Chapter 8

Analysis of Covariance

Analysis of Covariance

The **analysis of covariance** (ANCOVA) is a general linear **model** with one continuous explanatory variable and one or more factors. ANCOVA is a merger of ANOVA and regression for continuous variables. ANCOVA tests whether certain factors have an effect after removing the variance for which quantitative predictors (covariates) account. The inclusion of covariates can increase statistical power because it accounts for some of the variability.

The purpose of ANCOVA to compare two or more linear regression lines. It is a way of comparing the Y variable among groups while statistically controlling for variation in Y caused by variation in the X variable

one-way analysis of covariance (ANCOVA)

A **one-way analysis of covariance (ANCOVA)** evaluates whether
population means on the dependent variable are the same across levels of a factor
(independent variable), adjusting for differences on the covariate, or more simply stated, whether the adjusted group means differ significantly from each other. With a one-way analysis of covariance, each individual or case must have scores on three variables: a factor or independent variable, a covariate, and a dependent variable. The factor divides individuals into two or more groups or levels, while the covariate and the dependent variable differentiate individuals on quantitative dimensions. The one-way ANCOVA is used to analyze data from several types of studies; including studies with a pretest and random assignment of subjects to factor levels, studies with a pretest and assignment to factor levels based on the pretest, studies with a pretest, matching based on the pretest, and random assignment to factor levels, and studies with potential confounding (Green &Salkind, 2003).

The **analysis of covariance (ANCOVA)** is typically used to adjust or control for differences between the groups based on another, typically interval level, variable called the covariate. The ANCOVA is an extension of ANOVA that typically provides a way of statistically controlling for the effects of continuous or scale variables that you are concerned about but that are not the focal point or independent variable's) in the study. For example, imagine that we found that boys and girls differ on math achievement. However, this could be due to the fact that boys take more math courses in high school. ANCOVA allows us to adjust the math achievement scores based on the relationship between number of math courses taken and math achievement. We can then determine if boys and girls still have different math achievement scores after making the adjustment (Leech,Barrett, & Morgan, 2005).

STATISTICAL CONTROL USING ANCOVA

Controlling and explaining variation in the dependent variable can be accomplished with either experimental control, using research design, or statistical control, using analysis of covariance. Analysis of covariance is used primarily as a procedure for the statistical control of an extraneous variable. ANCOVA, which combines regression analysis and analysis of variance (ANOVA), controls for the effects of this extraneous variable, called a **covariate**, by partitioning out the variation attributed to this additional variable. In this way, the researcher is better able to investigate the effects of the primary independent variable. The ANCOVA *F* test evaluates whether the population means on the dependent variable, adjusted for differences on the covariate, differ across levels of a factor. If a factor has more than two levels and the *F* is significant, follow-up tests should be conducted to determine where there are differences on the adjusted means between groups. For example, if a factor has three levels, three pairwise comparisons among the adjusted means can be conducted: Group 1 versus Group 2, Group 1 versus Group 3, and Group 2 versus Group 3.

Covariate

Covariate – (also called a “concomitant” or “confound” variable) a variable that a researcher seeks to control for (statistically subtract the effects of) by using such techniques as multiple regression analysis (MRA) or analysis of covariance (ANCOVA) (Leech, Barrett, & Morgan, 2005; Vogt, 1999).

Extraneous variable – (sometimes called “nuisance variable.”) any condition not part of a study (that is, one in which researchers have no interest) but that could have an effect on the study’s dependent variable. (Note that, in this context, extraneous does not mean unimportant.) Researchers usually try to control for extraneous variables by experimental isolation, by randomization, or by statistical techniques such as analysis of covariance (Vogt, 1999).

Uses of ANCOVA

-Account (adjust) for “pre-existing” condition

- initial weight
- soil mousture/fertility at planting
- baseline value (y at start of experiment)

-Convenient alternative to regression contrasts (orthogonal polynomials) in treatment design with quantitative treatment levels

The assumptions underlying the ANCOVA had a slight modification from those for the ANOVA, however, conceptually, they are the same.

Assumption 1: The cases represent a random sample from the population, and the scores on the dependent variable are independent of each other, known as the assumption of independence.

