Chapter 10

10. Linear Regression and Correlation

Introduction

Linear regression and correlation is studying and measuring the linear relationship among two or more variables. When only two variables are involved, the analysis is referred to as simple correlation and simple linear regression analysis, and when there are more than two variables the term multiple regression and partial correlation is used.

Regression Analysis: is a statistical technique that can be used to develop a mathematical equation showing how variables are related.

10.1 The covariance and the correlation coefficient

Covariance:

- The covariance between two random variables is a measure of the nature of between the two.
- The *sign* of the covariance indicates whether the relationship between two dependent random variables is positive or negative.
- Covariance of X and Y measures the co-variability of X and Y together. It is denoted by S_{XY} and given by

$$S_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1} = \frac{\sum XY - n\overline{X}\overline{Y}}{n-1}$$

Correlation Analysis: deals with the measurement of the closeness of the relationship which are described in the regression equation.

We say there is correlation if the two series of items vary together directly or inversely.

Simple Correlation: Suppose we have two variables $X = (X_1, X_2, ...X_n)$ and

$$Y = (Y_1, Y_2, ...Y_n)$$

- When higher values of X are associated with higher values of Y and lower values of X are associated with lower values of Y, then the correlation is said to be positive or direct.

  Examples:

  - Income and expenditure

- Number of hours spent in studying and the score obtained
- Height and weight
- Distance covered and fuel consumed by car.

- When higher values of X are associated with lower values of Y and lower values of X are associated with higher values of Y, then the correlation is said to be negative or inverse.

  Examples:

  - Demand and supply
  - Income and the proportion of income spent on food.

The correlation between X and Y may be one of the following:

1. Perfect positive (slope=1)
2. Positive (slope between 0 and 1)
3. No correlation (slope=0)
4. Negative (slope between -1 and 0)
5. Perfect negative (slope=-1)

The presence of correlation between two variables may be due to three reasons:

1. One variable being the cause of the other. The cause is called "subject" or "independent" variable, while the effect is called "dependent" variable.
2. Both variables being the result of a common cause. That is, the correlation that exists between two variables is due to their being related to some third force.

   Example:

Let X_1= ESLCE result

Y_1= rate of surviving in the University

Y_2= the rate of getting a scholar ship.

Both X_1&Y_1 and X_1&Y_2 have high positive correlation, likewiseY_1 & Y_2 have positive correlation but they are not directly related, but they are related to each other via X_1.

3. Chance: The correlation that arises by chance is called spurious correlation.

   Examples:

   - Price of teff in Addis Ababa and grade of students in USA.
   - Weight of individuals in Ethiopia and income of individuals in Kenya.

Therefore, while interpreting correlation coefficient, it is necessary to see if there is any likelihood of any relationship existing between variables under study.

The correlation coefficient between X and Y denoted by r is given by

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad and \quad the\, short\, cut\, formula\, is$$

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2]\,[n\sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{[\sum X^2 - n\bar{X}^2]\,[\sum Y^2 - n\bar{Y}^2]}}$$

Remark: Always this r lies between -1 and 1 inclusively and it is also symmetric.

Interpretation of r

1.Perfect positive linear relationship ($if\ r = 1$)

2.Some Positive linear relationship ($if\ r\ \ is$ between 0 and 1)

3.No linear relationship ($if\ r = 0$)

4.Some Negative linear relationship ($if\ r\ \ is$ between -1 and 0)

5.Perfect negative linear relationship ($if\ r = -1$)

Examples:

1. Calculate the simple correlation between mid semester and final exam scores of 10 students (both out of 50)

Student	Mid Sem. Exam (X)	Final Sem. Exam (Y)
1	31	31
2	23	29
3	41	34
4	32	35
5	29	25

6	33	35
7	28	33
8	31	42
9	31	31
10	33	34

Solution:

$n = 10, \quad \overline{X} = 31.2, \quad \overline{Y} = 32.9, \quad \overline{X}^2 = 973.4, \quad \overline{Y}^2 = 1082.4$

$\sum XY = 10331, \qquad \sum X^2 = 9920, \qquad \sum Y^2 = 11003$

$$\Rightarrow r = \frac{\sum XY - n\overline{X}\overline{Y}}{\sqrt{[\sum X^2 - n\overline{X}^2]\,[\sum Y^2 - n\overline{Y}^2]}}$$

$$= \frac{10331 - 10(31.2)(32.9)}{\sqrt{(9920 - 10(973.4))\,(11003 - 10(1082.4))}}$$

$$= \frac{66.2}{182.5} = 0.363$$

This means mid semester exam and final exam scores have a slightly positive correlation.

Exercise The following data were collected from a certain household on the monthly income (X) and consumption (Y) for the past 10 months. Compute the simple correlation coefficient.

