# Linear Regression and Correlation

The simple linear regression consider a single regressor variable or predictor variable X and a dependent or response variable Y.

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad -(1)$$

Where, the intercept $\beta_0$ and the ~~$\beta$~~ slope $\beta_1$ are unknown regression coefficients. $\varepsilon$ is a random error.

## Least Square Method-

The method of least square uses to estimate the regression coefficient.

Using equation (1), we may express the n observations in the sample as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

and the sum of the squares of the ~~individual~~ deviations of the observations from the true regression line is

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators of $\beta_0$ and $\beta_1$, say $\hat{\beta}_0$ and $\hat{\beta}_1$ must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Simplifying the above two equation gives

$$\left.\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned}\right\} \quad (2)$$

Equations (2) are called the least square normal equations. The Solution to the normal eqn results in the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Least Squares Estimates-

The least squares estimates of the intercept and slope in the simple linear regresion model are

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

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

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

The fitted estimated regression line -

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Q find the line of regression on the basis of the following data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| X: | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Y: | 2 | 4 | 7 | 6 | 5 | 6 | 5 |

Soln

| X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|
| 1 | 2 | 1 | 4 | 2 |
| 2 | 4 | 4 | 16 | 8 |
| 3 | 7 | 9 | 49 | 21 |
| 4 | 6 | 16 | 36 | 24 |
| 5 | 5 | 25 | 25 | 25 |
| 6 | 6 | 36 | 36 | 36 |
| 7 | 5 | 49 | 25 | 35 |
| $\Sigma X = 28$ | $\Sigma Y = 35$ | $\Sigma X^2 = 140$ | $\Sigma Y^2 = 191$ | $\Sigma XY = 151$ |

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{7} y_i = \frac{1}{7} \times 35 = 5$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{7} x_i = \frac{1}{7} \times 28 = 4$$

$$\hat{\beta}_1 = \frac{\Sigma y_i x_i - \frac{(\Sigma y_i)(\Sigma x_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

$$= \frac{151 - \frac{(35)\times(28)}{7}}{140 - \frac{(28)^2}{7}}$$

$$= \frac{151 - 140}{140 - 112}$$

$$= \frac{11}{28}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$= 5 - \frac{11}{28} \times 4$$

$$= 5 - \frac{44}{28}$$

$$= \frac{140 - 44}{28}$$

$$= \frac{\cancel{96}\ \cancel{48}\ 24}{\cancel{28}\ \cancel{14}\ 7}$$

$$= \frac{24}{7}$$

The regression line is

$$\hat{y} = \frac{24}{7} + \frac{11}{28}x \quad \text{or} \quad \hat{y} = 3.42 + 0.39x$$

Q Fit a least square quadratic fit to the following data set.

X : 0 1 2 3 4
Y : 1.0 1.5 1.5 2.5 3.5

Soln

We have to fit the quadratic fit $y = a + bx + cx^2$

The three normal eqn for determining the value of numerical Constants a, b and c are :

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

| $x$ | $y$ | $x^2$ | $x^3$ | $x^4$ | $xy$ | $x^2y$ |
|---|---|---|---|---|---|---|
| 0 | 1.0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1.5 | 1 | 1 | 1 | 1.5 | 1.5 |
| 2 | 1.5 | 4 | 8 | 16 | 3 | 6.0 |
| 3 | 2.5 | 9 | 27 | 81 | 7.5 | 22.5 |
| 4 | 3.5 | 16 | 64 | 256 | 14 | 56.0 |
| 10 | 10 | 30 | 100 | 354 | 26 | 86 |

Substituting the values obtained from the in the normal eqn

$$10 = 5a + 10b + 30c \quad -(1)$$
$$26 = 10a + 30b + 100c \quad -(2)$$
$$86 = 30a + 100b + 354c \quad -(3)$$

Multiplying eqn (1) by 2 and then subtract from eqn(2) we get

$$6 = 10b + 40c \quad -(4)$$

Next multiplying eqn (2) by 3 and then subtracting from eqn(3)

$$8 = 10b + 54c \quad -(5)$$

Now subtracting equ$^n$ (4) from equ$^n$ (5) we get

$$2 = 14C$$

$$\Rightarrow C = \frac{2}{14} = \frac{1}{7} = 0.143$$

Substituting $c = \frac{1}{7}$ in equ$^n$ (4) we get

$$6 = 10b + 40 \times \frac{1}{7}$$

$$b = 0.0286$$

Substituting the values of b and c in equ$^n$ (1) we get

$$10 = 5a + \frac{10}{35} + \frac{30}{7}$$

$$\Rightarrow 5a = 10 - \frac{2}{7} - \frac{30}{7}$$

$$\Rightarrow a = \frac{38}{35} = 1.0854$$

Thus, $a = 1.0857$, $b = 0.0286$ and $c = 0.143$

Hence, the non-linear fitted line is

$$y = 1.0857 + 0.0286x + 0.143x^2$$

## Regression On Transformed Variable -

We occasionally find that the straight-line regression model $Y = \beta_0 + \beta_1 x + \varepsilon$ is inappropriate because the true regression function is nonlinear. In some ~~oth~~ of these situations, a nonlinear function can be expressed as a straight line by using a suitable transformation. Such nonlinear models are called intrinsically linear.

