\mathrm{y}'(1/n)\mathrm{J}\mathrm{y}$ to both sides of (4.12), we obtain the decomposition $\mathrm{y}'\mathrm{y} = \mathrm{y}'(\mathrm{I}-\mathrm{H})\mathrm{y} + \mathrm{y}'(\mathrm{H}-\mathrm{H}_1)\mathrm{y} + \mathrm{y}'\mathrm{H}_1\mathrm{y}$. The matrices $\mathrm{I}-\mathrm{H}$, $\mathrm{H}-\mathrm{H}_1$ and H_1 were shown to be idempotent in Theorem 4.2a. Hence all parts of the theorem follow. See Problem 4.9 for the derivation of λ_1.

If $\lambda_1 = 0$ in Theorem 4.2b(ii), then $\mathrm{y}'(\mathrm{H}-\mathrm{H}_1)\mathrm{y}/\sigma^2$ has the central chi-square distribution $\chi^2_{(h)}$. Since $\mathrm{X}_2'\mathrm{X}_2 - \mathrm{X}_2'\mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{X}_2$ is positive definite (see Problem 8), $\lambda_1 = 0$ if and only if $\beta_2 = 0$.

An F test for $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$ is given in the following theorem.

Theorem 8.2c: Let **y** be $N_n(X\beta, \sigma^2 I)$ and define an *F* statistic as follows:

$$F = \frac{y'(H - H_1)y/h}{y'(I - H)y/(n-k-1)} = \frac{SS(\beta_2|\beta_1)/h}{SSE/(n-k-1)} \qquad (4.16)$$

$$= \frac{\left(\hat{\beta}'X'y - \hat{\beta}_1^{*'}X_1'y\right)/h}{\left(y'y - \hat{\beta}'X'y\right)/(n-k-1)} \qquad (4.17)$$

where $\hat{\beta} = (X'X)^{-1}X'y$ is from the full model $y = X\beta + \varepsilon$ and $\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'y$ is from the reduced model $y = X_1\beta_1^* + \varepsilon^*$. The distribution of *F* in (4.17) is as follows:

(i) If $H_0 : \beta_2 = 0$ is false, then

F is distributed as $F_{(h, n-k-1, \lambda_1)}$,

Where $\lambda_1 = \beta_2'[X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2]\beta_2/2\sigma^2$

(ii) If $H_0 : \beta_2 = 0$ is true, then $\lambda_1 = 0$ and

F is distributed as $F_{(h, n-k-1)}$,

Proof:

(i) This result follows from Theorem 4.2b.

(ii) This result follows from Theorem 4.2b.

The test for $H_0 : \beta_2 = 0$ is carried out as follows: Reject H_0 if $F \geq F_{(\alpha, h, n-k-1)}$, where $F_{(\alpha,h,n-k-1)}$ is the upper a percentage point of the (central) *F* distribution. Alternatively, we reject H_0 if $p < \alpha$, where *p* is the *p*-value. Since $X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2$ is positive definite (see Problem 4.10), $\lambda_1 > 0$ if $H_0 : \beta_2 = 0$ is false. This justifies rejection of H_0 for large values of *F*.

Results and calculations leading to this *F* test are summarized in the ANOVA table (Table 4.3), where β_1 is $(k-h+1)\times 1$, β_2 is $h \times 1$, X_1 is $n \times (k-h+1)$, and X_2 is $n \times h$.

The entries in the column for expected mean squares are $E[SS(\beta_2|\beta_1)/h]$ and $E[SSE/(n-k-1)]$. For $E[SS(\beta_2|\beta_1)/h]$, see Problem 4.11. Note that if

H_0 is true, both expected mean squares (Table 4.3) are equal to σ^2, and if H_0 is false, $E[SS(\beta_2|\beta_1)/h] > E[SSE/(n-k-1)]$. since $X_2'X_2 - X_2'X_1 (X_1'X_1)^{-1} X_1'X_2$ is positive definite. This inequality provides another justification for rejecting H_0 for large values of *F*.

TABLE 4.3: ANOVA Table for *F*-Test of $H_0: \beta_2 = 0$

Source of Variation	df	Sum of Squares	Mean Square	Expected Mean Square
Due to β_2 adjusted for β_1	h	$SS(\beta_2\|\beta_1) = \hat{\beta}'X'y - \hat{\beta}_1^{*'} X_1'y$	$SS(\beta_2\|\beta_1)/h$	$\sigma^2 + \frac{1}{h}\beta_2'[X_2'X_2 - X_2' \times X_1(X_1'X_1)^{-1} X_1'X_2]\beta_2$
Error	n-k-1	$SSE = y'y - \hat{\beta}'X'y$	$SSE/(n-k-1)$	σ^2
Total	n-1	$SST = y'y - n\bar{y}^2$		

Example 4.2a: Consider the dependent variable y_2 in the chemical reaction data in Table 3.4 (see Problem 3.52 for a description of the variables). In order to check the usefulness of second-order terms in predicting y_2, we use as a full model, $y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3 + \varepsilon$, and test $H_0: \beta_4 = \beta_5 = \cdots = \beta_9 = 0$. For the full model, we obtain $\hat{\beta}'X'y - n\bar{y}^2 = 339.7888$, and for the reduced model $y_2 = \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3 + \varepsilon^*$, we have $\hat{\beta}_1^{*'} X_1'y - n\bar{y}^2 = 151.0022$. The difference is $\hat{\beta}'X'y - \hat{\beta}_1^{*'} X_1'y = 188.7866$. The error sum of squares is *SSE* = 60.6755, and the *F* statistic is given by (4.16) or Table 4.3 as

$$F = \frac{188.7866/6}{60.6755/9} = \frac{31.4644}{6.7417} = 4.6671$$

which has a *p*-value of 0.0198. Thus the second-order terms are useful in prediction of y_2. In fact, the overall *F* in (4.5) for the reduced model is 3.027 with *p* = 0.0623, so that x_1, x_2, and x_3 are inadequate for predicting y_2. The overall *F* for the full model is 5.600 with *p* = 0.0086.

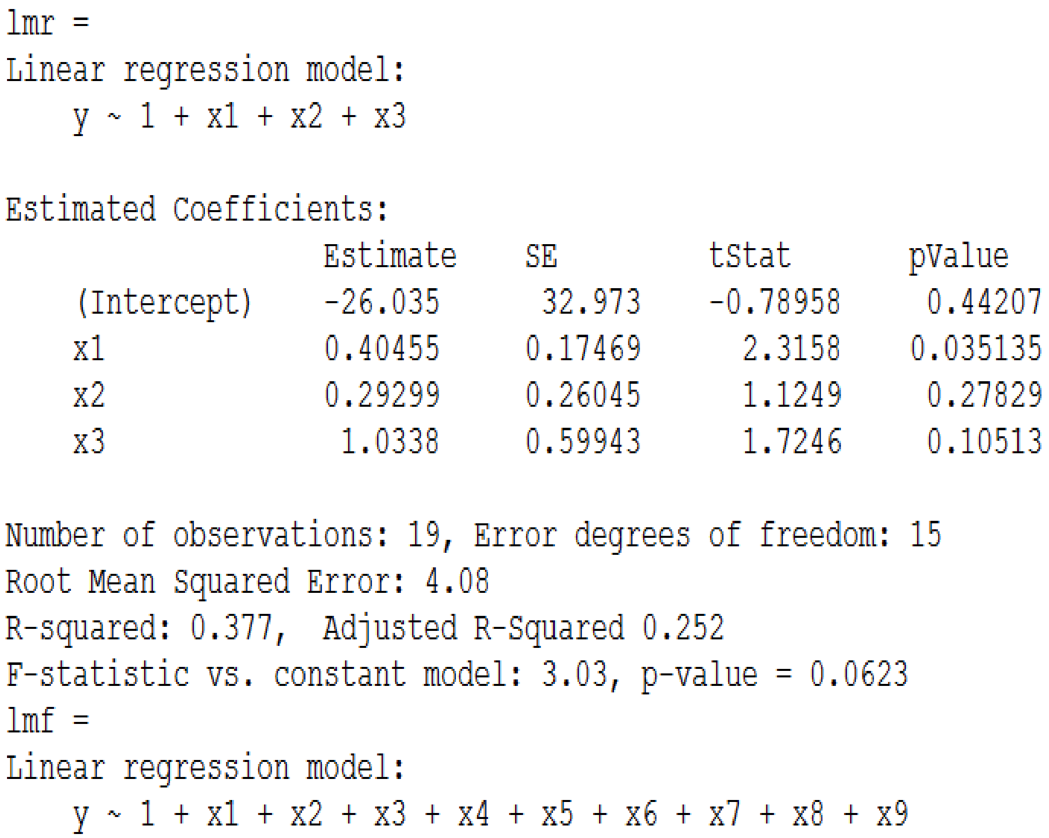
In the following theorem, we express $SS(\beta_2|\beta_1)$ as a quadratic form in $\hat{\beta}_2$ that corresponds to λ_1 in Theorem 4.2b(ii).

Example 4.2a [The program name ta12.m] | Applications using MATLAB

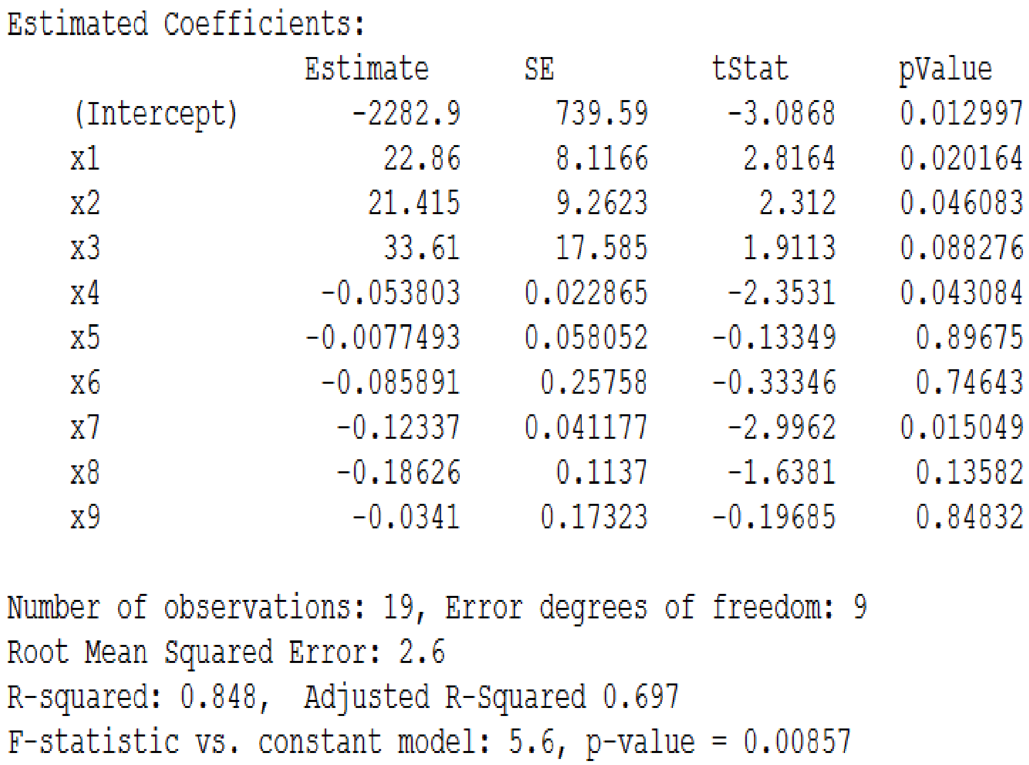
```
clc
data=[41.5 45.9 162 23 3;33.8 53.3 162 23 8;27.7 57.5 162 30 5
    21.7 58.8 162 30 8;19.9 60.6 172 25 5;15.0 58.0 172 25 8
    12.2 58.6 172 30 5;4.3 52.4 172 30 8;19.3 56.9 167 27.5 6.5
    6.4 55.4 177 27.5 6.5;37.6 46.9 157 27.5 6.5;18 57.3 167 32.5 6.5
    26.3 55.0 167 22.5 6.5;9.9 58.9 167 27.5 9.5;25.0 50.3 167 27.5 3.5
    14.1 61.1 177 20 6.5;15.2 62.9 177 20 6.5;15.9 60.0 160 34 7.5
    19.6 60.6 160 34 7.5];y2=data(:,2);x1=data(:,3);x2=data(:,4);
    x3=data(:,5);
% full model
x=[ones(size(x1)) x1 x2 x3 x1.^2 x2.^2 x3.^2 x1.*x2 x1.*x3 x2.*x3];
beta=x\y2;n=length(x1);
betaxy2corefac=beta'*x'*y2-n*mean(y2)^2
SSE=y2'*y2-beta'*x'*y2
Ff=(betaxy2corefac/9)/(SSE/9)
% reduced model
xr=[ones(size(x1)) x1 x2 x3];betar=xr\y2
betarxy2corefac=betar'*xr'*y2-n*mean(y2)^2
SSEr=y2'*y2-betar'*xr'*y2
Fr=(betarxy2corefac/3)/(SSEr/15)
% difference
SSD=betaxy2corefac-betarxy2corefac
% F-test
F=(SSD/6)/(SSE/9)
% test of R square for full model
SSR=betaxy2corefac
SST=SSR+SSE,RS=SSR/SST
% test of R square for redused model
SSRr=betarxy2corefac,SSTr=SSRr+SSEr
RSr=SSRr/SSTr
FR=((RS-RSr)/6)/((1-RS)/9)
% direect method for reduced model
d=[x1 x2 x3];lmr=LinearModel.fit(d,y2)
% direect method for full model
d=[x1    x2    x3    x1.^2    x2.^2    x3.^2    x1.*x2    x1.*x3    x2.*x3];
lmf=LinearModel.fit(d,y2),
```

Ans.

```
betaxy2corefac =    SSE =              Ff =
            339.79       60.675
                                       5.6001
betar =             betarxy2corefac=   SSEr =
      -26.035                   151       249.46
      0.40455
      0.29299
       1.0338
Fr =                SSD =              F =
       3.0266              188.79         4.6671
SSR =               SST =              RS =
       339.79              400.46        0.84849
SSRr =              SSTr =             RSr =
          151              400.46        0.37707
FR =
       4.6671
```



```
lmr =
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate     SE         tStat       pValue
    (Intercept)    -26.035      32.973    -0.78958      0.44207
    x1             0.40455     0.17469      2.3158     0.035135
    x2             0.29299     0.26045      1.1249      0.27829
    x3              1.0338     0.59943      1.7246      0.10513

Number of observations: 19, Error degrees of freedom: 15
Root Mean Squared Error: 4.08
R-squared: 0.377,  Adjusted R-Squared 0.252
F-statistic vs. constant model: 3.03, p-value = 0.0623
lmf =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
```



```
Estimated Coefficients:
                   Estimate        SE         tStat       pValue
    (Intercept)      -2282.9      739.59     -3.0868     0.012997
    x1                 22.86      8.1166      2.8164     0.020164
    x2                21.415      9.2623       2.312     0.046083
    x3                 33.61      17.585      1.9113     0.088276
    x4             -0.053803    0.022865     -2.3531     0.043084
    x5            -0.0077493    0.058052    -0.13349      0.89675
    x6             -0.085891     0.25758    -0.33346      0.74643
    x7              -0.12337    0.041177     -2.9962     0.015049
    x8              -0.18626      0.1137     -1.6381      0.13582
    x9               -0.0341     0.17323    -0.19685      0.84832

Number of observations: 19, Error degrees of freedom: 9
Root Mean Squared Error: 2.6
R-squared: 0.848,  Adjusted R-Squared 0.697
F-statistic vs. constant model: 5.6, p-value = 0.00857
```

Theorem 4.2d: If the model is partitioned as in (4.7), then $SS(\beta_2|\beta_1) = \hat{\beta}'X'y - \hat{\beta}_1^{*\prime} X_1'y$ can be written as

$$SS(\beta_2|\beta_1) = \hat{\beta}_2'[X_2'X_2 - X_2'X_1(X_1'X_1)^{-1} X_1'X_2]\hat{\beta}_2 \tag{4.18}$$

where $\hat{\beta}_2$ is from a partitioning of $\hat{\beta}$ in the full model:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1} X'y \tag{4.19}$$

Proof: We can write $X\hat{\beta}$ in terms of $\hat{\beta}_1$ and $\hat{\beta}_2$ as $X\hat{\beta} = (X_1, X_2)\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2$. To write $\hat{\beta}_1^*$ in terms of $\hat{\beta}_1$ and $\hat{\beta}_2$, we note that by (3.80), $E(\hat{\beta}_1^*) = \beta_1 + A\beta_2$, where $A = (X_1'X_1)^{-1} X_1'X_2$ is the alias matrix defined in Theorem 3.9a. This can be estimated by $\hat{\beta}_1^* = \hat{\beta}_1 + A\hat{\beta}_2$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are from the full model, as in (4.19). Then $SS(\beta_2|\beta_1)$ in (4.10) or Table 4.3 can be written as

$$SS(\beta_2|\beta_1) = \hat{\beta}'X'y - \hat{\beta}_1^{*'}X_1'y$$

$$= \hat{\beta}'X'X\hat{\beta} - \hat{\beta}_1^{*'}X_1'X_1\hat{\beta}_1^* \text{ [by (3.8)]}$$

$$= (\hat{\beta}_1'X_1' + \hat{\beta}_2'X_2')(X_1\hat{\beta}_1 + X_2\hat{\beta}_2) - (\hat{\beta}_1 + \hat{\beta}_2'A')X_1'X_1(\hat{\beta}_1 + A\hat{\beta}_2)$$

Multiplying this out and substituting $(X_1'X_1)^{-1}X_1'X_2$ for **A**, we obtain (4.18).

In (4.18), it is clear that $SS(\beta_2|\beta_1)$ is due to β_2. We also see in (4.18) a direct correspondence between $SS(\beta_2|\beta_1)$ and the non-centrality parameter λ_1 in Theorem 4.2b (ii) or the expected mean square in Table 4.3.

Example 4.2b: The full–reduced-model test of $H_0:\beta_2=0$ in Table 4.3 can be used to test for significance of a single $\hat{\beta}_j$. To illustrate, suppose that we wish to test $H_0:\beta_k=0$, where β is partitioned as

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \\ \hline \beta_k \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_k \end{pmatrix}$$

Then **X** is partitioned as $X = (X_1, x_k)$, where x_k is the last column of **X** and X_1 contains all columns except x_k. The reduced model is $y = X_1\beta_1^* + \varepsilon^*$, and β_1^* is estimated as $\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'y$. In this case, *h* = 1, and the *F* statistic in (4.17) becomes

$$F = \frac{\hat{\beta}'X'y - \hat{\beta}_1^{*'}X_1'y}{(y'y - \hat{\beta}'X'y)/(n-k-1)} \qquad (4.20)$$

which is distributed as $F_{(1,n-k-1)}$ if $H_0:\beta_k=0$ is true.

Example 4.2c: The test in Section 4.1 for overall regression can be obtained as a full-reduced-model test. In this case, the partitioning of **X** and of β is $X = (j, X_1)$ and

$$\beta = \begin{pmatrix} \beta_0 \\ \hline \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

The reduced model is $\mathrm{y} = \beta_0^* \mathrm{j} + \varepsilon^*$, for which we have

$$\hat{\beta}_0^* = \bar{y} \quad and \quad SS\left(\beta_0^*\right) = n\bar{y}^2 \tag{4.21}$$

(see Problem 4.13). Then $SS(\beta_1|\beta_0) = \hat{\beta}'\mathrm{X}'\mathrm{y} - n\bar{y}^2$, which is the same as (4.6).

4.3: *F* Test in terms of R^2

The *F* statistics in Sections 4.1 and 4.2 can be expressed in terms of R^2 as defined in (3.56).

Theorem 4.3: The *F* statistics in (4.5) and (4.17) for testing $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$, respectively, can be written in terms of R^2 as

$$F = \frac{\left(\hat{\beta}'\mathrm{X}'\mathrm{y} - n\bar{y}^2\right)/k}{\left(\mathrm{y}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\,\mathrm{y}\right)/(n-k-1)} \tag{4.22}$$

$$= \frac{R^2/k}{\left(1-R^2\right)/(n-k-1)} \tag{4.23}$$

And

$$F = \frac{\left(\hat{\beta}'\mathrm{X}'\mathrm{y} - \hat{\beta}_1^{*\prime}\mathrm{X}_1'\mathrm{y}\right)/h}{\left(\mathrm{y}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\mathrm{y}\right)/(n-k-1)} \tag{4.24}$$

$$= \frac{\left(R^2 - R_r^2\right)/h}{\left(1-R^2\right)/(n-k-1)} \tag{4.25}$$

where R^2 for the full model is given in (3.56) as $R^2 = \left(\hat{\beta}'\mathrm{X}'\mathrm{y} - n\bar{y}^2\right)/\left(\mathrm{y}'\mathrm{y} - n\bar{y}^2\right)$ and R_r^2 for the reduced model $\mathrm{y} = \mathrm{X}_1\beta_1^* + \varepsilon^*$ in (4.8) is similarly defined as

$$R_r^2 = \frac{\hat{\beta}_1^{*\prime}\mathrm{X}_1'\mathrm{y} - n\bar{y}^2}{\mathrm{y}'\mathrm{y} - n\bar{y}^2} \tag{4.26}$$

Proof: Adding and subtracting $n\bar{y}^2$ in the denominator of (4.22) gives

$$F = \frac{\left(\hat{\beta}'\mathrm{X}'\mathrm{y} - n\bar{y}^2\right)/k}{\left[\mathrm{y}'\mathrm{y} - n\bar{y}^2 - \left(\hat{\beta}'\mathrm{X}'\mathrm{y} - n\bar{y}^2\right)\right]/(n-k-1)}$$

Dividing numerator and denominator by $\mathrm{y}'\mathrm{y} - n\bar{y}^2$ yields (4.23). For (4.25), see Problem 4.15.

In (4.25), we see that the F test for $H_0 : \beta_2 = 0$ is equivalent to a test for significant reduction in R^2. Note also that since $F \geq 0$ in (4.25), we have $R^2 \geq R_r^2$, which is an additional confirmation of property 3 in Section 3.7, namely, that adding an x to the model increases R^2.

Example 4.3: For the dependent variable y_2 in the chemical reaction data in Table 3.4, a full model with nine x's and a reduced model with three x's were considered in Example 4.2a. The values of R^2 for the full model and reduced model are 0.8485 and 0.3771, respectively. To test the significance of the increase in R^2 from 0.3771 to 0.8485, we use (4.25)

$$F = \frac{(R^2 - R_r^2)/h}{(1 - R^2)/(n - k - 1)} = \frac{(0.8485 - 0.3771)/6}{(1 - 0.8485)/9}$$

$$= \frac{0.07857}{0.01683} = 4.6671$$

which is the same as the value obtained for F in Example 4.2a.

Example 4.3 [The program name ta13.m]	Applications using MATLAB

```
clc
data=[41.5 45.9 162 23 3;33.8 53.3 162 23 8;27.7 57.5 162 30 5
    21.7 58.8 162 30 8;19.9 60.6 172 25 5;15.0 58.0 172 25 8
    12.2 58.6 172 30 5;4.3 52.4 172 30 8;19.3 56.9 167 27.5 6.5
    6.4 55.4 177 27.5 6.5;37.6 46.9 157 27.5 6.5;18 57.3 167 32.5 6.5
    26.3 55.0 167 22.5 6.5;9.9 58.9 167 27.5 9.5;25.0 50.3 167 27.5 3.5
    14.1 61.1 177 20 6.5;15.2 62.9 177 20 6.5;15.9 60.0 160 34 7.5
    19.6 60.6 160 34 7.5];y2=data(:,2);x1=data(:,3);x2=data(:,4);
    x3=data(:,5);
% full model
x=[ones(size(x1)) x1 x2 x3 x1.^2 x2.^2 x3.^2 x1.*x2 x1.*x3 x2.*x3];
beta=x\y2;n=length(x1);betaxy2corefac=beta'*x'*y2-n*mean(y2)^2;
SSE=y2'*y2-beta'*x'*y2;
% reduced model
xr=[ones(size(x1)) x1 x2 x3];betar=xr\y2;
betarxy2corefac=betar'*xr'*y2-n*mean(y2)^2;
SSEr=y2'*y2-betar'*xr'*y2;
% Compute R square for full model
SSR=betaxy2corefac;SST=SSR+SSE;RS=SSR/SST
```

```
% F-test by using R square for redused model
SSRr=betarxy2corefac;SSTr=SSRr+SSEr;RSr=SSRr/SSTr
FR=((RS-RSr)/6)/((1-RS)/9)
```

Ans.

```
        RS =

                0.84849

        RSr =

                0.37707

        FR =

                4.6671
```

4.4: The General Linear Hypothesis Test for $H_0 : \mathrm{C}\beta = 0$ and $H_0 : \mathrm{C}\beta = \mathrm{t}$

We discuss a test for $H_0 : \mathrm{C}\beta = 0$ in Section 4.4.1 and a test for $H_0 : \mathrm{C}\beta = \mathrm{t}$ in Section 4.4.2.

4.4.1: The Test for $H_0 : \mathrm{C}\beta = 0$

The hypothesis $H_0 : \mathrm{C}\beta = 0$, where C is a known $q \times (k+1)$ coefficient matrix of rank $q \leq k+1$, is known as the *general linear hypothesis*. The alternative hypothesis is $H_1 : \mathrm{C}\beta \neq 0$. The formulation $H_0 : \mathrm{C}\beta = 0$ includes as special cases the hypotheses in Sections 4.1 and 4.2. The hypothesis $H_0 : \beta_1 = 0$ in Section 4.1 can be expressed in the form $H_0 : \mathrm{C}\beta = 0$ as follows

$$H_0 : \mathrm{C}\beta = (0, \mathrm{I}_k)\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \beta_1 = 0$$

where **0** is a $k \times 1$ vector. Similarly, the hypothesis $H_0 : \beta_2 = 0$ in Section 4.2 can be expressed in the form $H_0 : \mathrm{C}\beta = 0$:

$$H_0 : \mathrm{C}\beta = (\mathrm{O}, \mathrm{I}_h)\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \beta_2 = 0$$

where the matrix O is $h \times (k - h + 1)$ and the vector **0** is $h \times 1$.

The formulation $H_0 : \mathrm{C}\beta = 0$ also allows for more general hypotheses such as

$$H_0 : 2\beta_1 - \beta_2 = \beta_2 - 2\beta_3 + 3\beta_4 = \beta_1 - \beta_4 = 0$$

which can be expressed as follows:

$$H_0 : \begin{pmatrix} 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 3 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

As another illustration, the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ can be expressed in terms of three differences, $H_0 : \beta_1 - \beta_2 = \beta_2 - \beta_3 = \beta_3 - \beta_4 = 0$, or, equivalently, as $H_0 : \mathrm{C}\beta = 0$:

$$H_0 : \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

In the following theorem, we give the sums of squares used in the test of $H_0 : \mathrm{C}\beta = 0$ versus $H_1 : \mathrm{C}\beta \neq 0$, along with the properties of these sums of squares. We denote the sum of squares due to $\mathrm{C}\beta$ (due to the hypothesis) as *SSH*.

Theorem 4.4a: If **y** is distributed $N_n(\mathrm{X}\beta, \sigma^2 \mathrm{I})$ and **C** is $q \times (k+1)$ of rank $q \leq k+1$, then

(i) $\mathrm{C}\hat{\beta}$ is $N_q[\mathrm{C}\,\beta, \sigma^2\, \mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{C}']$.

(ii) $SSH/\sigma^2 = (\mathrm{C}\hat{\beta})'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{C}']^{-1}\,\mathrm{C}\hat{\beta}/\sigma^2$ is $\chi^2_{(q,\lambda)}$,

where $\lambda = (\mathrm{C}\beta)'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{C}']^{-1}\,\mathrm{C}\beta/2\sigma^2$.

(iii) $SSE/\sigma^2 = \mathrm{y}'[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{X}']\mathrm{y}/\sigma^2$ is $\chi^2_{(n-k-1)}$.

(iv) *SSH* and *SSE* are independent.

Proof:

(i) By Theorem 3.6b (i), $\hat{\beta}$ is $N_{k+1}[\beta, \sigma^2(X'X)^{-1}]$.

(ii) Since $Cov(C\hat{\beta}) = \sigma^2 C(X'X)^{-1} C'$ and $\sigma^2[C(X'X)^{-1} C']^{-1} C(X'X)^{-1} C'/\sigma^2 = I$.

(iii) This was established in Theorem 4.1b(ii).

(iv) Since $\hat{\beta}$ and *SSE* are independent [see Theorem 3.6b(iii)], $SSH = \hat{\beta}'C'[C(X'X)^{-1} C']^{-1} C\hat{\beta}$ and *SSE* are also independent (Seber 1977, pp. 17, 33–34). For a more formal proof, see Problem 4.16.

The *F* test for $H_0 : C\beta = 0$ versus $H_1 : C\beta \neq 0$ is given in the following theorem.

Theorem 4.4b: Let **y** be $N_n(X\beta, \sigma^2 I)$ and define the statistic

$$F = \frac{SSH/q}{SSE/(n-k-1)}$$

$$= \frac{(C\hat{\beta})'[C(X'X)^{-1} C']^{-1} C\hat{\beta}/q}{SSE/(n-k-1)} \qquad (4.27)$$

where **C** is $q \times (k+1)$ of rank $q \le k+1$, and $\hat{\beta} = (X'X)^{-1} X'y$. The distribution of *F* in (4.27) is as follows:

(i) If $H_0 : C\beta = 0$ is false, then

F is distributed as $F_{(q, n-k-1, \lambda)}$,

Where $\lambda = (C\beta)'[C(X'X)^{-1} C']^{-1} C\beta/2\sigma^2$

(ii) If $H_0 : C\beta = 0$ is true, then *F* is distributed as $F_{(q, n-k-1)}$.

Proof:

(i) This result follows from Theorem 4.4a.

(ii) This result follows from Theorem 4.4a.

The *F* test for $H_0 : C\beta = 0$ in Theorem 4.4b is usually called the *general linear hypothesis test*. The degrees of freedom *q* is the number of linear combinations in $C\beta$. The test for $H_0 : C\beta = 0$ is carried out as follows. Reject H_0 if $F \geq F_{(\alpha, q, n-k-1)}$, where *F* is as given in (4.27) and

$F_{(\alpha, q, n-k-1)}$ is the upper α percentage point of the (central) F distribution. Alternatively, we can reject H_0 if $p \le \alpha$ where p is the p-value for F. [The p-value is the probability that $F_{(q, n-k-1)}$ exceeds the observed F value.] Since $\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}'$ is positive definite (see Problem 4.17), $\lambda > 0$ if H_0 is false, where $\lambda = (\mathrm{C}\beta)'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']^{-1}\mathrm{C}\beta/2\sigma^2$. Hence we reject $H_0 : \mathrm{C}\beta = 0$ for large values of F.

In Theorems 4.4a and 4.4b, *SSH* could be written as $(\mathrm{C}\hat{\beta} - 0)'$ $[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']^{-1}(\mathrm{C}\hat{\beta} - 0)'$, which is the squared distance between $\mathrm{C}\hat{\beta}$ and the hypothesized value of $\mathrm{C}\beta$. The distance is standardized by the covariance matrix of $\mathrm{C}\hat{\beta}$. Intuitively, if H_0 is true, $\mathrm{C}\hat{\beta}$ tends to be close to **0** so that the numerator of F in (4.27) is small. On the other hand, if $\mathrm{C}\beta$ is very different from **0**, the numerator of F tends to be large.

