زانكۆی سەلاحەدین – هەولێر
Salahaddin Universty-Erbil

Correlation Used to Describe the Relation Between Variables

Research Project
Submitted to the Department of Mathematics in partial fulfillment of the requirements for the degree of BSc. In MATHEMATICS.

Prepared by:
Mobin Bahjat Dasko

Supervised by:
Assist. prof. Dr. Ivan Subhi Letif

April-2023

Certification of the supervisors

I certify that this report was prepared under my supervision at the Department of Mathematics / College of Education / Salahaddin University-Erbil in partial fulfillment of the requirements for the degree of Bachelor of philosophy of Science in Mathematics.

Signature

Supervisor: **Dr. Ivan Subhi Latif**

Scientific grade: **Assistant Professor**

Date:07/04/2023

In view of the available recommendations, I forward this report for debate by the examining committee.

Signature:

Name: **Dr. Rashad Raseed Haji**

Scientific grade: **Assistant Professor**

Chairman of the Mathematics Department

Date:07/04/2023

Acknowledgment

I express my deep sense of gratitude and thanks to the almighty ALLAH for providing me with strength, health, patience, wiling and self-confidence to accomplish this study.

I would like give special thanks to my supervisor "Dr. Ivan" for her constant and valuable guidance and encouragement during my research work, her attention, support and timely suggestion were useful and the most needed in the preparation of my bachelor thesis.

My sincere thanks and appreciation are extended to presidency of Salahaddin University, especially the deanery of the college of Education for their facilities to carry out my research work. My deepest thanks go to "Dr. Rashad Rashid" the head of Mathematics Department of the college of Education, further more I wish to thank the staff members of the College of Education, especially the library staff to the College of Education.

I would like give special thanks to family; they always supported me.

Abstract

Liner regression is a simple and versatile analysis, Simple liner regression shows the relationship between a dependent variable and an independent variable, Multiple regression controls for and models the effects of additional independent variables, in to theoretical meaning fullbacks. A useful function of linear regression is to predicted values of the dependent variable: through, such prediction does not causation. linear regression is available for a number of quantitative methods, including cross-sectional surveys, longitudinal surveys, and experimental designs.

Table of Contents

List of Figure

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the “statistical significance” of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship. (Alan, 1993)

Regression techniques have long been central to the field of economic statistics (“econometrics”). Increasingly, they have become important to lawyers and legal policy makers as well. Regression has been offered as evidence of liability of the Civil Rights Act of 1964, as evidence of racial bias in death penalty litigation, as evidence of damages in contract actions, as evidence of violations under the Voting Rights Act, and as evidence of damages in antitrust litigation, among other things. (Alan, 1993)

I will provide an overview of the most basic techniques of regression analysis how they work, what they assume,

and how they may go awry when key assumptions do not hold. To make the discussion concrete. Generalized least squares. The research is limited to the assumptions, mechanics, and common difficulties with single equation, ordinary least squares regression. (Alan, 1993)

2.1 History of Regression

The earliest form of regression was the method of least squares, which was published by Legendre in 1805, and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss-Markov theorem. (A.M.Legendre, 1805) (Garolo, 1825) (Angrist & Pischke, 2008)

The term "regression" was coined by Francis Galton in the 19th century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean. For Galton, regression had only this biological meaning, but his work was later extended by Yule and Karl Pearson to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821. (Fisher, 1922) (Ronald, 1954) (Aldirch, 2005)

In the 1950s and 1960s, economists used electromechanical desk "calculators" to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression. (Ramcharan, 2006)

Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor (independent variable) or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression. (Ramcharan, 2006)

CHAPTER TWO

BACKGROUND

Definition of Regression 2.1.1: (Brian, 2023)

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

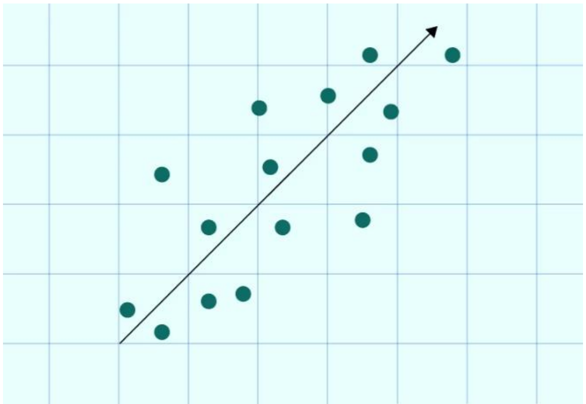

Figure 1 regression between dependent and independent variable

Definition of Dependent Variable 2.1.2: (Brian, 2023)

In a cause-and-effect relationship between two variables, the dependent variable is the effect.