As an example of a nonlinear model that is intrinsically linear, consider the exponential function

$$Y = \beta_0 e^{\beta_1 x} \varepsilon$$

The function is intrinsically linear, since it can be transformed to a straight line by a logarithmic transformation.

$$\log Y = \log \beta_0 + \beta_1 x + \log \varepsilon$$

Q Fit an exponential curve of the form $Y = Ae^{Bx}$ for the following data:

| | | | | |
|---|---|---|---|---|
| X : | 1 | 2 | 3 | 4 |
| Y : | 7 | 11 | 17 | 27 |

Nok –

1. Fitting of a power curve $Y = ax^b$ to a set n points.
$\llcorner(1)$

Taking logarithm of each side, we get

$$\log Y = \log a + b \log X$$

$$\Rightarrow U = A + bV$$

Where, $U = \log Y$, $A = \log a$ and $V = \log X$

This is a linear eqn in V and U

Normal equations for estimating A and B are;

$$\Sigma U = nA + b\Sigma V$$

and $\Sigma UV = A\Sigma V + b\Sigma V^2$

These equations can be solved for A and b and consequently, we get

$a = \text{antilog}(A)$.

With the values of 'a' and 'b' so obtained (1) is the curve of best fit to the set of n points.

2. Fitting of Exponential Curves (i) $Y = ab^X$, (ii) $ae^{bX}$ to a set of n points.

(i) $Y = ab^X$

Taking logarithm of each side, we get

$$\log Y + \log a + X \log b$$

$$\Rightarrow U + A + BX$$

Where $U = \log Y$, $A = \log a$ and $B = \log b$

This is a linear eqn in X and U

The normal equations for estimating A and B are:

$$\Sigma U = nA + B\Sigma X$$

$$\text{and } \Sigma XU + A\Sigma X + B\Sigma X^2$$

Solving these equations for A and B, we finally get

$$a = \text{antilog}(A) \text{ and } B = \text{antilog}(B)$$

With these values of 'a' and 'b' (i) is the curve of the best fit to the given set of n points.

(ii) $y = ae^{bx}$

$$\log y = \log a + bx\log e$$

$$= \log a + (b\log e)x$$

$$\Rightarrow U = A + BX$$

Where, $U = \log y$, $A = \log a$ and $B = b\log e$

This is a linear equation in X and U, and the normal equations are:

$$\Sigma U = nA + B\Sigma X$$

$$\text{and } \Sigma XU = A\Sigma X + B\Sigma X^2$$

From these we find A and B consequently $a = \text{Antilog } A$ and $b = \frac{B}{\log e}$.

Soln Taking logarithm of each side we get

$$\log Y = \log A + BX \log e$$

$$\Rightarrow U = a + bX$$

Where, $U = \log Y$, $a = \log A$ and $b = B \log e$

This is a linear equation in X and U and the normal equations are:

$$\left.\begin{aligned} \sum U &= na + b\sum X \\ \sum XU &= a\sum X + b\sum X^2 \end{aligned}\right\} - (1)$$

| X | Y | $U=\log Y$ | XU | $X^2$ |
|---|---|---|---|---|
| 1 | 7 | 0.845 | 0.845 | 1 |
| 2 | 11 | 1.041 | 2.082 | 4 |
| 3 | 17 | 1.230 | 3.690 | 9 |
| 4 | 27 | 1.431 | 5.724 | 16 |
| 10 | | 4.547 | 13.341 | 30 |

Substitute these values in equn (1) we get

$$4.547 = 4a + 10b \quad -(3$$

$$13.341 = 10a + 30b \quad -(4)$$

After solving equn (3) and (4) we get

$$a = 0.15 \quad \text{and} \quad b = 0.3947$$

$$A = \text{antilog}(a) = \text{antilog}(0.15) = 1.413$$

$$B = \frac{b}{\log e} = \frac{0.3947}{0.4342} = 0.9090$$

The fitting curve is $Y = 1.413\, e^{(0.9090)X}$

# Correlation-

If the change in one variable affects a change in the other variable, the variables are said to be correlated.

If the two variables deviate in the same direction, i.e., if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive.