The test will yield inaccurate results if the independence assumption is violated.

This is a design issue that should be addressed prior to data collection. Using random sampling is the best way of ensuring that the observations are independent; however, this is not always possible. The most important thing to avoid is having known relationships among participants in the study.

Assumption 2: The dependent variable is normally distributed in the population for any specific value of the covariate and for any one level of a factor (independent variable), known as the assumption of normality.

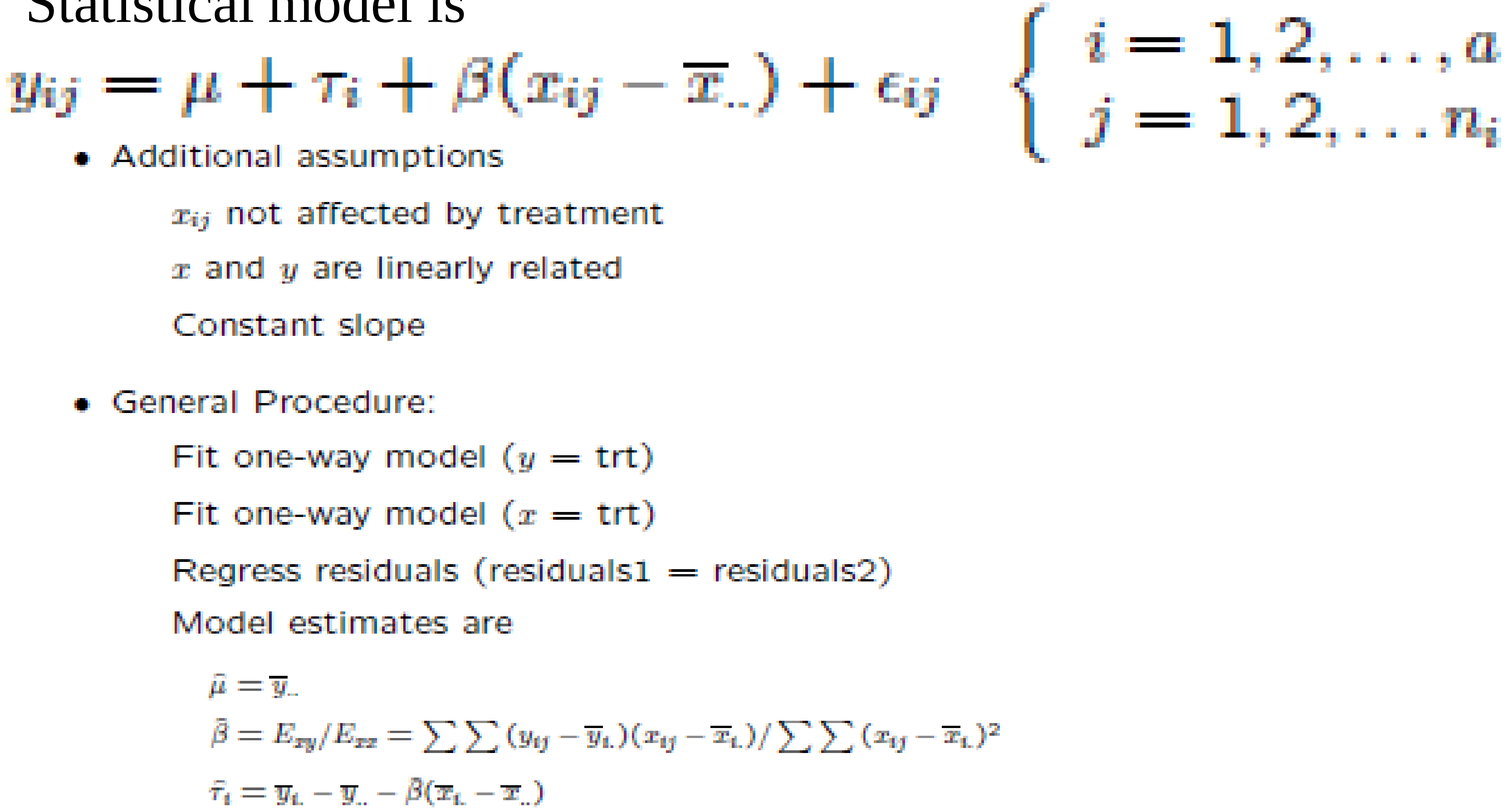
This assumption describes multiple conditional distributions of the dependent variable, one for every combination of values of the covariate and levels of the factor, and requires them all to be normally distributed. To the extent that population distributions are not normal and sample sizes are small, *p* values may be invalid. In addition, the power of ANCOVA tests may be reduced considerably if the population distributions are non-normal and, more specifically, thick-tailed or heavily skewed. The assumption of normality can be checked with skewness values (e.g., within +3.29 standard deviations).