Definition Independent Variable 2.1.3: (Brian, 2023)

In a cause-and-effect relationship between two variables, the dependent variable is the cause.

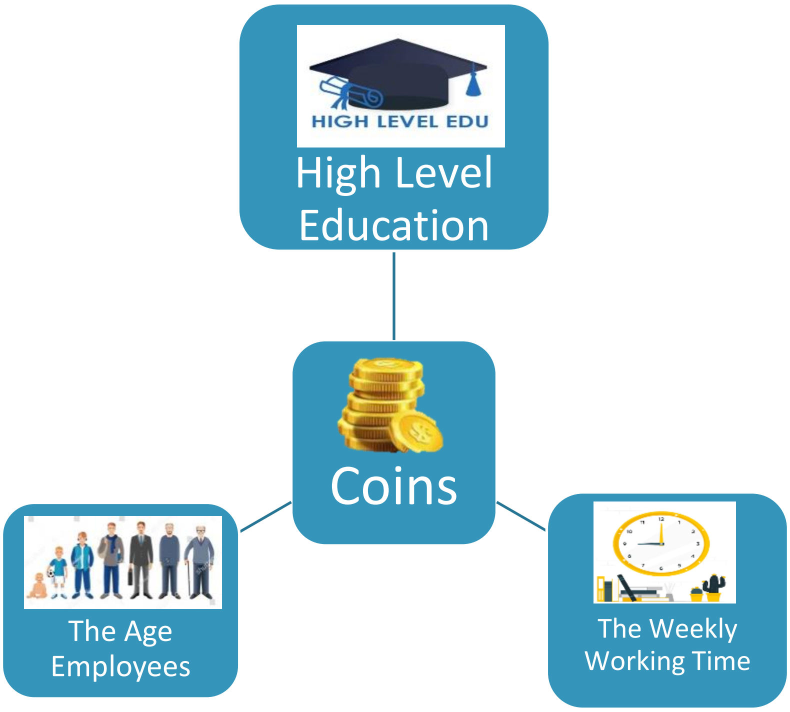

Figure 2 regression between dependent criterion and independent predictors variable

- **Coins** — **The dependent variable**
  Criterion

- **High Level Education**
- **The Weekly Working Time**
- **The Age Employees**

The independent variable
Predictors

Types of Regression (Andriy, 2022)

There are two basic types of regression:

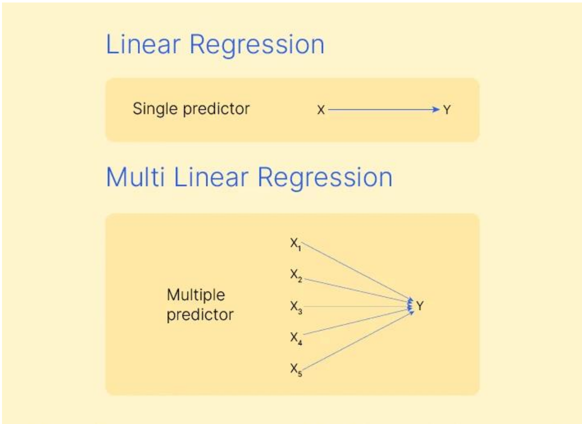

Figure 3 type of regression

1. **Simple Linear Regression:** In this type of regression, there is only one x and one y variable.
2. **Multiple Linear Regression**: In this type of regression, there is one y variable and two or more x variables.

Definition (Simple linear regression)2.1.4: (Andriy, 2022)

Involves using one independent variable (x) to explain the outcome of the dependent variable (y).

The formula for simple linear regression is:

$Y = a + bX + u$ {Equation 2.1.1}

Where,

1. Y = the variable that you are trying to predict (dependent variable).
2. X = the variable that you are using to predict Y (independent variable).

3. a = the y-intercept.

