Lecture 6

Linear Regression and Correlation

Simple Regression

- A regression model is a mathematical equation that describes the relationship between two or more variables.
- A *simple regression* model includes only two variables: one independent and one dependent.
- The dependent variable is the one being explained, and the independent variable is the one used to explain the variation in the dependent variable.

Linear Regression

- The relationship between two variables in a regression analysis is expressed by a mathematical equation called a **regression equation** or **model**.
- A regression equation, when plotted, may assume one of many possible shapes, including a straight line.
- A regression equation that gives a straight-line relationship between two variables is called a **linear regression model**; otherwise, the model is called a **nonlinear regression model**.

- The two diagrams in Figure 1 show a linear and a nonlinear relationship between the dependent variable food expenditure and the independent variable income.
- A linear relationship between income and food expenditure, shown in Figure 1*a*, indicates that as income increases, the food expenditure always increases at a constant rate.
- A nonlinear relationship between income and food expenditure, as depicted in Figure 1*b*, shows that as income increases, the food expenditure increases, although, after a point, the rate of increase in food expenditure is lower for every subsequent increase in income.

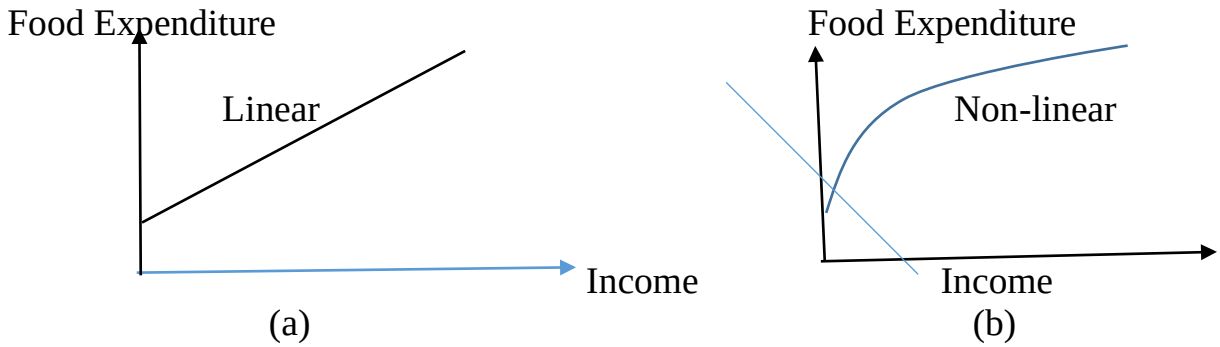


Figure:1 Relationship between food expenditure and income. (a) Linear relationship. (b) Nonlinear relationship.

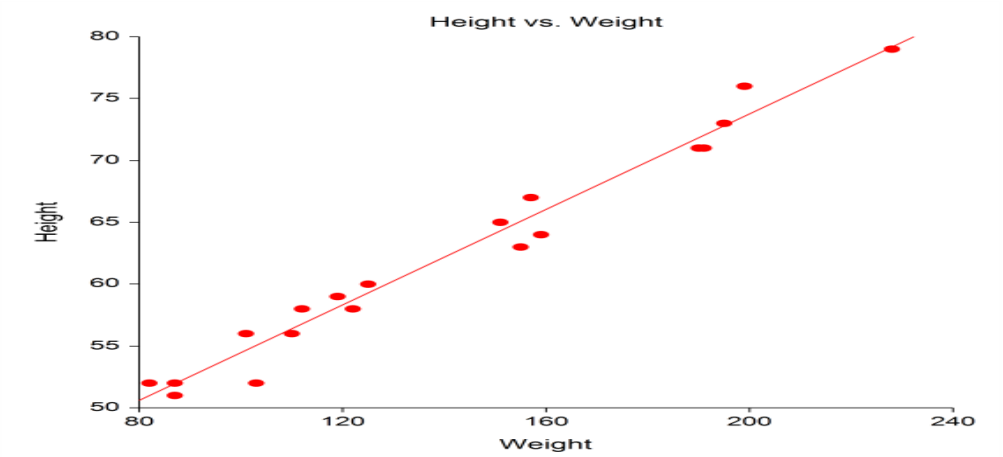
Simple Linear Regression Analysis

In a regression model, the independent variable is usually denoted by *x*, and the dependent variable is usually denoted by *y*. Simple linear regression model is written as

Constant term or *y*-intercept — Slope — Random error term

$$y = \beta_0 + \beta_1 x + \epsilon \text{ ----------(1)}$$

Dependent variable — Independent variable



- In model (1), β_0 and β_1 are the **population parameters**.
- The regression line obtained for model (1) by using the population data is called the **population regression line**.
- The values of β_0 and β_1 in the population regression line are called the **true values of the *y*-intercept and slope**, respectively.
- However, population data are difficult to obtain.
- As a result, we almost always use sample data to estimate model (1).
- The values of the *y*-intercept and slope calculated from sample data on *x* and *y* are called the **estimated values of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$** and are denoted by *a* and *b*, respectively.

