PAPER IV: Regression Analysis with R software

UNIT II: Simple Regression Model

The term Regression was introduced by Francis Galton and was confirmed by his friend Karl Pearson.

Regression analysis is a statistical technique for investigating and modelling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique.

Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, with a view to estimating and predicting the population mean or average value of the former in terms of the known or fixed values of the latter.

As an example of a problem in which regression analysis may be helpful, suppose that an industrial engineer employed by a soft drink beverage bottler is analysing the product delivery and service operations for vending machines. He suspects that the time required by a route deliveryman to load and service a machine is related to the number of cases of product delivered. The engineer visits 25 randomly chosen retail outlets having vending machines, and the in-outlet delivery time (in minutes) and the volume of product delivered (in cases) are observed for each. The 25 observations are plotted in Figure a. This graph is called a scatter diagram. This display clearly suggests a relationship between delivery time and delivery volume; in fact, the impression is that the data points generally, but not exactly, fall along a straight line. Figure b illustrates this straight-line relationship. If we let y represent delivery time and x represent delivery volume, then the equation of a straight line relating these two variables is

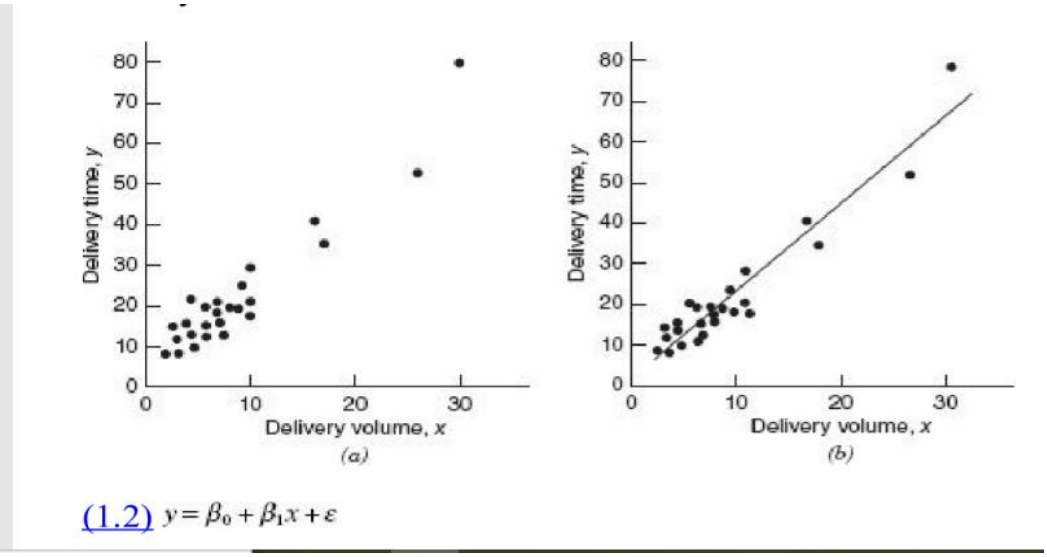
$y = \beta_0 + \beta_1 x$……Eq. 1.1

where β_0 is the intercept and β_1 is the slope. Now the data points do not fall exactly on a straight line, so Eq. (1.1) should be modified to account for this. Let the difference between the observed value of y and the straight line ($\beta_0 + \beta_1 x$) be an error ε. It is convenient to think of ε as a statistical error; that is, it is a random variable that accounts for the failure of the model to fit the data exactly. The error may be made up of the effects of other variables on delivery time, measurement errors, and so forth. Thus, a more plausible model for the delivery time data is

$y = \beta_0 + \beta_1 x + \epsilon$……Eq. 1.2

Scatter diagram for delivery

Figure a: Figure b:



Equation 1.2 is called a linear regression model. Customarily x is called the independent variable and y is called the dependent variable. However, this often causes confusion with the concept of statistical independence, so we refer to x as the predictor or regressor variable and y as the response or dependent

variable. Because Eq. (1.2) involves only one regressor variable, it is called a **simple linear regression model.**

The meaning of the term Linear:

Since we are concerned with Linear Regression model, it is essential to know the meaning of the term "Linear". Linearity means the variable y should be a linear function of the parameters, the $\beta's$.