4. b = (beta coefficient) is slope of the explanatory variables.

5. u = the regression residual or error term.

To understand when the appropriate use of linear regression, let's consider the following example:

If we were to assume height as the singular determinant of body weight, we could use the simple linear regression model to predict or explain the impact of a change in height on weight.

Definition (Multiple linear regression) 2.1.5: (Andriy, 2022; Brian, 2023)

Involves using two or more independent variables (x) to explain the outcome of the dependent variable (y).

The formula for multiple linear regression is as follows:

$Y = a + b_1X_1 + b_2X_2 + b_tX_t + u$ {Equation 2.1.2}

1. Y = the variable that you are trying to predict (dependent variable).

2. X = the variable that you are using to predict Y (independent variable).

3. a = the y-intercept.

4. b = (beta coefficient) is slope of the explanatory variables.

5. u = the regression residual or error term.

Multiple linear regression is used when simple linear regression is not enough to account for the multiple real-life factors that influence the outcome of a dependent variable. Let's continue with the previous example involving height and weight. Realistically, height is not the only determinant of weight. There are a lot of different factors that influence a person's weight, such as diet and exercise, and therefore a more realistic model would contain multiple x variables (independent variable).

2.2 The Simple Regression Model

Definition (Regression Line) 2.2.1: (Xin & Xiao, 2009)

Ordinarily, the data set we work with is a sample from a larger population. Because of random fluctuations from sample to sample, exact linearity (conditional means falling exactly in a straight line) hardly ever holds, though we often assume that it holds for the larger population. In fact, in a sample there may not even be any two people with exactly the same measurements on X, so the very concept of conditional means may have little or no meaning for the sample. Therefore, we need a way to derive a line and its equation that does not rely on sample values of conditional means.

The solution to this problem relies on the fact that the sum of squared residuals constructed from {Equation 2.1.1} is defined even when no two cases in the data have the same value of X. There is always one straight line that has a smaller SSresidual than any other straight line. That line is called the regression line, and its equation called the regression equation. The regression equation consists of the regression constant b , also called the Y-intercept, and the regression coefficient by, which is the slope of the line. Because the derivation of the regression equation is based on minimizing the sum of the squared residuals, this method is called ordinary least squares regression or just OLS regression.

2.3 Variance, Covariance, and Correlation

Regression lines can be computed from covariances. Covariances are not usually interpreted, but they are useful for computing both regression coefficients and correlations. Define x_i a deviation score as:

$x_i = x_i - \bar{x}$ {Equation 2.3.1}

meaning that x_i is the deviation of person i's X measurement from the mean of X. A comparable deviation score, y_i , equals $Y_i - \bar{y}$, or the deviation of person i's Y measurement from the mean of Y. The product $x_i y_i$ is the cross-product for person i . The cross-product is positive if person i is above the mean on both X and Y, or below on both. The cross-product is negative if person i is above the mean on one variable and below the mean on the other. The covariance between X and Y is the mean of the cross-products. We denote it $cov(xy)$ thus,

$$cov(xy) = \frac{\sum_{i=1}^{N}(x_i y_i)}{N} \qquad \text{\{Equation 2.3.2\}}$$

where N is the sample size. An alternative formula uses the original values of X and Y rather than deviation scores:

$$cov(xy) = \frac{\Sigma_{i=1}^{N}(X_i Y_i) - \Sigma_{i=1}^{N}(X_i)\Sigma_{i=1}^{N}(Y_i)}{N} \qquad \text{\{Equation 2.3.3\}}$$

The covariance of any variable with itself is the variable's variance. The variance of X we denote by $var(x)$. We have:

$$var(x) = \frac{\Sigma_{i=1}^{N}(x_i\, x_i)}{N} = \Sigma_{i=1}^{N} \frac{x_i^2}{N} \qquad \text{\{Equation 2.3.4\}}$$

or in terms of original values of X rather than deviation

N scores:

$$var(x) = \frac{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}{N^2} \qquad \text{\{Equation 2.3.5\}}$$

Like the covariance, the variance is not usually interpreted. But the variance is the square of an inherently interpretable statistic called the standard deviation. It is the square root of the variance:

$$S_{x=}\sqrt{Var(X)} \qquad \text{\{Equation 2.3.6\}}$$

The standard deviation is a widely used measure of a distribution's variability or spread. As its name implies, the standard deviation is the "standard " measure of spread, because theorists have shown that in normal distributions it is less susceptible than other measures of spread to random fluctuation from sample to sample.