The expected mean squares for the F test are given by

$$\left.\begin{aligned} E\left(\frac{SSH}{q}\right) &= \sigma^2 + \frac{1}{q}(\mathrm{C}\beta)'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']^{-1}\mathrm{C}\beta \\ E\left(\frac{SSE}{n-k-1}\right) &= \sigma^2 \end{aligned}\right\} \tag{4.28}$$

These expected mean squares provide additional motivation for rejecting H_0 for large values of F. If H_0 is true, both expected mean squares are equal to σ^2; if H_0 is false, $E(SSH/q) > E[SSE/(n-q-1)]$.

The F statistic in (4.27) is invariant to full-rank linear transformations on the x's or on y.

Theorem 4.4c: Let $\mathrm{z} = \mathrm{c}\mathrm{y}$ and $\mathrm{W} = \mathrm{X}\mathrm{K}$, where K is nonsingular (see Corollary 1 to Theorem 3.3e for the form of K). The F statistic in (4.27) is unchanged by these transformations.

Proof: See Problem 4.18.

In the first paragraph of this section, it was noted that the hypothesis $H_0 : \beta_2 = 0$ can be expressed in the form $H_0 : \mathrm{C}\beta = 0$. Since we used a full–reduced-model approach to develop the test for $H_0 : \beta_2 = 0$, we expect that the general linear hypothesis test is also a full–reduced-model test. This is confirmed in the following theorem.

Theorem 4.4d: The *F* test in Theorem 4.4b for the general linear hypothesis $H_0 : \mathrm{C}\beta = 0$ is a full–reduced-model test.

Proof: The reduced model under H_0 is

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon \quad \text{subject to} \quad \mathrm{C}\beta = 0 \tag{4.29}$$

Using Lagrange multipliers, it can be shown (see Problem 4.19) that the estimator for β in this reduced model is

$$\hat{\beta}_c = \hat{\beta} - (\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']^{-1}\mathrm{C}\hat{\beta} \tag{4.30}$$

where $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ is estimated from the full model unrestricted by the hypothesis and the subscript *c* in $\hat{\beta}_c$ indicates that β is estimated subject to the constraint $\mathrm{C}\beta = 0$. In (4.29), the **X** matrix for the reduced model is unchanged from the full model, and the regression sum of squares for the reduced model is therefore $\hat{\beta}_c'\mathrm{X}'\mathrm{y}$ (for a more formal justification of $\hat{\beta}_c'\mathrm{X}'\mathrm{y}$, see Problem 4.20). Hence, the regression sum of squares due to the hypothesis is

$$SSH = \hat{\beta}'\mathrm{X}'\mathrm{y} - \hat{\beta}_c'\mathrm{X}'\mathrm{y} \tag{4.31}$$

By substituting $\hat{\beta}_c$ [as given by (4.30)] into (4.31), we obtain

$$SSH = (\mathrm{C}\hat{\beta})'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']^{-1}\mathrm{C}\hat{\beta} \tag{4.32}$$

(see Problem 4.21), thus establishing that the *F* test in Theorem 4.4b for $H_0 : \mathrm{C}\beta = 0$, is a full–reduced-model test.

Example 4.4.1a: In many cases, the hypothesis can be incorporated directly into the model to obtain the reduced model. Suppose that the full model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

and the hypothesis is $H_0 : \beta_1 = 2\beta_2$. Then the reduced model becomes

$$\begin{aligned} y_i &= \beta_0 + 2\beta_2 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \\ &= \beta_{c0} + \beta_{c2}(2x_{i1} + x_{i2}) + \beta_{c3} x_{i3} + \varepsilon_i \end{aligned}$$

where β_{ci} indicates a parameter subject to the constraint $\beta_1 = 2\beta_2$.
The full model and reduced model could be fit, and the difference

$SS(\beta_2|\beta_1) = \hat{\beta}'X'y - \hat{\beta}_1^{*\prime} X_1'y$ would be the same as *SSH* in (4.32).

If $C\beta \neq 0$, the estimator $\hat{\beta}_c$ in (4.30) is a biased estimator of β, but the variances of the $\hat{\beta}_{cj}$'s in $\hat{\beta}_c$ are reduced, as shown in the following theorem.

Theorem 4.4e: The mean vector and covariance matrix of $\hat{\beta}_c$ in (4.30) are as follows:

$$(i) \quad E(\hat{\beta}_c) = \beta - (X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}C\beta \tag{4.33}$$

$$(ii) \quad Cov(\hat{\beta}_c) = \sigma^2(X'X)^{-1} - \sigma^2(X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}C(X'X)^{-1} \tag{4.34}$$

Proof: See Problem 4.22.

Since the second matrix on the right side of (4.34) is positive semi-definite, the diagonal elements of $Cov(\hat{\beta}_c)$ are less than those of $Cov(\hat{\beta}_c) = \sigma^2(X'X)^{-1}$; that is, $Var(\hat{\beta}_{cj}) \leq Var(\hat{\beta}_j)$ for *j* = 0, 1, 2, … , *k*, where $\hat{\beta}_{cj}$ is the *jth* diagonal element of $Cov(\hat{\beta}_c)$ in (4.34). This is analogous to the inequality $Var(\hat{\beta}_j^*) < Var(\hat{\beta}_j)$ in Theorem 3.9c, where $\hat{\beta}_j^*$ is from the reduced model.

Example 4.4.1b: Consider the dependent variable y_1 in the chemical reaction data in Table 3.4. For the model $y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, we test $H_0 : 2\beta_1 = 2\beta_2 = \beta_3$ using (4.27) in Theorem 4.4b. To express H_0 in the form $C\beta = 0$, the matrix **C** becomes

$$C = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 2 & -1 \end{pmatrix}$$

and we obtain

$$C\hat{\beta} = \begin{pmatrix} -0.1214 \\ -0.6118 \end{pmatrix}$$

$$C(X'X)^{-1}C' = \begin{pmatrix} 0.003366 & -0.006943 \\ -0.006943 & 0.044974 \end{pmatrix}$$

$$F = \frac{\begin{pmatrix} -0.1214 \\ -0.6118 \end{pmatrix}' \begin{pmatrix} 0.003366 & -0.006943 \\ -0.006943 & 0.044974 \end{pmatrix} \begin{pmatrix} -0.1214 \\ -0.6118 \end{pmatrix} \Big/ 2}{5.3449}$$

$$= \frac{28.62301/2}{5.3449} = 2.6776$$

Which has $p = 0.101$

Example 4.4.1b [The program name ta14.m]	Applications using MATLAB

```
clc
data=[41.5 45.9 162 23 3;33.8 53.3 162 23 8;27.7 57.5 162 30 5
     21.7 58.8 162 30 8;19.9 60.6 172 25 5;15.0 58.0 172 25 8
     12.2 58.6 172 30 5;4.3 52.4 172 30 8;19.3 56.9 167 27.5 6.5
     6.4 55.4 177 27.5 6.5;37.6 46.9 157 27.5 6.5;18 57.3 167 32.5 6.5
     26.3 55.0 167 22.5 6.5;9.9 58.9 167 27.5 9.5;25.0 50.3 167 27.5 3.5
     14.1 61.1 177 20 6.5;15.2 62.9 177 20 6.5;15.9 60.0 160 34 7.5
     19.6 60.6 160 34 7.5];y2=data(:,2);x1=data(:,3);x2=data(:,4);
     x3=data(:,5);
y1=data(:,1);x1=data(:,3);x2=data(:,4);x3=data(:,5);
% Test of H0: 2beta1=2beta2=beta3
x=[ones(size(x1)) x1 x2 x3];beta=x\y1
C=[0 1 -1 0;0 0 2 -1],Cbeta=C*beta,A=C*inv(x'*x)*C'
SSE=y1'*y1-beta'*x'*y1,MSE=SSE/15
F=(Cbeta'*inv(A)*Cbeta/2)/MSE,p=1-fcdf(F,2,15)
```

```
beta =              C =                             Cbeta =

      332.11           0    1   -1    0                    -0.1214

      -1.546           0    0    2   -1                    -0.61175

      -1.4246

      -2.2374

SSE =               A =                             MSE =

      80.174           0.0033664  -0.0069425               5.3449

                      -0.0069425    0.044974

F =                 p =

     2.6776               0.1013
```

4.4.2: The Test for $H_0 : \mathrm{C}\beta = \mathrm{t}$

The test for $H_0 : \mathrm{C}\beta = \mathrm{t}$ is a straightforward extension of the test for $H_0 : \mathrm{C}\beta = 0$.With the additional flexibility provided by t, we can test hypotheses such as $H_0 : \beta_2 = \beta_1 + 5$. We assume that the system of equations $\mathrm{C}\beta = \mathrm{t}$ is consistent, that is, that *rank*(C) = *rank*(C, t). The requisite sums of squares and their properties are given in the following theorem, which is analogous to Theorem 4.4a.

Theorem 4.4f: If **y** is $N_n(\mathrm{X}\beta, \sigma^2 \mathrm{I})$ and C is $q \times (k+1)$ of rank $q \le k+1$, then

(i) $\mathrm{C}\hat{\beta} = \mathrm{t}$ is $N_n[\mathrm{C}\beta - \mathrm{t}, \sigma^2 \mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']$.

(ii) $SSH/\sigma^2 = (\mathrm{C}\hat{\beta} - \mathrm{t})'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']^{-1}(\mathrm{C}\hat{\beta} - \mathrm{t})/\sigma^2$ is $\chi^2_{(q,\lambda)}$

where $\lambda = (\mathrm{C}\beta - \mathrm{t})'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']^{-1}(\mathrm{C}\beta - \mathrm{t})/2\sigma^2$.

(iii) $SSE/\sigma^2 = \mathrm{y}'[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\mathrm{y}/\sigma^2$ is $\chi^2_{(n-k-1)}$.

(iv) *SSH* and *SSE* are independent.

Proof:

(i) By Theorem 3.6b (i), $\hat{\beta}$ is $N_{k+1}[\beta, \sigma^2(\mathrm{X}'\mathrm{X})^{-1}]$.

(ii) By part (i), $Cov(\mathrm{C}\hat{\beta} - \mathrm{t}) = \sigma^2 \mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}'$. The result follows as in the proof of Theorem 4.4a (ii).

(iii) See Theorem 4.1b (ii).

(iv) Since $\hat{\beta}$ and *SSE* are independent [see Theorem 3.6b (iii)], *SSH* and *SSE* are independent [see Seber (1977, pp. 17, 33–34)]. For a more formal proof, see Problem 4.23.

An *F* test for $H_0 : \mathrm{C}\beta = \mathrm{t}$ versus $H_0 : \mathrm{C}\beta = \mathrm{t}$ is given in the following theorem, which is analogous to Theorem 4.4b.

Theorem 4.4g: Let **y** be $N_n(\mathrm{X}\beta, \sigma^2 \mathrm{I})$ and define an *F* statistic as follows:

$$F = \frac{SSH/q}{SSE/(n-k-1)}$$

$$= \frac{(\mathrm{C}\hat{\beta} - \mathrm{t})'[\mathrm{C}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{C}']^{-1}(\mathrm{C}\hat{\beta} - \mathrm{t})/q}{SSE/(n-k-1)} \quad (4.35)$$

where $\hat{\beta} = (X'X)^{-1} X'y$. The distribution of *F* in (4.35) is as follows:

(i) If $H_0 : C\beta = t$ is false, then

$$F \text{ is distributed as } F_{(q,n-k-1,\lambda)},$$

$$\text{Where } \lambda = (C\beta - t)' [C(X'X)^{-1} C']^{-1} (C\beta - t) / 2\sigma^2$$

(ii) If $H_0 : C\beta = t$ is true, then $\lambda = 0$ and

$$F \text{ is distributed as } F_{(q,n-k-1)}.$$

Proof:

(i) This result follows from Theorem 4.4f.
(ii) This result follows from Theorem 4.4f.

The test for $H_0 : C\beta = t$ is carried out as follows. Reject H_0 if $F \geq F_{(\alpha,q,n-k-1)}$, where $F_{(\alpha,q,n-k-1)}$ is the upper α percentage point of the central *F* distribution. Alternatively, we can reject H_0 if $p \leq \alpha$, where *p* is the *p*-value for *F*.

The expected mean squares for the *F* test are given by

$$\left.\begin{aligned} E\left(\frac{SSH}{q}\right) &= \sigma^2 + \frac{1}{q}(C\beta - t)' [C(X'X)^{-1} C']^{-1} (C\beta - t) \\ E\left(\frac{SSE}{n-k-1}\right) &= \sigma^2 \end{aligned}\right\} \qquad (4.36)$$

By extension of Theorem 4.4d, the *F* test for $H_0 : C\beta = t$ in Theorem 4.4g is a full–reduced-model test (see Problem 4.24 for a partial result).

4.5: Tests on β_j and $a'\beta$

We consider tests for a single β_j or a single linear combination $a'\beta$ in Section 4.5.1 and tests for several β_j's or several $a'\beta$'s in Section 4.5.2.

4.5.1: Testing One β_j or One $a'\beta$

Tests for an individual β_j can be obtained using either the full–reduced-model approach in Section 4.2 or the general linear hypothesis approach in Section 4.4 The test statistic for $H_0 : \beta_k = 0$ using a full–reduced–model is given in (4.20) as

$$F = \frac{\hat{\beta}'X'y - \hat{\beta}_1^{*'} X_1'y}{SSE/(n-k-1)} \qquad (4.37)$$

which is distributed as $F_{(1,n-k-1)}$ if H_0 is true. In this case, β_k is the last β, so that b is partitioned as $\beta = \begin{pmatrix} \beta_1 \\ \beta_k \end{pmatrix}$ and **X** is partitioned as $X = (X_1, x_k)$, where x_k is the last column of **X**. Then X_1 in the reduced model $y = X_1 \beta_1^* + \varepsilon^*$ contains all the columns of **X** except the last.

To test $H_0 : \beta_j = 0$ by means of the general linear hypothesis test of $H_0 : C\beta = 0$ (Section 4.4.1), we first consider a test of $H_0 : a'\beta = 0$ for a single linear combination, for example, $a'\beta = (0,\ 2,\ -2,\ 3,\ 1)\beta$. Using a' in place of the matrix **C** in $C\beta = 0$, we have *q* = 1, and (4.27) becomes

$$F = \frac{(a'\hat{\beta})'[a'(X'X)^{-1}a]^{-1}a'\hat{\beta}}{SSE/(n-k-1)} = \frac{(a'\hat{\beta})^2}{S^2 a'(X'X)^{-1}a} \qquad (4.38)$$

where $S^2 = SSE/(n-k-1)$.The *F* statistic in (4.38) is distributed as $F_{(1,n-k-1)}$ if $H_0 : a'\beta = 0$ is true. To test $H_0 : \beta_j = 0$ using (4.38), we define $a' = (0,\ \cdots\ ,0,\ 1,\ 0,\ \cdots\ ,0)$, where the 1 is in the *jth* position. This gives

$$F = \frac{\hat{\beta}_j^2}{S^2 g_{jj}} \qquad (4.39)$$

where g_{jj} is the *jth* diagonal element of $(X'X)^{-1}$. If $H_0 : \beta_j = 0$ is true, *F* in (4.39) is distributed as $F_{(1,n-k-1)}$. We reject $H_0 : \beta_j = 0$ if $F \geq F_{(\alpha,1,n-k-1)}$ or, equivalently, if $p \leq \alpha$, where *p* is the *p*-value for F.

By Theorem 4.4d (see also Problem 4.25), the *F* statistics in (4.37) and (4.39) are the same (for *j* = *k*). This confirms that (4.39) tests $H_0 : \beta_j = 0$ adjusted for the other β's.

Since the *F* statistic in (4.39) has 1 and *n*-*k*-1 degrees of freedom, we can equivalently use the *t* statistic

$$t_j = \frac{\hat{\beta}_j}{S\sqrt{g_{jj}}} \qquad (4.40)$$

to test the effect of β_j above and beyond the other β's . We reject $H_0 : \beta_j = 0$ if $|t_j| \geq t_{\alpha/2,n-k-1}$ or, equivalently, if $p \leq \alpha$, where *p* is the *p*-value. For a two-tailed *t* test such as this one, the *p*-value is twice the probability that $t_{(n-k-1)}$ exceeds the absolute value of the observed *t*.

For *j* = 1, (4.40) becomes $t_1 = \hat{\beta}_1 / S\sqrt{g_{11}}$, which is not the same as $t = \hat{\beta}_1 \Big/ \left[S \Big/ \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$ in (2.14). Unless the *x*'s are orthogonal, $g_{11}^{-1} \neq \sum_{i=1}^{n} (x_i - \bar{x})^2$.

4.5.2: Testing Several β_j's or $\mathrm{a}_i' \beta$'s

We sometimes want to carry out several separate tests rather than a single joint test of the hypotheses. For example, the test in (4.40) might be carried out separately for each β_i, *i* = 1, . . . , *k* rather than the joint test of $H_0 : \mathrm{\beta}_1 = 0$ in (4.5). Similarly, we might want to carry out separate tests for several (say, *d*) $\mathrm{a}_i \beta$'s using (4.38) rather than the joint test of $H_0 : \mathrm{C\beta} = 0$ using (4.27), where

$$\mathrm{C} = \begin{pmatrix} \mathrm{a}_1 \\ \mathrm{a}_2 \\ \vdots \\ \mathrm{a}_d \end{pmatrix}$$

In such situations there are two different α levels, the overall or *familywise* α level (α_f) and the α level for each test or *comparisonwise* α level (α_c). In some cases researchers desire to control (α_c) when doing several tests (Saville 1990), and so no changes are needed in the testing procedure. In other cases, the desire is to control (α_f). In yet other cases, especially those involving thousands of separate tests (e.g., microarray data), it makes sense to control other quantities such as the false discovery rate (Benjamini and Hochberg 1995, Benjamini and Yekutieli 2001). This will not be discussed further here. We consider two ways to control (α_f) when several tests are made.

The first of these methods is the Bonferroni approach (Bonferroni 1936), which reduces ac for each test, so that (α_f) is less than the desired level of α^*. As an example, suppose that we carry out the *k* tests of $H_{0j}:\beta_j=0$, j = 1, 2, . . . , k. Let E_j be the event that the *jth* test rejects H_{0j} when it is true, where $P(E_j)=\alpha_c$. The overall (α_f) can be defined as

$$\begin{aligned}\alpha_f &= P(\mathit{reject\ \ at\ \ least\ \ one}\ \ H_{0j}\ \ \mathit{when\ \ all}\ \ H_{0j}\ \ \mathit{are\ \ true}) \\ &= P(E_1\ \ or\ \ E_2\ \ \cdots\ \ or\ \ E_k)\end{aligned}$$

Expressing this more formally and applying the Bonferroni inequality, we obtain

$$\left.\begin{aligned}\alpha_f &= P(E_1\ \cup\ E_2\ \cup\ \cdots\ \cup\ E_k) \\ &\le \textstyle\sum_{j=1}^{k} P(E_j) = \sum_{j=1}^{k} \alpha_c = k\alpha_c\end{aligned}\right\} \tag{4.41}$$

We can thus ensure that (α_f) is less than or equal to the desired (α^*) by simply setting $\alpha_c=\alpha^*/k$. Since (α_f) in (4.41) is at most (α^*), the Bonferroni procedure is a conservative approach.

To test $H_{0j}:\beta_j=0$, j = 1, 2, . . . , k, with $\alpha_f\le\alpha^*$, we use (4.40)

$$t_j=\hat{\beta}_j\big/S\sqrt{g_{jj}} \tag{4.42}$$

and reject H_{0j} if $|t_j|\ge t_{\alpha^*/2k,n-k-1}$. Bonferroni critical values $t_{\alpha^*/2k,v}$ are available in Bailey (1977). See also Rencher (2002, pp. 562–565). The critical values $t_{\alpha^*/2k,v}$ can also be found using many software packages. Alternatively, we can carry out the test by the use of p-values and reject H_{0j} if $p\le\alpha^*/k$.

More generally, to test $H_{0i}:\mathrm{a}_i'\beta=0$ for $i=1,2,\cdots,d$ with $\alpha_f\le\alpha^*$, we use (4.38)

$$F_i=\frac{(\mathrm{a}_i'\hat{\beta})'[\mathrm{a}_i'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{a}_i]^{-1}\mathrm{a}_i'\hat{\beta}}{S^2} \tag{4.43}$$

and reject H_{0i} if $F_i\ge F_{\alpha^*/d,1,n-k-1}$. The critical values $F_{\alpha^*/d}$ are available in many software packages. To use p-values, reject H_{0i} if $p\le\alpha^*/d$.

The above Bonferroni procedures do not require independence of the $\hat{\beta}_j$'s; they are valid for any covariance structure on the $\hat{\beta}_j$'s. However, the logic of the Bonferroni procedure for testing $H_{0i}: \mathrm{a}_i'\beta = 0$ for $i = 1, 2, \cdots, d$ requires that the coefficient vectors a_1, a_2, . . . , a_d be specified before seeing the data. If we wish to choose values of a_i after looking at the data, we must use the Scheffe′ procedure described below. Modifications of the Bonferroni approach have been proposed that are less conservative but still control α_f. For examples of these modified procedures, see Holm (1979), Shaffer (1986), Simes (1986), Holland and Copenhaver (1987), Hochberg (1988), Hommel (1988), Rom (1990), and Rencher (1995, Section 3.4.4). Comparisons of these procedures have been made by Holland (1991) and Broadbent (1993).

A second approach to controlling α_f due to Scheffe′ (1953; 1959, p. 68) yields simultaneous tests of $H_0: \mathrm{a}'\beta = 0$ for all possible values of **a** including those chosen after looking at the data. We could also test $H_0: \mathrm{a}'\beta = \mathrm{t}$ for arbitrary t. For any given **a**, the hypothesis $H_0: \mathrm{a}'\beta = 0$ is tested as usual by (4.38)

$$F = \frac{\left(\mathrm{a}'\hat{\beta}\right)' [\mathrm{a}'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{a}]^{-1}\mathrm{a}'\hat{\beta}}{S^2}$$

$$= \frac{\left(\mathrm{a}'\hat{\beta}\right)^2}{S^2\,\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}} \qquad (4.44)$$

But the test proceeds by finding a critical value large enough to hold for all possible **a**. Accordingly, we now find the distribution of $\max_{\mathrm{a}} F$.

Theorem 4.5:

(i) The maximum value of F in (4.44) is given by

$$\max_{\mathrm{a}} \frac{\left(\mathrm{a}'\hat{\beta}\right)^2}{S^2\,\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}} = \frac{\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}}{S^2} \qquad (4.45)$$

(ii) If **y** is $N_n\left(\mathrm{X}\beta, \sigma^2\mathrm{I}\right)$, then $\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}/(k+1)S^2$ is distributed as $F_{(k+1, n-k-1)}$. Thus

$$\max_{\mathrm{a}} \frac{\left(\mathrm{a}'\hat{\beta}\right)^2}{S^2\,\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}(k+1)}$$

is distributed as $F_{(k+1, n-k-1)}$.

Proof:

(i) Using the quotient rule, chain rule, we differentiate $\left(\mathrm{a}'\hat{\beta}\right)^2 /$ $S^2\,\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}$ with respect to **a** and set the result equal to 0:

$$\frac{\partial}{\partial_{\mathrm{a}}}\frac{\left(\mathrm{a}'\hat{\beta}\right)^2}{\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}}=\frac{[\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}]2\left(\mathrm{a}'\hat{\beta}\right)\hat{\beta}-\left(\mathrm{a}'\hat{\beta}\right)^2 2\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}}{[\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}]^2}=0$$

Multiplying by $[\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}]^2/2\mathrm{a}'\hat{\beta}$ and treating 1×1 matrices as scalars, we obtain

$$[\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}]\hat{\beta}-\mathrm{a}'\hat{\beta}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}=0$$

$$\mathrm{a}=\frac{\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}}{\mathrm{a}'\hat{\beta}}\mathrm{X}'\mathrm{X}\hat{\beta}=c\,\mathrm{X}'\mathrm{X}\hat{\beta}$$

where $c=\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}/\mathrm{a}'\hat{\beta}$. Substituting $\mathrm{a}=c\,\mathrm{X}'\mathrm{X}\hat{\beta}$ into (4.44) gives

$$\max_{\mathrm{a}}\frac{\left(\mathrm{a}'\hat{\beta}\right)^2}{S^2\,\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}}=\frac{\left(c\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}\right)^2}{S^2 c\hat{\beta}'\mathrm{X}'\mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}c\,\mathrm{X}'\mathrm{X}\hat{\beta}}=\frac{c^2\left(\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}\right)^2}{S^2c^2\,\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}}=\frac{\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}}{S^2}$$

(ii) Using $\mathrm{C}=\mathrm{I}_{k+1}$ in (4.27), we have, by Theorem 4.4b (ii), that

$F=\dfrac{\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}}{(k+1)S^2}$ is distributed as $F_{(k+1,n-k-1)}$.

By Theorem 4.5(ii), we have

$$P\Big[\max_{\mathrm{a}}\frac{\left(\mathrm{a}'\hat{\beta}\right)^2}{S^2\,\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}\left(k+1\right)}\geq F_{\alpha^*,k+1,n-k-1}\Big]=\alpha^*$$

$$P\Big[\max_{\mathrm{a}}\frac{\left(\mathrm{a}'\hat{\beta}\right)^2}{S^2\,\mathrm{a}'\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{a}}\geq (k+1)F_{\alpha^*,k+1,n-k-1}\Big]=\alpha^*$$

Thus, to test $H_0:\mathrm{a}'\beta=0$ for any and all **a** (including values of **a** chosen after seeing the data) with $\alpha_f\leq\alpha^*$, we calculate F in (4.44) and reject H_0 if $F\geq(k+1)F_{\alpha^*,k+1,n-k-1}$.

To test for individual β_j's using Scheffe´'s procedure, we set $\mathrm{a}'=(0,\cdots,0,1,0,\cdots,0)$ with a_1 in the *jth* position. Then *F* in (4.44) reduces to $F = \hat{\beta}_j^2 / S^2 g_{jj}$ in (4.39), and the square root is $t_j = \hat{\beta}_j / S\sqrt{g_{jj}}$ in (4.42). By Theorem 4.5, we reject $H_0 : \mathrm{a}'\beta = \beta_j = 0$ if $|t_j| \geq \sqrt{(k+1)F_{\alpha^*, k+1, n-k-1}}$.

For practical purposes $[k \leq (n-3)]$, we have

$$t_{\alpha^*/2k, n-k-1} < \sqrt{(k+1)F_{\alpha^*, k+1, n-k-1}}$$

and thus the Bonferroni tests for individual β_j's in (4.42) are usually more powerful than the Scheffe′ tests. On the other hand, for a large number of linear combinations $\mathrm{a}'\beta$, the Scheffe′ test is better since $(k+1)F_{\alpha^*, k+1, n-k-1}$ is constant, while the critical value $F_{\alpha^*/d, n-k-1}$ for Bonferroni tests in (4.43) increases with the number of tests *d* and eventually exceeds the critical value for Scheffe′ tests.

It has been assumed that the tests in this section for $H_0 : \beta_j = 0$ are carried out without regard to whether the overall hypothesis $H_0 : \beta_1 = 0$ is rejected. However, if the test statistics $t_j = \hat{\beta}_j / S\sqrt{g_{jj}}$, *j* = 1, 2, . . . , *k*, in (4.42) are calculated only if $H_0 : \beta_1 = 0$ is rejected using *F* in (4.5), then clearly α_f is reduced and the conservative critical values $t_{\alpha^*/2k, n-k-1}$ and $\sqrt{(k+1)F_{\alpha^*, k+1, n-k-1}}$ become even more conservative. Using this protected testing principle (Hocking 1996, p. 106), we can even use the critical value $t_{\alpha^*/2, n-k-1}$ for all *k* tests and α_f will still be close to α^*. [For illustrations of this familywise error rate structure, see Hummel and sligo (1971) and Rencher and Scott (1990).] A similar statement can be made for testing the overall hypothesis $H_0 : \mathrm{C}\beta = 0$ followed by *t* tests or *F* tests of $H_0 : c_i'\beta = 0$ using the rows of **C**.

Example 4.5.2: We test $H_{01} : \beta_1 = 0$ and $H_{02} : \beta_2 = 0$ for the data in Table 3.1. Using (4.42) and the results in Examples 3.3.1(a), 3.33 and 4.1, we have

$$t_1 = \frac{\hat{\beta}_1}{S\sqrt{g_{11}}} = \frac{3.0118}{\sqrt{2.8288}\sqrt{0.16207}} = \frac{3.0118}{0.67709} = 4.448$$

$$t_2 = \frac{\hat{\beta}_2}{S\sqrt{g_{22}}} = \frac{-1.2855}{\sqrt{2.8288}\sqrt{0.08360}} = \frac{-1.2855}{0.48629} = -2.643$$

Using $\alpha = 0.05$ for each test, we reject both H_{01} and H_{02} because $t_{0.025,9} = 2.262$. The (two-sided) *p*-values are 0.00160 and 0.0268, respectively. If we use $\alpha = 0.05/2 = 0.025$ for a Bonferroni test, we would not reject H_{02} since $p = 0.0268 > 0.025$. However, using the protected testing principle, we would reject H_{02} because the overall regression hypothesis $H_0 : \beta_1 = 0$ was rejected in Example 4.1.

Example 4.5.2 [The program name ta15.m]	Applications using MATLAB

```
clc
y=[2 3 2 7 6 8 10 7 8 12 11 14]';
x1=[0 2 2 2 4 4 4 6 6 6 8 8]';
x2=[2 6 7 5 9 8 7 10 11 9 15 13]';
n=length(x1);k=2;x=[ones(size(x1)) x1 x2];
beta=x\y,Y=x*beta;e=Y-y;MSE=e'*e/(n-3);
% test of t
d=sqrt(diag(inv(x'*x)))
S=sqrt(MSE)
g11=d(2)^2,g22=d(3)^2
t1=beta(2)/(S*sqrt(g11))
t2=beta(3)/(S*sqrt(g22))
p1=(1-tcdf(abs(t1),n-k-1))*2
p2=(1-tcdf(abs(t2),n-k-1))*2
```

```
beta =          d =              S =
       5.3754          0.9873          1.6819
       3.0118         0.40257
      -1.2855         0.28913

g11 =                   g22 =
      0.16207                0.083596
t1 =                    t2 =
       4.4482                 -2.6435
p1 =                    p2 =
    0.0016044                0.026761
```

4.6: Confidence Intervals and Prediction Intervals

In this section we consider a confidence region for β, confidence intervals for β_j, $a'\hat{\beta}$, $E(y)$, and σ^2, and prediction intervals for future observations. We assume throughout Section 4.6 that **y** is $N_n(X\beta, \sigma^2 I)$.

4.6.1 Confidence Region for β

If **C** is equal to **I** and **t** is equal to β in (4.35), q becomes k +1, we obtain a central F distribution, and we can make the probability statement

$$P\left[\left(\hat{\beta}-\beta\right)' X'X\left(\hat{\beta}-\beta\right) \big/ (k+1)S^2 \le F_{\alpha,k+1,n-k-1}\right] = 1-\alpha$$

where $S^2 = SSE/(n-k-1)$. From this statement, a $100(1-\alpha)\%$ joint confidence region for $\beta_0, \beta_1, \cdots, \beta_k$ in β is defined to consist of all vectors β that satisfy

$$\left(\hat{\beta}-\beta\right)' X'X\left(\hat{\beta}-\beta\right) \le (k+1)S^2 F_{\alpha,k+1,n-k-1} \qquad (4.46)$$

For k = 1, this region can be plotted as an ellipse in two dimensions. For $k > 1$, the ellipsoidal region in (4.46) is unwieldy to interpret and report, and we therefore consider intervals for the individual β_j's.