This simple linear regression coefficients of $y = \beta_0 + \beta_1 x + \epsilon$ eq.(1.2) where the intercept β_0 and the slope β_1 are unknown constants and ε is a random error component. The parameters β_0 and β_1 are usually called regression coefficients. These coefficients have a simple and often useful interpretation. The slope β_1 is the change in the mean of the distribution of y produced by a unit change in x. If the range of data on x includes x = 0, then the intercept β_0 is the mean of the distribution of the response y when x = 0. If the range of x does not include zero, then β_0 has no practical interpretation.

Least - Squares Estimation of the Parameters:

The method of ordinary least squares is used to estimate β_0 and β_1. This OLS method is attributed to Carl Friedrich Gauss, a German Mathematician. The parameters β_0 and β_1 are unknown and must be estimated using sample data. Suppose that we have n pairs of data, say (y1, x1), (y2, x2), …, (yn, xn). These data may result either from a controlled experiment designed specifically to collect the data, from an observational study, or from existing historical records (a retrospective study).

Assumptions of the model:

1. The regression model is linear in the parameters.

2. Values taken by the regressor X are considered fixed in repeated samples.

3. Given the value of X, the mean or expected value of the random disturbance term ϵ $is\ zero. Technically\ the\ conditional\ mean$ value of ϵ is zero. Symbolically we have

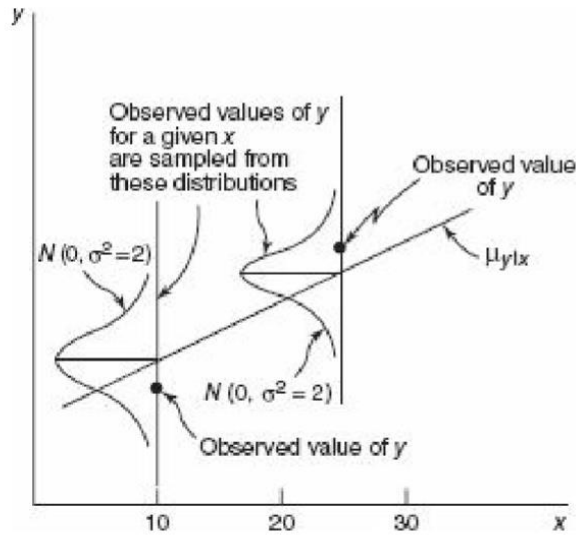
E $(\epsilon_i/X_i) = 0.$

$\Rightarrow$ E $(Y_i / X_i) = \beta_0 + \beta_1 x$

4. Homoscedasticity or equal variance of ϵ_i:

Given the values of X, the variance of ϵ_i is same for all observations. That is the conditional variances of ϵ_i are identical. Symbolically

Var $(\epsilon_i/X_i) = \quad \sigma^2$



5. No autocorrelation between the error terms.

6. Zero covariance between $X_i\ and\ \epsilon_i$.

7. The number of observations n must be greater than the number of parameters to be estimated.

8. The X values in a given sample must not all be same. Technically Var (X) must be a positive number.

9. The regression model should be correctly specified.

10. There is no perfect multicollinearity. That is, there are no relationship among X_i's.

Estimation of β_0 and β_1

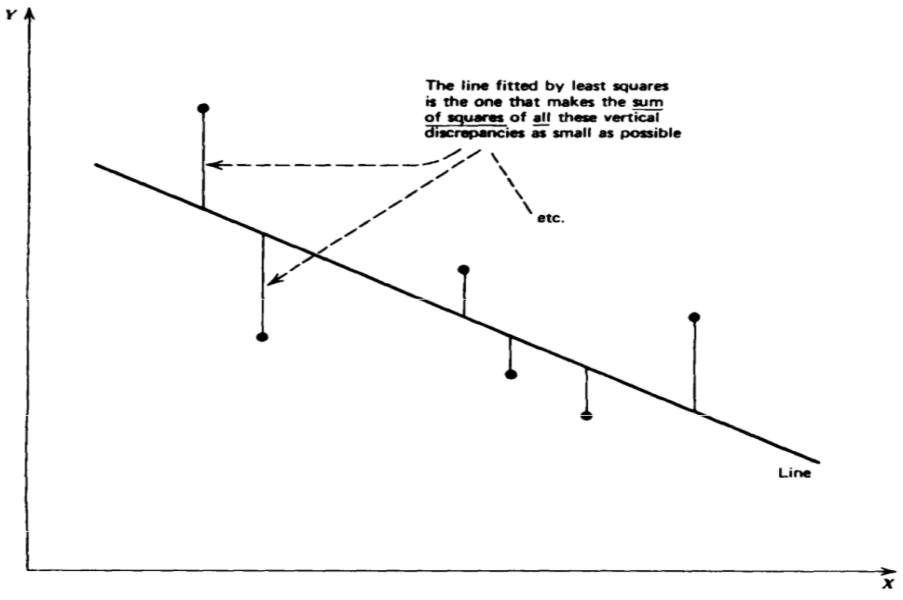
The method of least squares is used to estimate β_0 and β_1. That is, we estimate β_0 and β_1 so that the sum of the squares of the differences between the observations yi and the straight line is a minimum.