The Pearson correlation coefficient, or simply the correlation, between X and Y is defined as

$$r_{xy} = \frac{Cov(XY)}{\text{sXsY}} \qquad \{\text{Equation 2.3.7}\}$$

The correlation measures the strength of the association between X and Y; there is perfect linear association between X and Y if $r_{xy} = 1$ or $r_{xy} = -1$,

where as $r_{xy} = 0$ if X and Y are linearly independent. The sign of r conveys

the direction of association. If r_{xy} is positive, that means cases above the mean on X tend to be above the mean on Y , and cases below the mean on X tend to be below the mean on Y. If r_{xy} is negative, then cases above the mean on one variable tend to be below the mean on the other.

Since a variance is a type of covariance and a standard deviation is the square root of a variance, {Equation 2.3.7} shows that a correlation is determined entirely by covariances. (Abdi, et al., 2002)

Finding the Regression Line:

Covariances also define the regression coefficient bi. The formula is

$$b_1 = \frac{Cov(XY)}{\text{sXsY}} \qquad \{\text{Equation 2.3.8}\}$$

An alternative formula is

$b_1 = r_{xy} \frac{s_Y}{s_X}$ {Equation 2.3.9}

We can call {Equation 2.3.8} the computing formula and equation {Equation 2.3.9} the definitional formula. {Equation 2.3.9} shows more clearly how by relates to the familiar concepts of correlation and standard deviations, while {Equation 2.3.8} allows us to compute by without taking any square roots. (Abdi, et al., 2002)

2.4 Linear Regression Analysis Using SPSS Statistics

Linear regression is the next step up after correlation. It is used when we want to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable (or sometimes, the outcome variable). The variable we are using to predict the other variable's value is called the independent variable (or sometimes, the predictor variable). For example, you could use linear regression to understand whether exam performance can be predicted based on revision time; whether cigarette consumption can be predicted based on smoking duration; and so forth. If you have two or more independent variables, rather than just one, you need to use multiple regression. (Gallo, 2019)

CHAPTER THREE

Some Examples of Difference Between Dependent and

Independent Variable

Example 3.1: From the given data calculate regression equation taking deviation of item from the mean of x and y series

x	1	3	5	7	9	4	6
y	2	4	3	7	5	6	1

Solution:

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
1	2	-4	-2	16	4	8
3	4	-2	0	4	0	0
5	3	0	-1	0	1	0
7	7	2	3	4	9	6
9	5	4	1	16	1	4
4	6	-2	2	4	4	-4
6	1	1	-3	1	9	-3
Σx =35	Σy =28			ΣX^2 =42	ΣY^2 =42	ΣXY =13