X:	650	654	720	456	536	853	735	650	536	666
Y:	450	523	235	398	500	632	500	635	450	360

10.2 The rank correlation coefficient

- The above formula and procedure is only applicable on quantitative data, but when we have qualitative data like efficiency, honesty, intelligence, etc we calculate what is called Spearman's rank correlation coefficient as follows:

Steps

i. Rank the different items in X and Y.

ii. Find the difference of the ranks in a pair , denote them by D_i

iii. Use the following formula

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

$$Where \;\; r_s = coefficient \;\; of\ rank\ correlation$$

$$D = the\ difference\ between\ paired\ ranks$$

$$n = the\ number\ of\ pairs$$

Example:

Aster and Almaz were asked to rank 7 different types of lipsticks, see if there is correlation between the tests of the ladies.

Lipstick types	A	B	C	D	E	F	G
Aster	2	1	4	3	5	7	6
Almaz	1	3	2	4	5	6	7

Solution:

X **(R_1)**	**Y** **(R_2)**	**R_1-R_2** **(D)**	**D^2**
2	1	1	1
1	3	-2	4
4	2	2	4
3	4	-1	1
5	5	0	0
7	6	1	1
6	7	-1	1
Total			*12*

$$\Rightarrow r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)} = 1 - \frac{6(12)}{7(48)} = 0.786$$

Yes, there is positive correlation.

10.3 Simple Linear Regression

- Simple linear regression refers to the linear relationship between two variables
- We usually denote the dependent variable by Y and the independent variable by X.
- A simple regression line is the line fitted to the points plotted in the scatter diagram, which would describe the average relationship between the two variables. Therefore, to see the type of relationship, it is

advisable to prepare scatter plot before fitting the model.

- The linear model is:

$$Y = \alpha + \beta X + \varepsilon$$

$$Where: Y = Dependent\ variable$$

$$X = independent\ variable$$

$$\alpha = Regression\ constant$$

$$\beta = regression\ slope$$

$$\varepsilon = random\ disturbance\ term$$

$$Y \sim N(\alpha + \beta X, \sigma^2)$$

$$\varepsilon \sim N(0, \sigma^2)$$

- To estimate the parameters ($\alpha\ and\ \beta$) we have several methods:

  - The free hand method
  - The semi-average method
  - The least square method
  - The maximum likelihood method
  - The method of moments
  - Bayesian estimation technique.

- The above model is estimated by: $\hat{Y} = a + bX$

Where a is a constant which gives the value of Y when X=0 .It is called the Y-intercept. b is a constant indicating the slope of the regression line, and it gives a measure of the change in Y for a unit change in X. It is also regression coefficient of Y on X.

- a and b are found by minimizing $SSE = \sum \varepsilon^2 = \sum (Yi - \hat{Y}_i)^2$

$$Where: Y_i = observed\ value$$

$$\hat{Y}_i = estimated\ value = a + bX_i$$

And this method is known as OLS (ordinary least square)

- Minimizing $SSE = \sum \varepsilon^2$ gives

$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$a = \overline{Y} - b\overline{X}$$

Example 1: The following data shows the score of 12 students for Accounting and Statistics examinations.

a) Calculate a simple correlation coefficient

b) Fit a regression equation of Statistics on Accounting using least square estimates.

c) Predict the score of Statistics if the score of accounting is 85.

	Accounting X	**Statistics Y**	**X^2**	**Y^2**	**XY**
1	74.00	81.00	5476.00	6561.00	5994.00
2	93.00	86.00	8649.00	7396.00	7998.00
3	55.00	67.00	3025.00	4489.00	3685.00
4	41.00	35.00	1681.00	1225.00	1435.00
5	23.00	30.00	529.00	900.00	690.00
6	92.00	100.00	8464.00	10000.00	9200.00
7	64.00	55.00	4096.00	3025.00	3520.00
8	40.00	52.00	1600.00	2704.00	2080.00
9	71.00	76.00	5041.00	5776.00	5396.00
10	33.00	24.00	1089.00	576.00	792.00
11	30.00	48.00	900.00	2304.00	1440.00
12	71.00	87.00	5041.00	7569.00	6177.00
Total	687.00	741.00	45591.00	52525.00	48407.00

Mean	57.25	61.75			

a)

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2} \times \sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r = \frac{12 \times 48407 - 687 \times 741}{\sqrt{12 \times 45591 - 687^2} \times \sqrt{12 \times 52525 - 741^2}}$$

$$r = \mathbf{0.9194}$$

The Coefficient of Correlation (r) has a value of 0.92. This indicates that the two variables are positively correlated (Y increases as X increases).