From Eq. $y = \beta_0 + \beta_1 x + \epsilon$……Eq. 1.2

we may write $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$……Eq. 1.3

Equation 1.2 may be viewed as a population regression model while Eq.1.3 is a sample regression model, written in terms of the n pairs of data (yi, xi) (i = 1, 2, …, n). Thus, the least-squares criterion is to fix sample regression function in such a way that

$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ is as small as possible.



Let f = $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

The least-squares estimators of β_0 and β_1 must satisfy

$\frac{\partial f}{\partial \beta_0}$ = -2 $\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) = 0$ and

$\frac{\partial f}{\partial \beta_1}$ = -2 $\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) x_i = 0$ and

Here $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are estimated values of β_0 *and* β_1 *from the sample*.

Simplifying these two equations we get,

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) x_i = 0$$

$\Rightarrow$ n $\widehat{\beta_0}$ + $\widehat{\beta_1}$ $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$…….Eq. 1.4(a)

$\widehat{\beta_0}$ $\sum_{i=1}^{n} x_i + \widehat{\beta_1} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$…….Eq. 1.4(b)

Equations 1.4(a and b) are called the least-squares normal equations. The solution to the normal equations is

Multiplying Eq. 1.4 (a) by $\sum_{i=1}^{n} x_i$ *and Eq.* 1.4(b)*by n we get,*

n $\widehat{\beta_0}$ $(\sum_{i=1}^{n} x_i)$+ $\widehat{\beta_1}$ $(\sum_{i=1}^{n} x_i)^2 = (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)$……..1.4(c)

n $\widehat{\beta_0}$ $(\sum_{i=1}^{n} x_i) + n\widehat{\beta_1}$ $(\sum_{i=1}^{n} x_i^2) = n \sum_{i=1}^{n} x_i y_i$…….1.4(d)

Subtracting 1.4(d) from 1.4(c), we get,

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

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

$\widehat{\beta_1} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$

$= \frac{(n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i))/n}{(n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2)/n}$

$= \frac{S_{xy}}{S_{xx}}$ where $S_{xy} = \ (n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i))$/n

$= \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$

And $S_{xx} = (n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2\ )$/n

$= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

⇒ From Eq. 1.4(a), we get,

n $\widehat{\beta_0}$ + $\widehat{\beta_1}\ \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$

⇒ n $\widehat{\beta_0} = \sum_{i=1}^{n} y_i - \widehat{\beta_1}\ \sum_{i=1}^{n} x_i$

⇒ $\widehat{\beta_0} = (\sum_{i=1}^{n} y_i)/n\ - \widehat{\beta_1}\frac{\sum_{i=1}^{n} x_i}{n}$

⇒ $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$

Where $\bar{x} = \frac{1}{n}(\sum_{i=1}^{n} x_i\ )$, $\bar{y} = \frac{1}{n}(\sum_{i=1}^{n} y_i\ )$,

are the averages of yi and xi, respectively. Therefore, $\widehat{\beta_0}\ and \widehat{\beta_1}$ are the least-squares estimators of the intercept and slope, respectively The fitted simple linear regression model is

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}\, x \ldots\ldots Eq.\ 1.5$$

Equation (1.5) gives a point estimate of the mean of y for a particular x.

Eq. 1.5 can also be written as

$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}\, x$

$\hat{y} = \bar{y} - \widehat{\beta_1}\,\bar{x} + \widehat{\beta_1}\, x$ [Since $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$]

$\hat{y} - \bar{y} = \widehat{\beta_1}(x - \bar{x})$ Eq. 1.6

The difference between the observed value y_i and the corresponding fitted value of $\hat{y_i}$ is a residual. Mathematically the ith residual is

$e_i = y_i - \hat{y_i} = y_i - \widehat{\beta_0} + \widehat{\beta_1}\, x_i$; i= 1, 2, 3, …….n

Eq. 1.7

Properties of the Least-Squares Estimators and the Fitted Regression Model:

The least-squares estimators have several important properties.