$$\bar{x} = \frac{\Sigma x}{n(x)} = \frac{34}{7} = 5$$

$$\bar{y} = \frac{\Sigma y}{n(y)} = \frac{28}{7} = 4$$

1-regression of x on y

$$x - \bar{x} = bx.y\ (y - \bar{y})$$

$$bx.y = \frac{\Sigma xy}{\Sigma y^2} = \frac{13}{28} = 0.46$$

$$x - \bar{x} = bx.y\ (y - \bar{y})$$

$$x - 5 = 0.46\ (y - 4)$$

$$x = 0.46y - 1.84 + 5$$

$$x = 0.46y + 3.16$$

2-Regress on equation of y on x

$$y - \bar{y} = byx\ (x - \bar{x})$$

$$byx = \frac{\Sigma x.y}{\Sigma x^2} = \frac{13}{42} = 0.31$$

$$y - 4 = 0.31 - (x - 5)$$

$$y = 0.31x - 1.55 + 4$$

$$y = 0.31x + 2.45$$

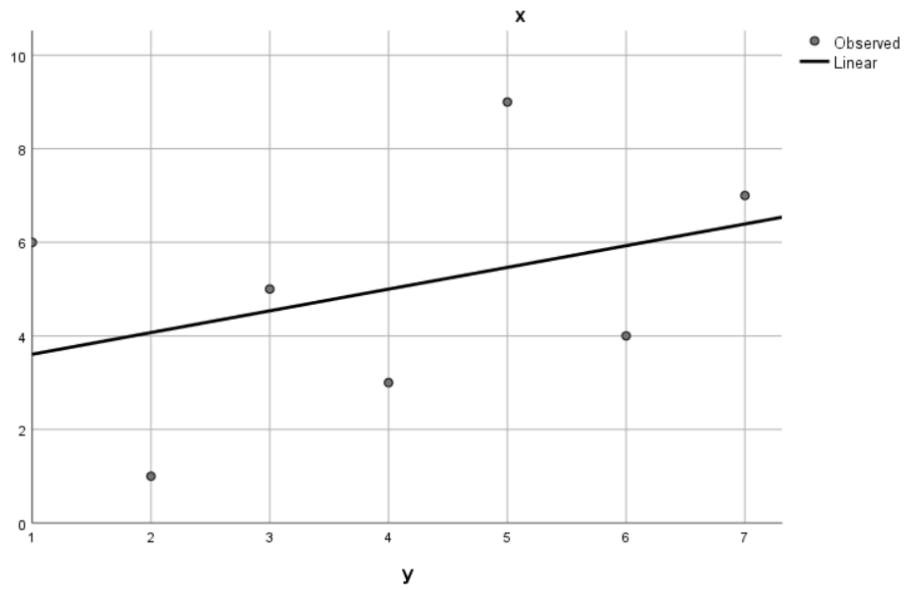

Figure 4 difference between x and y for example 3.1

Example 3.2: From the given data calculate regression equation taking devotion of item from the mean of x and y series.

x	2	5	6	8	9
y	3	9	5	7	6

Solution:

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
2	3	-4	-3	16	9	12
5	9	-1	3	1	9	-3
6	5	0	-1	0	1	0
8	7	2	1	4	1	2
9	6	3	0	9	0	0

$\Sigma x = 30$	$\Sigma y = 30$			$\Sigma X^2 = 30$	$\Sigma Y^2 = 20$	$\Sigma XY = 11$