If the two variables deviate in the opposite direction i.e, if the increase (or decrease) in one result in a corresponding decrease (or increase) in the other, correlation is said to be diverse or negative.

For example - The income and expenditure is positivly correlated.
The price and demand is negatively correlated.

## Karl Pearson's Coefficient of Correlation-

$$r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

where, $Cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}$

$$\sigma_x^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2$$

$$\sigma_y^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2$$

<u>Note</u>:- Correlation coefficient always lies between -1 and +1.
If $r = +1$, the correlation is perfect and positive
If $r = -1$, the correlation is perfect and negative.

Q Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| X: | 65 | 66 | 67 | 67 | 68 | 69 | 70 | 72 |
| Y: | 67 | 68 | 65 | 68 | 72 | 72 | 69 | 71 |

Soln

| X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|
| 65 | 67 | 4225 | 4489 | 4355 |
| 66 | 68 | 4356 | 4624 | 4488 |
| 67 | 65 | 4489 | 4225 | 4355 |
| 67 | 68 | 4489 | 4624 | 4556 |
| 68 | 72 | 4624 | 5184 | 4896 |
| 69 | 72 | 4761 | 5184 | 4968 |
| 70 | 69 | 4900 | 4761 | 4830 |
| 72 | 71 | 5184 | 5041 | 5112 |
| 544 | 552 | 37028 | 38132 | 37560 |

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{544}{8} = 68 \quad , \quad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{552}{8} = 69$$

$$r(XY) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma XY - \bar{X}\bar{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \bar{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8}\times 37560 - 68\times 69}{\sqrt{\left\{\frac{37028}{8} - (68)^2\right\}\left\{\frac{38132}{8} - (69)^2\right\}}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628.5 - 4624)(4766.5 - 4761)}}$$

$$= \frac{3}{\sqrt{4.5\times 5.5}} = 0.603$$

# Confidence Intervals

Under the assumption that the observations are normally and independent distributed, a $100(1-\alpha)\%$ confidence interval on the slope $\beta_1$ in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2, n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2, n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Similarly, a $100(1-\alpha)\%$ confidence interval on the intercept $\beta_0$ is

$$\hat{\beta}_0 - t_{\alpha/2, n-2}\sqrt{\sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2, n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

Q We will find a 95% confidence interval on the slope of the regression line using data $\hat{\beta}_1 = 14.947$, $S_{xx} = 0.68088$ and $\hat{\sigma}^2 = 1.18$, $n = 20$.

Soln From eqn -

$$\hat{\beta}_1 - t_{\alpha/2, n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2, n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$14.947 - t_{0.025, 18}\sqrt{\frac{1.18}{0.68088}} \le \hat{\beta}_1 \le 14.947 + t_{0.025, 18}\sqrt{\frac{1.18}{0.68088}}$$

$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \le \beta_1 \le 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

This simplifies to

$$12.181 \le \beta_1 \le 17.713$$

Interpretation - The CI does not include zero, so there is strong evidence (at $\alpha = 0.05$) that the slope is not zero.

# Hypothesis Tests in Simple Linear Regression -

Suppose we wish to test the hypothesis that the slope equals a constant, say $\beta_{1,0}$. The appropriate hypotheses are

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_1 : \beta_1 \neq \beta_{1,0}$$

Test statistic -

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \qquad \text{where, } S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

follow the t distribution with n-2 degrees of freedom under $H_0 : \beta_1 = \beta_{1,0}$. We would reject $H_0 : \beta_1 = \beta_{1,0}$ if

$$|t_0| > t_{\alpha/2, n-2}$$

A similar procedure can be used to test hypotheses about the intercept.

$$H_0 : \beta_0 = \beta_{0,0}$$
$$H_1 : \beta_0 \neq \beta_{0,0}$$

Test statistic -
$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$$

and reject the null hypothesis if the computed value of this test statistic, $t_0$, is such that $|t_0| > t_{\alpha/2, n-2}$.

and $S_{xx} = 4152.18$ $\hat{\sigma}^2 = 3.2295$

Q Using the estimated value $\hat{\beta}_1 = 0.903643$ ; test the hypothesis $\beta_1 = 1.0$ against the alternative that $\beta < 1.0$.

Soln

Given - $S_{xx} = 4152.18$

$\sigma = 3.2295$

$n = 33$

$t_{0.05,31} = 1.648$

Null Hypothesis $\beta_1 = 1.0$

Alternative Hypothesis $\beta_1 < 1.0$

Test Statistic $t = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}}$

$$t = \frac{0.903643 - 1.0}{3.2295 / \sqrt{4152.18}}$$

$t = -1.92$

with $n-2 = 31$ degree of freedom

Decision: Since $t > t_{\alpha, n-2}$. We Reject the null hypothesis, Suggesting that it is a strong evidence that $\beta < 1.0$