4.6.2: Confidence Interval for β_j

If $\beta_j \ne 0$, we can subtract β_j in (4.40) so that $t_j = \left(\hat{\beta}_j - \beta_j\right) \big/ S\sqrt{g_{jj}}$ has the central t distribution, where g_{jj} is the *jth* diagonal element of $(X'X)^{-1}$. Then

$$P\left[-t_{\alpha/2,n-k-1} \le \frac{\left(\hat{\beta}_j - \beta_j\right)}{S\sqrt{g_{jj}}} \le t_{\alpha/2,n-k-1}\right] = 1-\alpha$$

Solving the inequality for β_j gives

$$P\left(\hat{\beta}_j - t_{\alpha/2,n-k-1} S\sqrt{g_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-k-1} S\sqrt{g_{jj}}\right) = 1-\alpha$$

Before taking the sample, the probability that the random interval will contain β_j is $1-\alpha$. *After* taking the sample, the $100(1-\alpha)\%$ confidence interval for β_j

$$\hat{\beta}_j \pm t_{\alpha/2,n-k-1} S \sqrt{g_{jj}} \qquad (4.47)$$

is no longer random, and thus we say that we are $100(1-\alpha)\%$ confident that the interval contains β_j.

Note that the confidence coefficient $1-\alpha$ holds only for a single confidence interval for one of the β_j's. For confidence intervals for all k + 1 of the β's that hold simultaneously with overall confidence coefficient $1-\alpha$, see Section 4.6.7.

Example 4.6.2: We compute a 95% confidence interval for each β_j using y_2 in the chemical reaction data in Table 3.4 (see Example 4.2a). The matrix $(X'X)^{-1}$ (see the answer to Problem 3.52) and the estimate $\hat{\beta}$ have the following values:

$$(X'X)^{-1} = \begin{pmatrix} 65.37550 & -0.33885 & -0.31252 & -0.02041 \\ -0.33885 & 0.00184 & 0.00127 & -0.00043 \\ -0.31252 & 0.00127 & 0.00408 & -0.00176 \\ -0.02041 & -0.00043 & -0.00176 & 0.02161 \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} -26.0353 \\ 0.4046 \\ 0.2930 \\ 1.0338 \end{pmatrix}$$

For β_1, we obtain by (4.47),

$$\hat{\beta}_1 \pm t_{0.025,15} S \sqrt{g_{11}}$$
$$0.4046 \pm (2.1314)(4.0781)\sqrt{0.00184}$$
$$0.4046 \pm 0.3723$$

For the other β_j's, we have

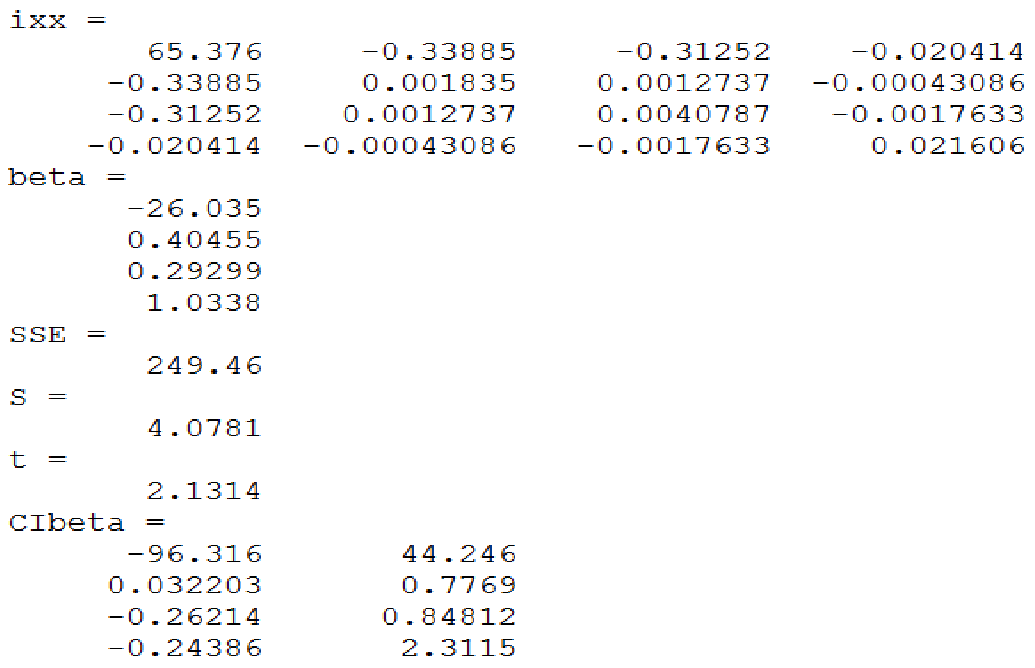
$$\beta_0 : -26.0353 \pm 70.2812 \qquad \beta_2 : 0.2930 \pm 0.5551 \qquad \beta_3 : 1.0338 \pm 1.27777$$
$$(-96.3165, \quad 44.2459) \qquad (-0.2621, \quad 0.8481) \qquad (-0.2439, \quad 2.3115)$$

The confidence coefficient 0.95 holds for only one of the four confideence intervals. For more than one interval, see Example 4.6.7.

Example 4.6.2 [The program name ta16.m]	Applications using MATLAB

```
clc
clear all
data=[41.5 45.9 162 23 3;33.8 53.3 162 23 8;27.7 57.5 162 30 5
    21.7 58.8 162 30 8;19.9 60.6 172 25 5;15.0 58.0 172 25 8
    12.2 58.6 172 30 5;4.3 52.4 172 30 8;19.3 56.9 167 27.5 6.5
    6.4 55.4 177 27.5 6.5;37.6 46.9 157 27.5 6.5;18 57.3 167 32.5 6.5
    26.3 55.0 167 22.5 6.5;9.9 58.9 167 27.5 9.5;25.0 50.3 167 27.5 3.5
    14.1 61.1 177 20 6.5;15.2 62.9 177 20 6.5;15.9 60.0 160 34 7.5
    19.6 60.6 160 34 7.5];y2=data(:,2);x1=data(:,3);x2=data(:,4);
    x3=data(:,5);y1=data(:,1);x1=data(:,3);x2=data(:,4);x3=data(:,5);
% C.I. for parameter beta
n=length(y1);k=3;V=n-k-1;
x=[ones(size(x1)) x1 x2 x3];ixx=inv(x'*x),beta=x\y2
SSE=y2'*y2-beta'*x'*y2,S=sqrt(SSE/(n-4)),alfa=0.05/2;
t=abs(tinv(alfa,V))
clbeta=beta-t*S*sqrt(diag(ixx));
cubeta=beta+t*S*sqrt(diag(ixx));
CIbeta=[clbeta cubeta]
```

Ans.



```
ixx =
       65.376      -0.33885      -0.31252     -0.020414
     -0.33885      0.001835     0.0012737   -0.00043086
     -0.31252     0.0012737     0.0040787    -0.0017633
    -0.020414   -0.00043086    -0.0017633      0.021606
beta =
      -26.035
      0.40455
      0.29299
       1.0338
SSE =
       249.46
S =
       4.0781
t =
       2.1314
CIbeta =
      -96.316        44.246
     0.032203        0.7769
     -0.26214       0.84812
     -0.24386        2.3115
```

4.6.3: Confidence Interval for $\mathrm{a}'\beta$

If $\mathrm{a}'\beta \neq 0$, we can subtract $\mathrm{a}'\beta$ from $\mathrm{a}'\hat{\beta}$ in (4.44) to obtain

$$F = \frac{\left(\mathrm{a}'\hat{\beta} - \mathrm{a}'\beta\right)^2}{S^2\,\mathrm{a}'(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{a}}$$

Which is distributed as $F_{(1,n-k-1)}$. And

$$t = \frac{\mathrm{a}'\hat{\beta} - \mathrm{a}'\beta}{S\sqrt{\mathrm{a}'(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{a}}} \qquad (4.48)$$

is distributed as $t_{(n-k-1)}$, and a $100(1-\alpha)\%$ confidence interval for a single value of $\mathrm{a}'\beta$ is given by

$$\mathrm{a}'\hat{\beta} \pm t_{(\alpha/2,n-k-1)} S\sqrt{\mathrm{a}'(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{a}} \qquad (4.49)$$

4.6.4: Confidence Interval for $E(y)$

Let $\mathrm{x}_0 = (1, x_{01}, x_{02}, \cdots, x_{0k})'$ denote a particular choice of $\mathrm{x} = (1, x_1, x_2, \cdots, x_k)'$. Note that x_0 need not be one of the **x**'s in the sample; that is, x_0' need not be a row of **X**. If x_0 is very far outside the area covered by the sample however, the prediction may be poor. Let y_0 be an observation corresponding to x_0. Then

$$y_0 = \mathrm{x}_0'\,\beta + \varepsilon$$

And [assuming that the model is correct so that $E(\varepsilon) = 0$]

$$E(y_0) = \mathrm{x}_0'\beta \qquad (4.50)$$

We wish to find a confidence interval for $E(y_0)$, that is, for the mean of the distribution of y-values corresponding to x_0.

By Corollary 1 to Theorem 3.6d, the minimum variance unbiased estimator of $E(y_0)$ is given by

$$E(\hat{y}_0) = \mathrm{x}_0'\hat{\beta} \qquad (4.51)$$

Since (4.50) and (4.51) are of the form $\mathrm{a}_0'\beta$ and $\mathrm{a}_0'\hat{\beta}$, respectively, we obtain a $100(1-\alpha)\%$ confidence interval for $E(y_0) = \mathrm{x}_0'\beta$ from (4.49):

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm t_{(\alpha/2,n-k-1)} S \sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \qquad (4.52)$$

The confidence coefficient $1-\alpha$ for the interval in (4.52) holds only for a single choice of the vector $\mathbf{x}_0$. For intervals covering several values of $\mathbf{x}_0$ or all possible values of $\mathbf{x}_0$, see Section 4.6.7.

We can express the confidence interval in (4.52) in terms of the centered model in Section 3.5, $y_i = \alpha + \boldsymbol{\beta}_1'(\mathbf{x}_{01} - \bar{\mathbf{x}}_1)$, where $\mathbf{x}_{01} = (x_{01}, x_{02}, \cdots, x_{0k})'$ and $\bar{\mathbf{x}}_1 = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k)'$. [We use the notation $\mathbf{x}_{01}$ to distinguish this vector from $\mathbf{x}_0 = (1, x_{01}, x_{02}, \cdots, x_{0k})'$ above.] For the centered model, (4.50), (4.51), and (4.52) become

$$E(y_0) = \alpha + \boldsymbol{\beta}_1'(\mathbf{x}_{01} - \bar{\mathbf{x}}_1) \qquad (4.53)$$

$$E(\hat{y}_0) = \bar{y} + \hat{\boldsymbol{\beta}}_1'(\mathbf{x}_{01} - \bar{\mathbf{x}}_1) \qquad (4.54)$$

$$\bar{y} + \hat{\boldsymbol{\beta}}_1'(\mathbf{x}_{01} - \bar{\mathbf{x}}_1) \pm t_{(\alpha/2,n-k-1)} S \sqrt{\frac{1}{n} + (\mathbf{x}_{01} - \bar{\mathbf{x}}_1)'(\mathbf{X}_c'\mathbf{X}_c)^{-1}(\mathbf{x}_{01} - \bar{\mathbf{x}}_1)} \qquad (4.55)$$

Note that in the form shown in (4.55), it is clear that if $\mathbf{x}_{01}$ is close to $\bar{\mathbf{x}}_1$ the interval is narrower; in fact, it is narrowest for $\mathbf{x}_{01} = \bar{\mathbf{x}}$. The width of the interval increases as the distance of $\mathbf{x}_{01}$ from $\bar{\mathbf{x}}$ increases.

For the special case of simple linear regression, (4.50), (4.51), and (4.55) reduce to

$$E(y_0) = \beta_0 + \beta_1 x_0 \qquad (4.56)$$

$$E(\hat{y}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 \qquad (4.57)$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{(\alpha/2,n-2)} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (4.58)$$

where S is given by (2.11). The width of the interval in (4.58) depends on how far x_0 is from $\bar{x}$.

Example 4.6.4: For the grades data in Example 4.2, we find a 95% confidence interval for $E(y_0)$, where $x_0 = 80$. Using (4.58), we obtain

$$\hat{\beta}_0+\hat{\beta}_1(80)\pm t_{(0.025,16)}S\sqrt{\frac{1}{18}+\frac{(80-58.056)^2}{19530.944}}$$

$$80.5386\ \pm 2.1199\,(13.8547\,)(0.2832\,)$$

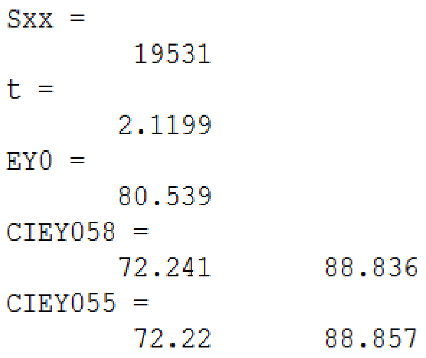
$$80.5386\ \pm 8.3183$$

$$(72.2204,\quad 88.8569)$$

Example 4.6.4 [The program name ta17.m]	Applications using MATLAB

```
clc
clear all
y=[95 80 0 0 79,77 72 66 98 90 0 95 35 50 72 55 75 66]';
x=[96 77 0 0 78 64 89 47 90 93 18 86 0 30 59 77 74 67]';
n=length(x);E=[ones(size(x)) x];beta=E\y
Yhad=E*beta;e=y-Yhad;MSE=e'*e/(n-2),S=sqrt(MSE)
Sxx=sum((x-mean(x)).^2),alfa=0.05/2;V=n-2;
% C. I. E(Y0)by using formula (4.55)
x0=[1 80]';x01=80;t=abs(tinv(alfa,V))
EY0=x0'*beta
EY0l=x0'*beta-t*S*sqrt(x0'*inv(x'*x)*x0);
EY0u=x0'*beta+t*S*sqrt(x0'*inv(x'*x)*x0);
CIEY058=[EY0l EY0u]
% C. I. E(Y0)by using formula (4.58)
EY0=beta(1)+beta(2)*80;
EY0l=EY0-t*S*sqrt((1/n)+(x01-mean(x))^2/Sxx);
EY0u=EY0+t*S*sqrt((1/n)+(x01-mean(x))^2/Sxx);
CIEY055=[EY0l EY0u]
```

Ans.



```
Sxx =
        19531
t =
       2.1199
EY0 =
       80.539
CIEY058 =
       72.241       88.836
CIEY055 =
        72.22       88.857
```

4.6.5: Prediction Interval for a Future Observation

A "confidence interval" for a future observation y_0 corresponding to x_0 is called a prediction interval. We speak of a *prediction interval* rather than a confidence interval because y_0 is an individual observation and is thereby a random variable rather than a parameter. To be $100(1-\alpha)\%$ confident that the interval contains y_0, the prediction interval will clearly have to be wider than a confidence interval for the parameter $E(y_0)$.

Since $y_0 = \mathrm{x}_0'\beta + \varepsilon_0$, we predict y_0 by $\hat{y}_0 = \mathrm{x}_0'\hat{\beta}$, which is also the estimator of $E(y_0) = \mathrm{x}_0'\beta$. The random variables y_0 and $\hat{y}_0$ are independent because y_0 is a future observation to be obtained independently of the *n* observations used to compute $\hat{y}_0 = \mathrm{x}_0'\hat{\beta}$. Hence the variance of $y_0 - \hat{y}_0$ is

$$Var(y_0 - \hat{y}_0) = Var\left(y_0 - \mathrm{x}_0'\hat{\beta}\right) = Var\left(\mathrm{x}_0'\beta + \varepsilon_0 - \mathrm{x}_0'\hat{\beta}\right)$$

Since $\mathrm{x}_0'\beta$ is a constant, this becomes

$$Var(y_0 - \hat{y}_0) = Var(\varepsilon_0) + Var\left(\mathrm{x}_0'\hat{\beta}\right) = \sigma^2 + \sigma^2\,\mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0$$

$$= \sigma^2[1 + \mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0] \qquad (4.59)$$

which is estimated by $S^2[1 + \mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0]$. It can be shown that $E(y_0 - \hat{y}_0) = 0$ and that S^2 is independent of both y_0 and $\hat{y}_0 = \mathrm{x}_0'\hat{\beta}$. Therefore, the *t* statistic

$$t = \frac{y_0 - \hat{y}_0 - 0}{S\sqrt{1 + \mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0}} \qquad (4.60)$$

is distributed as $t_{(n-k-1)}$, and

$$P\left[-t_{(\alpha/2,n-k-1)} \le \frac{y_0 - \hat{y}_0}{S\sqrt{1 + \mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0}} \le t_{(\alpha/2,n-k-1)}\right] = 1 - \alpha$$

The inequality can be solved for y_0 to obtain the $100(1-\alpha)\%$ prediction interval

$$\hat{y}_0 - t_{(\alpha/2,n-k-1)}S\sqrt{1 + \mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0} \le y_0 \le \hat{y}_0 + t_{(\alpha/2,n-k-1)}S\sqrt{1 + \mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0}$$

or, using $\hat{y}_0 = \mathrm{x}_0'\hat{\beta}$, we have

$$\mathrm{x}_0'\hat{\beta} \pm t_{(\alpha/2,n-k-1)} S\sqrt{1+\mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0} \qquad (4.61)$$

Note that the confidence coefficient $1-\alpha$ for the prediction interval in (4.61) holds for only one value of x_0.

In $1+\mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0$, the second term, $\mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0$, is typically much smaller than 1 (provided *k* is much smaller than *n*) because the variance of $\hat{y}_0 = \mathrm{x}_0'\hat{\beta}$ is much less than the variance of y_0. [To illustrate, if $\mathrm{X}'\mathrm{X}$ were diagonal and x_0 were in the area covered by the rows of **X**, then $\mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_0$ would be a sum with *k* +1 terms, each of the form $x_{0j}^2 / \sum_{i=1}^n x_{ij}^2$, which is of the order of 1/*n*.] Thus prediction intervals for y_0 are generally much wider than confidence intervals for $E(y_0) = \mathrm{x}_0'\beta$.

In terms of the centered model in Section 3.5, the $100(1-\alpha)\%$ predict-tion interval in (4.61) becomes

$$\bar{y} + \hat{\beta}_1'(\mathrm{x}_{01} - \bar{\mathrm{x}}_1) \pm t_{(\alpha/2,n-k-1)} S\sqrt{1+\frac{1}{n}+(\mathrm{x}_{01}-\bar{\mathrm{x}}_1)'(\mathrm{X}_c'\mathrm{X}_c)^{-1}(\mathrm{x}_{01}-\bar{\mathrm{x}}_1)} \qquad (4.62)$$

For the case of simple linear regression, (4.61) and (4.62) reduce to

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{(\alpha/2,n-2)} S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}} \qquad (4.63)$$

where *S* is given by (2.11). In (4.63), it is clear that the second and third terms within the square root are much smaller than 1 unless x_0 is far removed from the interval bounded by the smallest and largest *x*'s.

For a prediction interval for the mean of *q* future observations, see Problem 4.30.

Example 4.6.5: Using the data from Example 2.2, we find a 95% prediction interval for y_0 when $\mathrm{x}_0 = 80$. Using (4.63), we obtain

$$\hat{\beta}_0 + \hat{\beta}_1(80) \pm t_{(0.025,16)} S\sqrt{1+\frac{1}{18}+\frac{(80-58.056)^2}{19530.944}}$$

$$80.5386 \pm 2.1199\,(13.8547)(1.0393)$$

$$80.5386 \pm 30.5258$$

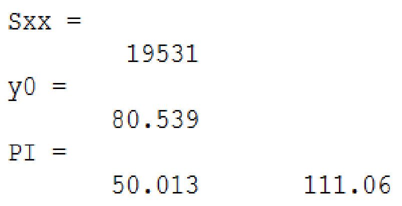
$$(50.0128,\quad 111.0644)$$

Note that the prediction interval for y_0 here is much wider than the confidence interval for $E(y_0)$ in Example 4.6.4.

Example 4.6.5 [The program name ta18.m]	Applications using MATLAB

```
clc
clear all
y=[95 80 0 0 79,77 72 66 98 90 0 95 35 50 72 55 75 66]';
x=[96 77 0 0 78 64 89 47 90 93 18 86 0 30 59 77 74 67]';
n=length(x);E=[ones(size(x)) x];beta=E\y;
Yhad=E*beta;e=y-Yhad;MSE=e'*e/(n-2);S=sqrt(MSE);
Sxx=sum((x-mean(x)).^2);alfa=0.05/2;V=n-2;
t=abs(tinv(alfa,V));
% Prediction Interval for y0
x0=80;Sxx=sum((x-mean(x)).^2),y0=beta(1)+x0*beta(2)
PIL=y0-t*S*sqrt(1+1/n+(x0-mean(x))^2/Sxx);
PIU=y0+t*S*sqrt(1+1/n+(x0-mean(x))^2/Sxx);
PI=[PIL PIU]
```

Ans.



```
Sxx =
        19531
y0 =
       80.539
PI =
       50.013       111.06
```

4.6.6: Confidence Interval for σ^2

By Theorem 3.6b(ii), $(n-k-1)S^2/\sigma^2$ is $\chi^2_{(n-k-1)}$. Therefore

$$P[\chi^2_{(1-\alpha/2,n-k-1)} \le \frac{(n-k-1)S^2}{\sigma^2} \le \chi^2_{(\alpha/2,n-k-1)}] = 1-\alpha \qquad (4.64)$$

where $\chi^2_{(\alpha/2,n-k-1)}$ is the upper $\alpha/2$ percentage point of the chi-square distribution and $\chi^2_{(1-\alpha/2,n-k-1)}$ is the lower $\alpha/2$ percentage point. Solving the inequality for σ^2 yields the $100(1-\alpha)\%$ confidence interval

$$\frac{(n-k-1)S^2}{\chi^2_{(\alpha/2,n-k-1)}} \le \sigma^2 \le \frac{(n-k-1)S^2}{\chi^2_{(1-\alpha/2,n-k-1)}} \qquad (4.65)$$

A $100(1-\alpha)\%$ confidence interval for σ is given by

$$\sqrt{\frac{(n-k-1)S^2}{\chi^2_{(\alpha/2,n-k-1)}}} \le \sigma \le \sqrt{\frac{(n-k-1)S^2}{\chi^2_{(1-\alpha/2,n-k-1)}}} \qquad (4.66)$$

Example 4.6.6: Using the data from Example 2.2, we find a 95% predi confidence interval for σ^2 and σ. Using (4.65), we obtain

$$\frac{(16)S^2}{\chi^2_{(0.025,16)}} \le \sigma^2 \le \frac{(16)S^2}{\chi^2_{(0.975,16)}}$$

$$\frac{(16)(191.95)}{28.845} \le \sigma^2 \le \frac{(16)(191.95)}{6.9077}$$

$$106.47 \le \sigma^2 \le 444.61$$

Using (4.66), we obtain

$$\sqrt{\frac{(16)S^2}{\chi^2_{(0.025,16)}}} \le \sigma \le \sqrt{\frac{(16)S^2}{\chi^2_{(0.975,16)}}}$$

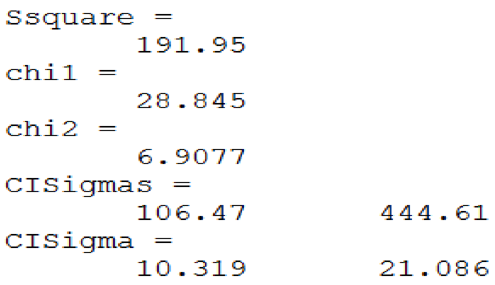
$$\sqrt{\frac{(16)(191.95)}{28.845}} \le \sigma \le \sqrt{\frac{(16)(191.95)}{6.9077}}$$

$$10.319 \le \sigma^2 \le 21.086$$

Example 4.6.6 [The program name ta19.m]	Applications using MATLAB

```
clc
clear all
y=[95 80 0 0 79,77 72 66 98 90 0 95 35 50 72 55 75 66]';
x=[96 77 0 0 78 64 89 47 90 93 18 86 0 30 59 77 74 67]';
n=length(x);k=1;E=[ones(size(x)) x];beta=E\y;Yhad=E*beta;
e=y-Yhad;MSE=e'*e/(n-2);Ssquare=MSE,alfa=0.05/2;V=n-k-1;
% Confidence Interval for Sigma^2
chi1=chi2inv(1-alfa,16),chi2=chi2inv(alfa,16)
SigmasL=(n-2)*Ssquare/chi1;SigmasU=(n-2)*Ssquare/chi2;
CISigmas=[SigmasL SigmasU]
% Confidence Interval for Sigma
SigmaL=sqrt((n-2)*Ssquare/chi1);SigmaU=sqrt((n-2)*Ssquare/chi2);
CISigma=[SigmaL SigmaU]
```

Ans.



```
Ssquare =
        191.95
chi1 =
        28.845
chi2 =
        6.9077
CISigmas =
        106.47        444.61
CISigma =
        10.319        21.086
```

4.6.7: Simultaneous Intervals

By analogy to the discussion of testing several hypotheses (Section 4.5.2), when several intervals are computed, two confidence coefficients can be considered: familywise confidence $(1-\alpha_f)$ and individual confidence $(1-\alpha_c)$. Familywise confidence of $(1-\alpha_f)$ means that we are $100(1-\alpha_f)\%$ confident that every interval contains its respective parameter.

In some cases, our goal is simply to control $(1-\alpha_c)$ for each one of several confidence or prediction intervals so that no changes are needed to expressions (4.47), (4.49), (4.52), and (4.61). In other cases the desire is to control $(1-\alpha_f)$. To do so, both the Bonferroni and Scheffe′ methods can be adapted to the situation of multiple intervals. In yet other cases we may want to control other properties of multiple intervals (Benjamini and Yekutieli 2005).

The Bonferroni procedure increases the width of each individual interval so that $(1-\alpha_f)$ for the set of intervals is greater than or equal to the desired value $(1-\alpha^*)$. As an example suppose that it is desired to calculate the k confidence intervals for $\beta_1, \beta_2, \cdots, \beta_k$. Let E_j be the event that the *jth* interval includes β_j, and E_j^c be the complement of that event. Then by definition

$$\begin{aligned} 1-\alpha_f &= P(E_1 \cap E_2 \cap \cdots \cap E_k) \\ &= 1 - P\left(E_1^c \cup E_2^c \cup \cdots \cup E_k^c\right) \end{aligned}$$

Assuming that $P(E_j^c) = \alpha_c$ for j = 1, . . . , k, the Bonferroni inequality now implies that

$$1-\alpha_f \geq 1-k\alpha_c$$

Hence we can ensure that $1-\alpha_f$ is greater than or equal to the desired $1-\alpha^*$ by setting $1-\alpha_c = 1-\alpha^*/k$ for the individual intervals.

Using this approach, Bonferroni confidence intervals for $\beta_1, \beta_2, \cdots, \beta_k$ are given by

$$\hat{\beta}_j \pm t_{\alpha^*/2k, n-k-1} S\sqrt{g_{jj}}, \quad j = 1,2,\cdots,k \tag{4.67}$$

where g_{jj} is the *jth* element of $(\mathrm{X}'\mathrm{X})^{-1}$. Bonferroni *t* values $t_{\alpha^*/2k}$ are available in Bailey (1977) and can also be obtained in many software programs. For example, a probability calculator for the *t*, the *F*, and other distributions is available free from NCSS (download at www.ncss.com).

Similarly for *d* linear functions $\mathrm{a}_1'\beta, \mathrm{a}_2'\beta, \cdots, \mathrm{a}_d'\beta$ (chosen before seeing the data), Bonferroni confidence intervals are given by

$$\mathrm{a}_i'\hat{\beta} \pm t_{\left(\alpha^*/2d, n-k-1\right)} S\sqrt{\mathrm{a}_i'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{a}_i}, \quad i = 1,2,\cdots,d \tag{4.68}$$

These intervals hold simultaneously with familywise confidence of at least $1-\alpha^*$.

Bonferroni confidence intervals for $E(y_0) = \mathrm{x}_0'\beta$ for a few values of x_0, say, $\mathrm{x}_{01}, \mathrm{x}_{02}, \cdots, \mathrm{x}_{0d}$ are given by

$$\mathrm{x}_{0i}'\hat{\beta} \pm t_{\left(\alpha^*/2d, n-k-1\right)} S\sqrt{\mathrm{x}_{0i}'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_{0i}}, \quad i = 1,2,\cdots,d \tag{4.69}$$

[Note that x_{01} here differs from x_{01} in (4.53)–(4.55).]

For simultaneous prediction of d new observations $y_{01}, y_{02}, \cdots, y_{0d}$ at *d* values of x_0, say, $\mathrm{x}_{01}, \mathrm{x}_{02}, \cdots, \mathrm{x}_{0d}$, we can use the Bonferroni prediction intervals

$$\mathrm{x}_{0i}'\hat{\beta} \pm t_{\left(\alpha^*/2d, n-k-1\right)} S\sqrt{1+\mathrm{x}_{0i}'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_{0i}}, \quad i = 1,2,\cdots,d \tag{4.70}$$

[see (4.61) and (4.69)].

Simultaneous Scheffe′ confidence intervals for all possible linear functions $\mathrm{a}'\beta$ (including those chosen after seeing the data) can be based on the distribution of $\max_{\mathrm{a}} F$ [Theorem 4.5(ii)]. Thus a conservative confidence interval for any and all $\mathrm{a}'\beta$ is

$$\mathrm{a}'\hat{\beta} \pm S\sqrt{(k+1)F_{\left(\alpha^*,k+1,n-k-1\right)}\,\mathrm{a}'(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{a}} \qquad (4.71)$$

The (potentially infinite number of) intervals in (4.71) have an overall confidence coefficient of at least $1-\alpha^*$. For a few linear functions, the intervals in (4.68) will be narrower, but for a large number of linear functions, the intervals in (4.71) will be narrower. A comparison of $t_{\left(\alpha^*/2d,n-k-1\right)}$ and $\sqrt{(k+1)F_{\left(\alpha^*,k+1,n-k-1\right)}\,\mathrm{a}'(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{a}}$ will show which is preferred in a given case.

For confidence limits for $E(y_0)=\mathrm{a}'\beta$ for all possible values of x_0, we use (4.71):

$$\mathrm{x}_0'\hat{\beta} \pm S\sqrt{(k+1)F_{\left(\alpha^*,k+1,n-k-1\right)}\,\mathrm{x}_0'(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{x}_0} \qquad (4.72)$$

These intervals hold simultaneously with a confidence coefficient of $1-\alpha^*$. Thus, (4.72) becomes a confidence region that can be applied to the entire regression surface for all values of x_0. The intervals in (4.71) and (4.72) are due to Scheffe′ (1953; 1959, p. 68) and Working and Hotelling (1929).