$$\bar{x} = \frac{\Sigma x}{n(x)} = \frac{30}{5} = 6$$

$$\bar{y} = \frac{\Sigma y}{n(y)} = \frac{30}{5} = 6$$

1-regression of x on y

$x - \bar{x} = bx.y\ (y - \bar{y})$

$$bx.y = \frac{\Sigma xy}{\Sigma y^2} = \frac{11}{20} = 0.55$$

$x - 6 = 0.55(y - 6)$

$x - 5 = 0.55y - 3.3 + 6$

$x = 0.55 + 2.7$

2-Regression of y on x

$y - \bar{y} = byx\ (x - \bar{x})$

$$byx = \frac{\Sigma x.y}{\Sigma x^2} = \frac{11}{30} = 0.55$$

$y - 6 = 0.55(x - 6)$

$y = 0.31x - 1.55 + 4$

$y = 0.55x + 2.7$

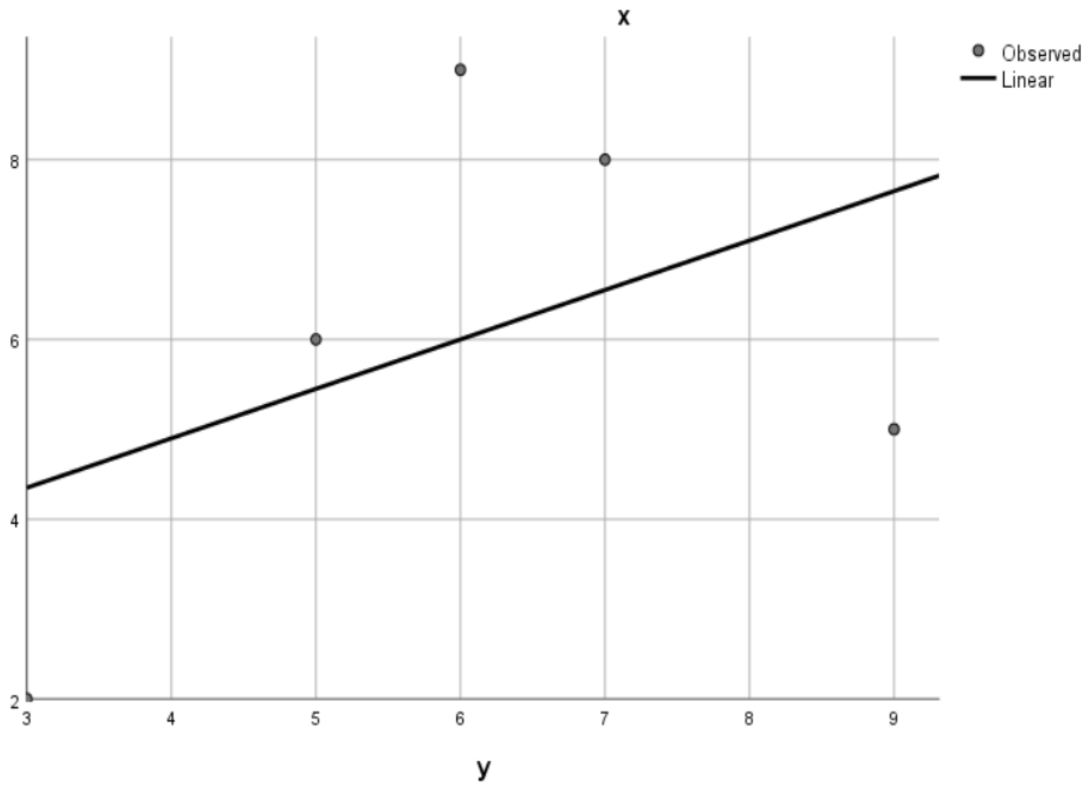

Figure 5 difference between x and y for example 3.2

Example 3.3: From the given data calculate regression equation taking deviation of item from the mean of degree of Mathematics and Average in student 12 grade.

Degree of Math	Average
98	74
90	62
54	68
76	73
60	68
76	67
78	72
98	72
54	63
54	67
84	75
88	75
84	74
88	80
64	66
90	77
60	68
68	63
70	67
66	68
66	69
54	69

66	64
62	63
60	70
86	76
68	76
66	67
78	69

Solution:

Finding regression using spss program

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT degreeofmath
  /METHOD=ENTER average.
```

Regression

[DataSet0]

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	average[b]	.	Enter

a. Dependent Variable: degreeofmath