1. $\widehat{\beta_0}\ and\ \widehat{\beta_1}$ are linear combinations of the observations yi. For example,

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

$$and\ \widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

2. The least-squares estimators $\widehat{\beta_0}\ and\ \widehat{\beta_1}$ and are unbiased estimators of the model parameters β_0 and β_1.

3. $\widehat{\beta_0}$ and $\widehat{\beta_1}$ have the minimum variance in the class of all such linear unbiased estimators, an unbiased estimator with the least variance is known as an Efficient Estimator.

4. Another important result concerning the quality of the least-squares estimators is the Gauss-Markov theorem, which states that for the regression model y = $\beta_0 + \beta_1 x + \epsilon$……Eq. 1.2

with the assumptions

E(ε) = 0, Var(ε) = σ^2, and uncorrelated errors, the least-squares estimators are unbiased and have minimum variance when compared with all other unbiased estimators that are linear combinations of the yi. We often say that the least-squares estimators are best linear unbiased estimators, where "best" implies minimum variance.

5. The sum of the residuals in any regression model that contains an intercept β_0 is always zero, that is,

$\sum_{i=1}^{n}(y_i - \widehat{y_i}) = \sum_{i=1}^{n} e_i = 0$

6. The sum of the observed values yi equals the sum of the fitted values $\widehat{y_i}$, or

$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y_i}$

7. The least-squares regression line always passes through the centroid [the point $(\bar{x}, \bar{y})$] of the data.

8. The sum of the residuals weighted by the corresponding value of the regressor variable always equals zero, that is,

$$\sum_{i=1}^{n} x_i e_i = 0$$

9. The sum of the residuals weighted by the corresponding fitted value always equals zero, that is,

$$\sum_{i=1}^{n} \hat{y}_i \, e_i = 0$$

Co-efficient of Determination R^2 and adjusted R^2:

We now consider the Goodness of Fit of the fitted regression line to a set of data; i.e. we will find out how well the sample regression line fits the data. If all the observations were to lie on the regression line, we would obtain a perfect fit, but this is rarely the case. Generally there will be some positive and some negative e_i. These residuals around the regression line are as small as possible. The **Co-efficient of Determination R^2** is a summary measure that tells how well the sample regression fits the data.

$e_i = y_i - \hat{y}_i = y_i - \widehat{\beta_0} + \widehat{\beta_1} x_i$; i= 1, 2, 3, …….n Eq.1.7

Since $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} \bar{x}$, from Eq. 1.7 we get

$e_i = y_i - \hat{y}_i = y_i - \bar{y} + \widehat{\beta_1} \bar{x} + \widehat{\beta_1} x_i$;

$\Rightarrow e_i = (y_i - \bar{y}) - \widehat{\beta_1}(x_i - \bar{x})$

$\Rightarrow (y_i - \bar{y}) = e_i + \widehat{\beta_1}(x_i - \bar{x})$ Eq.1.8

Squaring Eq. 1.8 on both sides and summing over the sample, we obtain

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left(\widehat{\beta_1}(x_i - \bar{x})\right)^2 + \sum_{i=1}^{n} e_i^2 + 2\widehat{\beta_1}\sum_{i=1}^{n}(x_i - \bar{x})e_i$$

$$= \sum_{i=1}^{n}\left(\widehat{\beta_1}(x_i - \bar{x})\right)^2 + \sum_{i=1}^{n} e_i^2$$

[Since

$\sum_{i=1}^{n} x_i e_i = 0 \ and \ \sum_{i=1}^{n} e_i = 0$]

Here, $\sum_{i=1}^{n}(y_i - \bar{y})^2$ = total variation of the actual y values about their sample mean, which may be called the total sum of squares (TSS)

Now, $\sum_{i=1}^{n}\left(\widehat{\beta_1}(x_i - \bar{x})\right)^2 = \widehat{\beta_1}^2 \sum_{i=1}^{n}\left((x_i - \bar{x})\right)^2$

$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$……(From Eq.1.6 $\hat{y} - \bar{y} = \widehat{\beta_1}(x - \bar{x})$)

= Sum of squares due to regression or explained by regression or simply called explained sum of squares = ESS

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ = residual or unexplained variation of the y values about the regression line or simply called residual sum of squares = RSS.