Using *a* and *b*, we write the estimated regression model as

$$\hat{y} = a + bx \text{----------(3)}$$

where $\hat{y}$ (read as *y hat*) is the **estimated** or **predicted value of *y*** for a given value of *x*.

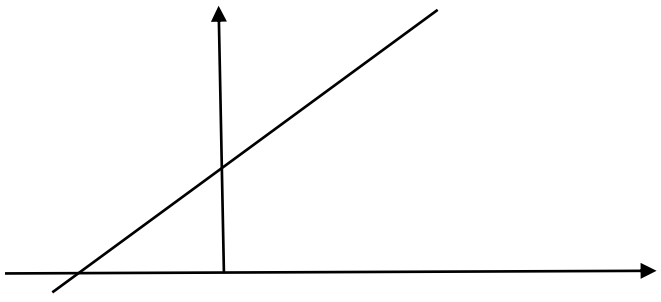
- Equation (3) is called the **estimated regression model**; it gives the **regression of *y* on *x***.

Least Squares Line

For the least squares regression line $\hat{y} = a + bx$

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and $a = \bar{y} - b\bar{x}$



EXAMPLE
Find the least squares regression line for the data on incomes and food expenditures on the seven households given in the following Table. Use income as an independent variable and food expenditure as a dependent variable.

Income, x	55	83	38	61	33	49	67
Food expenditure, y	14	24	13	16	9	15	17

Solution We are to find the values of *a* and *b* for the regression model. Table below shows the calculations required for the computation of *a* and *b*.

Income, x	Food expenditure, y	xy	x^2
55	14	770	3025
83	24	1992	6889
38	13	494	1444
61	16	976	3721
33	9	297	1089
49	15	735	2401
67	17	1139	4489
$\sum x = 386$	$\sum y = 108$	$\sum xy$=6403	$\sum x^2 = 23{,}058$

Thus,

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{6403 - \frac{386 \times 108}{7}}{23058 - \frac{(386)^2}{7}} = \frac{447.5714}{1772.8571} = 0.2525$$

$$\bar{x} = \frac{386}{7} = 55.1429 \qquad \bar{y} = \frac{108}{7} = 15.4286$$

$$a = \bar{y} - b\bar{x} = 15.4286 - (.25252)(55.1429) = 1.5050$$

Thus, our estimated regression model $\hat{y} = a + bx$=1.5050+0.2525x

This regression line is called the least squares regression line. It gives the *regression of food expenditure on income*.

SPSS Result

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.507	2.174		.693	.519
	X	.252	.038	.948	6.664	.001

a. Dependent Variable: y

Interpretation of *a* and *b*

Interpretation of *a*

Consider a household with zero income. Using the estimated regression line obtained in Example, we get the predicted value of *y* for *x* = 0 as

$$\hat{y} = 1.5050 + 0.2525(0) = \$1.5050\ hundreded = 150.50$$

Thus, we can state that a household with no income is expected to spend $150.50 per month on food.

Interpretation of *b*

The value of *b* in a regression model gives the change in *y* (dependent variable) due to a change of one unit in *x* (independent variable).

Note that when *b* is positive, an increase in *x* will lead to an increase in *y*, and a decrease in *x* will lead to a decrease in *y*. In other words, when *b* is positive, the movements in *x* and *y* are in the same direction. Such a relationship between *x* and *y* is called a **positive linear relationship**. The regression line in this case slopes upward from left to right.

On the other hand, if the value of *b* is negative, an increase in *x* will lead to a decrease in *y*, and a decrease in *x* will cause an increase in *y*. The changes in *x* and *y* in this case are in opposite directions. Such a relationship between *x* and *y* is called a **negative linear relationship**. The regression line in this case slopes downward from left to right.

Reliability Measures of Estimating Equation

Standard Error of Estimate

To measure the reliability of the estimating equation, statisticians have developed the standard error of estimate. This standard error is symbolized s_e and is similar to the standard deviation, in that both are measures of dispersion. The standard deviation is used to measure the dispersion of a set of observations about the mean. The standard error of estimate, on the other hand, measures the variability, or scatter, of the observed values around the regression line.

The standard error may be defined as follows:

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

where

- y = values of the dependent variable
- $\hat{y}$ = estimated values from the estimating equation that correspond to each y valueP
- number of data points used to fit the regression line.