Scheffe′-type prediction intervals for $y_{01}, y_{02}, \cdots, y_{0d}$ are given by

$$\mathrm{x}_{0i}'\hat{\beta} \pm S\sqrt{\mathrm{d}\,F_{\left(\alpha^*,d,h-k-1\right)}[1+\mathrm{x}_{0i}'(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{x}_{0i}]}, \quad i=1,2,\cdots,d \qquad (4.73)$$

(see Problem 4.32). These d prediction intervals hold simultaneously with overall confidence coefficient at least $1-\alpha^*$, but note that $\mathrm{d}\,F_{\left(\alpha^*,d,h-k-1\right)}$ is not constant. It depends on the number of predictions.

Example 4.6.7: We compute 95% Bonferroni confidence limits for β_1, β_2, and β_3, using y_2 in the chemical reaction data in Table 3.4; see Example 4.6.2 for $(\mathrm{X}'\mathrm{X})^{-1}$ and $\hat{\beta}$. By (4.67), we have

$$\hat{\beta}_1 \pm t_{0.05/2(3),15}\, S\sqrt{g_{jj}}$$

$$0.4056 \quad \pm (2.6937\ )(4.0781\ )\sqrt{0.00184}$$

$$0.4056 \quad \pm 0.4706$$

$$(-0.0660, \quad 0.8751)$$

$$\beta_2 : 0.2930 \quad \pm \quad 0.7016$$

$$(-0.4086, \quad 0.9946)$$

$$\beta_3 : 1.0338 \quad \pm \quad 1.6147$$

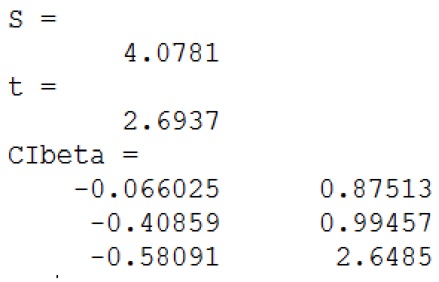
$$(-0.5809, \quad 2.6485)$$

These three intervals hold simultaneously with confidence coefficient at least .95.

Example 4.6.7 [The program name ta20.m]	Applications using MATLAB

```
clc
clear all
data=[41.5 45.9 162 23 3;33.8 53.3 162 23 8;27.7 57.5 162 30 5
     21.7 58.8 162 30 8;19.9 60.6 172 25 5;15.0 58.0 172 25 8
     12.2 58.6 172 30 5;4.3 52.4 172 30 8;19.3 56.9 167 27.5 6.5
     6.4 55.4 177 27.5 6.5;37.6 46.9 157 27.5 6.5;18 57.3 167 32.5 6.5
     26.3 55.0 167 22.5 6.5;9.9 58.9 167 27.5 9.5;25.0 50.3 167 27.5 3.5
     14.1 61.1 177 20 6.5;15.2 62.9 177 20 6.5;15.9 60.0 160 34 7.5
     19.6 60.6 160 34 7.5];x1=data(:,3);x2=data(:,4);x3=data(:,5);
  y1=data(:,1);y2=data(:,2);
n=length(x1);k=3;alfa=0.05;
% C.I. for beta
x=[ones(size(x1)) x1 x2 x3];
ixx=inv(x'*x);
beta=x\y2;
SSE=y2'*y2-beta'*x'*y2;
S=sqrt(SSE/(n-4))
t=abs(tinv(alfa/(2*k),n-k-1))
clbeta=beta(2:4)-t*S*sqrt(diag(ixx(2:4,2:4)));
cubeta=beta(2:4)+t*S*sqrt(diag(ixx(2:4,2:4)));
CIbeta=[clbeta cubeta]
```

Ans.



```
S =
       4.0781
t =
       2.6937
CIbeta =
    -0.066025      0.87513
     -0.40859      0.99457
     -0.58091       2.6485
```

4.7: Likelihood Ratio Tests

The tests in Sections 4.1, 4.2, and 4.4 were derived using informal methods based on finding sums of squares that have chi-square distributions and are independent. These same tests can be obtained more formally by the likelihood ratio approach. Likelihood ratio tests have some good properties and sometimes have optimal properties.

We describe the likelihood ratio method in the simple context of testing $H_0:\beta=0$ versus $H_1:\beta\neq 0$. The likelihood function $L(\beta,\sigma^2)$ was defined in Section 3.6.2 as the joint density of the y's. For a random sample $\mathrm{y}=(y_1,y_2,\cdots,y_n)'$ with density $N_n(\mathrm{X}\beta,\sigma^2\mathrm{I})$, the likelihood function is given by (3.50) as

$$L(\beta,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{n/2}}e^{-(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)/2\sigma^2} \qquad (3.74)$$

The likelihood ratio method compares the maximum value of $L(\beta,\sigma^2)$ restricted by $H_0:\beta=0$ to the maximum value of $L(\beta,\sigma^2)$ under $H_1:\beta\neq0$, which is essentially unrestricted. We denote the maximum value of $L(\beta,\sigma^2)$ restricted by $\beta=0$ as $\max_{H_0} L(\beta,\sigma^2)$ and the unrestricted maximum as $\max_{H_1} L(\beta,\sigma^2)$. If β is equal (or close) to 0, then $\max_{H_0} L(\beta,\sigma^2)$ should be close to $\max_{H_1} L(\beta,\sigma^2)$. If $\max_{H_0} L(\beta,\sigma^2)$ is not close to $\max_{H_1} L(\beta,\sigma^2)$, we would conclude that $\mathrm{y}=(y_1,y_2,\cdots,y_n)'$ apparently did not come from $N_n(\mathrm{X}\beta,\sigma^2\mathrm{I})$ with $\beta=0$.

In this illustration, we can find $\max_{H_0} L(\beta,\sigma^2)$ by setting $\beta=0$ and then estimating σ^2 as the value that maximizes $L(0,\sigma^2)$. Under $H_1:\beta\neq0$, both β and σ^2 are estimated without restriction as the values that maximize $L(\beta,\sigma^2)$. [In designating the unrestricted maximum as $\max_{H_1} L(\beta,\sigma^2)$, we are ignoring the restriction in H_1 that $\beta \neq 0$.]

It is customary to describe the likelihood ratio method in terms of maximizing *L* subject to ω, the set of all values of β and σ^2 satisfying H_0, and subject to Ω, the set of all values of β and σ^2 without restrictions (other than natural restrictions such as $\sigma^2 > 0$). However, to simplify notation in cases such as this in which H_1 includes all values of β except 0, we refer to maximizing *L* under H_0 and H_1.

We compare the restricted maximum under H_0 with the unrestricted maximum under H_1 by the *likelihood ratio*

$$LR = \frac{\max_{H_0} L(\beta,\sigma^2)}{\max_{H_1} L(\beta,\sigma^2)} = \frac{\max L(0,\sigma^2)}{\max L(\beta,\sigma^2)} \qquad (4.75)$$

It is clear that $0 \le LR \le 1$, because the maximum of *L* restricted to $\beta = 0$ cannot exceed the unrestricted maximum. Smaller values of *LR* would favor H_1, and larger values would favor H_0. We thus reject H_0 if $LR \le c$, where *c* is chosen so that $P(LR \le c) = \alpha$ if H_0 is true.

Wald (1943) showed that, under H_0

$$-2\,\mathrm{Ln}\,LR \text{ is approximately } \chi^2_{(\nu)}$$

For large *n*, where *n* is the number of parameters estimated under H_1 minus the number estimated under H_0. In the case of $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, we have $\nu = k + 2 - 1 = k + 1$ because β and σ^2 are estimated under H_1 while only σ^2 is estimated under H_0. In some cases, the χ^2 approximation is not needed because *LR* turns out to be a function of a familiar test statistic, such as *t* or *F*, whose exact distribution is available.

We now obtain the likelihood ratio test for $H_0 : \beta = 0$. The resulting likelihood ratio is a function of the *F* statistic obtained in Problem 4.6 by partitioning the total sum of squares.

Theorem 4.7a: If **y** is $N_n(\mathrm{X}\beta, \sigma^2 \mathrm{I})$, the likelihood ratio test for $H_0 : \beta = 0$ can be based on

$$F = \frac{\hat{\beta}'\mathrm{X}'\mathrm{y}/(k+1)}{(\mathrm{y}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\mathrm{y})/(n-k-1)}$$

We reject H_0 if $F > F_{\alpha, k+1, n-k-1}$.

Proof: To find $\max_{H_1} L(\beta, \sigma^2) = \max L(\beta, \sigma^2)$, we use the maximum likelihood estimators $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ and $\hat{\sigma}^2 = \left(\mathrm{y} - \mathrm{X}\hat{\beta}\right)'\left(\mathrm{y} - \mathrm{X}\hat{\beta}\right)/n$ from Theorem 3.6a. Substituting these in (4.74), we obtain

$$\max_{H_1} L(\beta, \sigma^2) = \max L(\beta, \sigma^2) = L(\hat{\beta}, \hat{\sigma}^2)$$

$$= \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} e^{-(\mathrm{y} - \mathrm{X}\hat{\beta})'(\mathrm{y} - \mathrm{X}\hat{\beta})/2\hat{\sigma}^2}$$

$$= \frac{n^{n/2} e^{-n/2}}{(2\pi)^{n/2}\left[(\mathrm{y} - \mathrm{X}\hat{\beta})'(\mathrm{y} - \mathrm{X}\hat{\beta})\right]^{n/2}} \qquad (4.76)$$

To find $\max_{H_0} L(\beta, \sigma^2) = \max L(0, \sigma^2)$, we solve $\partial \mathrm{Ln} L(0, \sigma^2)/\partial\sigma^2 = 0$ to obtain

$$\hat{\sigma}_0^2 = \frac{\mathrm{y}'\mathrm{y}}{n} \qquad (4.77)$$

Then

$$\max_{H_0} L(\beta, \sigma^2) = \max L(0, \sigma^2) = L(0, \hat{\sigma}_0^2)$$

$$= \frac{1}{(2\pi\hat{\sigma}_0^2)^{n/2}} e^{-\mathrm{y}'\mathrm{y}/2\hat{\sigma}_0^2}$$

$$= \frac{n^{n/2} e^{-n/2}}{(2\pi)^{n/2}(\mathrm{y}'\mathrm{y})^{n/2}} \qquad (4.78)$$

Substituting (4.76) and (4.78) into (4.75), we obtain

$$LR = \frac{\max_{H_0} L(\beta, \sigma^2)}{\max_{H_1} L(\beta, \sigma^2)} = \left[\frac{(\mathrm{y} - \mathrm{X}\hat{\beta})'(\mathrm{y} - \mathrm{X}\hat{\beta})}{\mathrm{y}'\mathrm{y}}\right]^{n/2}$$

$$=\left[\frac{1}{1+(k+1)F/(n-k-1)}\right]^{n/2} \qquad (4.79)$$

Where

$$F=\frac{\hat{\beta}'X'y/(k+1)}{(y'y-\hat{\beta}'X'y)/(n-k-1)}$$

Thus, rejecting $H_0:\beta=0$ for a small value of LR is equivalent to rejecting H_0 for a large value of F.

We now show that the F test in Theorem 4.4b for the general linear hypothesis $H_0:C\beta=0$ is a likelihood ratio test.

Theorem 4.7b: If **y** is $N_n(X\beta,\sigma^2 I)$, then the F test for $H_0:C\beta=0$ in Theorem 4.4b is equivalent to the likelihood ratio test.

Proof: Under $H_1:C\beta\neq 0$, which is essentially unrestricted, $\max_{H_1} L(\beta,\sigma^2)$ is given by (4.76). To find $\max_{H_0} L(\beta,\sigma^2)=\max L(\beta,\sigma^2)$ subject to $C\beta=0$, we use the method of Lagrange multipliers and work with $L(\beta,\sigma^2)$ to simplify the differentiation:

$$v=\mathrm{Ln}L(\beta,\sigma^2)+\lambda'(C\beta-0)$$

$$=-\frac{n}{2}\mathrm{Ln}(2\pi)-\frac{n}{2}\mathrm{Ln}(\sigma^2)-\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}+\lambda'C\beta$$

Expanding $(y-X\beta)'(y-X\beta)$ and differentiating with respect to β, λ, and σ^2, we obtain

$$\frac{\partial v}{\partial\beta}=(2X'y-2X'X\beta)/2\sigma^2+C'\lambda=0 \qquad (4.80)$$

$$\frac{\partial v}{\partial\lambda}=C\beta=0 \qquad (4.81)$$

$$\frac{\partial v}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}(y-X\beta)'(y-X\beta)=0 \qquad (4.82)$$

Eliminating λ and solving for β and σ^2 gives

$$\hat{\beta}_0=\hat{\beta}-(X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}C\hat{\beta} \qquad (4.83)$$

$$\hat{\sigma}_0^2 = \frac{1}{n}\left(\mathrm{y} - \mathrm{X}\,\hat{\beta}_0\right)'\left(\mathrm{y} - \mathrm{X}\,\hat{\beta}_0\right) \tag{4.84}$$

$$= \hat{\sigma}^2 + \frac{1}{n}\left(\mathrm{C}\hat{\beta}\right)'[\mathrm{C}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{C}']^{-1}\,\mathrm{C}\hat{\beta} \tag{4.85}$$

(Problems 4.35 and 4.36), where $\hat{\sigma}^2 = \left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)'\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)\big/n$ and $\hat{\beta} = \left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathrm{y}$ are the maximum likelihood estimates from Theorem 3.6a. Thus

$$\max_{H_0} L\left(\beta, \sigma^2\right) = L\left(\hat{\beta}_0, \hat{\sigma}_0^2\right)$$

$$= \frac{1}{\left(2\pi\right)^{n/2}\left(\hat{\sigma}_0^2\right)^{n/2}}\, e^{-\left(\mathrm{y}-\mathrm{X}\,\hat{\beta}_0\right)'\left(\mathrm{y}-\mathrm{X}\,\hat{\beta}_0\right)\big/2\hat{\sigma}_0^2}$$

$$= \frac{n^{n/2}e^{-n/2}}{\left(2\pi\right)^{n/2}\left\{SSE + \left(\mathrm{C}\hat{\beta}\right)'[\mathrm{C}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{C}']^{-1}\,\mathrm{C}\hat{\beta}\right\}^{n/2}}$$

And

$$LR = \frac{\max_{H_0} L\left(\beta, \sigma^2\right)}{\max_{H_1} L\left(\beta, \sigma^2\right)}$$

$$= \left[\frac{SSE}{SSE + \left(\mathrm{C}\hat{\beta}\right)'[\mathrm{C}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{C}']^{-1}\,\mathrm{C}\hat{\beta}}\right]^{n/2}$$

$$= \left[\frac{1}{1 + SSH/SSE}\right]^{n/2} = \left[\frac{1}{1 + qF/\left(n-k-1\right)}\right]^{n/2}$$

where $SSH = \left(\mathrm{C}\hat{\beta}\right)'[\mathrm{C}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{C}']^{-1}\,\mathrm{C}\hat{\beta}$, $SSE = \left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)'\left(\mathrm{y}-\mathrm{X}\hat{\beta}\right)$, and F is given in (4.27).

PROBLEMS

4.1: Show that $SSR = \hat{\beta}_1'\mathrm{X}_c'\mathrm{X}_c\hat{\beta}_1$ in (4.1) becomes $\mathrm{y}'\mathrm{X}_c\left(\mathrm{X}_c'\mathrm{X}_c\right)^{-1}\mathrm{X}_c'\mathrm{y}$ as in (4.2).

4.2: (a) Show that $\mathrm{H}_c\left[\mathrm{I} - \left(1/n\right)\mathrm{J}\right] = \mathrm{H}_c$, as in (4.3) in Theorem 4.1a(i), where $\mathrm{H}_c = \mathrm{X}_c\left(\mathrm{X}_c'\mathrm{X}_c\right)^{-1}\mathrm{X}_c'$.

(b) Prove Theorem 4.1a(ii).

(c) Prove Theorem 4.1a(iii).

(d) Prove Theorem 4.1a(iv).

4.3: Show that $\lambda_1 = \beta_1' X_c' X_c \beta_1 / 2\sigma^2$ as in Theorem 4.1b(i).

4.4: Prove Theorem 4.1b(ii).

4.5: Show that $E(SSR/k) = \sigma^2 + (1/k)\beta_1' X_c' X_c \beta_1$, as in the expected mean square column of Table 4.1. Employ the following two approaches:

4.6: Develop a test for $H_0 : \beta = 0$ in the model $y = X\beta + \varepsilon$, where **y** is $N_n(X\beta, \sigma^2 I)$. (It was noted at the beginning of Section 4.1 that this hypothesis is of little practical interest because it includes $\beta_0 = 0$.) Use the partitioning $y'y = (y'y - \hat{\beta}'X'y) + \hat{\beta}'X'y$, and proceed as follows:

(a) Show that $\hat{\beta}'X'y = y'X(X'X)^{-1}X'y$ and $y'y - \hat{\beta}'X'y = y'[I - X(X'X)^{-1}X']y$.

(b) Let $H = X(X'X)^{-1}X'$: Show that **H** and **I** - **H** are idempotent of rank $k+1$ and $n - k$ - 1, respectively.

(c) Show that $y'Hy/\sigma^2$ is $\chi^2_{(k+1, \lambda_1)}$, where $\lambda_1 = \beta'X'X\beta/2\sigma^2$, and that $y'(1 - H)y/\sigma^2$ is $\chi^2_{(n-k-1)}$.

(d) Show that $y'Hy$ and $y'(1 - H)y$ are independent.

(e) Show that

$$\frac{\hat{\beta}'X'y}{(k+1)S^2} = \frac{y'Hy/(k+1)}{y'(1-H)y/(n-k-1)}$$

is distributed as $F_{(k+1, n-k-1, \lambda_1)}$.

3.7: Show that $HH_1 = H_1$ and $H_1H = H_1$, as in (4.15), where H and H_1 are as defined in (4.11) and (4.12).

3.8: Show that satisfied for the sum of quadratic forms in (4.12), as noted in the proof of Theorem 4.2b.

3.9: Show that $\lambda_1 = \beta_2'[X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2]\beta_2/2\sigma^2$ as in Theorem 4.2b(ii).

4.10: Show that $X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2$ is positive definite, as noted below Theorem 4.2b.

4.11: Show that $E[SS(\beta_2|\beta_1)/h] = \sigma^2 + \beta_2'[X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2]\beta_2/h$ as in Table 4.3.

4.12: Find the expected mean square corresponding to the numerator of the F statistic in (4.20) in Example 4.2b.

4.13: Show that $\hat{\beta}_0^* = \bar{y}$ and $SS\left(\beta_0^*\right) = n\bar{y}^2$, as in (4.21) in Example 4.2c.

4.14: In the proof of Theorem 4.2d, show that $\left(\hat{\beta}_1'X_1' + \hat{\beta}_2'X_2'\right)\left(X_1\hat{\beta}_1 + X_2\hat{\beta}_2\right) -$ $\left(\hat{\beta}_1' + \hat{\beta}_2'\,A'\right)X_1'\,X_1\left(\hat{\beta}_1 + A\hat{\beta}_2\right) = \hat{\beta}_2'[X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}\,X_1'X_2]\hat{\beta}_2$.

4.15 Express the test for $H_0 : \beta_2 = 0$ in terms of R^2, as in (4.25) in Theorem 4.3.

4.16: Prove Theorem 4.4a(iv).

4.17: Show that $C(X'X)^{-1}\,C'$ is positive definite, as noted following Theorem 4.4b.

4.18: Prove Theorem 4.4c.

4.19: Show that in the model $y = X\beta + \varepsilon$ subject to $C\beta = 0$ in (4.29), the estimator of β is $\hat{\beta}_c = \hat{\beta} - (X'X)^{-1}\,C'[C(X'X)^{-1}\,C']^{-1}\,C\hat{\beta}$ as in (4.30), where $\hat{\beta} = (X'X)^{-1}\,X'y$. Use a Lagrange multiplier λ and minimize $u = (y - X\beta)'(y - X\beta) + \lambda'(C\beta - 0)$ with respect to β and λ as follows:

(a) Differentiate *u* with respect to λ and set the result equal to 0 to obtain $C\hat{\beta}_c = 0$.

(b) Differentiate *u* with respect to β and set the result equal to 0 to obtain

$$\hat{\beta}_c = \hat{\beta} - \tfrac{1}{2}(X'X)^{-1}\,C'\lambda \qquad (1)$$

where $\hat{\beta} = (X'X)^{-1}\,X'y$.

(c) Multiply (1) in part (b) by **C**, use $C\hat{\beta}_c = 0$ from part (a), solve for λ, and substitute back into (1).

4.20: Show that $\hat{\beta}_c'\,X'X\hat{\beta}_c = \hat{\beta}_c'X'\,y$, thus demonstrating directly that the sum of squares due to the reduced model is $\hat{\beta}_c'X'y$ and that (4.31) holds.

4.21: Show that for the general linear hypothesis $H_0 : C\beta = 0$ in Theorem 4.4d, we have $\hat{\beta}'X'y - \hat{\beta}_c'\,X'\,y = \left(C\hat{\beta}\right)'[C(X'X)^{-1}\,C']^{-1}\,C\hat{\beta}$ as in (4.32), where $\hat{\beta}_c$ is as given in (4.30).

4.22: Prove Theorem 4.4e.

4.23: Prove Theorem 4.4f(iv) by expressing *SSH* and *SSE* as quadratic forms in the same normally distributed random vector.

4.24: Show that the estimator for β in the reduced model $y = X\beta + \varepsilon$ subject to $C\beta = t$ is given by $\hat{\beta}_c = \hat{\beta} - (X'X)^{-1} C'[C(X'X)^{-1} C']^{-1}(C\hat{\beta} - t)$, where $\hat{\beta} = (X'X)^{-1} X'y$.

4.25: Show that $\hat{\beta}'X'y - \hat{\beta}_1^{*\prime} X_1'y$ in (4.37) is equal to $\hat{\beta}_k^2 / g_{kk}$ in (4.39) (for $j = k$), as noted below (4.39).

4.26: Obtain the confidence interval for $a'\beta$ in (4.49) from the *t* statistic in (4.48).

4.27: Show that the confidence interval for $a_0'\beta$ in (4.52) is the same as that for the centered model in (4.55).

4.28: Show that the confidence interval for $\beta_0 + \beta_1 x_0$ in (4.58) follows from (4.55).

4.29: Show that $t = (y_0 - \hat{y}_0) \big/ S\sqrt{1 + x_0'(X'X)^{-1} x_0}$ in (4.60) is distributed as $t_{(n-k-1)}$.

4.30: (a) Given that $\bar{y}_0 = \sum_{i=1}^{q} y_{0i}/q$ is the mean of *q* future observations at x_0, show that a $100(1-\alpha)\%$ prediction interval for $\bar{y}_0$ is given by $x_0'\hat{\beta} \pm t_{(\alpha/2, n-k-1)} S \sqrt{1/q + x_0'(X'X)^{-1} x_0}$.

(b) Show that for simple linear regression, the prediction interval for $\bar{y}_0$ in part (a) reduces to $\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{(\alpha/2, n-2)} S \times \sqrt{1/q + 1/n + (x_0 - \bar{x})^2 \big/ \sum_{i=1}^{n} (x_i - \bar{x})^2}$.

4.31: Obtain the confidence interval for σ^2 in (4.65) from the probability statement in (4.64).

4.32: Show that the Scheffe′ prediction intervals for d future observations are given by (4.73).

4.33: Verify (4.76)–(4.79) in the proof of Theorem 4.7a.

4.34: Verify (4.80), $\partial v / \partial \beta = (2X'y - 2X'X\beta)/2\sigma^2 + C'\lambda$.

4.35: Show that the solution to (4.80)–(4.82) is given by $\hat{\beta}_0$ and $\hat{\sigma}_0^2$ in (4.83) and (4.84).

4.36: Show that $\left(y-X\hat{\beta}_0\right)'\left(y-X\hat{\beta}_0\right)=n\hat{\sigma}^2+\left(C\hat{\beta}\right)'[C(X'X)^{-1}C']^{-1}C\hat{\beta}$ as in (4.85).

4.37: Use the gas vapor data in Table 4.3.

(a) Test the overall regression hypothesis $H_0:\beta_1=0$ using (4.5) [or (4.22)] and (4.23).

(b) Test $H_0:\beta_1=\beta_3=0$, that is, that x_1 and x_3 do not significantly contribute above and beyond x_2 and x_4.

(c) Test $H_0:\beta_j=0$ for j = 1, 2, 3, 4 using t_j in (4.40). Use $t_{0.05/2}$ for each test and also use a Bonferroni approach based on $t_{0.05/8}$ (or compare the p value to 0.05/4).

(d) Using general linear hypothesis tests, test $H_0:\beta_1=\beta_2=12\beta_3=12\beta_4$, $H_{01}:\beta_1=\beta_2$, $H_{02}:\beta_2=12\beta_3$, $H_{03}:\beta_3=\beta_4$ and $H_{04}:\beta_1=\beta_2$ and $\beta_3=\beta_4$.

(e) Find confidence intervals for β_1, β_2, β_3 and β_4 using both (4.47) and (4.67).

4.38: Use the land rent data in Table 3.5.

(a) Test the overall regression hypothesis $H_0:\beta_1=0$ using (4.5) [or (4.22)] and (4.23).

(b) Test $H_0:\beta_j=0$ for j = 1, 2, 3 using t_j in (4.40). Use $t_{0.05/2}$ for each test and also use a Bonferroni approach based on $t_{0.05/6}$ (or compare the p value to 0.05/3).

(c) Find confidence intervals for β_1, β_2, β_3 using both (4.47) and (4.67).

(d) Using (4.52), find a 95% confidence interval for $E(y_0)=x_0'\beta$, where $x_0'=(1,\ \ 15,\ \ 30,\ \ 0.5)$.

(e) Using (4.61), find a 95% prediction interval for $y_0=x_0'\beta+\varepsilon$, where $x_0'=(1,\ \ 15,\ \ 30,\ \ 0.5)$.

4.39: Use y_2 in the chemical reaction data in Table 3.4.

(a) Using (4.52), find a 95% confidence interval for $E(y_0) = \mathrm{x}_0'\beta$, where $\mathrm{x}_0' = (1,\ 165,\ 32,\ 5)$.

(b) Using (4.61), find a 95% prediction interval for $y_0 = \mathrm{x}_0'\beta + \varepsilon$, where $\mathrm{x}_0' = (1,\ 165,\ 32,\ 5)$.

(c) Test $H_0 : 2\beta_1 = 2\beta_2 = \beta_3$ using (4.27). (This was done for y_1 in Example 4.4.b.)

4.40: Use y_1 in the chemical reaction data in Table 3.4. The full model with second order terms and the reduced model with only linear terms were fit in Problem 3.52.

(a) Test $H_0 : \beta_4 = \beta_5 = \cdots = \beta_9 = 0$, that is, that the second-order terms are not useful in predicting y_1. (This was done for y_2 in Example 4.2a.)

(b) Test the significance of the increase in R^2 from the reduced model to the full model. (This was done for y_2 in Example 4.3. See Problem 3.52 for values of R^2.)

(c) Find a 95% confidence interval for each of β_0, β_1, β_2, β_3 using (4.47).

(d) Find Bonferroni confidence intervals for β_1, β_2, β_3 using (4.67).

(e) Using (4.52), find a 95% confidence interval for $E(y_0) = \mathrm{x}_0'\beta$, where $\mathrm{x}_0' = (1,\ 165,\ 32,\ 5)$.

(f) Using (4.61), find a 95%, prediction interval for $y_0 = \mathrm{x}_0'\beta + \varepsilon$, where $\mathrm{x}_0' = (1,\ 165,\ 32,\ 5)$.

Chapter Five

Multiple Regression: Model Validation and Diagnostics

5: Introduction

In Sections 3.8.2 and 3.9 we discussed some consequences of misspecification of the model. In this chapter we consider various approaches to checking the model and the attendant assumptions for adequacy and validity. Some properties of the residuals [see (3.11)] and the hat matrix are developed in Sections 5.1 and 5.2.We discuss outliers, the influence of individual observations, and leverage in Sections 5.3 and 5.4.

For additional reading, see Snee (1977), Cook (1977), Belsley et al. (1980), Draper and Smith (1981, Chapter 6), Cook and Weisberg (1982), Beckman and Cook (1983), Weisberg (1985, Chapters 5, 6), Chatterjee and Hadi (1988), Myers (1990, Chapters 5–8), Sen and Srivastava (1990, Chapter 8), Montgomery and Peck (1992, pp. 67–113, 159–192), Jørgensen (1993, Chapter 5), Graybill and Iyer (1994, Chapter 5), Hocking (1996, Chapter 9), Christensen (1996, Chapter 13), Ryan (1997, Chapters 2, 5), Fox (1997, Chapters 11–13) and Kutner et al. (2005, Chapter 10).

5.1: Residuals

The usual model is given by (3.4) as $\mathrm{y} = \mathrm{X}\beta + \varepsilon$ with assumptions $E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 \mathrm{I}$, where **y** is $n \times 1$, **X** is $n \times (k+1)$ of rank $k+1 < n$, and β is $(k+1) \times 1$. The error vector ε is unobservable unless β is known. To estimate ε for a given sample, we use the residual vector

$$\hat{\varepsilon} = \mathrm{y} - \mathrm{X}\hat{\beta} = \mathrm{y} - \hat{\mathrm{y}} \qquad (5.1)$$

As defined in (3.11). The n residuals in (5.1), $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n$, are used in various plots and procedures for checking on the validity or adequacy of the model.

We first consider some properties of the residual vector $\hat{\varepsilon}$. Using the least-squares estimator $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ in (3.6), the vector of predicted values $\hat{\mathrm{y}} = \mathrm{X}\hat{\beta}$ can be written as

$$\begin{aligned} \hat{\mathrm{y}} &= \mathrm{X}\hat{\beta} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y} \\ &= \mathrm{H}\mathrm{y} \qquad (5.2) \end{aligned}$$

where $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ (see Section 4.2). The $n \times n$ matrix **H** is called the hat matrix because it transforms **y** to $\hat{\mathrm{y}}$. We also refer to **H** as a *projection matrix* for essentially the same reason; geometrically it projects **y** (perpendicularly) onto $\hat{\mathrm{y}}$ (see Fig. 3.4). The hat matrix **H** is symmetric and idempotent.

Multiplying **X** by **H**, we obtain

$$\mathrm{HX} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X} = \mathrm{X} \qquad (5.3)$$

Writing **X** in terms of its columns, we can write (5.3) as

$$\mathrm{HX} = \mathrm{H}(\mathrm{j}, \mathrm{x}_1, \cdots, \mathrm{x}_k) = (\mathrm{Hj}, \mathrm{Hx}_1, \cdots, \mathrm{Hx}_k)$$

so that

$$\mathrm{j} = \mathrm{H}\,\mathrm{j}, \quad \mathrm{x}_i = \mathrm{Hx}_i, \quad i = 1,2,\cdots,k \qquad (5.4)$$

Using (5.2), the residual vector $\hat{\varepsilon}$ (5.1) can be expressed in terms of **H**:

$$\begin{aligned}\hat{\varepsilon} &= \mathrm{y} - \hat{\mathrm{y}} = \mathrm{y} - \mathrm{Hy} \\ &= (\mathrm{I} - \mathrm{H})\mathrm{y} \qquad (5.5)\end{aligned}$$

We can rewrite (5.5) to express the residual vector $\hat{\varepsilon}$ in terms of ε:

$$\begin{aligned}\hat{\varepsilon} &= (\mathrm{I} - \mathrm{H})\mathrm{y} = (\mathrm{I} - \mathrm{H})(\mathrm{X}\beta + \varepsilon) \\ &= (\mathrm{X}\beta - \mathrm{HX}\,\beta) + (\mathrm{I} - \mathrm{H})\varepsilon \\ &= (\mathrm{X}\beta - \mathrm{X}\beta) + (\mathrm{I} - \mathrm{H})\varepsilon \text{ [by (5.3)]} \\ &= (\mathrm{I} - \mathrm{H})\varepsilon \qquad (5.6)\end{aligned}$$

In terms of the elements h_{ij} of **H**, we have $\hat{\varepsilon}_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j$, $i = 1,2,\cdots,n$. Thus, if the h_{ij}'s are small (in absolute value), $\hat{\varepsilon}$ is close to ε.