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.334	.309	11.282

a. Predictors: (Constant), average

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1724.046	1	1724.046	13.544	.001[b]
	Residual	3436.781	27	127.288		
	Total	5160.828	28			

a. Dependent Variable: degreeofmath

b. Predictors: (Constant), average

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-42.195	31.268		-1.349	.188
	average	1.647	.447	.578	3.680	.001

a. Dependent Variable: degreeofmath

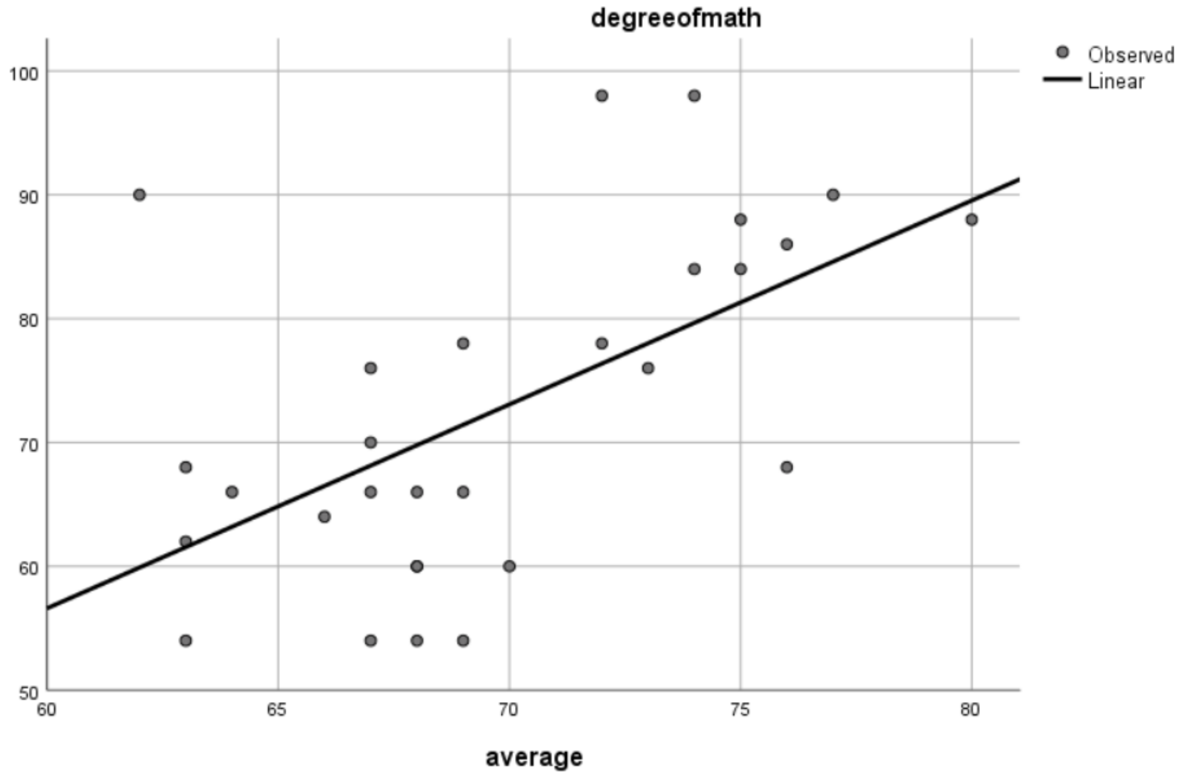

Figure 6 difference between degree of math and average

Example 3.4: From the given data calculate regression equation taking devotion of item from the mean of price of dollar and gold.

Dollar	Gold
153000	395000
153150	400000
152750	398000
155800	405000
156100	410000
156300	409000
156700	410000
155700	407000
157200	415000
159600	420000
159900	421000
157250	420000
157850	430000
157800	430000
157950	430000
157500	425000
160400	430000
160900	430000
162600	435000
164800	437000
165200	450000
163600	450000

165500	450000
159250	445000
162250	448000
162600	445000
162850	445000
162800	445000
164500	445000
167500	450000
167500	450000

Solution:

Finding regression using spss program

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT dollar
  /METHOD=ENTER gold.
```