Thus TSS = ESS + RSS…….Eq.1.9

So, it shows that the total variation in the observed y values about their mean value can be partitioned into two parts.

Now dividing by TSS in Eq. 1.9, we get

$$1= \frac{ESS}{TSS} + \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} + \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

We now define

Co-efficient of Determination R^2 as

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

Since TSS is a measure of the variability in y without considering the effect of the regressor variable x and ESS is a measure of the variability in y remaining after x has been considered, R^2 is often called the proportion of variation explained by the regressor x.

Properties of R^2

1. Since $0 \leq ESS \leq TSS$, it follows that $0 \leq R^2 \leq 1$. Values of R^2 that are close to 1 imply that most of the variability in y is explained by the regression model.

2. Sometimes, a low value of R^2 is a result of a poorly specified model. In these cases the model can often be improved by the addition of one or more predictor or regressor variables.

3. Sometimes, a low value of R^2 results from having a lot of variability in the measurements of the response

4. The statistic R^2 should be used with caution, since it is always possible to make R^2 large by adding enough terms to the model. For example, if there are no repeat points (more than one y value at the same x value), a polynomial of degree $n - 1$ will give a "perfect" fit ($R^2 = 1$) to n data points. When there are repeat points, R^2 can never be exactly equal to 1 because the model cannot explain the variability related to "pure" error.

We don't necessarily discard a model based on a low R-Squared value. Its a better practice to look at the AIC and prediction accuracy on validation sample when deciding on the efficacy of a model.

What about adjusted R-Squared?

As you add more *X* variables to your model, the R-Squared value of the new bigger model will always be greater than that of the smaller subset. This is because, since all the variables in the original model is also present, their contribution to explain the dependent variable will be present in the super-set as well, therefore, whatever new variable we add can only add (if not significantly) to the variation that was already explained. It is here, the adjusted R-Squared value comes to help. Adj R-Squared penalizes total value for the number of

terms (read predictors) in your model. Therefore when comparing nested models, it is a good practice to look at adj-R-squared value over R-squared.

$$R^2_{adj} = 1 - \frac{MSR}{MST}$$

here, MSR is the mean squared error given by MSR = $\frac{RSS}{n-2}$

and MST = $\frac{TSS}{n-1}$ is the *mean squared total*, where *n* is the number of observations

Therefore, by moving around the numerators and denominators, the relationship between R^2 and R^2_{adj} becomes:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n-2}$$

Procedure of Testing

Overall Significance of the models

Significance of individual coefficients

We are often interested in testing hypotheses and constructing confidence intervals about the model parameters.

1. **Normality Assumption:**

These procedures require that we make the additional assumption that the model errors ϵ_i are normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with mean 0 and variance σ^2, Thus ϵ_i ~ NID (0, σ^2).

Why Normal Distribution?

ϵ_i represents the combined influence of a large number of independent variables that are not introduced in in the regression model. Now by CLT, if there are

large number of independent variables, the distribution of their sum tends to Normal Distribution. So ϵ_i ~ NID (0, σ^2).

2. **Expectation of Least Square Estimators:**

(i) E $(\widehat{\beta_0}) = \beta_0$. and E $(\widehat{\beta_1}) = \beta_1$

Proof: $\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$$E(\widehat{\beta_1}) = E\left(\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

$$= \frac{\sum_{i=1}^{n} E(y_i)(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(\beta_0+\beta_1 x_i)(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad [\text{Since } E(y_i) = \beta_0 + \beta_1 x_i \,]$$

$$= \beta_0 \frac{\sum_{i=1}^{n}((x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \beta_1 \frac{\sum_{i=1}^{n}(x_i)(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \beta_1 \frac{\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad [Since \sum_{i=1}^{n}((x_i-\bar{x}) = 0]$$

$$= \beta_1 \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \beta_1 \frac{(\bar{x})\sum_{i=1}^{n}(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \beta_1 \times 1 + 0 \quad [Since \sum_{i=1}^{n}((x_i-\bar{x}) = 0]$$

$$= \beta_1$$

$$E(\widehat{\beta_0}) = E(\bar{y} - \widehat{\beta_1}\,\bar{x})$$

$$= E(\bar{y}) - (\bar{x}\,)E(\widehat{\beta_1}\,)$$

$$= \mathrm{E}(\beta_0 + \beta_1\bar{x}) - \bar{x}\,\beta_1$$

$$= \beta_0 + \beta_1\bar{x} - \bar{x}\,\beta_1$$

$$= \beta_0$$

3. **Variance of Least Square Estimators:**

Var $(\widehat{\beta_0}) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$

Var $(\widehat{\beta_1}\,) = \frac{\sigma^2}{S_{xx}}$

Proof:

[Note: ϵ_i ~ NID (0, σ^2), Since y_i *is a function of* ϵ_i, y_i also follows Normal distribution with $E(y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i$ and Var $(y_i) = \sigma^2$

$\Rightarrow y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$]

Var $(\widehat{\beta_1}) =$ Var $\left(\frac{\sum_{i=1}^{n} y_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

$= \frac{\sum_{i=1}^{n} V(y_i)\,(x_i-\bar{x})^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}$

$= \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2\,\sigma^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}$ [Since V(cX) = $c^2V(X)$]

$= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$= \frac{\sigma^2}{S_{xx}}$

Var $(\widehat{\beta_0}) = Var(\bar{y} - \widehat{\beta_1}\,\bar{x})$

= V$(\bar{y}) + V(\hat{\beta}_1\bar{x})$ [Since V(ax+by) = $a^2V(x) + b^2V(y), if\ x\ and\ y\ are\ independent\ variables.$]

= V$(\bar{y}) + \bar{x}^2\,V(\hat{\beta}_1)$

$= \frac{\sigma^2}{n} + \bar{x}^2\,\frac{\sigma^2}{S_{xx}}$

$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

4. $\widehat{\beta_0} \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$

$\widehat{\beta_1} \sim N\left(\beta_1, \left(\frac{\sigma^2}{S_{xx}}\right)\right)$

5. RSS/(n-2) = MSR is an unbiased estimator of σ^2

Where RSS = $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = Residual\ sum\ of\ squres$

e_i ~ NID (0, σ^2)

$\Rightarrow \frac{e_i}{\sigma}$ ~ NID (0, 1)

$\Rightarrow \left(\frac{e_i}{\sigma}\right)^2 \sim \chi^2$ distribution with (1) degrees of freedom.

$\Rightarrow \sum_{i=1}^{n} \left(\frac{e_i}{\sigma}\right)^2 = \frac{RSS}{\sigma^2} \sim \chi^2$ distribution with (n-2) degrees of freedom.

[Since RSS has (n-2) degrees of freedom. (Explanation given in ANOVA)]

$E\left(\frac{RSS}{\sigma^2}\right) = n - 2$

$\Rightarrow \sigma^2 = E\left(\frac{RSS}{n-2}\right)$

$\Rightarrow \hat{\sigma}^2 = \frac{RSS}{n-2} = \text{MSR}$

⇒ MSR is an unbiased estimator of σ^2

6. **Use of Z and t Tests**

Suppose that we wish to test the hypothesis that the slope equals a constant, say β_1'. The appropriate hypotheses are

$H_0: \beta_1 = \beta_1'$ against $H_1: \beta_1 \neq \beta_1'$

Since the errors ϵ_i are NID $(0, \sigma^2)$, the observations y_i are NID $(\beta_0 + \beta_1 x_i, \sigma^2)$. Now as $\widehat{\beta_1}$ is a linear combination of the observations, so it is normally distributed with

mean β_1 and variance $\frac{\sigma^2}{S_{xx}}$.

a. If σ^2 is known

The test statistic is $Z = \frac{\widehat{\beta_1} - \beta_1'}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$

This procedure rejects the null hypothesis at $\alpha\%\ level\ of\ significance$ if $|Z| > Z_{\alpha/2}$.

Alternatively, a p-value approach could also be used for decision making.

b. If σ^2 is unknown.

RSS/(n-2) = MSR is an unbiased estimator of σ^2 is

The quantity MSR is called the residual mean square. The square root of it is sometimes called the standard error of regression, and it has the same units as the response variable y.

The test statistic is t = $\frac{\widehat{\beta_1}-\beta_1'}{\sqrt{\frac{MSR}{S_{xx}}}}$

which follows a t_{n-2} distribution.