Example: Let the estimated regression equation is

$$\hat{y} = 1.5050 + 0.2525x$$

To calculate s_e for this problem, we must determine the value of $\sum(y-\hat{y})^2$. We have done this in the following table:

x	y	$\hat{y}$=1.5050+ .2525x	$(y-\hat{y})^2$
55	14	15.39	1.94
83	24	22.46	2.36
38	13	11.1	3.61
61	16	16.91	0.82
33	9	9.84	0.70
49	15	13.88	1.26
67	17	18.42	2.02
			12.72

Thus, $s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{12.72}{7-2}} = \sqrt{2.54} = 1.59$

Interpreting the Standard Error of Estimate

As was true of the standard deviation, the larger the standard error of estimate, the greater the scattering (or dispersion) of points around the regression line. Conversely, if $s_e = 0$, we expect the estimating equation to be a "perfect" estimator of the dependent variable. In that case, all the data points lie directly on the regression line, and no points would be scattered around it.

Coefficient of Determination

The coefficient of determination is the primary way we can measure the extent, or strength, of the association that exists between two variables, X and Y. Statisticians interpret the coefficient of determination by looking at the amount of the variation in Y that is explained by the regression line. The coefficient of determination is defined by

$$r^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

We can use the following Table to calculate coefficient of determination:

x	y	$\hat{y}$=1.5050+ .2525x	$(y - \hat{y})^2$	$(y - \bar{y})^2$
55	14	15.39	1.94	2.04
83	24	22.46	2.36	73.44
38	13	11.10	3.61	5.90
61	16	16.91	0.82	0.32
33	9	9.84	0.70	41.34
49	15	13.88	1.26	0.18
67	17	18.42	2.02	2.46
Total	108		12.72	125.71

$$\bar{y} = \frac{\sum y}{n} = \frac{108}{7} = 15.43$$

$$r^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{12.72}{125.71} = 1 - 0.10 = 0.90$$

Thus, we can conclude that the variation in income (the independent variable X) explains 90 percent of the variation in food expenditure (the dependent variable Y).

SPSS Result

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	Sig. F Change
1	.948[a]	.899	.879	1.595	.899	44.410	1	.001

Correlation

- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
- Only concerned with strength of the relationship
- No causal effect is implied

Correlation Coefficient

- The population correlation coefficient ρ (rho) measures the strength of the association between the variables
- The sample correlation coefficient r is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

Features of ρ and r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}}$$

The value of correlation coefficient r lies between -1 to +1.

Interpretation of Correlation Coefficient

+ r values	Positive	-r values	Negative
1.0	**Perfect**	**-1.0**	**Perfect**
0.8 to 0.99	**Very Strong**	**-0.8 to -0.99**	**Very Strong**
0.6 to 0.79	**Strong**	**-0.6 to -0.79**	**Strong**
0.4 to 0.59	**Moderate**	**-0.4 to -0.59**	**Moderate**
0.2 to 0.39	**Weak**	**-0.2 to -0.39**	**Weak**
0.01 to 0.19	**Very Week**	**-0.01 to -0.19**	**Very Weak**
0	**No Linear Relationship**		

EXAMPLE

Find the correlation between incomes and food expenditures on the seven households given in the following Table.

Income, x	55	83	38	61	33	49	67
Food expenditure, y	14	24	13	16	9	15	17

Solution

The computing formula for Karl Pearson 's correlation coefficient is

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}}$$

Let us make a table to calculate correlation coefficient

x	y	x^2	y^2	xy
55	14	3025.00	196.00	770.00
83	24	6889.00	576.00	1992.00
38	13	1444.00	169.00	494.00
61	16	3721.00	256.00	976.00
33	9	1089.00	81.00	297.00
49	15	2401.00	225.00	735.00
67	17	4489.00	289.00	1139.00
$\sum x$ =386	$\sum y$ =108	$\sum x^2$ =23058.00	$\sum y^2$ =1792.00	$\sum xy$ =6403.00

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}}$$

$$= \frac{6403 - \frac{386*108}{7}}{\sqrt{\{23058 - \frac{(386)^2}{7}\}\{1792 - \frac{(108)^2}{7}\}}}$$

$$= \frac{6403-5955}{\sqrt{(23058-21285)(1792-1666)}} = \frac{448}{\sqrt{1773*126}} = \frac{448}{\sqrt{223398}} = \frac{448}{472.65} = 0.95$$

Conclusion: There exist a very strong positive relationship between x and y.

Reliability Measures of Estimating Equation

Standard Error of Estimate

To measure the reliability of the estimating equation, statisticians have developed the standard error of estimate. This standard error is symbolized s_e and is similar to the standard deviation, in that both are measures of dispersion. The standard deviation is used to measure the dispersion of a set of observations about the mean. The standard error of estimate, on the other hand, measures the variability, or scatter, of the observed values around the regression line.