Assumption 3: The variances of the dependent variable for the conditional distributions are equal, known as the assumption of homogeneity of variance.

To the extent that this assumption is violated and the group sample sizes differ, the validity of the results of the one-way ANCOVA should be questioned. Even with equal sample sizes, the results of the standard post hoc tests should be mistrusted if the population variances differ. The assumption of homogeneity of variance can be checked with the Levene’s *F* test.

Model Description

- Consider single covariate in CRD
- Statistical model is



$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \overline{x}_{..}) + \epsilon_{ij} \qquad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots n_i \end{cases}$$

- Additional assumptions

  x_{ij} not affected by treatment

  x and y are linearly related

  Constant slope

- General Procedure:

  Fit one-way model (y = trt)

  Fit one-way model (x = trt)

  Regress residuals (residuals1 = residuals2)

  Model estimates are

$$\hat{\mu} = \overline{y}_{..}$$

$$\hat{\beta} = E_{xy}/E_{xx} = \sum\sum(y_{ij} - \overline{y}_{i.})(x_{ij} - \overline{x}_{i.})/\sum\sum(x_{ij} - \overline{x}_{i.})^2$$

$$\hat{\tau}_i = \overline{y}_{i.} - \overline{y}_{..} - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{..})$$

ANCOVA TABLE

S.O.V	d.f	Sum of squares and sum of cross Products(SS and SCP)				Adjusted SS
		xx	xy	yy	df'’	SS’
Treatments	t-1	t_{xx}	t_{xy}	t_{yy}	t-1	$t'_{yy} = (t_{yy} + e_{yy})' - e'_{yy}$
Error	t(r-1)	e_{xx}	e_{xy}	e_{yy}	t(r-1)-1	$e'_{yy} = e_{yy} - \frac{(e_{xy})^2}{e_{xx}}$
Total	rt-1	T_{xx}	T_{xy}	T_{yy}		
Treat+Error	rt-1	$t_{xx}+e_{xx}$	$t_{xy}+e_{xy}$	$t_{yy}+e_{yy}$	(tr-1)-1	$(t_{yy} + e_{yy})' = (t_{yy} + e_{yy}) - \frac{(t_{xy} + e_{xy})^2}{(t_{xx} + e_{xx})}$

Adjusted M.S	F	
$\mathrm{MSt}' = \frac{t'_{yy}}{t-1}$	$F = \frac{MSt'}{MSe'}$	
$\mathrm{MSe}' = \frac{e'_{yy}}{t(r-1)-1}$		

$$T_{yy} = \sum y_{ij}^2 - \frac{(Y..)^2}{rt}$$

$$t_{yy} = \frac{\sum Y_{i.}^2}{r} - \frac{(Y..)^2}{rt}$$

$$e_{yy} = T_{yy} - t_{yy}$$

$$T_{xx} = \sum x_{ij}^2 - \frac{(X..)^2}{rt}$$

$$t_{xx} = \frac{\sum X_{i.}^2}{r} - \frac{(X..)^2}{rt}$$

$$e_{xx} = T_{xx} - t_{xx}$$

$$T_{xy} = \sum x_{ij} y_{ij} - \frac{(X..)(Y..)}{rt}$$

$$t_{xy} = \frac{\sum X_{i.} Y_{i.}}{r} - \frac{(X..)(Y..)}{rt}$$

$$e_{xy} = T_{xy} - t_{xy}$$

To test the hypothesis that there is no difference (null hypothesis Null Hypothesis) between the Mean of the treatments is adjusted for the variable y, the F-test will be as usual

$$F = \frac{t_{yy} / t - 1}{e_{yy} / t(r-1)} = \frac{MSt \ for \ \ y}{MSe \ for \ y}$$