The following are some of the properties of $\hat{\varepsilon}$ (see Problem 5.1). For the first four, we assume that $E(\mathrm{y}) = \mathrm{X}\beta$ and $Cov(\mathrm{y}) = \sigma^2\mathrm{I}$:

$$E(\hat{\varepsilon}) = 0 \qquad (5.7)$$

$$Cov(\hat{\varepsilon}) = \sigma^2[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'] = \sigma^2(\mathrm{I} - \mathrm{H}) \qquad (5.8)$$

$$Cov(\hat{\varepsilon}, \mathrm{y}) = \sigma^2[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'] = \sigma^2(\mathrm{I} - \mathrm{H}) \qquad (5.9)$$

$$Cov(\hat{\varepsilon}, \hat{\mathrm{y}}) = \mathrm{O} \qquad (5.10)$$

$$\bar{\hat{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i / n = \hat{\varepsilon}'\mathrm{j}/n = 0 \qquad (5.11)$$

$$\hat{\varepsilon}'\mathrm{y} = SSE = \mathrm{y}'[\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\mathrm{y} = \mathrm{y}'(\mathrm{I} - \mathrm{H})\mathrm{y} \qquad (5.12)$$

$$\hat{\varepsilon}'\hat{\mathrm{y}} = 0 \qquad (5.13)$$

$$\hat{\varepsilon}'\mathrm{X} = 0' \qquad (5.14)$$

In (5.7), the residual vector $\hat{\varepsilon}$ has the same mean as the error term ε, but in (5.8) $Cov(\hat{\varepsilon}) = \sigma^2(\mathrm{I} - \mathrm{H})$ differs from the assumption $Cov(\varepsilon) = \sigma^2 \mathrm{I}$. Thus the residuals $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n$, are not independent. However, in many cases, especially if *n* is large, the h_{ij}'s tend to be small (for $i \neq j$), and the dependence shown in $\sigma^2(\mathrm{I}-\mathrm{H})$ does not unduly affect plots and other techniques for model validation. Each $\hat{\varepsilon}_i$ is seen to be correlated with each y_j in (5.9), but in (5.10) the $\hat{\varepsilon}_i$'s are uncorrelated with the $\hat{y}_j$'s.

Some sample properties of the residuals are given in (5.11)–(5.14). The sample mean of the residuals is zero, as shown in (5.11). By (5.12), it can be seen that $\hat{\varepsilon}$ and **y** are correlated in the sample since $\hat{\varepsilon}'\mathrm{y}$ is the numerator of

$$r_{\hat{\varepsilon}\mathrm{y}} = \frac{\hat{\varepsilon}'(\mathrm{y} - \bar{y}\,\mathrm{j})}{\sqrt{(\hat{\varepsilon}'\hat{\varepsilon})(\mathrm{y} - \bar{y}\,\mathrm{j})'(\mathrm{y} - \bar{y}\,\mathrm{j})}} = \frac{\hat{\varepsilon}'\mathrm{y}}{\sqrt{(\hat{\varepsilon}'\hat{\varepsilon})(\mathrm{y} - \bar{y}\,\mathrm{j})'(\mathrm{y} - \bar{y}\,\mathrm{j})}}$$

However, $\hat{\varepsilon}$ and $\hat{\mathrm{y}}$ are orthogonal by (5.13), and therefore

$$r_{\hat{\varepsilon}\hat{\mathrm{y}}} = 0 \qquad (5.15)$$

Similarly, by (5.14), $\hat{\varepsilon}$ is orthogonal to each column of **X** and

$$r_{\hat{\varepsilon} x_i} = 0, \quad i = 1, 2, \cdots, k \qquad (5.16)$$

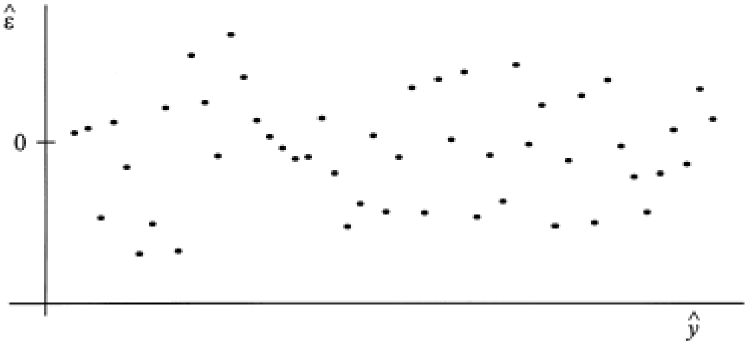


Figure 5.1: Ideal residual plot when model is correct.

If the model and attendant assumptions are correct, then by (5.15), a plot of the residuals versus predicted values, $(\hat{\varepsilon}_1, \hat{y}_1)$, $(\hat{\varepsilon}_2, \hat{y}_2)$, …, $(\hat{\varepsilon}_n, \hat{y}_n)$, should show no systematic pattern. Likewise, by (5.16), the *k* plots of the residuals versus each of $x_1, x_2, \cdots, x_k$ should show only random variation. These plots are therefore useful for checking the model. A typical plot of this type is shown in Figure 5.1. It may also be useful to plot the residuals on normal probability paper and to plot residuals in time sequence (Christensen 1996, Section 13.2).

If the model is incorrect, various plots involving residuals may show departures from the fitted model such as outliers, curvature, or non-constant variance. The plots may also suggest remedial measures to improve the fit of the model. For example, the residuals could be plotted versus any of the x_i's, and a simple curved pattern might suggest the addition of x_i^2 to the model. We will consider various approaches for detecting outliers in Section 5.3 and for finding influential observations in Section 5.4. Before doing so, we discuss some properties of the hat matrix in Section 5.2.

5.2: The Hat Matrix

It was noted following (9.2) that the hat matrix $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ is symmetric and idempotent. We now present some additional properties of this matrix. These properties will be useful in the discussion of outliers and influential observations in Sections 5.3 and 5.4.

For the centered model

$$\mathrm{y} = \alpha\,\mathrm{j} + \mathrm{X}_c\,\beta_1 + \varepsilon \tag{5.17}$$

In (3.32), $\hat{\mathrm{y}}$ becomes

$$\hat{\mathrm{y}} = \hat{\alpha}\,\mathrm{j} + \mathrm{X}_c\hat{\beta}_1 \tag{5.18}$$

And the hat matrix is $\mathrm{H}_c = \mathrm{X}_c(\mathrm{X}_c'\mathrm{X}_c)^{-1}\mathrm{X}_c'$, where

$$\mathrm{X}_c = \left(\mathrm{I} - \frac{1}{n}\mathrm{J}\right)\mathrm{X}_1 = \begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \cdots & x_{1k} - \overline{x}_k \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{2k} - \overline{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 & \cdots & x_{nk} - \overline{x}_k \end{pmatrix}$$

By (3.36) and (3.37), we can write (5.18) as

$$\hat{\mathrm{y}} = \bar{y}\,\mathrm{j} + \mathrm{X}_c\left(\mathrm{X}_c'\mathrm{X}_c\right)^{-1}\mathrm{X}_c'\mathrm{y} = \left(\frac{1}{n}\mathrm{j}'\mathrm{y}\right)\mathrm{j} + \mathrm{H}_c\,\mathrm{y}$$

$$= \left(\frac{1}{n}\mathrm{J} + \mathrm{H}_c\right)\mathrm{y} \qquad (5.19)$$

Comparing (5.19) and (5.2), we have

$$\mathrm{H} = \frac{1}{n}\mathrm{J} + \mathrm{H}_c = \frac{1}{n}\mathrm{J} + \mathrm{X}_c\left(\mathrm{X}_c'\mathrm{X}_c\right)^{-1}\mathrm{X}_c' \qquad (5.20)$$

We now examine some properties of the elements h_{ij} of **H**.

Theorem 5.2: If **X** is $n \times (k+1)$ of rank $k+1 < n$, and if the first column of **X** is **j**, then the elements h_{ij} of $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ have the following properties:

(i) $(1/n) \le h_{ii} \le 1$ for $i = 1, 2, \ldots, n$.

(ii) $-0.5 \le h_{ij} \le 0.5$ for all $j \ne i$.

(iii) $h_{ii} = (1/n) + (\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)'(\mathrm{X}_c'\mathrm{X}_c)^{-1}(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)$, where $\mathrm{x}_{1i}' = (x_{i1}, x_{i2}, \cdots, x_{ik})$, $\bar{\mathrm{x}}_1' = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k)$, and $(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)'$ is the *i*th row of the centered matrix X_c.

(iv) $tr(\mathrm{H}) = \sum_{i=1}^{n} h_{ii} = k + 1$.

Proof

(i) The lower bound follows from (5.20), since $\mathrm{X}_c'\mathrm{X}_c$ is positive definite. Since **H** is symmetric and idempotent, we use the relationship $\mathrm{H} = \mathrm{H}^2$ to find an upper bound on h_{ii}. Let h_i' be the *i*th row of **H**. Then

$$h_{ii} = \mathrm{h}_i'\mathrm{h}_i = (h_{i1}, h_{i2}, \cdots, h_{in})\begin{pmatrix} h_{i1}, \\ h_{i2}, \\ \vdots \\ h_{in} \end{pmatrix} = \sum_{j=1}^{n} h_{ij}^2$$

$$= h_{ii}^2 + \sum_{j \ne i} h_{ij}^2 \qquad (5.21)$$

Dividing both sides of (5.21) by h_{ii} [which is positive since $h_{ii} \geq (1/n)$], we obtain

$$1 = h_{ii} + \frac{\sum_{j \neq i} h_{ij}^2}{h_{ii}} \tag{5.22}$$

which implies $h_{ii} \leq 1$.

(ii) (Chatterjee and Hadi 1988, p. 18.) We can write (5.21) in the form

$$h_{ii} = h_{ii}^2 + h_{ij}^2 + \sum_{r \neq i, j} h_{ir}^2$$

or

$$h_{ii} - h_{ii}^2 = h_{ij}^2 + \sum_{r \neq i, j} h_{ir}^2$$

Thus, $h_{ij}^2 \leq h_{ii} - h_{ii}^2$, and since the maximum value of $h_{ii} - h_{ii}^2$ is $1/4$, we have $h_{ij}^2 \leq 1/4$ for $j \neq i$.

(iii) This follows from (5.20); see Problem 5.2b.

(iv) See Problem 5.2c.

By Theorem 5.2(iv), we see that as *n* increases, the values of h_{ii} will tend to decrease.

The function $(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)'(\mathrm{X}_c'\mathrm{X}_c)^{-1}(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)$ in Theorem 5.2(iii) is a standardized distance. The standardized distance (Mahalanobis distance) is for a population covariance matrix. The matrix $\mathrm{X}_c'\mathrm{X}_c$ is proportional to a sample covariance matrix [see (3.44)]. Thus, $(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)'(\mathrm{X}_c'\mathrm{X}_c)^{-1}(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)$ is an estimated standardized distance and provides a good measure of the relative distance of each x_{1i} from the center of the points as represented by $\bar{\mathrm{x}}_1$.

5.3: Outliers

In some cases, the model appears to be correct for most of the data, but one residual is much larger (in absolute value) than the others. Such an outlier may be due to an error in recording or may be from another population or may simply be an unusual observation from the assumed distribution. For example, if the errors ε_i are distributed as $N(0, \sigma^2)$, a value of ε_i greater than 3σ or less than -3σ would occur with frequency .0027.

If no explanation for an apparent outlier can be found, the dataset could be analyzed both with and without the outlying observation. If the results differ sufficiently to affect the conclusions, then both analyses could be maintained until additional data become available. Another alternative is to discard the outlier, even though no explanation has been found. A third possibility is to use robust methods that accommodate the outlying observation (Huber 1973, Andrews 1974, Hampel 1974, Welsch 1975, Devlin et al. 1975, Mosteller and Turkey 1977, Birch 1980, Krasker and Welsch 1982).

One approach to checking for outliers is to plot the residuals $\hat{\varepsilon}_i$ versus $\hat{y}_i$ or versus *i*, the observation number. In our examination of residuals, we need to keep in mind that by (5.8), the variance of the residuals is not constant:

$$Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}) \qquad (5.23)$$

By Theorem 5.2(i), $h_{ii} \le 1$; hence, $Var(\hat{\varepsilon}_i)$ will be small if h_{ii} is near 1. By Theorem 5.2(iii), h_{ii} will be large if x_{1i} is far from $\overline{\mathrm{x}}_1$, where x_{1i} = $(x_{i1}, x_{i2}, \cdots, x_{ik})'$ and $\overline{\mathrm{x}}_1 = (\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_k)'$. By (9.23), such observations will tend to have small residuals, which seem unfortunate because the model is less likely to hold far from $\overline{\mathrm{x}}_1$. A small residual at a point where x_{1i} is far from $\overline{\mathrm{x}}_1$ may result because the fitted model will tend to pass close to a point isolated from the bulk of the points, with a resulting poorer fit to the bulk of the data. This may mask an inadequacy of the true model in the region of x_{1i}.

An additional verification that large value of h_{ii} are accompanied by small residuals is provided by the following inequality (see Problem 5.4):

$$\frac{1}{n} \le h_{ii} + \frac{\hat{\varepsilon}_i^2}{\hat{\varepsilon}'\hat{\varepsilon}} \le 1 \qquad (5.24)$$

For the reasons implicit in (5.23) and (5.24), it is desirable to scale the residuals so that they have the same variance. There are two common (and related) methods of scaling.

For the first method of scaling, we use $Var(\hat{\varepsilon}_i) = \sigma^2(1-h_{ii})$ in (9.23) to obtain the standardized residuals $\hat{\varepsilon}_i / \sigma\sqrt{(1-h_{ii})}$, which have mean 0 and variance 1. Replacing σ by *S* yields the *studentized residual*

$$r_i = \frac{\hat{\varepsilon}_i}{S\sqrt{(1-h_{ii})}} \tag{5.25}$$

where $S^2 = SSE/(n-k-1)$ is as defined in (3.24). The use of r_i in place of $\hat{\varepsilon}_i$ eliminates the location effect (due to h_{ii}) on the size of residuals, as discussed following (5.23). A second method of scaling the residuals uses an estimate of σ that excludes the *i*th observation

$$t_i = \hat{\varepsilon}_i / S_{(i)}\sqrt{(1-h_{ii})} \tag{5.26}$$

Where $S_{(i)}$ is the standard error computed with the $n-1$ observations remaining after omitting $(y_i, \mathrm{x}'_i) = (y_{i1}, x_{i1}, \cdots x_{ik})$, in which y_i is the *i*th element of **y** and x'_i is the *i*th row of **X**. If the *i*th observation is an outlier, it will more likely show up as such with the standardization in (5.26), which is called the *externally studentized residual* or the *studentized deleted residual* or *R student*.

Another option is to examine the *deleted residuals*. The *i*th deleted residual, $\varepsilon_{(i)}$, is computed with $\hat{\beta}_{(i)}$ on the basis of $n-1$ observations with (y_i, x'_i) deleted:

$$\hat{\varepsilon}_{(i)} = y_i - \hat{y}_{(i)} = y_i - \mathrm{x}'_i\,\hat{\beta}_{(i)} \tag{5.27}$$

By definition

$$\hat{\beta}_{(i)} = \left(\mathrm{X}'_{(i)}\mathrm{X}_{(i)}\right)^{-1}\mathrm{X}'_{(i)}\,y_{(i)} \tag{5.28}$$

Where $\mathrm{X}_{(i)}$ is the $(n-1)\times(k+1)$ matrix obtained by deleting $\mathrm{x}'_i = (1, x_{i1}, x_{i2}, \cdots, x_{ik})'$, the *i*th row of **X**, and $y_{(i)}$ is the corresponding $(n-1)\times 1$ **y** vector after deleting y_i. The deleted vector $\hat{\beta}_{(i)}$ can also be found without actually deleting (y_i, x'_i) since

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{\hat{\varepsilon}_i}{1-h_{ii}}(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{x}_i \tag{5.29}$$

(see Problem 5.5).

The deleted residual $\hat{\varepsilon}_{(i)} = y_i - \mathrm{x}_i' \hat{\beta}_{(i)}$ in (5.27) can be expressed in terms of $\hat{\varepsilon}_i$ and h_{ii} as

$$\hat{\varepsilon}_{(i)} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}} \tag{5.30}$$

(see Problem 5.6). Thus the *n* deleted residuals can be obtained without computing *n* regressions. The scaled residual t_i in (5.26) can be expressed in terms of $\hat{\varepsilon}_{(i)}$ in (5.30) as

$$t_i = \frac{\hat{\varepsilon}_{(i)}}{\sqrt{\hat{V}ar\left(\hat{\varepsilon}_{(i)}\right)}} \tag{5.31}$$

(see Problem 5.7).

The deleted sample variance $s_{(i)}^2$ used in (5.26) is defined as $s_{(i)}^2$ = $SSE_{(i)}/(n-k-1)$, where $SSE_{(i)} = \mathrm{y}_{(i)}'\mathrm{y}_{(i)} - \hat{\beta}_{(i)}' \mathrm{X}_{(i)}' \mathrm{y}_{(i)}$. This can be found without excluding the *i*th observation as

$$s_{(i)}^2 = \frac{SSE_{(i)}}{n-k-2} = \frac{SSE - \hat{\varepsilon}_i^2/(1-h_{ii})}{n-k-2} \tag{5.32}$$

(see Problem 5.8).

Another option for outlier detection is to plot the ordinary residuals $\hat{\varepsilon}_i = y_i - \mathrm{x}_i'\hat{\beta}$ against the deleted residuals $\hat{\varepsilon}_{(i)}$ in (5.27) or (5.30). If the fit does not change substantially when the *i*th observation is deleted in computation of $\hat{\beta}$, the plotted points should approximately follow a straight line with a slope of 1. Any points that are relatively far from this line are potential outliers.

If an outlier is from a distribution with a different mean, the model can be expressed as $E(y_i) = \mathrm{x}_i'\beta + \theta$, where x_i' is the *i*th row of **X**. This is called the *mean-shift outlier model*. The distribution of t_i in (5.26) or (5.31) is $t_{(n-k-1)}$, and t_i can therefore be used in a test of the hypothesis $H_0: \theta = 0$. Since *n* tests will be made, a Bonferroni adjustment to the critical values can be used, or we can simply focus on the largest t_i values.

The *n* deleted residuals in (5.30) can be used for model validation or selection by defining the *prediction sum of squares* (PRESS):

$$\text{PRESS} = \sum\nolimits_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum\nolimits_{i=1}^{n} \left(\frac{\hat{\varepsilon}_i}{1-h_{ii}} \right)^2 \qquad (5.33)$$

Thus, a residual $\hat{\varepsilon}_i$ that corresponds to a large value of h_{ii} contributes more to **PRESS**. For a given dataset, **PRESS** may be a better measure than *SSE* of how well the model will predict future observations. To use **PRESS** to compare alternative models when the objective is prediction, preference would be shown to models with small values of **PRESS**.

5.4: Influential Observations and Leverage

In Section 5.3, we emphasized a search for outliers that did not fit the model. In this section, we consider the effect that deletion of an observation (y_i, x_i') has on the estimates $\hat{\beta}$ and $\mathrm{X}\hat{\beta}$. An observation that makes a major difference on these estimates is called an *influential observation*. A point (y_i, x_i') is potentially influential if it is an outlier in the **y** direction or if it is unusually far removed from the center of the *x*'s.

We illustrate influential observations for the case of one *x* in Figure 5.2. Points 1 and 3 are extreme in the *x* direction; points 2 and 3 would likely appear as outliers in the *y* direction. Even though point 1 is extreme in *x*, it will not unduly influence the slope or intercept. Point 3 will have a dramatic influence on the slope and intercept since the regression line would pass near point 3. Point 2 is also influential, but much less so than point 3.

Thus, influential points are likely to be found in areas where little or no other data were collected. Such points may be fitted very well, sometimes to the detriment of the fit to the other data.

To investigate the influence of each observation, we begin with $\hat{\mathrm{y}} = \mathrm{Hy}$ in (5.2), the elements of which are

$$\hat{y}_i = \sum\nolimits_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum\nolimits_{j \neq i} h_{ij} y_i \qquad (5.34)$$

By (5.22), if h_{ii} is large (close to 1), then the h_{ij} 's, $j \neq i$, are all small, and y_i contributes much more than the other *y*'s to $\hat{y}_i$. Hence, h_{ii} is called the *leverage* of y_i. Points with high leverage have high potential for influencing regression results. In general, if an observation (y_i, x_i')

has a value of h_{ii} near 1, then the estimated regression equation will be close to y_i; that is, $\hat{y}_i - y_i$ will be small.

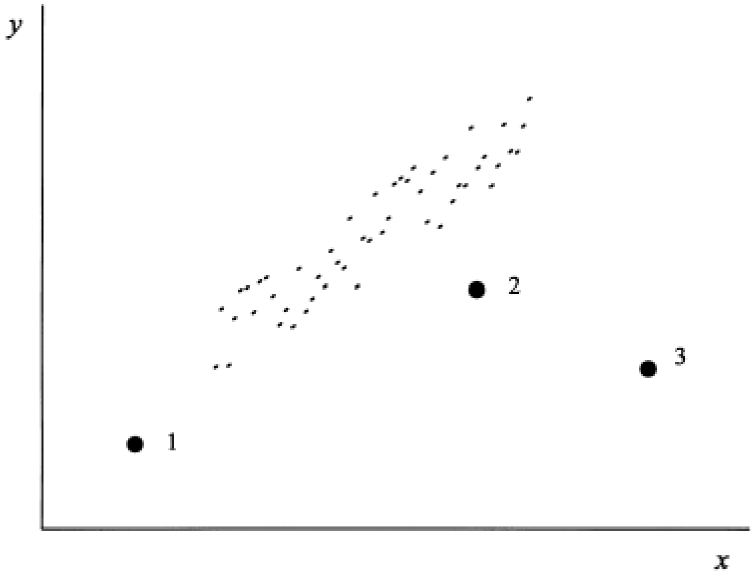


Figure 5.2: Simple linear regression showing three outliers.

By Theorem 5.2(iv), the average value of the h_{ii}'s is $(k+1)/n$. Hoaglin and Welsch (1978) suggest that a point with $h_{ii} > 2(k+1)/n$ is a high leverage point. Alternatively, we can simply examine any observation whose value of h_{ii} is unusually large relative to the other values of h_{ii}.

In terms of fitting the model to the bulk of the data, high leverage points can be either good or bad, as illustrated by points 1 and 3 in Figure 5.2. Point 1 may reduce the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$. On the other hand, point 3 will drastically alter the fitted model. If point 3 is not the result of a recording error, then the researcher must choose between two competing fitted models. Typically, the model that fits the bulk of the data might be preferred until additional points can be observed in other areas.

To formalize the influence of a point (y_i, x_i'), we consider the effect of its deletion on β and $\hat{\mathrm{y}} = \mathrm{X}\hat{\beta}$. The estimate of β obtained by deleting the *i*th observation (y_i, x_i') is defined in (5.28) as $\hat{\beta}_{(i)} = \left(\mathrm{X}_{(i)}'\mathrm{X}_{(i)}\right)^{-1} \mathrm{X}_{(i)}' y_{(i)}$. We can compare $\hat{\beta}_{(i)}$ to $\hat{\beta}$ by means of *Cook's distance*, defined as

$$D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)' (X'X)\left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{(k+1)S^2} \tag{5.35}$$

This can be rewritten as

$$D_i = \frac{\left(X\hat{\beta}_{(i)} - X\hat{\beta}\right)' \left(X\hat{\beta}_{(i)} - X\hat{\beta}\right)}{(k+1)S^2}$$

$$= \frac{\left(\hat{y}_{(i)} - \hat{y}\right)' \left(\hat{y}_{(i)} - \hat{y}\right)}{(k+1)S^2} \tag{5.36}$$

In which D_i is proportional to the ordinary Euclidean distance between $\hat{y}_{(i)}$ and $\hat{y}$. Thus if D_i is large, the observation (y_i, x_i') has substantial influence on both $\hat{\beta}$ and $\hat{y}$. A more computationally convenient form of D_i is given by

$$D_i = \frac{r_i^2}{k+1}\left(\frac{h_{ii}}{1-h_{ii}}\right) \tag{5.37}$$

TABLE 5.1: Residuals and Influence Measures for the Chemical Data with Dependent Variable y_1

Observation	y_i	$\hat{y}_i$	$\hat{\varepsilon}_i$	h_{ii}	r_i	t_i	D_i
1	41.5	42.19	-0.688	0.430	-0.394	-0.383	0.029
2	33.8	31.00	2.798	0.310	1.457	1.520	0.239
3	27.7	27.74	-0.042	0.155	-0.020	-0.019	0.000
4	21.7	21.03	0.670	0.139	0.313	0.303	0.004
5	19.9	19.40	0.495	0.129	0.230	0.222	0.002
6	15	12.69	2.307	0.140	1.076	1.083	0.047
7	12.2	12.28	-0.082	0.228	-0.040	-0.039	0.000
8	4.3	5.57	-1.270	0.186	-0.610	-0.596	0.021
9	19.3	20.22	-0.917	0.053	-0.408	-0.396	0.002
10	6.40	4.758	1.642	0.233	0.811	0.801	0.050
11	37.6	35.68	1.923	0.240	0.954	0.951	0.072
12	18	13.09	4.906	0.164	2.320	2.800	0.264
13	26.3	27.34	-1.040	0.146	-0.487	-0.474	0.010
14	9.9	13.51	-3.605	0.245	-1.795	-1.956	0.261
15	25	26.93	-1.929	0.250	-0.964	-0.961	0.077
16	14.1	15.44	-1.342	0.258	-0.674	-0.661	0.039
17	15.2	15.44	-0.242	0.258	-0.121	-0.117	0.001
18	15.9	19.54	-3.642	0.217	-1.780	-1.937	0.220
19	19.6	19.54	0.058	0.217	0.028	0.027	0.000

(see Problem 5.9). Muller and Mok (1997) discuss the distribution of D_i and provide a table of critical values.

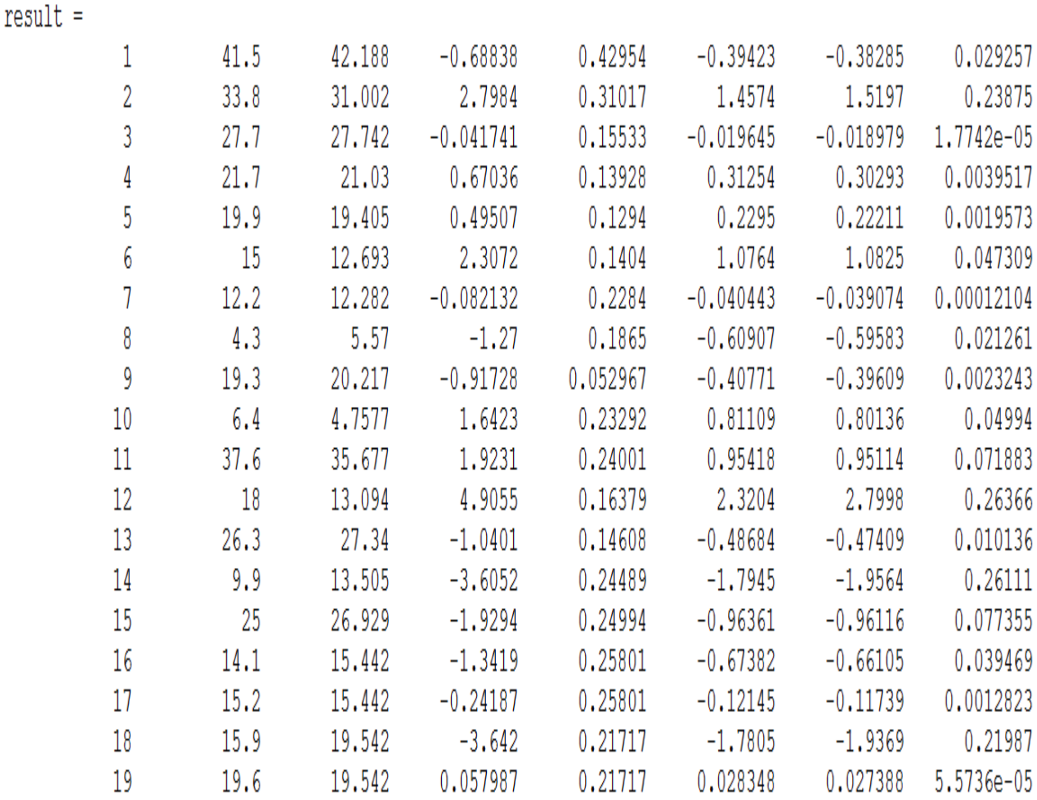
Example 5.4: We illustrate several diagnostic tools for the chemical reaction data of Table 3.4 using y_1. In Table 5.1, we give $\hat{\varepsilon}_i$, h_{ii}, and some functions of these from Sections 5.3 and 5.4.

The guideline for h_{ii} in Section 5.4 is 2(*k*+1)/*n* = 2(4)/19 = 0.421. The only value of h_{ii} that exceeds 0.421 is the first, $h_{11} = 0.430$. Thus the first observation has potential for influencing the model fit, but this influence does not appear in $t_1 = -0.383$ and $D_1 = 0.029$. Other relatively large values of h_{ii} are seen for observations 2, 11, 14, 15, 16, and 17. Of these only observation 14 has a very large (absolute) value of t_i. Observation 12 has large values of $\hat{\varepsilon}_i$, r_i, t_i and D_i and is a potentially influential outlier.

The value of **PRESS** as defined in (5.33) is **PRESS** = 130.76, which can be compared to *SSE* = 80.17.