b)

$$b = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$b = \frac{48407 - 12 \times 57.25 \times 61.75}{45591 - 12 \times (57.25)^2}$$

$$b = 0.9560$$

$$a = \overline{Y} - b\overline{X}$$

$$a = 61.75 - 0.9560 \times 57.25$$

where: $a = 7.0194$

$$\Rightarrow \quad \hat{Y} = 7.0194 + 0.9560X \quad \textit{is the estimated regression line.}$$

c) Insert X=85 in the estimated regression line.

$$\hat{Y} = 7.0194 + 0.9560X$$

$$= 7.0194 + 0.9560(85) = 88.28$$

Exercise: A car rental agency is interested in studying the relationship between the distance driven in kilometer (Y) and the maintenance cost for their cars (X in birr). The following summarized information is given based on samples of size 5.

$$\sum_{i=1}^{5} X_i^{\,2} = 147{,}000{,}000 \qquad \sum_{i=1}^{5} Y_i^2 = 314$$

$$\sum_{i=1}^{5} X_i = 23{,}000\,, \qquad \sum_{i=1}^{5} Y_i = 36, \qquad \sum_{i=1}^{5} X_i\, Y_i = 212{,}000$$

a) Find the least squares regression equation of Y on X

b) Compute the correlation coefficient and interpret it.

c) Estimate the maintenance cost of a car which has been driven for 6 km

- To know how far the regression equation has been able to explain the variation in Y we use a measure called coefficient of determination (r^2)

$$i.e \quad r^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

$$Where \;\; r = the\, simple\, correlation\;\; coefficient.$$

R- Square

- r^2 -value measures the percentage of variation in the values of the dependent variable that can be explained by the variation in the independent variable.
- r^2 -value varies from 0 to 1.
- A value of 0.7654 means that 76.54% of the variance in Y can be explained by the changes in X. the remaining 23.46% of the variation in Y is presumed to be due to random variability.

- r^2 gives the proportion of the variation in Y explained by the regression of Y on X.
- $1 - r^2$ gives the unexplained proportion and is called coefficient of indetermination.

Example: For the above problem (example 1): $r = 0.9194$

$\Rightarrow r^2 = 0.8453 \Rightarrow$ 84.53% of the variation in Y is explained and only 15.47% remains unexplained and it will be accounted by the random term.

10.4 Multiple linear regression and correlations (three-VLRM)

Multiple Linear Regression

So far, we have seen the concept of simple linear regression where a single predictor variable X was used to model the response variable Y. In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables.

Multiple regression analysis is a statistical method or tool used to predict or drive the value a dependent variable based on the values of two or more independent or predictor variables. It is

the simultaneous combination of multiple factors to assess how and to what extent they affect a certain outcome.

The value being predicted is termed dependent variable because its outcome or value depends on the behavior of other variables. The independent variables' value is usually ascertained from the population or sample.

The Model

The primary objective of regression is to develop a regression model, to explain the relationship between two or more variables in a given population.

The multiple linear regression model with k predictor variables $x_1, x_2, ...,x_k$ and a response Y, can be written as:

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ...+ \beta_k x_k + \varepsilon$$

Where

Y=the dependent Variables

$\boldsymbol{x_k}$=the independent Variable

$\boldsymbol{\beta_o}$=coefficients of the slope

β_k=Coefficients independent variables

The above equation has one key feature. It assumes that all individuals are drawn from a single population with common population parameters. The term ε is the residual or random error for individual i and represents the deviation of the observed value of the response for this individual from that expected by the model. These error terms are assumed to have a normal distribution with mean zero and variance $\sigma^{2.}$

$\varepsilon_i = Y_i - \hat{Y}$ Is normally distributed with mean zero and variance σ^2

Assumptions of multiple regression model:

- Multiple Regression model has linear relationship between dependent variable and explanatory variables. Multiple regression technique does not test whether data are linear. On the contrary, it proceeds by assuming that the relationship between the Y and each of X_i's is linear. Hence as a rule, it is prudent to always look at the scatter plots of (Y, X_i), i= 1,

2,…,k. If any plot suggests non-linearity, one may use a suitable transformation to attain linearity.

- Another important assumption is non-existence of multicollinearity the independent variables are not related among themselves. At a very basic level, this can be tested by computing the correlation coefficient between each pair of independent variables.
- The error terms follow normally distribution, i.e.$\boldsymbol{\varepsilon_{i\sim N(0,\delta^2)}}$.homoscedasticity.
- The values of explanatory variable is fixed
- $\epsilon(\varepsilon_i,\varepsilon_{i-1}) = 0$, no autocorrelation among error term
- The error terms and independent variables are independent ,i.e $\epsilon(\varepsilon_i,x_i)$=0
- The rank of explanatory variables is k and where k is number of the parameters or number of column and it should be less than the number of observation (n).

Examples:

• The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage of the lot and a number of other factors.

• The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.