To test the hypothesis that there is no difference (null hypothesis Null Hypothesis) between the mean of the treatments of the variable X, the F-test will be as usual

$$F = \frac{t_{xx} / t - 1}{e_{xx} / t(r-1)} = \frac{MSt \ for \ \ x}{MSe \ for \ x}$$

To test the hypothesis that there is no difference (null hypothesis Null Hypothesis) between the mean of the treatments of the variable y adjusted to the regression in y on x is F-test on the basis of adjusted variance

$$F = \frac{MSt'}{MSe'}$$

Example
Given the following data in the table below use CRD Design

Treat ti	Variant	Observation							TOTAL
t_1	X_{ij}	30	27	20	21	33	29	X1.	160
	Y_{ij}	165	170	130	156	167	151	Y1.	939
t_2	X_{ij}	24	31	20	26	20	25	X2.	146
	Y_{ij}	180	169	171	161	180	170	Y2.	1031
t_3	X_{ij}	34	32	35	35	30	29	X3.	195
	Y_{ij}	156	189	138	190	160	172	Y3.	1005
t_4	X_{ij}	41	32	30	35	28	36	X4.	202
	Y_{ij}	201	173	200	193	142	189	Y4.	1098
									X..=703 Y..=4073

$$T_{XX} = (30)^2 + (27)^2 + \ldots + (36)^2 - \frac{(703)^2}{(6)(4)} = 726.96$$

$$t_{XX} = \frac{(160)^2 + (146)^2 + (195)^2 + (202)^2}{6} - \frac{(703)^2}{(6)(4)} = 365.46$$

$$e_{xx} = T_{XX} - t_{XX}$$

$$= 361.50$$

$$T_{yy} = (165)^2 + (170)^2 + \ldots + (189)^2 - \frac{(4073)^2}{(6)(4)} = 8100.96$$

$$t_{yy} = \frac{(939)^2 + (1031)^2 + (1005)^2 + (1089)^2}{6} - \frac{(4073)^2}{(6)(4)} = 2163.13$$

$$e_{yy} = T_{yy} - t_{yy}$$

$$= 5937.83$$

$$T_{Xy} = (30)(165) + (27)(170) + \ldots + (36)(189) - \frac{(703)(4073)}{(6)(4)} = 8100.96$$

$$t_{Xy} = \frac{(160)(939) + (146)(1031) + \ldots. + (202)(1089)}{6} - \frac{(703)(4073)}{(6)(4)} = 451.21$$

$$e_{Xy} = T_{Xy} - t_{Xy}$$

$$= 496.83$$

$$t_{xx} + e_{xx} = 726.96$$

$$t_{xy} + e_{xy} = 948.04$$

$$t_{yy} + e_{yy} = 8100.96$$

$$(t_{yy} + e_{yy})' = (t_{yy} + e_{yy}) - \frac{(t_{xy} + e_{xy})^2}{(t_{xx} + e_{xx})} = 8100.96 - \frac{(948.04)^2}{726.96}$$

$$= 6864.61$$

$$e'_{yy} = e_{yy} - \frac{(e_{xy})^2}{e_{xx}} = 5937.83 - \frac{(496.83)^2}{361.50} = 5255.01$$

$$t'_{yy} = (t_{yy} + e_{yy})' - e'_{yy} = 6864.61 - 5255.01 = 1609.60$$

$$MSt' = \frac{t'_{yy}}{t-1} = \frac{1609.60}{3} = 536.53$$

$$MSe' = \frac{e'_{yy}}{t(r-1)-1} = \frac{5255.01}{19} = 276.58$$

ANCOVA TABEL

S.O.V	d.f	Sum of squares and sum of cross Products(SS and SCP)				Adjusted SS	MS	F
		xx	xy	yy	df'’	SS’		
Treatments	3	365.46	451.21	2163.13	3	1609.60	536.53	1.94
Error	20	361.50	496.83	5937.83	19	276.58	276.58	
Total	23	726.96	948.04	8100.96				
Treat+Error	23	726.96	948.04	8100.96	22			