Example 5.4 [The program name ta21.m]	Applications using MATLAB

```
clc
clear all
data=[41.5 45.9 162 23 3;33.8 53.3 162 23 8;27.7 57.5 162 30 5
     21.7 58.8 162 30 8;19.9 60.6 172 25 5;15.0 58.0 172 25 8
     12.2 58.6 172 30 5;4.3 52.4 172 30 8;19.3 56.9 167 27.5 6.5
     6.4 55.4 177 27.5 6.5;37.6 46.9 157 27.5 6.5;18 57.3 167 32.5 6.5
     26.3 55.0 167 22.5 6.5;9.9 58.9 167 27.5 9.5;25.0 50.3 167 27.5 3.5
     14.1 61.1 177 20 6.5;15.2 62.9 177 20 6.5;15.9 60.0 160 34 7.5
     19.6 60.6 160 34 7.5];x1=data(:,3);x2=data(:,4);x3=data(:,5);
  y1=data(:,1);n=length(x1);x=[ones(size(x1)) x1 x2 x3];
ixx=inv(x'*x);beta=x\y1;SSE=y1'*y1-beta'*x'*y1;S=sqrt(SSE/(n-4));
Y1=x*beta;E=y1-Y1;observation=(1:19)';H=x*ixx*x';h=diag(H);
r=E./(S*sqrt(1-h));
for i=1:n
   x(i,:)=[];y1(i)=[];beta=x\y1;SSE=y1'*y1-beta'*x'*y1;
   s(i)=sqrt(SSE/(n-5));y1=data(:,1);x1=data(:,3);x2=data(:,4);
   x3=data(:,5);x=[ones(size(x1)) x1 x2 x3];t(i)=E(i)/(s(i)*sqrt(1-h(i)));
   D(i)=(r(i)^2/(4))*(h(i)/(1-h(i)));
end
t=t';D=D';
result=[observation y1 Y1 E h r t D],PRESS=sum((E./(1-h)).^2),SSE
```



```
result =
           1        41.5      42.188    -0.68838     0.42954    -0.39423    -0.38285    0.029257
           2        33.8      31.002      2.7984     0.31017      1.4574      1.5197     0.23875
           3        27.7      27.742   -0.041741     0.15533   -0.019645   -0.018979  1.7742e-05
           4        21.7       21.03     0.67036     0.13928     0.31254     0.30293   0.0039517
           5        19.9      19.405     0.49507      0.1294      0.2295     0.22211   0.0019573
           6          15      12.693      2.3072      0.1404      1.0764      1.0825    0.047309
           7        12.2      12.282   -0.082132      0.2284   -0.040443   -0.039074  0.00012104
           8         4.3        5.57       -1.27      0.1865    -0.60907    -0.59583    0.021261
           9        19.3      20.217    -0.91728    0.052967    -0.40771    -0.39609   0.0023243
          10         6.4      4.7577      1.6423     0.23292     0.81109     0.80136     0.04994
          11        37.6      35.677      1.9231     0.24001     0.95418     0.95114    0.071883
          12          18      13.094      4.9055     0.16379      2.3204      2.7998     0.26366
          13        26.3       27.34     -1.0401     0.14608    -0.48684    -0.47409    0.010136
          14         9.9      13.505     -3.6052     0.24489     -1.7945     -1.9564     0.26111
          15          25      26.929     -1.9294     0.24994    -0.96361    -0.96116    0.077355
          16        14.1      15.442     -1.3419     0.25801    -0.67382    -0.66105    0.039469
          17        15.2      15.442    -0.24187     0.25801    -0.12145    -0.11739   0.0012823
          18        15.9      19.542      -3.642     0.21717     -1.7805     -1.9369     0.21987
          19        19.6      19.542    0.057987     0.21717    0.028348    0.027388  5.5736e-05
```

PRESS =

130.76

SSE =

80.169

PROBLEMS

5.1: Verify the following properties of the residual vector $\hat{\varepsilon}$ as given in (5.7)–(5.14):

(a) $E(\hat{\varepsilon}) = 0$

(b) $Cov(\hat{\varepsilon}) = \sigma^2(\mathrm{I} - \mathrm{H})$

(c) $Cov(\hat{\varepsilon}, \mathrm{y}) = \sigma^2(\mathrm{I} - \mathrm{H})$

(d) $Cov(\hat{\varepsilon}, \hat{\mathrm{y}}) = \mathrm{O}$

(e) $\bar{\hat{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i / n = 0$

(f) $\hat{\varepsilon}'\mathrm{y} = \mathrm{y}'(\mathrm{I} - \mathrm{H})\mathrm{y}$

(g) $\hat{\varepsilon}'\hat{\mathrm{y}} = 0$

(h) $\hat{\varepsilon}'\mathrm{X} = 0'$

5.2: (a) In the proof of Theorem 5.2(ii), verify that the maximum value of $h_{ii} - h_{ii}^2$ is 14

(b) Prove Theorem 5.2(iii). (c) Prove Theorem 5.2(iv).

5.3: Show that an alternative expression for h_{ii} in Theorem 5.2(iii) is the following:

$$h_{ii} = \frac{1}{n} + (\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)'(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)\sum_{r=1}^{k}\frac{1}{\lambda_r}\cos^2\theta_{ir}$$

Where θ_{ir} is the angle between $(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)$ and a_r, the *r*th eigenvector of $\mathrm{X}_c'\mathrm{X}_c$ (Cook and Weisberg 1982, p. 13). Thus h_{ii} is large if $(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)'(\mathrm{x}_{1i} - \bar{\mathrm{x}}_1)$ is large or if θ_{ir} is small for some *r*.

5.4: Show that $\frac{1}{n} \le h_{ii} + \hat{\varepsilon}_i^2 / \hat{\varepsilon}'\hat{\varepsilon} \le 1$ as in (5.24). The following steps are suggested:

(a) Let H_ be the hat matrix corresponding to the augmented matrix (**X**, **y**). Then

$$\mathrm{H}^* = (\mathrm{X}, \mathrm{y})[(\mathrm{X}, \mathrm{y})'(\mathrm{X}, \mathrm{y})]^{-1}(\mathrm{X}, \mathrm{y})'$$

$$= (\mathrm{X}, \mathrm{y})\begin{pmatrix} \mathrm{X}'\mathrm{X} & \mathrm{X}'\mathrm{y} \\ \mathrm{y}'\mathrm{X} & \mathrm{y}'\mathrm{y} \end{pmatrix}^{-1}\begin{pmatrix} \mathrm{X}' \\ \mathrm{y}' \end{pmatrix}$$

Use the inverse of a partitioned matrix with $\mathrm{a}_{11} = \mathrm{X}'\mathrm{X}$, $\mathrm{a}_{12} = \mathrm{X}'\mathrm{y}$, and $\mathrm{a}_{22} = \mathrm{y}'\mathrm{y}$ to obtain

$$\mathrm{H}^* = \mathrm{H} + \frac{1}{b}[\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}\mathrm{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \mathrm{y}\mathrm{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$$
$$- \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}\mathrm{y}' + \mathrm{y}\mathrm{y}']$$

$$\mathrm{H}^* = \mathrm{H} + \frac{1}{b}[\mathrm{H}\mathrm{y}\mathrm{y}'\mathrm{H} - \mathrm{y}\mathrm{y}'\mathrm{H} - \mathrm{H}\mathrm{y}\mathrm{y}' + \mathrm{y}\mathrm{y}']$$

Where $b = \mathrm{y}'\mathrm{y} - \mathrm{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$.

(b) Show that the above expression factors into

$$\mathrm{H}^* = \mathrm{H} + \frac{(\mathrm{I} - \mathrm{H})\mathrm{y}\mathrm{y}'(\mathrm{I} - \mathrm{H})}{\mathrm{y}'(\mathrm{I} - \mathrm{H})\mathrm{y}} = \mathrm{H} + \frac{\hat{\varepsilon}\hat{\varepsilon}'}{\hat{\varepsilon}'\hat{\varepsilon}}$$

Which gives $h_{ii}^{*} = h_{ii} + \hat{\varepsilon}_i^2 / \hat{\varepsilon}'\hat{\varepsilon}$.

(c) The proof is easily completed by noting that H^{*} is a hat matrix and therefore $(1/n) \le h_{ii}^{*} \le 1$ by Theorem 5.2(i).

5.5: Show that $\hat{\beta}_{(i)} = \hat{\beta} - \hat{\varepsilon}_i (\mathrm{X}'\mathrm{X})^{-1} \mathrm{x}_i / (1 - h_{ii})$ as in (5.29). The following steps are suggested:

(a) Show that $\mathrm{X}'\mathrm{X} = \mathrm{X}'_{(i)} \mathrm{X}_{(i)} + \mathrm{x}_i \mathrm{x}'_i$ and that $\mathrm{X}'\mathrm{y} = \mathrm{X}'_{(i)} \mathrm{y}_{(i)} + \mathrm{x}_i\, y_i$.

(b) Show that $(\mathrm{X}'\mathrm{X})^{-1} \mathrm{X}'_{(i)} \mathrm{y}_{(i)} = \hat{\beta} - (\mathrm{X}'\mathrm{X})^{-1} \mathrm{x}_i\, y_i$.

(c) Using the following adaptation of equation:

$$(\mathrm{B} - \mathrm{cc}')^{-1} = \mathrm{B}^{-1} + \frac{\mathrm{B}^{-1}\mathrm{cc}'\mathrm{B}^{-1}}{1 - \mathrm{c}'\mathrm{B}^{-1}\mathrm{c}}$$

show that

$$\hat{\beta}_{(i)} = [(\mathrm{X}'\mathrm{X})^{-1} + \frac{(\mathrm{X}'\mathrm{X})^{-1} \mathrm{x}_i \mathrm{x}'_i (\mathrm{X}'\mathrm{X})^{-1}}{1 - h_{ii}}] \mathrm{X}'_{(i)} \mathrm{y}_{(i)}$$

(d) Using the result of parts (b) and (c), show that

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{\hat{\varepsilon}_i}{1 - h_{ii}} (\mathrm{X}'\mathrm{X})^{-1} \mathrm{x}_i$$

5.6: Show that $\hat{\varepsilon}_{(i)} = \hat{\varepsilon}_i / (1 - h_{ii})$ as in (5.30).

5.7: Show that $t_i = \hat{\varepsilon}_{(i)} \big/ \sqrt{\hat{Var}(\hat{\varepsilon}_{(i)})}$ in (5.31) is the same as $t_i = \hat{\varepsilon}_i \big/ S_{(i)} \sqrt{(1 - h_{ii})}$ in (5.26). The following steps are suggested:

(a) Using $\hat{\varepsilon}_{(i)} = \hat{\varepsilon}_i / (1 - h_{ii})$ in (5.30), show that $Var(\hat{\varepsilon}_{(i)}) = \sigma^2 / (1 - h_{ii})$.

(b) If $Var(\hat{\varepsilon}_{(i)})$ in part (a) is estimated by $\hat{Var}(\hat{\varepsilon}_{(i)}) = S_{(i)}^2 / (1 - h_{ii})$, show that $\hat{\varepsilon}_{(i)} \big/ \sqrt{\hat{Var}(\varepsilon_{(i)})} = \hat{\varepsilon}_i \big/ S_{(i)} \sqrt{(1 - h_{ii})}$.

5.8: Show that $SSE_{(i)} = \mathrm{y}'_{(i)} \mathrm{y}_{(i)} - \mathrm{y}'_{(i)} \mathrm{X}_{(i)} \hat{\beta}_{(i)}$ can be written in the form

$$SSE_{(i)} = SSE - \hat{\varepsilon}_i^2 / (1 - h_{ii})$$

as in (5.32). One way to do this is as follows:

(a) Show that $\mathrm{y}'_{(i)} \mathrm{y}_{(i)} = \mathrm{y}'\mathrm{y} - y_i^2$.

(b) Using Problem 5.5a,d, we have

$$\mathrm{y}'_{(i)}\mathrm{X}_{(i)}\hat{\beta}_{(i)} = (\mathrm{y}'\mathrm{X} - y_i\,\mathrm{x}'_i)\left[\hat{\beta} - \frac{\hat{\varepsilon}_i}{1-h_{ii}}(\mathrm{X}'\mathrm{X})^{-1}\,\mathrm{x}_i\right]$$

Show that this can be written as

$$\mathrm{y}'_{(i)}\mathrm{X}_{(i)}\hat{\beta}_{(i)} = \mathrm{y}'\mathrm{X}\hat{\beta} - y_i^2 + \frac{\hat{\varepsilon}_i^2}{1-h_{ii}}$$

(c) Show that

$$SSE_{(i)} = SSE - \hat{\varepsilon}_i^2/(1-h_{ii})$$

5.9: Show that $D_i = r_i^2 h_{ii}/(k+1)(1-h_{ii})$ in (5.37) is the same as D_i in (5.35). This may be done by substituting (5.29) into (5.35).

5.10: For the gas vapor data in Table 3.3, compute the diagnostic measures $\hat{y}_i$, $\hat{\varepsilon}_i$, h_{ii}, r_i, t_i and D_i. Display these in a table similar to Table 5.1. Are there outliers or potentially influential observations? Calculate **PRESS** and compare to *SSE*.

5.11: For the land rent data in Table 3.5, compute the diagnostic measures $\hat{y}_i$, $\hat{\varepsilon}_i$, h_{ii}, r_i, t_i and D_i. Display these in a table similar to Table 5.1. Are there outliers or potentially influential observations? Calculate **PRESS** and compare to *SSE*.

5.12: For the chemical reaction data of Table 3.4 with dependent variable y_2, compute the diagnostic measures $\hat{y}_i$, $\hat{\varepsilon}_i$, h_{ii}, r_i, t_i and D_i. Display these in a table similar to Table 5.1. Are there outliers or potentially influential observations? Calculate **PRESS** and compare to *SSE*.

Chapter Six

Multiple Regression: Random x's

6: Introduction

Throughout Chapters 3–5 we assumed that the x variables were fixed; that is, that they remain constant in repeated sampling. However, in many regression applications, they are random variables. In this chapter we obtain estimators and test statistics for a regression model with random x variables. Many of these estimators and test statistics are the same as those for fixed x's, but their properties are somewhat different.

In the random-x case, $k+1$ variables $y, x_1, x_2, \cdots, x_k$ are measured on each of the n subjects or experimental units in the sample. These n observation vectors yield the data

$$\begin{matrix} y_1 & x_{11} & x_{12} & \cdots & x_{1k} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{nk} \end{matrix} \tag{6.1}$$

The rows of this array are random vectors of the second type. The variables $y, x_1, x_2, \cdots, x_k$ in a row are typically correlated and have different variances; that is, for the random vector $(y, x_1, x_2, \cdots, x_k) = (y, \mathrm{x}')$, we have

$$Cov\begin{pmatrix} y \\ x_1 \\ \vdots \\ x_k \end{pmatrix} = Cov\begin{pmatrix} y \\ \mathrm{x} \end{pmatrix} = \Sigma$$

where Σ is not a diagonal matrix. The vectors themselves [rows of the array in (6.1)] are ordinarily mutually independent (uncorrelated) if they arise from a random sample.

In Sections 6.1–6.5 we assume that y and the x variables have a multivariate normal distribution. Many of the results in Sections 6.6–6.8 do not require a normality assumption.

6.1: Multivariate Normal Regression Model

The estimation and testing results in Sections 6.1–6.5 are based on the assumption that $(y, x_1, x_2, \cdots, x_k) = (y, \mathrm{x}')$ is distributed as $N_{k+1}(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} \mu_y \\ \hline \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \tag{6.2}$$

$$\Sigma = \left(\begin{array}{c|ccc} \sigma_{yy} & \sigma_{y1} & \cdots & \sigma_{yk} \\ \hline \sigma_{1y} & \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \vdots & \ddots & \\ \sigma_{ky} & \sigma_{k1} & \cdots & \sigma_{kk} \end{array}\right) = \begin{pmatrix} \sigma_{yy} & \sigma'_{yx} \\ \sigma_{yx} & \Sigma_{xx} \end{pmatrix} \tag{6.3}$$

where μ_x is the mean vector for the x's, σ_{yx} is the vector of covariances between y and the x's, and Σ_{xx} is the covariance matrix for the x's.

We have

$$E(y \mid \mathrm{x}) = \mu_y + \sigma'_{yx}\Sigma_{xx}^{-1}(\mathrm{x} - \mu_x) \tag{6.4}$$

$$= \beta_0 + \beta_1' \, \mathrm{x} \tag{6.5}$$

Where

$$\beta_0 = \mu_y - \sigma'_{yx}\Sigma_{xx}^{-1}\mu_x \tag{6.6}$$

$$\beta_1 = \Sigma_{xx}^{-1}\sigma_{yx} \tag{6.7}$$

We also obtain

$$Var(y \mid \mathrm{x}) = \sigma_{yy} - \sigma'_{yx}\Sigma_{xx}^{-1}\sigma_{yx} = \sigma^2 \tag{6.8}$$

The mean, $E(y \mid \mathrm{x}) = \mu_y + \sigma'_{yx}\Sigma_{xx}^{-1}(\mathrm{x} - \mu_x)$, is a linear function of **x**, but the variance, $\sigma^2 = \sigma_{yy} - \sigma'_{yx}\Sigma_{xx}^{-1}\sigma_{yx}$, is not a function **x**. Thus under the multivariate normal assumption, (6.4) and (6.8) provide a linear model with constant variance, which is analogous to the fixed-x case. Note, however, that $E(y \mid \mathrm{x}) = \beta_0 + \beta_1'\mathrm{x}$ in (6.5) does not allow for curvature such as $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$. Thus $E(y \mid \mathrm{x}) = \beta_0 + \beta_1'\mathrm{x}$ represents a model that is linear in the x's as well as the β 's. This differs from the linear model in the fixed-x case, which requires only linearity in the β 's.

6.2: Estimation and testing in Multivariate Normal Regression

Before obtaining estimators of β_0, β_1, and σ^2 in (6.6)–(6.8), we must first estimate μ and Σ. Maximum likelihood estimators of μ and Σ are given in the following theorem.

Theorem 6.2a: If (y_1, x_1'), (y_2, x_2'), . . . , (y_n, x_n') [rows of the array in (6.1)] is a random sample from $N_{k+1}(\mu, \Sigma)$, with μ and Σ as given in (6.2) and (6.3), the maximum likelihood estimators are

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_y \\ \hat{\mu}_x \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \bar{\mathrm{x}} \end{pmatrix} \qquad (6.9)$$

$$\hat{\Sigma} = \frac{n-1}{n} S = \frac{n-1}{n} \begin{pmatrix} S_{yy} & S_{yx}' \\ S_{yx} & S_{xx} \end{pmatrix} \qquad (6.10)$$

where the partitioning of $\hat{\mu}$ and S is analogous to the partitioning of μ and Σ in (6.2) and (6.3). The elements of the sample covariance matrix S are defined in (3.40) and in (6.14).

Proof: Denote (y_i, x_i') by v_i', $i = 1, 2, \ldots, n$. As noted below (6.1), v_1, $\mathrm{v}_2, \ldots, \mathrm{v}_n$ are independent because they arise from a random sample. The likelihood function (joint density) is therefore given by the product

$$\begin{aligned} L(\mu, \Sigma) &= \prod_{i=1}^{n} f(\mathrm{v}_i; \mu, \Sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\left(\sqrt{2\pi}\right)^{(k+1)} |\Sigma|^{1/2}} e^{-(\mathrm{v}_i - \mu)' \Sigma^{-1} (\mathrm{v}_i - \mu)/2} \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^{n(k+1)} |\Sigma|^{n/2}} e^{-\sum_{i=1}^{n} (\mathrm{v}_i - \mu)' \Sigma^{-1} (\mathrm{v}_i - \mu)/2} \qquad (6.11) \end{aligned}$$

Note that $L(\mu, \Sigma) = \prod_{i=1}^{n} f(\mathrm{v}_i; \mu, \Sigma)$ is a product of n multivariate normal densities, each involving $k + 1$ random variables. Thus there are $n(k+1)$ random variables as compared to the likelihood $L(\beta, \sigma^2)$ in (3.50) that involves n random variables $y_1, y_2, \cdots, y_n$ [the x's are fixed in (3.50)].

To find the maximum likelihood estimator for μ, we expand and sum the exponent in (6.11) and then take the logarithm to obtain

$$\ln L(\mu,\Sigma) = -n(k+1)\ln\sqrt{2\pi} - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n} \mathrm{v}_i' \Sigma^{-1} \mathrm{v}_i$$
$$+ \mu'\Sigma^{-1}\sum_{i=1}^{n} \mathrm{v}_i - \frac{n}{2}\mu'\Sigma^{-1}\mu \qquad (6.12)$$

Differentiating (6.12) with respect to μ and setting the result equal to 0, we obtain

$$\frac{\partial \ln L(\mu,\Sigma)}{\partial\mu} = -0 - 0 - 0 + \Sigma^{-1}\sum_{i=1}^{n} \mathrm{v}_i - \frac{2n}{2}\Sigma^{-1}\mu = 0$$

which gives

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{v}_i = \overline{\mathrm{v}} = \begin{pmatrix} \overline{y} \\ \overline{\mathrm{x}} \end{pmatrix}$$

where $\overline{\mathrm{x}} = (\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_k)'$ is the vector of sample means of the x's. To find the maximum likelihood estimator of Σ, we rewrite the exponent of (6.11) and then take the logarithm to obtain

$$\ln L(\mu,\Sigma^{-1}) = -n(k+1)\ln\sqrt{2\pi} + \frac{n}{2}\ln|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{n} (\mathrm{v}_i - \overline{\mathrm{v}})' \Sigma^{-1} (\mathrm{v}_i - \overline{\mathrm{v}})$$
$$- \frac{n}{2}(\overline{\mathrm{v}} - \mu)' \Sigma^{-1} (\overline{\mathrm{v}} - \mu)$$
$$= -n(k+1)\ln\sqrt{2\pi} + \frac{n}{2}\ln|\Sigma^{-1}| - \frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^{n} (\mathrm{v}_i - \overline{\mathrm{v}})(\mathrm{v}_i - \overline{\mathrm{v}})'\right]$$
$$- \frac{n}{2}\mathrm{tr}\left[\Sigma^{-1}(\overline{\mathrm{v}} - \mu)(\overline{\mathrm{v}} - \mu)'\right]$$

Differentiating this with respect to Σ^{-1}, and setting the result equal to 0, we obtain

$$\frac{\partial \ln L(\mu,\Sigma^{-1})}{\partial\Sigma^{-1}} = n\Sigma - \frac{n}{2}\mathrm{diag}(\Sigma) - \sum_{i=1}^{n} (\mathrm{v}_i - \overline{\mathrm{v}})(\mathrm{v}_i - \overline{\mathrm{v}})' + \frac{1}{2}\mathrm{diag}\left[\sum_{i=1}^{n} (\mathrm{v}_i - \overline{\mathrm{v}})(\mathrm{v}_i - \overline{\mathrm{v}})'\right]$$
$$- n(\overline{\mathrm{v}} - \mu)(\overline{\mathrm{v}} - \mu)' + \frac{n}{2}\mathrm{diag}\left[(\overline{\mathrm{v}} - \mu)(\overline{\mathrm{v}} - \mu)'\right] = 0$$

Since $\hat{\mu} = \overline{\mathrm{v}}$, the last two terms disappear and we obtain

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{v}_i - \bar{\mathrm{v}})(\mathrm{v}_i - \bar{\mathrm{v}})' = \frac{n-1}{n}S \tag{6.13}$$

See Problem 6.1 for verification that $\sum_{i=1}^{n}(\mathrm{v}_i - \bar{\mathrm{v}})(\mathrm{v}_i - \bar{\mathrm{v}})' = (n-1)S$.

In partitioned form, the sample covariance matrix S can be written as in (6.10)

$$S = \begin{pmatrix} S_{yy} & S'_{yx} \\ S_{yx} & S_{xx} \end{pmatrix} = \left(\begin{array}{c|cccc} S_{yy} & S_{y1} & \cdots & S_{yk} \\ \hline S_{1y} & S_{11} & \cdots & S_{1k} \\ \vdots & \vdots & \ddots & \\ S_{ky} & S_{k1} & \cdots & S_{kk} \end{array}\right) \tag{6.14}$$

where S_{yx} is the vector of sample covariances between y and the x's and S_{xx} is the sample covariance matrix for the x's. For example

$$S_{y1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{i1} - \bar{x}_1)}{n-1}$$

$$S_{11} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}{n-1}$$

$$S_{12} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{n-1}$$

[see (3.41)–(3.43)]. $E(S_{yy}) = \sigma_{yy}$ and $E(S_{jj}) = \sigma_{jj}$. $E(S_{yj}) = \sigma_{yj}$ and $E(S_{ij}) = \sigma_{ij}$. Thus $E(S) = \Sigma$, where Σ is given in (6.3). The maximum likelihood estimator $\hat{\Sigma} = (n-1)S/n$ is therefore biased.

In order to find maximum likelihood estimators of β_0, β_1, and σ^2 we first note the *invariance property* of maximum likelihood estimators.

Theorem 6.2b: The maximum likelihood estimator of a function of one or more parameters is the same function of the corresponding estimators; that is, if $\hat{\theta}$ is the maximum likelihood estimator of the vector or matrix of parameters θ, then $g(\hat{\theta})$ is the maximum likelihood estimator of $g(\theta)$.

Proof: See Hogg and Craig (1995, p. 265).

Example 6.2: We illustrate the use of the invariance property in Theorem 6.2b by showing that the sample correlation matrix **R** is the maximum likelihood estimator of the population correlation matrix P_ρ when sampling from the multivariate normal distribution. The relationship between P_ρ and Σ is given by $\mathrm{P}_\rho = \mathrm{D}_\sigma^{-1}\Sigma\mathrm{D}_\sigma^{-1}$, where $\mathrm{D}_\sigma = [\mathrm{diag}(\Sigma)]^{1/2}$, so that

$$\mathrm{D}_\sigma^{-1} = \mathrm{diag}\left(\frac{1}{\sqrt{\sigma_{11}}}, \frac{1}{\sqrt{\sigma_{22}}}, \cdots, \frac{1}{\sqrt{\sigma_{pp}}}\right)$$

The maximum likelihood estimator of $1/\sqrt{\sigma_{jj}}$ is $1/\sqrt{\hat{\sigma}_{jj}}$, where $\hat{\sigma}_{jj} = (1/n)\sum_{i=1}^n (y_{ij} - \bar{y}_j)^2$. Thus $\hat{\mathrm{D}}_\sigma^{-1} = \mathrm{diag}\left(1/\sqrt{\hat{\sigma}_{11}}, 1/\sqrt{\hat{\sigma}_{22}}, \cdots, 1/\sqrt{\hat{\sigma}_{pp}}\right)$, and we obtain

$$\hat{\mathrm{P}}_\rho = \hat{\mathrm{D}}_\sigma^{-1}\hat{\Sigma}\hat{\mathrm{D}}_\sigma^{-1} = \left(\frac{\hat{\sigma}_{jk}}{\sqrt{\hat{\sigma}_{jj}}\sqrt{\hat{\sigma}_{kk}}}\right)$$

$$= \left(\frac{\sum_{i=1}^n (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)/n}{\sqrt{\sum_{i=1}^n (y_{ij} - \bar{y}_j)^2/n}\sqrt{\sum_{i=1}^n (y_{ik} - \bar{y}_k)^2/n}}\right)$$

$$= \left(\frac{\sum_{i=1}^n (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{\sqrt{\sum_{i=1}^n (y_{ij} - \bar{y}_j)^2}\sqrt{\sum_{i=1}^n (y_{ik} - \bar{y}_k)^2}}\right)$$

$$= (r_{jk}) = \mathrm{R}$$

Maximum likelihood estimators of β_0, $\mathrm{\beta}_1$, and σ^2 are now given in the following theorem.

Theorem 6.2c: If (y_1, x_1'), (y_2, x_2'), . . . , (y_n, x_n'), is a random sample from $N_{k+1}(\mu, \Sigma)$, where μ and Σ are given by (6.2) and (6.3), the maximum likelihood estimators for β_0, $\mathrm{\beta}_1$, and σ^2 in (6.6)–(6.8) are as follows:

$$\hat{\beta}_0 = \bar{y} - S_{yx}' S_{xx}^{-1}\bar{\mathrm{x}} \tag{6.15}$$

$$\hat{\beta}_1 = S_{xx}^{-1} S_{yx} \qquad (6.16)$$

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2 \qquad where \qquad S^2 = S_{yy} - S'_{yx} S_{xx}^{-1} S_{yx} \qquad (6.17)$$

The estimator S^2 is a bias-corrected estimator of σ^2.

Proof: By the invariance property of maximum likelihood estimators (Theorem 6.2b), we insert (6.9) and (6.10) into (6.6), (6.7), and (6.8) to obtain the desired results (using the unbiased estimator S in place of $\hat{\Sigma}$).

The estimators $\hat{\beta}_0$, $\hat{\beta}_1$, and S^2 have a minimum variance property analogous to that of the corresponding estimators for the case of normal y's and fixed x's in Theorem 3.6d. It can be shown that $\hat{\mu}$ and S in (6.9) and (6.10) are jointly sufficient for μ and Σ (see Problem 6.2). Then, with some additional properties that can be demonstrated, it follows that $\hat{\beta}_0$, $\hat{\beta}_1$, and S^2 are minimum variance unbiased estimators for β_0, β_1, and σ^2 (Graybill 1976, p. 380).

The maximum likelihood estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in (6.15) and (6.16) are the same algebraic functions of the observations as the least-squares estimators given in (3.47) and (3.46) for the fixed-*x* case. The estimators in (6.15) and (6.16) are also identical to the maximum likelihood estimators for normal *y*'s and fixed *x*'s in Section 3.6.2 (see Problem 3.17). However, even though the estimators in the random-*x* case and fixed-*x* case are the same, their distributions differ. When *y* and the *x*'s are multivariate normal, $\hat{\beta}_1$ does not have a multivariate normal distribution as it does in the fixed-*x* case with normal *y*'s [Theorem 3.6b(i)]. For large *n*, the distribution is similar to the multivariate normal, but for small *n*, the distribution has heavier tails than the multivariate normal.

In spite of the non-normality of $\hat{\beta}_1$ in the random-*x* model, the *F* tests and *t* tests and associated confidence regions and intervals of Chapter 4 (fixed-*x* model) are still appropriate. To see this, note that since the conditional distribution of *y* for a given value of **x** is normal, the conditional distribution of the vector of observations $y = (y_1, y_2, \cdots, y_n)'$ for a given value of the **X** matrix is multivariate normal. Therefore, a test statistic such as (4.35) is distributed conditionally as an *F* for the given value of **X** when H_0 is true. However, the central *F* distribution depends only on degrees of freedom; it does not depend on **X**. Thus under H_0, the statistic has (unconditionally) an *F* distribution for all values of **X**, and so tests can be carried out exactly as in the fixed-*x* case.

The main difference is that when H_0 is false, the non-centrality parameter is a function of **X**, which is random. Hence the non-central *F* distribution does not apply to the random-*x* case. This only affects such things as power calculations.

Confidence intervals for the β_j's in Section 4.6.2 and for linear functions of the β_j's in Section 4.6.3 are based on the central *t* distribution [e.g., see (4.48)]. Thus they also remain valid for the random-*x* case. However, the expected width of the interval differs in the two cases (random *x*'s and fixed *x*'s) because of randomness in **X**.

In Section 6.5, we obtain the *F* test for $H_0 : \mathrm{β}_1 = 0$ using the likelihood ratio approach.