The standard error may be defined as follows:

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

where

- y = values of the dependent variable
- $\hat{y}$ = estimated values from the estimating equation that correspond to each y value
- number of data points used to fit the regression line.

Example: Let the estimated regression equation is

$$\hat{y} = 1.5050 + 0.2525x$$

To calculate s_e for this problem, we must determine the value of $\sum (y - \hat{y})^2$. We have done this in the following table:

x	y	$\hat{y}$=1.5050+ .2525x	$(y-\hat{y})^2$
55	14	15.39	1.94
83	24	22.46	2.36
38	13	11.1	3.61
61	16	16.91	0.82
33	9	9.84	0.70
49	15	13.88	1.26
67	17	18.42	2.02
			12.72

Thus, $s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{12.72}{7-2}} = \sqrt{2.54} = 1.59$

Interpreting the Standard Error of Estimate

As was true of the standard deviation, the larger the standard error of estimate, the greater the scattering (or dispersion) of points around the regression line. Conversely, if $s_e = 0$, we expect the estimating equation to be a "perfect" estimator of the dependent variable. In that case, all the data points lie directly on the regression line, and no points would be scattered around it.

Coefficient of Determination

The coefficient of determination is the primary way we can measure the extent, or strength, of the association that exists between two variables, X and Y. Statisticians interpret the coefficient of determination by looking at the amount of the variation in Y that is explained by the regression line. The coefficient of determination is defined by

$$r^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

We can use the following Table to calculate coefficient of determination:

x	y	$\hat{y}$=1.5050+ .2525x	$(y-\hat{y})^2$	$(y-\bar{y})^2$
55	14	15.39	1.94	2.04
83	24	22.46	2.36	73.44
38	13	11.10	3.61	5.90
61	16	16.91	0.82	0.32
33	9	9.84	0.70	41.34
49	15	13.88	1.26	0.18
67	17	18.42	2.02	2.46
Total	108		12.72	125.71

$$\bar{y} = \frac{\sum y}{n} = \frac{108}{7} = 15.43$$

$$r^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{12.72}{125.71} = 1 - 0.10 = 0.90$$

Thus, we can conclude that the variation in income (the independent variable X) explains 90 percent of the variation in food expenditure (the dependent variable Y).

SPSS Result

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	Sig. F Change
1	.948[a]	.899	.879	1.595	.899	44.410	1	.001

Correlation

- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
- Only concerned with strength of the relationship
- No causal effect is implied

Correlation Coefficient

- The population correlation coefficient ρ (rho) measures the strength of the association between the variables
- The sample correlation coefficient r is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

Features of ρ and r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}}$$

The value of correlation coefficient r lies between -1 to +1.

Interpretation of Correlation Coefficient

+ r values	Positive	-r values	Negative
1.0	Perfect	-1.0	Perfect
0.8 to 0.99	Very Strong	-0.8 to -0.99	Very Strong
0.6 to 0.79	Strong	-0.6 to -0.79	Strong
0.4 to 0.59	Moderate	-0.4 to -0.59	Moderate
0.2 to 0.39	Weak	-0.2 to -0.39	Weak
0.01 to 0.19	Very Week	-0.01 to -0.19	Very Weak
0	No Linear Relationship		

EXAMPLE

Find the correlation between incomes and food expenditures on the seven households given in the following Table.

Income, x	55	83	38	61	33	49	67
Food expenditure, y	14	24	13	16	9	15	17

Solution

The computing formula for Karl Pearson ‘s correlation coefficient is

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}}$$

Let us make a table to calculate correlation coefficient

x	y	x^2	y^2	xy
55	14	3025.00	196.00	770.00
83	24	6889.00	576.00	1992.00
38	13	1444.00	169.00	494.00
61	16	3721.00	256.00	976.00
33	9	1089.00	81.00	297.00
49	15	2401.00	225.00	735.00
67	17	4489.00	289.00	1139.00
$\sum x$ =386	$\sum y$ =108	$\sum x^2$ =23058.00	$\sum y^2$ =1792.00	$\sum xy$ =6403.00

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}}$$

$$= \frac{6403 - \frac{386*108}{7}}{\sqrt{\{23058 - \frac{(386)^2}{7}\}\{1792 - \frac{(108)^2}{7}\}}}$$

$$= \frac{6403 - 5955}{\sqrt{(23058 - 21285)(1792 - 1666)}} = \frac{448}{\sqrt{1773*126}} = \frac{448}{\sqrt{223398}} = \frac{448}{472.65} = 0.95$$

Conclusion: There exist a very strong positive relationship between x and y.