Regression

[DataSet0]

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	gold[b]	.	Enter

a. Dependent Variable: dollar

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.917[a]	.841	.835	1693.340

a. Predictors: (Constant), gold

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	438525056.8	1	438525056.8	152.935	.000[b]
	Residual	83154620.59	29	2867400.710		
	Total	521679677.4	30			

a. Dependent Variable: dollar

b. Predictors: (Constant), gold

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	67941.751	7441.927		9.130	.000
	gold	.215	.017	.917	12.367	.000

a. Dependent Variable: dollar

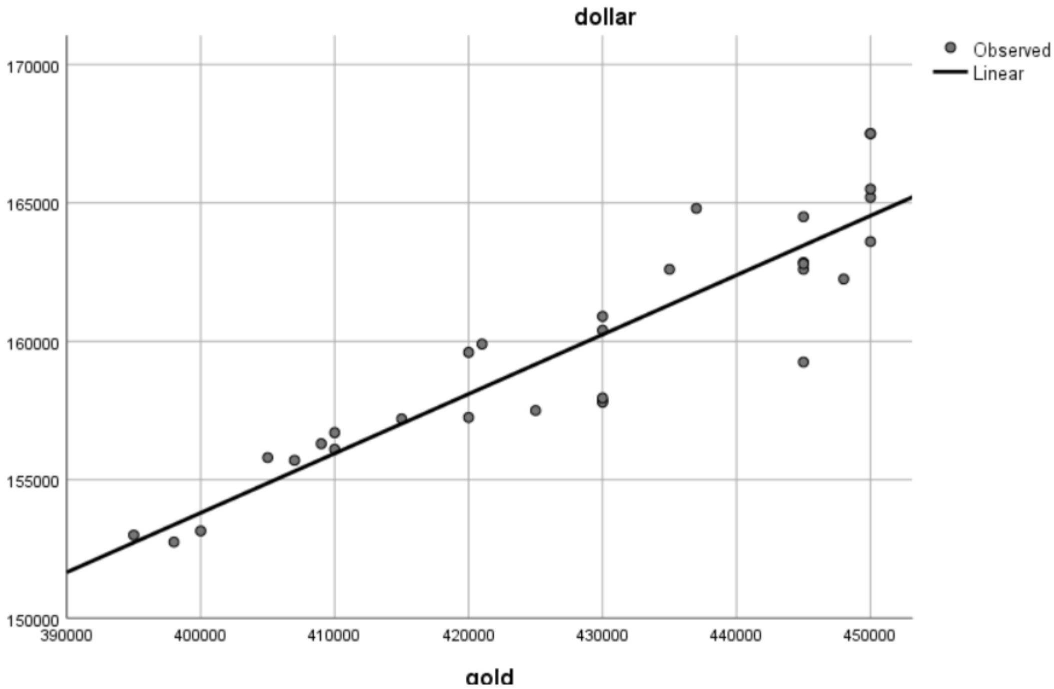

Figure 7 difference between dollar and gold

Reference

A.M.Legendre, 1805. *Nouvelles Methodes Pour la Determination Des Orbites Des Cometes.* Paris: Firmin Didot.

Abdi, H., W.J.Valentin, D.Edelman, B. & Posamentier, M., 2002. *Experimental Design and Research Methods.* Dallas: Richardson.

Alan, O., 1993. An Introduction to Regression Analysis. *Coase -Sandor Working Paper Series in Law and Economics ,* Volume 20, pp. 3-35.

Aldirch, J., 2005. Statistcal Science. *Fisher and Regression ,* 20(4), pp. 401-417.

Andriy, B., 2022. *Investopedia.* [Online]
Available at: https://www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp
[Accessed 21 12 2022].

Angrist, J. & Pischke, J., 2008. *Mostly Harmless Econometrics.* New Jersey : Princeton University Press.

Brian, B., 2023. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/r/regression.asp
[Accessed 31 March 2022].

Fisher, R., 1922. The Goodness of Fit of Regression Formulae and the Distribution of Regression Coefficients. *Journal of the Royal Statistical Society,* 85(4), pp. 597-612.

Gallo, J., 2019. *Aquick and Easy Guide in Using SPSS for Linear Regression Analysis:Learning SPSS for Linear Regression Analysis in 30 minutes Or Less!.* New York: PHD.

Garolo, F., 1825. *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae.* Belgium: H.Dieterich.

Ramcharan, R., 2006. Why Are Economists Obessessed With Them?. *Regressions,* 43(1), pp. 5-17.

Ronald, A., 1954. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd .

Xin, Y. & Xiao, G. S., 2009. *Linear Redression Analysis:Theory and Computing.* London: World Scientific.

پوختە

پاشەکشەی لاینەر شیکارییەکی سادە و گشتگیرە، پاشەکشەی لاینەری سادە پەیوەندی نێوان گۆڕاوێکی وابەستە و گۆڕاوێکی سەربەخۆ نیشان دەدات، پاشەکشەی فرە کاریگەرییەکانی گۆڕاوە سەربەخۆکانی زیادە کۆنترۆڵ دەکات و مۆدێل دەکات، لە بۆ مانای تیۆری فۆڵباکەکان. ئەرکێکی بەسوود لە پاشەکشەی هێڵی بریتییە لە بەها پێشبینیکراوەکانی گۆڕاوە وابەستەکراوەکە: لە ڕێگەی، پێشبینییەکی لەو شێوەیە هۆکار نییە. پاشەکشەی هێڵی بۆ کۆمەڵێک شێوازی چەندایەتی بەردەستە، لەوانە ڕووپێوی بڕبڕەیی، ڕووپێوی درێژخایەن و دیزاینی تاقیکاری.

ملخص

الانحدار الخطي هو تحليل بسيط ومتعدد الاستخدامات ، ويظهر الانحدار الخطي البسيط العلاقة بين متغير تابع ومتغير مستقل ، وضوابط الانحدار المتعدد لتأثيرات المتغيرات المستقلة الإضافية ونماذجها ، إلى المدافعين عن المعنى النظري. تتمثل إحدى الوظائف المفيدة للانحدار الخطي في القيم المتوقعة للمتغير التابع: من خلال ، مثل هذا التنبؤ ليس سببية. الانحدار الخطي متاح لعدد من الأساليب الكمية ، بما في ذلك المسوحات المقطعية والمسوحات الطولية والتصاميم التجريبية.