6.3: Standardized Regression Coefficients

We now show that the regression coefficient vector $\hat{\mathrm{β}}_1$ in (6.16) can be expressed in terms of sample correlations. By analogy to (6.14), the sample correlation matrix can be written in partitioned form as

$$\mathrm{R} = \begin{pmatrix} 1 & \mathrm{r}'_{yx} \\ \mathrm{r}_{yx} & \mathrm{R}_{xx} \end{pmatrix} = \left(\begin{array}{c|cccc} 1 & r_{y1} & r_{y2} & \cdots & r_{yk} \\ \hline r_{1y} & 1 & r_{12} & \cdots & r_{1k} \\ r_{2y} & r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{ky} & r_{k1} & r_{k2} & \cdots & 1 \end{array}\right) \tag{6.18}$$

where r_{yx} is the vector of correlations between *y* and the *x*'s and R_{xx} is the correlation matrix for the *x*'s. For example

$$r_{y2} = \frac{S_{y2}}{\sqrt{S_y^2 S_2^2}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2}}$$

$$r_{12} = \frac{S_{12}}{\sqrt{S_1^2 S_2^2}} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}}$$

R can be converted to S by

$$S = \mathrm{DRD}$$

where $\mathrm{D} = [\mathrm{diag}(S)]^{1/2}$, which can be written in partitioned form as

$$\mathrm{D} = \left(\begin{array}{c|cccc} S_y & 0 & 0 & \cdots & 0 \\ \hline 0 & \sqrt{S_{11}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{S_{22}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{S_{kk}} \end{array}\right) = \begin{pmatrix} S_y & 0' \\ 0 & \mathrm{D}_x \end{pmatrix}$$

Using the partitioned form of S in (6.14), $S = \mathrm{DRD}$ can be written as

$$S = \begin{pmatrix} S_{yy} & S'_{yx} \\ S_{yx} & S_{xx} \end{pmatrix} = \begin{pmatrix} S_y^2 & S_y\,\mathrm{r}'_{yx}\mathrm{D}_x \\ S_y\,\mathrm{D}_x\mathrm{r}_{yx} & \mathrm{D}_x\mathrm{R}_{xx}\mathrm{D}_x \end{pmatrix} \qquad (6.19)$$

So that

$$S_{xx} = \mathrm{D}_x \mathrm{R}_{xx} \mathrm{D}_x \qquad (6.20)$$

$$S_{yx} = S_y\,\mathrm{D}_x\,\mathrm{r}_{yx} \qquad (6.21)$$

where $\mathrm{D}_x = \mathrm{diag}(S_1, S_2, \cdots, S_k)$ and $S_y = \sqrt{S_y^2} = \sqrt{S_{yy}}$ is the sample standard deviation of y. When (6.20) and (6.21) are substituted into (6.16), we obtain an expression for $\hat{\beta}_1$ in terms of correlations:

$$\hat{\beta}_1 = S_y\,\mathrm{D}_x^{-1}\mathrm{R}_{xx}^{-1}\mathrm{r}_{yx} \qquad (6.22)$$

The regression coefficients $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ in $\hat{\beta}_1$can be standardized so as to show the effect of standardized x values (sometimes called *z scores*). We illustrate this for $k = 2$. The model in centered form [see (3.30) and an expression following (3.38)] is

$$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2)$$

This can be expressed in terms of standardized variables as

$$\frac{\hat{y}_i - \bar{y}}{S_y} = \frac{S_1}{S_y}\hat{\beta}_1\left(\frac{x_{i1} - \bar{x}_1}{S_1}\right) + \frac{S_2}{S_y}\hat{\beta}_2\left(\frac{x_{i2} - \bar{x}_2}{S_2}\right) \tag{6.23}$$

Where $S_j = \sqrt{S_{jj}}$ is the standard deviation of x_j. We thus define the standardized coefficients as

$$\hat{\beta}_j^* = \frac{S_j}{S_y}\hat{\beta}_j$$

These coefficients are often referred to as *beta weights* or *beta coefficients*. Since they are used with standardized variables $(x_{ij} - \bar{x}_j)/S_j$ in (6.23), the $\hat{\beta}_j^*$'s can be readily compared to each other, whereas the $\hat{\beta}_j$'s cannot be so compared. [Division by S_y in (6.23) is customary but not necessary; the relative values of $S_1\hat{\beta}_1$ and $S_2\hat{\beta}_2$ are the same as those of $S_1\hat{\beta}_1/S_y$ and $S_2\hat{\beta}_2/S_y$.]

The beta weights can be expressed in vector form as

$$\hat{\beta}_1^* = \frac{1}{S_y}\mathrm{D}_x\,\hat{\beta}_1$$

Using (6.22), this can be written as

$$\hat{\beta}_1^* = \mathrm{R}_{xx}^{-1}\mathrm{r}_{yx} \tag{6.24}$$

Note that $\hat{\beta}_1^*$ in (6.24) is not the same as $\hat{\beta}_1^*$ from the reduced model in (4.8). Note also the analogy of $\hat{\beta}_1^* = \mathrm{R}_{xx}^{-1}\mathrm{r}_{yx}$ in (6.24) to $\hat{\beta}_1 = S_{xx}^{-1}S_{yx}$ in (6.16). In effect, R_{xx} and r_{xy} are the covariance matrix and covariance vector for standardized variables.

Replacing S_{xx}^{-1} and S_{yx} by R_{xx}^{-1} and r_{xy} leads to regression coefficients for standardized variables.

Example 6.3: The following six hematology variables were measured on 51 workers (Royston 1983):

y = lymphocyte count
x_1 = hemoglobin concentration
x_2 = packed-cell volume
x_3 = white blood cell count $(\times 0.01)$
x_4 = neutrophil count
x_5 = serum lead concentration

The data are given in Table 6.1. For $\bar{y}$, $\bar{\mathrm{x}}$, S_{xx} and S_{yx}, we have

$$\bar{y} = 22.98, \ \bar{\mathrm{x}}' = \begin{pmatrix} 15.108 & 45.196 & 53.824 & 25.627 & 20.882 \end{pmatrix}$$

$$S_{xx} = \begin{pmatrix} 0.69074 & 1.49440 & 3.25540 & 0.35098 & -0.23506 \\ 1.49440 & 5.4008 & 10.155 & 1.3545 & 1.5235 \\ 3.25540 & 10.155 & 200.67 & 65.273 & 6.6788 \\ 0.35098 & 1.3545 & 65.273 & 58.158 & 1.3553 \\ -0.23506 & 1.5235 & 6.6788 & 1.3553 & 16.946 \end{pmatrix}$$

TABLE 6.1: Hematology Data

Observation Number	y	x_1	x_2	x_3	x_4	x_5
1	14	13.4	39	41	25	17
2	15	14.6	46	50	30	20
3	19	13.5	42	45	21	18
4	23	15	46	46	16	18
5	17	14.6	44	51	31	19
6	20	14	44	49	24	19
7	21	16.4	49	43	17	18
8	16	14.8	44	44	26	29
9	27	15.2	46	41	13	27
10	34	15.5	48	84	42	36
11	26	15.2	47	56	27	22
12	28	16.9	50	51	17	23
13	24	14.8	44	47	20	23
14	26	16.2	45	56	25	19
15	23	14.7	43	40	13	17
16	9	14.7	42	34	22	13
17	18	16.5	45	54	32	17
18	28	15.4	45	69	36	24
19	17	15.1	45	46	29	17
20	14	14.2	46	42	25	28
21	8	15.9	46	52	34	16
22	25	16	47	47	14	18
23	37	17.4	50	86	39	17
24	20	14.3	43	55	31	19
25	15	14.8	44	42	24	19
26	9	14.9	43	43	32	17
27	16	15.5	45	52	30	20
28	18	14.5	43	39	18	25
29	17	14.4	45	60	37	23
30	23	14.6	44	47	21	27
31	43	15.3	45	79	23	23
32	17	14.9	45	34	15	24
33	23	15.8	47	60	32	21
34	31	14.4	44	77	39	23
35	11	14.7	46	37	23	23
36	25	14.8	43	52	19	22
37	30	15.4	45	60	25	18
38	32	16.2	50	81	38	18

39	17	15	45	49	26	24
40	22	15.1	47	60	33	16
41	20	16	46	46	22	22
42	20	15.3	48	55	23	23
43	20	14.5	41	62	36	21
44	26	14.2	41	49	20	20
45	40	15	45	72	25	25
46	22	14.2	46	58	31	22
47	61	14.9	45	84	17	17
48	12	16.2	48	31	15	18
49	20	14.5	45	40	18	20
50	35	16.4	49	69	22	24
51	38	14.7	44	78	34	16

$$S_{yx} = \begin{pmatrix} 1.8782 \\ 5.6639 \\ 108.5 \\ 1.6725 \\ 5.0176 \end{pmatrix}$$

By (6.15) to (6.17), we obtain

$$\hat{\beta}_1 = S_{xx}^{-1} S_{yx} = \begin{pmatrix} -0.20772 \\ -0.29172 \\ 0.85902 \\ -0.92858 \\ 0.05515 \end{pmatrix}$$

$$\hat{\beta}_0 = \bar{y} - S'_{yx} S_{xx}^{-1}\,\bar{\mathrm{x}} = 22.98 - 7.2672 = 15.713$$

$$S^2 = S_{yy} - S'_{yx} S_{xx}^{-1} S_{yx} = 3.9376$$

$$\mathrm{R}_{xx} = \begin{pmatrix} 1 & 0.77373 & 0.27651 & 0.055376 & -0.068705 \\ 0.77373 & 1 & 0.30848 & 0.076427 & 0.15925 \\ 0.27651 & 0.30848 & 1 & 0.60421 & 0.11453 \\ 0.055376 & 0.076427 & 0.60421 & 1 & 0.043171 \\ -0.068705 & 0.15925 & 0.11453 & 0.043171 & 1 \end{pmatrix}$$

$$\mathbf{r}_{yx} = \begin{pmatrix} 0.23331 \\ 0.25162 \\ 0.79073 \\ 0.02264 \\ 0.12584 \end{pmatrix}$$

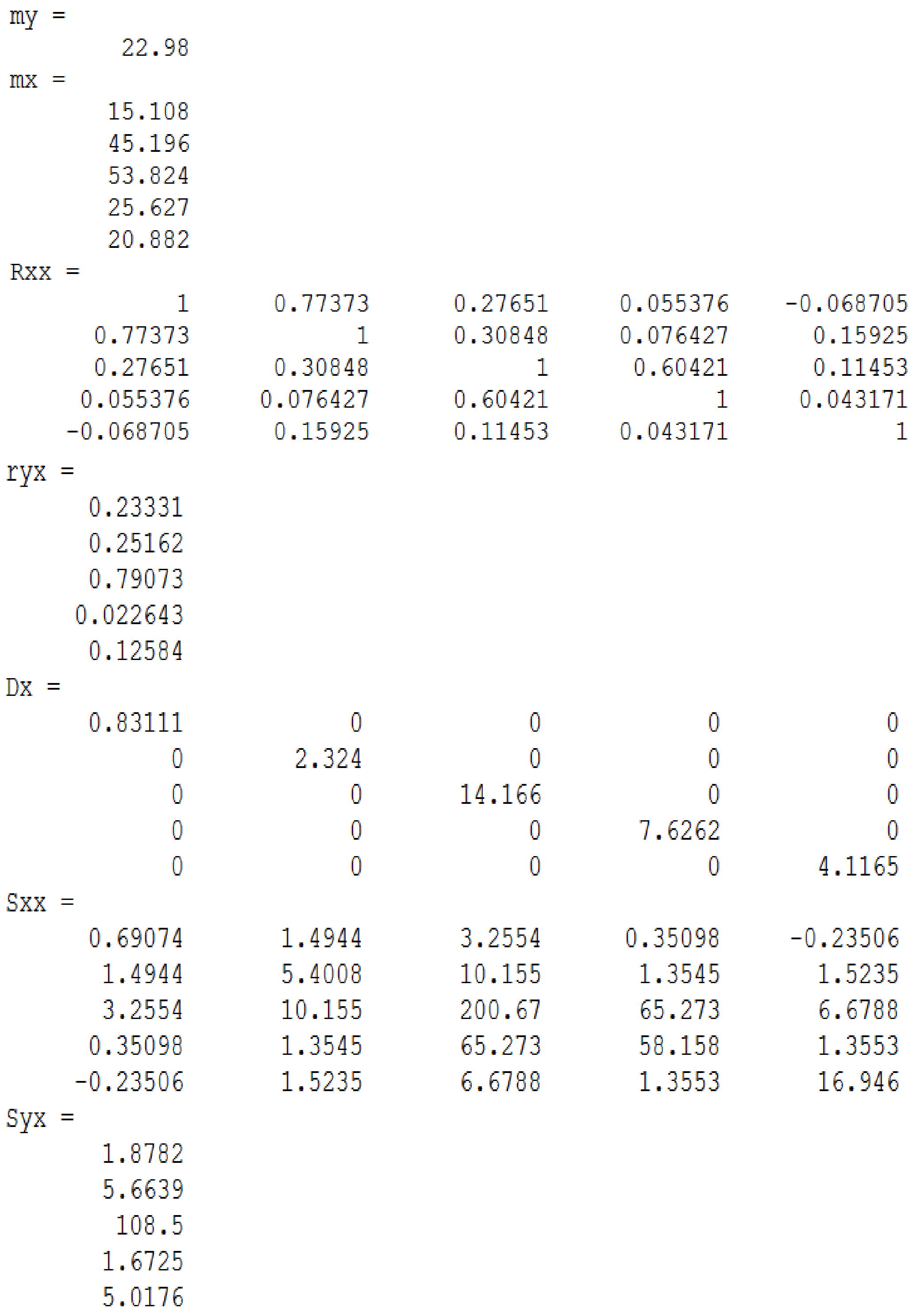
By (10.24), the standardized coefficient vector is given by

$$\hat{\beta}_1^* = \mathrm{R}_{xx}^{-1}\mathbf{r}_{yx} = \begin{pmatrix} -0.017823 \\ -0.069992 \\ 1.2563 \\ -0.73111 \\ 0.023437 \end{pmatrix}$$

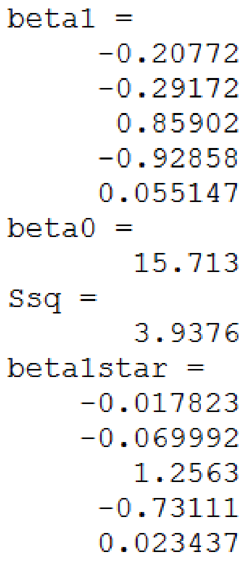
Example 6.3 [The program name ta22.m]	Applications using MATLAB

```
clc
clear all
data=[14   13.4   39  41  25  17; 15  14.6   46  50  30  20;19   13.5   42  45  21  18
   23  15  46  46  16  18; 17  14.6   44  51  31  19; 20  14  44  49  24  19; 21  16.4
49  43  17  18   16  14.8   44  44  26  29; 27  15.2   46  41  13  27;34   15.5   48
84  42  36;26   15.2   47  56  27  22   28  16.9   50  51  17  23; 24  14.8   44  47
20  23; 26  16.2   45  56  25  19; 23  14.7   43  40  13  17  9   14.7   42  34  22  13;
18  16.5   45  54  32  17; 28  15.4   45  69  36  24; 17  15.1   45  46  29  17   14
14.2   46  42  25  28; 8   15.9   46  52  34  16; 25  16  47  47  14  18; 37  17.4   50
86  39  17  20  14.3   43  55  31  19;15   14.8   44  42  24  19; 9   14.9   43  43  32
17;16   15.5   45  52  30  20    18  14.5   43  39  18  25; 17  14.4   45  60  37  23;
23  14.6   44  47  21  27; 43  15.3   45  79  23  23  17  14.9   45  34  15  24; 23
15.8   47  60  32  21; 31  14.4   44  77  39  23;11   14.7   46  37  23  23    25
14.8   43  52  19  22; 30  15.4   45  60  25  18; 32  16.2   50  81  38  18; 17  15  45
49  26  24  22  15.1   47  60  33  16; 20  16  46  46  22  22;20   15.3   48  55  23
23; 20  14.5   41  62  36  21;    26  14.2   41  49  20  20; 40  15  45  72  25  25; 22
14.2   46  58  31  22; 61  14.9   45  84  17  17  12  16.2   48  31  15  18;20   14.5
45  40  18  20;35   16.4   49  69  22  24;38   14.7   44  78  34  16];
y=data(:,1);x1=data(:,2);x2=data(:,3);x3=data(:,4);
x4=data(:,5);x5=data(:,6);n=length(x1);my=mean(y)
x=[x1 x2 x3 x4 x5];mx=mean(x)',Rxx=corr(x)
ryx=corr(y,x)',Dx=[diag(diag(cov(x)))].^.5,
Sxx=Dx*Rxx*Dx,Syx=std(y)*Dx*ryx,beta1=Sxx^(-1)*Syx
beta0=my-Syx'*Sxx^(-1)*mx,Ssq=var(y)-Syx'*Sxx^(-1)*Syx
beta1star=inv(Rxx)*ryx
```

Ans.



```
my =
         22.98
mx =
        15.108
        45.196
        53.824
        25.627
        20.882
Rxx =
             1       0.77373       0.27651      0.055376     -0.068705
       0.77373             1       0.30848      0.076427       0.15925
       0.27651       0.30848             1       0.60421       0.11453
      0.055376      0.076427       0.60421             1      0.043171
     -0.068705       0.15925       0.11453      0.043171             1
ryx =
       0.23331
       0.25162
       0.79073
      0.022643
       0.12584
Dx =
       0.83111             0             0             0             0
             0         2.324             0             0             0
             0             0        14.166             0             0
             0             0             0        7.6262             0
             0             0             0             0        4.1165
Sxx =
       0.69074        1.4944        3.2554       0.35098      -0.23506
        1.4944        5.4008        10.155        1.3545        1.5235
        3.2554        10.155        200.67        65.273        6.6788
       0.35098        1.3545        65.273        58.158        1.3553
      -0.23506        1.5235        6.6788        1.3553        16.946
Syx =
        1.8782
        5.6639
         108.5
        1.6725
        5.0176
```



```
beta1 =
     -0.20772
     -0.29172
      0.85902
     -0.92858
     0.055147
beta0 =
       15.713
Ssq =
       3.9376
beta1star =
    -0.017823
    -0.069992
       1.2563
     -0.73111
     0.023437
```

6.4: R^2 in multivariate Normal Regression

In the case of fixed x's, we defined R^2 as the proportion of variation in y due to regression [see (3.55)]. In the case of random x's, we obtain R as an estimate of a population multiple correlation between y and the x's. Then R^2 is the square of this sample multiple correlation.

The *population multiple correlation coefficient* $\rho_{y|\mathrm{x}}$ is defined as the correlation between y and the linear function $w = \mu_y + \sigma'_{yx}\Sigma_{xx}^{-1}(\mathrm{x} - \mu_x)$:

$$\rho_{y|x} = Corr(y, w) = \frac{\sigma_{yw}}{\sigma_y \sigma_w} \tag{6.25}$$

(We use the subscript $y|\mathrm{x}$ to distinguish $\rho_{y|\mathrm{x}}$ from ρ, the correlation between y and x in the bivariate normal case; see Sections 2.4, and 6.5). By (6.4), w is equal to $E(y|\mathrm{x})$, which is the population analogue of $\hat{y} = \hat{\beta}_0 + \hat{\beta}'_1\,\mathrm{x}_1$, the sample predicted value of y. As **x** varies randomly, the *population predicted value* $w = \mu_y + \sigma'_{yx}\Sigma_{xx}^{-1}(\mathrm{x} - \mu_x)$ becomes a random variable.

It is easily established that $Cov(y, w)$ and $Var(w)$ have the same value:

$$Cov(y, w) = Var(w) = \sigma'_{yx}\Sigma_{xx}^{-1}\sigma_{yx} \tag{6.26}$$

Then the population multiple correlation $\rho_{y|\mathrm{x}}$ in (6.25) becomes

$$\rho_{y|\mathrm{x}} = \frac{Cov(y,w)}{\sqrt{Var(y)Var(w)}} = \sqrt{\frac{\sigma'_{yx}\Sigma_{xx}^{-1}\sigma_{yx}}{\sigma_{yy}}}$$

And the *population coefficient of determination* or *population squared multiple correlation* $\rho^2_{y|\mathrm{x}}$ is given by

$$\rho^2_{y|\mathrm{x}} = \frac{\sigma'_{yx}\Sigma_{xx}^{-1}\sigma_{yx}}{\sigma_{yy}} \tag{6.27}$$

We now list some properties of $\rho_{y|\mathrm{x}}$ and $\rho^2_{y|\mathrm{x}}$.

1. $\rho_{y|\mathrm{x}}$ is the maximum correlation between *y* and any linear function of **x**:

   $$\rho_{y|\mathrm{x}} = \max_{\alpha} \rho_{y}, \alpha'\mathrm{x} \tag{6.28}$$

   This is an alternative definition of $\rho_{y|\mathrm{x}}$ that is not based on the multivariate normal distribution as is the definition in (6.25).

2. $\rho^2_{y|\mathrm{x}}$ can be expressed in terms of determinants:

   $$\rho^2_{y|\mathrm{x}} = 1 - \frac{|\Sigma|}{\sigma_{yy}|\Sigma_{xx}|} \tag{6.29}$$

   where Σ and Σ_{xx} are defined in (6.3).

3. $\rho^2_{y|\mathrm{x}}$ is invariant to linear transformations on *y* or on the *x*'s; that is, if *u* = *ay* and **v = Bx**, where **B** is nonsingular, then

   $$\rho^2_{u|\mathrm{v}} = \rho^2_{y|\mathrm{x}} \tag{6.30}$$

   (Note that **v** here is not the same as v_i used in the proof of Theorem 6.2a.)

4. Using $Var(w) = \sigma'_{yx}\Sigma_{xx}^{-1}\sigma_{yx}$ in (6.26), $\rho^2_{y|\mathrm{x}}$ in (6.27) can be written in the form

   $$\rho^2_{y|\mathrm{x}} = \frac{Var(w)}{Var(y)} \tag{6.31}$$

   Since $w = \mu_y + \sigma'_{yx}\Sigma_{xx}^{-1}(\mathrm{x} - \mu_x)$ is the population regression equation, $\rho^2_{y|\mathrm{x}}$ in (6.31) represents the proportion of the variance of *y* that can be attributed to the regression relationship with the variables in **x**. In this sense, $\rho^2_{y|\mathrm{x}}$ is analogous to R^2 in the fixed-*x* case in (3.55).

5. By (6.8) and (6.27), $Var(y|\mathrm{x})$ can be expressed in terms of $\rho^2_{y|\mathrm{x}}$:

$$\left.\begin{aligned} Var(y|\mathrm{x}) &= \sigma_{yy} - \sigma'_{yx}\Sigma^{-1}_{xx}\sigma_{yx} = \sigma_{yy} - \sigma_{yy}\rho^2_{y|\mathrm{x}} \\ &= \sigma_{yy}\left(1-\rho^2_{y|\mathrm{x}}\right) \end{aligned}\right\} \tag{6.32}$$

6. If we consider $y-w$ as a residual or error term, then $y-w$ is uncorrelated with the *x*'s

$$Cov(y-w,\mathrm{x}) = 0' \tag{6.33}$$

(see Problem 6.8).

We can obtain a maximum likelihood estimator for $\rho^2_{y|\mathrm{x}}$ by substituting estimators from (6.14) for the parameters in (6.27):

$$R^2 = \frac{S'_{yx}S^{-1}_{xx}S_{yx}}{S_{yy}} \tag{6.34}$$

We use the notation R^2 rather than $\hat{\rho}^2_{y|\mathrm{x}}$ because (6.34) is recognized as having the same form as R^2 for the fixed-*x* case in (3.59).We refer to R^2 as the sample coefficient of determination or as the sample squared multiple correlation. The square root of R^2

$$R = \sqrt{\frac{S'_{yx}S^{-1}_{xx}S_{yx}}{S_{yy}}} \tag{6.35}$$

is the *sample multiple correlation coefficient*.

We now list several properties of R and R^2, some of which are analogous to properties of $\rho^2_{y|\mathrm{x}}$ above.

1. *R* is equal to the correlation between *y* and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'_1\mathrm{x}$:

$$R = r_{y\hat{y}} \tag{6.36}$$

2. *R* is equal to the maximum correlation between *y* and any linear combination of the *x*'s, $\mathrm{a}'\mathrm{x}$:

$$R = \max_{\mathrm{a}} r_{y,\mathrm{a}'\mathrm{x}} \tag{6.37}$$

3. R^2 can be expressed in terms of correlations:

$$R^2 = \mathrm{r}'_{yx}\mathrm{R}_{xx}^{-1}\mathrm{r}_{yx} \tag{6.38}$$

Where r_{yx} and R_{xx} are from the sample correlation matrix R partitioned as in (6.18).

4. R^2 can be obtained from R^{-1}:

$$R^2 = 1 - \frac{1}{r^{yy}} \tag{6.39}$$

Where r^{yy} is the first diagonal element of R^{-1}. Using the other diagonal elements of R^{-1}, this relationship can be extended to give the multiple correlation of any x_j with the other x's and y. Thus from R^{-1} we obtain multiple correlations, as opposed to the simple correlations in **R**.

5. R^2 can be expressed in terms of determinants:

$$R^2 = 1 - \frac{|S|}{S_{yy}|S_{xx}|} \tag{6.40}$$

$$R^2 = 1 - \frac{|\mathrm{R}|}{|\mathrm{R}_{xx}|} \tag{6.41}$$

Where S_{xx} and R_{xx} are defined in (6.14) and (6.18).

6. From (6.24) and (6.38), we can express R^2 in terms of beta weights:

$$R^2 = \mathrm{r}'_{yx}\hat{\beta}_1^* \tag{6.42}$$

Where $\hat{\beta}_1^* = \mathrm{R}_{xx}^{-1}\mathrm{r}_{yx}$. This equation does not imply that R^2 is the sum of squared partial correlations (Section 6.8).

7. If $\rho^2_{y|\mathrm{x}} = 0$, the expected value of R^2 is given by

$$E(R^2) = \frac{k}{n-1} \tag{6.43}$$

Thus R^2 is biased when $\rho^2_{y|\mathrm{x}}$ is 0 [this is analogous to (5.57)].

8. $R^2 \ge \max_j r_{yj}^2$, where r_{yj} is an element of $\mathrm{r}'_{yx} = \left(r_{y1}, r_{y2}, \cdots, r_{yk}\right)$.
9. R^2 is invariant to full rank linear transformations on y or on the x's.

Example 6.4: For the hematology data in Table 6.1, S_{xx}, S_{yx}, R_{xx}, and r_{yx} were obtained in Example 6.3. Using either (6.34) or (6.38), we obtain

$$R^2 = 0.95803$$

Example 6.4 [The program name ta23.m] Add to ta22	Applications using MATLAB

```
Rs=Syx'*Sxx^(-1)*Syx/Syy
yhat=beta0+beta1'*x';ryyhat=corr(y,yhat')
```

6.5: Test and confidence intervals for R^2

Note that by (6.27), $\rho_{y|\mathrm{x}}^2 = 0$ becomes $\rho_{y|\mathrm{x}}^2 = \dfrac{\sigma'_{yx}\Sigma_{xx}^{-1}\sigma_{yx}}{\sigma_{yy}} = 0$

which leads to $\sigma_{yx} = 0$ since Σ_{xx} is positive definite. Then by (6.7), $\beta_1 = \Sigma_{xx}^{-1}\sigma_{yx}$, and $H_0 : \rho_{y|\mathrm{x}}^2 = 0$ is equivalent to $H_0 : \beta_1 = 0$.

The F statistic for fixed x's is given in (4.5), (4.22), and (4.23) as

$$F = \frac{\left(\hat{\beta}'\mathrm{X}'\mathrm{y} - n\bar{y}^2\right)/k}{\left(\mathrm{y}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\,\mathrm{y}\right)/(n-k-1)}$$

$$= \frac{R^2/k}{\left(1-R^2\right)/(n-k-1)} \tag{5.44}$$

The test statistic in (6.44) can be obtained by the likelihood ratio approach in the case of random x's (Anderson 1984, pp. 140–142):

Theorem 6.5: If (y_1, x'_1), (y_2, x'_2), . . . , (y_n, x'_n) is a random sample from $N_{k+1}(\mu, \Sigma)$, where μ and Σ are given by (6.2) and (6.3), the likelihood ratio test for $H_0 : \beta_1 = 0$ or equivalently $H_0 : \rho_{y|\mathrm{x}}^2 = 0$ can be based on F in (6.44). We reject H_0 if $F \ge F_{\alpha,k,n-k-1}$.

Proof: Using the notation $\mathrm{v}_i' = (y_i, \mathrm{x}_i')$, as in the proof of Theorem 6.2a, the likelihood function $L(\mu,\Sigma) = \prod_{i=1}^{n} f(\mathrm{v}_i; \mu, \Sigma)$ is given by (6.11), and the likelihood ratio is

$$LR = \frac{\max_{H_0} L(\mu,\Sigma)}{\max_{H_1} L(\mu,\Sigma)}$$

Under H_1, the parameters μ and Σ are essentially unrestricted, and we have

$$\max_{H_1} L(\mu,\Sigma) = \max L(\mu,\Sigma) = L\left(\hat{\mu},\hat{\Sigma}\right)$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the maximum likelihood estimators in (6.9) and (6.10). Since $(\mathrm{v}_i - \mu)' \Sigma^{-1} (\mathrm{v}_i - \mu)$ is a scalar, the exponent of $L(\mu,\Sigma)$ in (6.11) can be written as

$$\frac{\sum_{i=1}^{n} tr[(\mathrm{v}_i - \mu)' \Sigma^{-1} (\mathrm{v}_i - \mu)]}{2} = \frac{\sum_{i=1}^{n} tr[\Sigma^{-1} (\mathrm{v}_i - \mu)(\mathrm{v}_i - \mu)']}{2}$$

$$= \frac{tr[\Sigma^{-1} \sum_{i=1}^{n} (\mathrm{v}_i - \mu)(\mathrm{v}_i - \mu)']}{2}$$

Then substitution of $\hat{\mu}$ and $\hat{\Sigma}$ for μ and Σ in $L(\mu,\Sigma)$ gives

$$\max_{H_1} L(\mu,\Sigma) = L\left(\hat{\mu},\hat{\Sigma}\right) = \frac{1}{\left(\sqrt{2\pi}\right)^{n(k+1)} \left|\hat{\Sigma}\right|^{n/2}} e^{-tr\left(\hat{\Sigma}^{-1} n\hat{\Sigma}/2\right)}$$

$$= \frac{e^{-n(k+1)/2}}{\left(\sqrt{2\pi}\right)^{n(k+1)} \left|\hat{\Sigma}\right|^{n/2}}$$

Under $H_0 : \rho_{y|\mathrm{x}}^2 = 0$, we have $\sigma_{yx} = 0$, and Σ in (6.3) becomes

$$\Sigma_0 = \begin{pmatrix} \sigma_{yy} & 0' \\ 0 & \Sigma_{xx} \end{pmatrix} \tag{6.45}$$

Whose maximum likelihood estimator is

$$\hat{\Sigma}_0 = \begin{pmatrix} \hat{\sigma}_{yy} & 0' \\ 0 & \hat{\Sigma}_{xx} \end{pmatrix} \tag{6.46}$$

Using $\hat{\Sigma}_0$ in (6.46) and $\hat{\mu} = \bar{\text{v}}$ in (6.9), we have

$$\max_{H_0} L(\mu, \Sigma) = L(\hat{\mu}, \hat{\Sigma}_0) = \frac{1}{\left(\sqrt{2\pi}\right)^{n(k+1)} \left|\hat{\Sigma}_0\right|^{n/2}} e^{-tr\left(\hat{\Sigma}_0^{-1} n\hat{\Sigma}_0 / 2\right)}$$

This becomes

$$L(\hat{\mu}, \hat{\Sigma}_0) = \frac{e^{-n(k+1)/2}}{\left(\sqrt{2\pi}\right)^{n(k+1)} \hat{\sigma}_{yy}^{n/2} \left|\hat{\Sigma}_{xx}\right|^{n/2}} \tag{6.47}$$

Thus

$$LR = \frac{\left|\hat{\Sigma}\right|^{n/2}}{\hat{\sigma}_{yy}^{n/2} \left|\hat{\Sigma}_{xx}\right|^{n/2}} \tag{6.48}$$

Substituting $\hat{\Sigma} = (n-1)S/n$ and using (6.40), we obtain

$$LR = \left(1 - R^2\right)^{n/2} \tag{6.49}$$

We reject H_0 for $\left(1 - R^2\right)^{n/2}$, which is equivalent to

$$F = \frac{R^2/k}{\left(1 - R^2\right)/(n - k - 1)} \geq F_{\alpha, k, n-k-1}$$

since $R^2/\left(1 - R^2\right)$ is a monotone increasing function of R^2 and F is distributed as $F_{(\alpha, k, n-k-1)}$ when H_0 is true (Anderson 1984, pp. 138–139).

When $k = 1$, F in (6.44) reduces to $F = (n-2)r^2/\left(1 - r^2\right)$. Then

$$t = \frac{\sqrt{n-2}r}{\sqrt{1 - r^2}}$$

[see (2.20)] has a t distribution with n - 2 degrees of freedom (df) when (y, x) has a bivariate normal distribution with $\rho = 0$.

If (y, x) is bivariate normal and $\rho \neq 0$, then $Var(r) = \left(1 - \rho^2\right)^2/n$ and the function

$$u = \frac{\sqrt{n}(r - \rho)}{1 - \rho^2} \tag{6.50}$$

Is approximately standard normal for large n. However, the distribution of u approaches normality very slowly as n increases (Kendall and Stuart 1969, p. 236). Its use is questionable for $n < 500$.

Fisher (1921) found a function of r that approaches normality much faster than does (6.50) and can thereby be used with much smaller n than that required for (6.50). In addition, the variance is almost independent of r. Fisher's function is

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} = \tanh^{-1} r \qquad (6.51)$$

Where $\tanh^{-1} r$ is the inverse hyperbolic tangent of r. The approximate mean and variance of z are

$$E(z) \cong \frac{1}{2}\ln\frac{1+\rho}{1-\rho} = \tanh^{-1}\rho \qquad (6.52)$$

$$Var(z) \cong \frac{1}{n-3} \qquad (6.53)$$

We can use Fisher's z transformation in (6.51) to test hypotheses such as $H_0 : \rho = \rho_0$ or $H_0 : \rho_1 = \rho_2$. To test $H_0 : \rho = \rho_0$ vs. $H_1 : \rho \neq \rho_0$, we calculate

$$v = \frac{z - \tanh^{-1}\rho_0}{\sqrt{1/(n-3)}} \qquad (6.54)$$

Which is approximately distributed as the standard normal $N(0, 1)$. We reject H_0 if $|v| \geq z_{\alpha/2}$, where $z = \tanh^{-1} r$ and $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution. To test $H_1 : \rho_1 = \rho_2$ vs. $H_1 : \rho_1 \neq \rho_2$ for two independent samples of sizes n_1 and n_2 yielding sample correlations r_1 and r_2, we calculate

$$v = \frac{z_1 - z_2}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}} \qquad (6.55)$$

and reject H_0 if $|v| \geq z_{\alpha/2}$, where $z_1 = \tanh^{-1} r_1$ and $z_2 = \tanh^{-1} r_2$. To test $H_0 : \rho_1 = \rho_2 = \cdots = \rho_q$ for $q > 2$, see Problem 6.18.

To obtain a confidence interval for ρ, we note that since z in (6.51) is approximately normal, we can write

$$P\left(-z_{\alpha/2} \le \frac{z-\tanh^{-1}\rho}{1/\sqrt{n-3}} \le z_{\alpha/2}\right) \cong 1-\alpha \tag{6.56}$$

Solving the inequality for ρ, we obtain the approximate $100(1-\alpha)\%$ confidence interval

$$\tanh\left(z-\frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(z+\frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \tag{6.57}$$

A confidence interval for $\rho^2_{y|x}$ was given by Helland (1987).

Example 6.5a: For the hematology data in Table 6.1, we obtained R^2 in Example 6.4. The overall F test of $H_0:\beta_1=0$ or $H_0:\rho^2_{y|x}=0$ is carried out using F in (6.44):

$$F=\frac{R^2/k}{(1-R^2)/(n-k-1)}=\frac{0.95803/5}{(1-0.95803)/(45)}=205.44$$

The p value is zero.

Example 6.5a [The program name ta24.m]	Applications using MATLAB

```
clc
clear all
data=[14   13.4   39  41  25  17; 15  14.6   46  50  30  20;19   13.5   42  45  21  18
  23  15  46  46  16  18; 17  14.6   44  51  31  19; 20  14  44  49  24  19; 21  16.4
49  43  17  18    16  14.8   44  44  26  29; 27  15.2   46  41  13  27;34   15.5   48
84  42  36;26   15.2   47  56  27  22   28  16.9   50  51  17  23; 24  14.8   44  47
20  23; 26  16.2   45  56  25  19; 23  14.7   43  40  13  17   9   14.7   42  34  22
13; 18  16.5   45  54  32  17; 28  15.4   45  69  36  24; 17  15.1   45  46  29  17
  14  14.2   46  42  25  28; 8   15.9   46  52  34  16; 25  16  47  47  14  18; 37
17.4   50  86  39  17   20  14.3   43  55  31  19;15   14.8   44  42  24  19; 9   14.9
43  43  32  17;16   15.5   45  52  30  20   18  14.5   43  39  18  25; 17  14.4   45
60  37  23; 23  14.6   44  47  21  27; 43  15.3   45  79  23  23    17  14.9   45  34
15  24; 23  15.8   47  60  32  21; 31  14.4   44  77  39  23;11   14.7   46  37  23  23
  25  14.8   43  52  19  22; 30  15.4   45  60  25  18; 32  16.2   50  81  38  18; 17
15  45  49  26  24  22  15.1   47  60  33  16; 20  16  46  46  22  22;20   15.3   48
55  23  23; 20  14.5   41  62  36  21;   26  14.2   41  49  20  20; 40  15  45  72  25
25; 22  14.2   46  58  31  22; 61  14.9   45  84  17  17   12  16.2   48  31  15  18;20
14.5   45  40  18  20;35   16.4   49  69  22  24;38   14.7   44  78  34  16];
y=data(:,1);x1=data(:,2);x2=data(:,3);x3=data(:,4);
x4=data(:,5);x5=data(:,6);n=length(x1);k=5;my=mean(y);
x=[x1 x2 x3 x4 x5];mx=mean(x)';Rxx=corr(x);
```

```
ryx=corr(y,x)';Dx=[diag(diag(cov(x)))].^.5;Syy=var(y);
Sxx=Dx*Rxx*Dx;Syx=std(y)*Dx*ryx;beta1=Sxx^(-1)*Syx;
beta0=my-Syx'*Sxx^(-1)*mx;Ssq=var(y)-Syx'*Sxx^(-1)*Syx;
beta1star=inv(Rxx)*ryx; Rs=Syx'*Sxx^(-1)*Syx/Syy
F=(Rs/k)/((1-Rs)/(n-k-1)); p=1-fcdf(F,k,n-k-1)
```

```
Ans.
```

```
Rs =                    F =               p =
    0.95803                 205.44              0
```

Example 6.5b: To illustrate Fisher's z transformation in (6.51) and its use to compare two independent correlations in (6.55), we divide the hematology data in Table 6.1 into two sub samples of sizes $n_1 = 26$ and $n_2 = 25$ (the first 26 observations and the last 25 observations). For the correlation between y and x_1 in each of the two sub-samples, we obtain $r_1 = 0.49938$ and $r_2 = 0.05739$. The z transformation in (6.51) for each of these two values is given by

$$z_1 = \tanh^{-1} r_1 = 0.54849$$

$$z_2 = \tanh^{-1} r_2 = 0.05745$$

To test $H_0 : \rho_1 = \rho_2$, we use the approximate test statistic (6.55) to obtain

$$v = \frac{0.54849 - 0.05745}{\sqrt{1/(26-3) + 1/(25-3)}} = 1.7183$$

Since $1.7183 < z_{0.025} = 1.96$ 1.6969, we do not reject H_0.

To obtain approximate 95% confidence limits for ρ_1, we use (6.57):

$$\text{Lower limit for } \rho_1 : \tanh\left(0.54849 - \frac{1.96}{\sqrt{23}}\right) = 0.13889$$

$$\text{Upper limit for } \rho_1 : \tanh\left(0.54849 + \frac{1.96}{\sqrt{23}}\right) = 0.74301$$

For ρ_2, the limits are given by

$$\text{Lower limit for } \rho_2 : \tanh\left(0.05745 - \frac{1.96}{\sqrt{22}}\right) = -0.34558$$

$$\text{Upper limit for } \rho_2 \text{ : } \tanh\left(0.05745+\frac{1.96}{\sqrt{22}}\right)=0.44249$$

Example 6.5b [The program name ta25.m]	Applications using MATLAB

```
clc
clear all
data=[14   13.4   39  41  25  17; 15  14.6   46  50  30  20;19   13.5   42  45  21  18
   23  15  46  46  16  18; 17  14.6   44  51  31  19; 20  14  44  49  24  19; 21  16.4
49  43  17  18 16  14.8   44  44  26  29; 27  15.2   46  41  13  27;34   15.5   48  84
42  36;26   15.2   47  56  27  22  28  16.9   50  51  17  23; 24  14.8   44  47  20
23; 26  16.2   45  56  25  19; 23  14.7   43  40  13  17  9   14.7   42  34  22  13; 18
16.5   45  54  32  17; 28  15.4   45  69  36  24; 17  15.1   45  46  29  17 14  14.2
46  42  25  28; 8   15.9   46  52  34  16; 25  16  47  47  14  18; 37  17.4   50  86  39
17  20  14.3   43  55  31  19;15   14.8   44  42  24  19; 9   14.9   43  43  32  17;16
15.5   45  52  30  20 18  14.5  43  39  18  25; 17  14.4   45  60  37  23; 23  14.6
44  47  21  27; 43  15.3   45  79  23  23 17  14.9   45  34  15  24; 23  15.8   47  60
32  21; 31  14.4   44  77  39  23;11   14.7   46  37  23  23  25  14.8   43  52  19 22;
30  15.4   45  60  25  18; 32  16.2   50  81  38  18; 17  15  45  49  26  24 22  15.1
47  60  33  16; 20  16  46  46  22  22;20   15.3   48  55  23  23; 20  14.5   41  62
36  21; 26  14.2   41  49  20  20; 40  15  45  72  25  25; 22  14.2   46  58  31  22;
61  14.9   45  84  17  17  12  16.2   48  31  15  18;20   14.5   45  40  18  20;35
16.4   49  69  22  24;38   14.7   44  78  34  16];
y1=data(1:26,1);x1=data(1:26,2);y2=data(27:51,1);x2=data(27:51,2);
n1=length(x1);n2=length(x2);r1=corr(y1,x1),r2=corr(y2,x2),alfa=0.05;
z1=atanh(r1),z2=atanh(r2),v=(z1-z2)/sqrt(1/(n1-1)+1/(n2-1)),
z=abs(tinv(alfa/2,inf));
LowerLimitForr1 = tanh(z1-z/sqrt(n1-3)),UpperLimitForr1 = tanh(z1+z/sqrt(n1-3))
LowerLimitForr2 = tanh(z2-z/sqrt(n2-3)),UpperLimitForr2 = tanh(z2+z/sqrt(n2-3))
```

Ans.

```
r1 =                          r2 =

     0.49938                        0.057391

z1 =                          z2 =                          v =

     0.54849                        0.057454                     1.7183

LowerLimitForr1 =                       UpperLimitForr1 =

                  0.1389                                      0.74301

LowerLimitForr2 =                       UpperLimitForr2 =

                  -0.34558                                    0.44249
```

6.6: Effect of each variable on R^2

The contribution of a variable x_j to the multiple correlation *R* will, in general, be different from its bivariate correlation with *y*; that is, the increase in R^2 when x_j is added is not equal to $r^2_{yx_j}$. This increase in R^2 can be either more or less than $r^2_{yx_j}$. It seems clear that relationships with other variables can render a variable partially redundant and thereby reduce the contribution of x_j to R^2, but it is not intuitively apparent how the contribution of x_j to R^2 can exceed $r^2_{yx_j}$. The latter phenomenon has been illustrated numerically by Flury (1989) and Hamilton (1987).

In this section, we provide a breakdown of the factors that determine how much each variable adds to R^2 and show how the increase in R^2 can exceed $r^2_{yx_j}$ (Rencher 1993). We first introduce some notation. The variable of interest is denoted by *z*, which can be one of the *x*'s or a new variable added to the *x*'s. We make the following additional notational definitions:

$R^2_{yw} =$ squared multiple correlation between *y* and $\mathrm{w} = (x_1, x_2, \cdots, x_k, z)'$.

$R^2_{yx} =$ squared multiple correlation between *y* and $\mathrm{x} = (x_1, x_2, \cdots, x_k)'$.

$R^2_{zx} = S'_{zx} S^{-1}_{xx} S_{zx} \big/ S^2_z$ = squared multiple correlation between *y* and x.

$r_{yz} =$ simple correlation between *y* and *z*.

$\mathrm{r}_{yx} = \left(r_{yx_1}, r_{yx_2}, \cdots, r_{yx_k}\right)' =$ vector of correlations between *y* and x.

$\mathrm{r}_{zx} = \left(r_{zx_1}, r_{zx_2}, \cdots, r_{zx_k}\right)' =$ vector of correlations between *z* and x.

$\hat{\beta}^*_{zx} = \mathrm{R}^{-1}_{xx}\mathrm{r}_{zx}$ is the vector of standardized regression coefficients (beta weights) of *z* regressed on x [see (6.24)].

The effect of *z* on R^2 is formulated in the following theorem.

Theorem 6.6: The increase in R^2 due to *z* can be expressed as

$$R_{yw}^2 - R_{yx}^2 = \frac{(\hat{r}_{yz} - r_{yz})^2}{1 - R_{zx}^2} \qquad (6.58)$$

Where $\hat{r}_{yz} = \hat{\beta}_{zx}^{*\prime} \mathrm{r}_{yx}$ is a *predicted* value of r_{yz} based on the relationship of *z* to the *x*'s.

Proof: See Problem 6.19.

Since the right side of (6.58) is positive, R^2 cannot decrease with an additional variable, which is a verification of property 3 in Section 3.7. If *z* is orthogonal to x (i.e., if $r_{zx} = 0$), then $\hat{\beta}_{zx}^* = 0$, which implies that $\hat{r}_{yz} = 0$ and $R_{zx}^2 = 0$. In this case, (6.58) can be written as $R_{yw}^2 = R_{yx}^2 + r_{yz}^2$, which verifies property 5 of Section 3.7.

It is clear in Theorem 6.6 that the contribution of *z* to R^2 can either be less than or greater than r_{yz}^2. If $\hat{r}_{yz}$ is close to r_{yz}, the contribution of *z* is less than r_{yz}^2. There are three ways in which the contribution of *z* can exceed r_{yz}^2: (1) $\hat{r}_{yz}$ is substantially larger in absolute value than r_{yz}, (2) $\hat{r}_{yz}$ and r_{yz} are of opposite signs, and (3) R_{zx}^2 is large.

In many cases, the researcher may find it helpful to know why a variable contributed more than expected or less than expected. For example, admission to a university or professional school may be based on previous grades and the score on a standardized national test. An applicant for admission to a university with limited enrollment would submit high school grades and a national test score. These might be entered into a regression equation to obtain a predicted value of first-year grade-point average at the university. It is typically found that the standardized test increases R^2 only slightly above that based on high school grades alone. This small increase in R^2 would be disappointing to admissions officials who had hoped that the national test score might be a more useful predictor than high school grades. The designers of such standardized tests may find it beneficial to know precisely why the test makes such an unexpectedly small contribution relative to high school grades.

In Theorem 6.6, we have available the specific information needed by the designer of the standardized test. To illustrate the use of (6.58), let

y be the grade-point average for the first year at the university, let z be the score on the standardized test, and let $x_1, x_2, \cdots, x_k$ be high school grades in key subject areas. By (6.58), the increase in R^2 due to z is $(\hat{r}_{yz} - r_{yz})^2 / (1 - R^2_{zx})$, in which we see that z adds little to R^2 if $\hat{r}_{yz}$ is close to r_{yz}. We could examine the coefficients in $\hat{r}_{yz} = \hat{\beta}^{*\prime}_{zx} \mathrm{r}_{yx}$ to determine which of the $\hat{r}_{yx_j}$'s in r_{yx} have the most effect. This information could be used in redesigning the questions so as to reduce these particular $\hat{r}_{yx_j}$'s. It may also be possible to increase the contribution of z to R^2_{yw} by increasing R^2_{zx} (thereby reducing $1 - R^2_{zx}$). This might be done by designing the questions in the standardized test so that the test score z is more correlated with high school grades, $x_1, x_2, \cdots, x_q$.

Theil and Chung (1988) proposed a measure of the relative importance of a variable in multiple regression based on information theory.

Example 6.6: For the hematology data in Table 6.1, the overall R^2_{yw} was found in Example 6.4 to be 0.95803. From Theorem 6.6, the increase in R^2 due to a variable z has the breakdown $R^2_{yw} - R^2_{yx} = (\hat{r}_{yz} - r_{yz})^2 / (1 - R^2_{zx})$, where z represents any one of $x_1, x_2, \cdots, x_5$, and x represents the other four variables. The values of $\hat{r}_{yz}$, r_{yz}, R^2_{zx}, $R^2_{yw} - R^2_{yx}$, and F are given below for each variable in turn as z:

z	$\hat{r}_{yz}$	r_{yz}	R^2_{zx}	$R^2_{yw} - R^2_{yx}$	F	p-value
x_1	0.23995	0.23331	0.64078	0.000114	0.12235	0.72813
x_2	0.29014	0.25162	0.64939	0.001718	1.8416	0.18153
x_3	0.17303	0.79073	0.44327	0.878690	942.12	0
x_4	-0.05259	0.02264	0.38109	0.330820	354.7	0
x_5	0.27356	0.12584	0.12586	0.000480	0.51484	0.47676

The F value is from the partial F test in (4.25), (4.37), or (4.39) for the significance of the increase in R^2 due to each variable.

An interesting variable here is x_4, whose value of r_{yz} is .02264, the smallest among the five variables. Despite this small individual correlation with y, x_4 contributes much more to R^2_{yw} than do all other variables except x_3 because $\hat{r}_{yz}$ is much greater for x_4 than for the other variables. This illustrates how the contribution of a variable can be augmented in the presence of other variables as reflected in $\hat{r}_{yz}$.

The difference between the two major contributors x_3 and x_4 may be very revealing to the researcher. The contribution of x_3 to R^2_{yw} is due mostly to its own correlation with y, whereas virtually all the effect of x_4 comes from its association with the other variables as reflected in $\hat{r}_{yz}$.

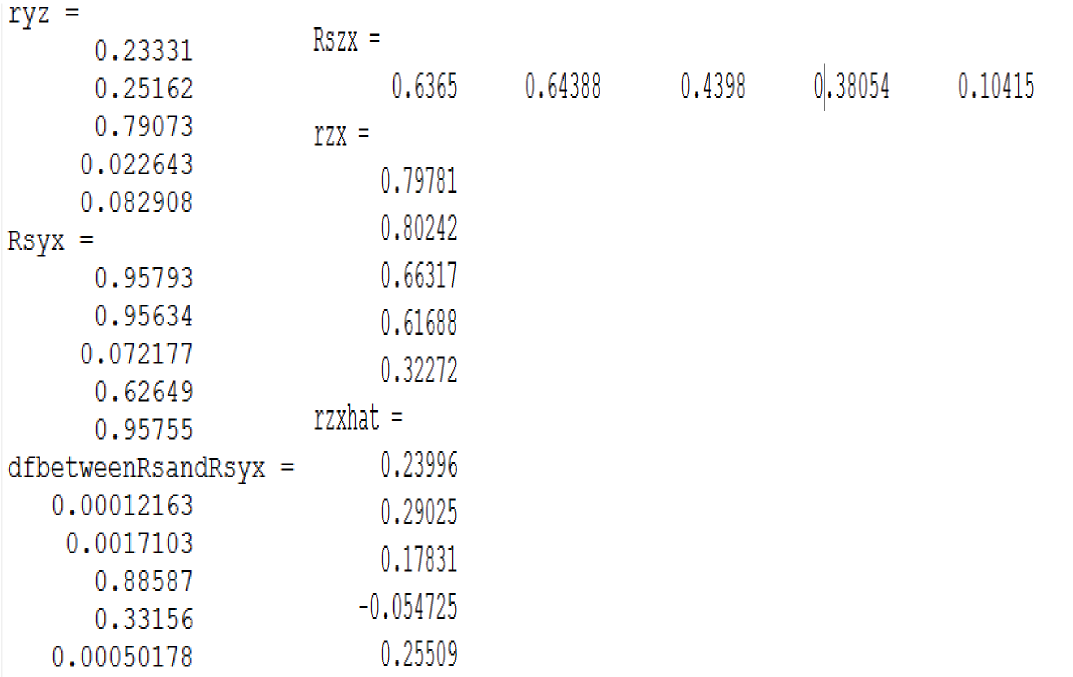
Example 6.6 [The program name ta26.m]	Applications using MATLAB

```
clc
clear all
data=[14   13.4   39  41  25  17; 15  14.6   46  50  30  20;19   13.5   42  45  21  18
  23  15  46  46  16  18; 17  14.6    44  51  31  19; 20  14  44  49  24  19; 21  16.4
49  43  17  18  16  14.8    44  44  26  29; 27  15.2   46  41  13  27;34   15.5   48  84
42  36;26   15.2    47  56  27  22  28  16.9    50  51  17  23; 24  14.8    44  47  20
23; 26  16.2    45  56  25  19; 23  14.7    43  40  13  17     9   14.7    42  34  22  13;
18  16.5    45  54  32  17; 28  15.4    45  69  36  24; 17  15.1    45  46  29  17  14
14.2    46  42  25  28; 8   15.9    46  52  34  16; 25  16  47  47  14  18; 37  17.4    50
86  39  17   20  14.3    43  55  31  19;15   14.8    44  42  24  29; 9   14.9    43  43  32
17;16   15.5    45  52  30  20      18  14.5    43  39  18  25; 17  14.4    45  60  37  23;
23  14.6    44  47  21  27; 43  15.3    45  79  23  23  17  14.9    45  34  15  24; 23
15.8    47  60  32  21; 31  14.4    44  77  39  23;11   14.7    46  37  23  23 25  14.8
43  52  19  22; 30  15.4    45  60  25  18; 32  16.2    50  81  38  18; 17  15  45  49
26  24 22  15.1    47  60  33  16; 20  16  46  46  22  22;20   15.3    48  55  23  23; 20
14.5    41  62  36  21; 26  14.2    41  49  20  20; 40  15  45  72  25  25; 22  14.2    46
58  31  22; 61  14.9    45  84  17  17 12  16.2    48  31  15  18;20   14.5    45  40  18
20;35   16.4    49  69  22  24;38   14.7    44  78  34  16];
y=data(:,1);x1=data(:,2); x2=data(:,3);x3=data(:,4); x4=data(:,5);
x5=data(:,6);n=length(x1);k=5;my=mean(y);
x=[x1 x2 x3 x4 x5]; mx=mean(x)';Rxx=corr(x); ryx=corr(y,x)';
Dx=[diag(diag(cov(x)))].^.5; Syy=var(y);Sxx=Dx*Rxx*Dx;Syx=std(y)*Dx*ryx;
Rs=Syx'*Sxx^(-1)*Syx/Syy
X1=[ones(size(x1)) x2 x3 x4 x5];X2=[ones(size(x1)) x1 x3 x4 x5];
X3=[ones(size(x1)) x1 x2 x4 x5];X4=[ones(size(x1)) x1 x2 x3 x5];
X5=[ones(size(x1)) x1 x2 x3 x4];ryz=corr(x,y)
[B,BL,R,RI,S]=regress(y,X1),Rsyx1=S(1);
[B,BL,R,RI,S]=regress(y,X2);Rsyx2=S(1);
[B,BL,R,RI,S]=regress(y,X3);Rsyx3=S(1);
[B,BL,R,RI,S]=regress(y,X4);Rsyx4=S(1);
```

```
[B,BL,R,RI,S]=regress(y,X5);Rsyx5=S(1);
Rsyx=[Rsyx1 Rsyx2 Rsyx3 Rsyx4 Rsyx5]'
dfbetweenRsandRsyx=Rs-Rsyx
[B1,BL,R,RI,S]=regress(x1,X1);Rsz1=S(1);
[B2,BL,R,RI,S]=regress(x2,X2);Rsz2=S(1);
[B3,BL,R,RI,S]=regress(x3,X3);Rsz3=S(1);
[B4,BL,R,RI,S]=regress(x4,X4);Rsz4=S(1);
[B5,BL,R,RI,S]=regress(x5,X5);Rsz5=S(1);
Rszx=[Rsz1 Rsz2 Rsz3 Rsz4 Rsz5]
rzx=[(Rszx).^(0.5)]'
% std regression
x1=(x1-mx(1))/std(x1);x2=(x2-mx(2))/std(x2);x3=(x3-mx(3))/std(x3);
x4=(x4-mx(4))/std(x4);x5=(x5-mx(5))/std(x5);
X1=[ones(size(x1)) x2 x3 x4 x5];X2=[ones(size(x1)) x1 x3 x4 x5];
X3=[ones(size(x1)) x1 x2 x4 x5];X4=[ones(size(x1)) x1 x2 x3 x5];
X5=[ones(size(x1)) x1 x2 x3 x4];betaxz1=regress(x1,X1);
betaxz2=regress(x2,X2);betaxz3=regress(x3,X3);
betaxz4=regress(x4,X4);betaxz5=regress(x5,X5);
betaxz=[betaxz1 betaxz2 betaxz3 betaxz4 betaxz5]; rzxhat=betaxz'*ryx
```

Ans.

```
Rs =
       0.95805
```



```
ryz =                    Rszx =
      0.23331                    0.6365    0.64388    0.4398    0.38054    0.10415
      0.25162            rzx =
      0.79073                   0.79781
     0.022643                   0.80242
     0.082908                   0.66317
Rsyx =                          0.61688
      0.95793                   0.32272
      0.95634            rzxhat =
     0.072177                   0.23996
      0.62649                   0.29025
      0.95755                   0.17831
dfbetweenRsandRsyx =           -0.054725
   0.00012163                   0.25509
    0.0017103
      0.88587
      0.33156
   0.00050178
```

6.7: Prediction for multivariate Normal or non-Normal data

In this section, we consider an approach to modeling and estimation in the random-*x* case that is somewhat reminiscent of least squares in the fixed-x case. Suppose that $(y, \mathrm{x}') = (y, x_1, x_2, \cdots, x_k)$ is not necessarily assumed to be multivariate normal and we wish to find a function *t*(x) for predicting *y*. In order to find a predicted value *t*(x) that is expected to be "close" to *y*, we will choose the function *t*(x) that minimizes the mean squared error $E[y - t(\mathrm{x})]^2$, where the expectation is in the joint distribution of $y, x_1, x_2, \cdots, x_k$. This function is given in the following theorem.

Theorem 6.7: For the random vector (y, x'), the function *t*(*x*) that minimizes the mean squared error $E[y - t(x)]^2$ is given by $E(y|\mathrm{x})$.

Proof: For notational simplicity, we use *k* = 1. The joint density $g(y, x)$ can be written as $g(y, x) = f(y|x)h(x)$. Then

$$\begin{aligned} E[y - t(x)]^2 &= \iint [y - t(x)]^2 g(y, x) dy dx \\ &= \iint [y - t(x)]^2 f(y|x) h(x) dy dx \\ &= \int h(x) \{ \int [y - t(x)]^2 f(y|x) dy \} dx \end{aligned}$$

To find the function *t*(*x*) that minimizes $E(y-t)^2$, we differentiate with respect to *t* and set the result equal to 0 [for a more general proof not involving differentiation, see Graybill (1976, pp. 432–434) or Christensen (1996, p. 119)]. Assuming that we can interchange integration and differentiation, we obtain

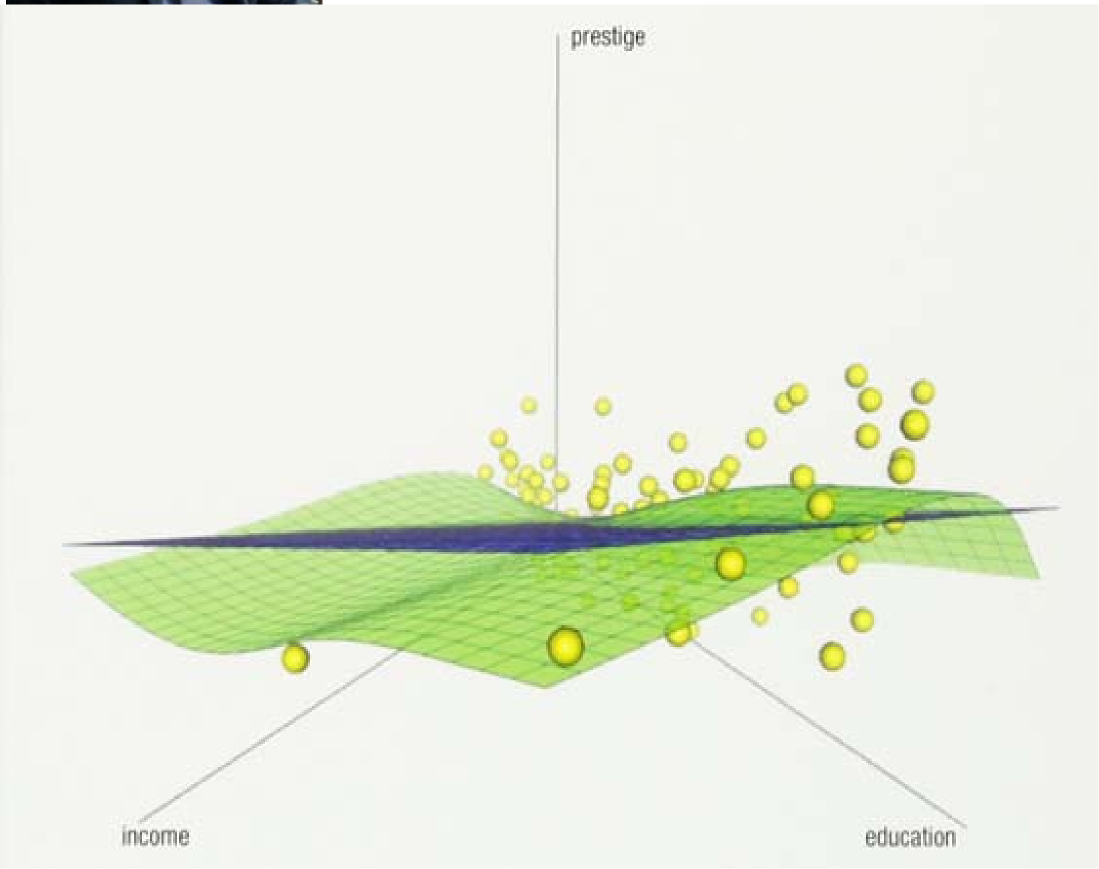
$$\frac{\partial E[y - t(x)]^2}{\partial t} = \int h(x) \{ \int 2(-1)[y - t(x)] f(y|x) dy \} dx = 0$$

which gives

$$2\int h(x) [\int y f(y|x) dy - \int t(x) f(y|x) dy] dx = 0$$

$$2\int h(x) [E(y|x) - t(x)] dx = 0$$

Linear Models With MATLAB



Prof. Dr. Taha Hussein Ali Alzubaydi

Department of Statistics, College of Administration and
Economics Salahaddin University, Erbil, Kurdistan Region, Iraq

2021