This procedure rejects the null hypothesis at $\alpha\%\ level\ of\ significance$ if | t | > $t_{\alpha/2,\ n-2}$

Alternatively, a p-value approach could also be used for decision making. The denominator of the test statistic, is often called the estimated standard error, or more simply, the standard error of the slope. That is,

$$\text{s.e. } (\beta_1') = \sqrt{\frac{MSR}{S_{xx}}}$$

Suppose that we wish to test the hypothesis that the intercept equals a constant, say β_0'. The appropriate hypotheses are

$H_0: \beta_0 = \beta_0'$ against $H_1: \beta_0 \neq \beta_0'$

Now as $\widehat{\beta_0}$ is a linear combination of the observations, so it is normally distributed with

mean β_0 and variance $\sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$

a. If σ^2 is known

The test statistic is Z = $\frac{\widehat{\beta_0}-\beta_0'}{\sqrt{\sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}})}}$

This procedure rejects the null hypothesis at $\alpha\%\ level\ of\ significance$if | Z | > $Z_{\alpha/2}$.

b. If σ^2 is unknown.

The test statistic is $t = \frac{\widehat{\beta_0} - \beta_0'}{\sqrt{MSR\ (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}}$

This procedure rejects the null hypothesis at $\alpha\%\ level\ of\ significance$ if | t | > $t_{\alpha/2,\ n-2}$

$$s.e.\ (\beta_0') = \sqrt{MSR\ (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$$

Testing Significance of Regression:

A very important special case of the hypotheses is

$H_0: \beta_1 = \beta_1'$ against $H_1: \beta_1 \neq \beta_1'$

This hypotheses relate to the significance of regression. Failing to reject H_0: $\beta_1 = 0$ implies that there is no linear relationship between x and y. Note that this may imply either that x is of little value in explaining the variation in y or that the true relationship between x and y is not linear. Therefore, failing to reject H_0: $\beta_1 = 0$ is equivalent to saying that there is no linear relationship between y and x.

Alternatively, if H_0: $\beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y.

However, rejecting H_0: $\beta_1 = 0$ could mean either that the straight-line model is adequate or that even though there is a linear effect of x, better results could be obtained with the addition of higher order polynomial terms in x.

Analysis of Variance

We may also use an **analysis-of-variance** approach to test significance of regression. The analysis of variance is based on a partitioning of total variability in the response variable y.

$$\text{TSS} = \text{ESS} + \text{RSS}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

We can use the usual **analysis-of-variance** *F* test to test the Hypothesis H_0: $\beta_1 = 0$.

Note: Degrees of Freedom refers to the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample. Degrees of Freedom are commonly discussed in relation to various forms of hypothesis testing in statistics, such as a Chi-Square

The **degree-of-freedom** breakdown is determined as follows. The total sum of squares, TSS, has df = $n - 1$ degrees of freedom because one degree of freedom is lost as a result of the constraint on the deviations $\sum(y_i - \bar{y}) = 0$. The model or regression sum of squares, ESS, has df = 1 degree of freedom because ESS is completely determined by one parameter β_1'. Finally, we note RSS has df = $n - 2$ degrees of freedom.

$\frac{RSS}{\sigma^2}$ follows χ^2 distribution with (n-2) degrees of freedom.

$\frac{ESS}{\sigma^2}$ follows χ^2 distribution with 1 degree of freedom.

RSS and ESS are independent. By the definition of an *F* statistic given, we get

F = $\frac{\frac{ESS}{\sigma^2}}{\frac{\frac{1}{RSS}}{\frac{\sigma^2}{n-2}}}$ follows $F_{1,n-2}$ distribution

Analysis of Variance Table:

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	F
Regression	ESS	1	MSE=ESS/1	F= MSE/MSR
Residual	RSS	n-2	MSR=RSS/(n-2)	
Total	TSS	n-1	MST=TSS/(n-1)	

Therefore, to test the hypothesis H_0: β_1 =0, compute the test statistic *F* and reject H_0 if F > $F_{\alpha,1,n-2}$.

Confidence Intervals on β_0, β_1:

100(1 − α) % CI on the intercept β_0 is

$\beta_0' - t_{\alpha/2,\, n-2}$ s.e. $(\beta_0') \leq \beta_0 \leq \beta_0' + t_{\alpha/2,\, n-2}$ s.e. (β_0')

100(1 − α) % CI on the intercept β_1 is

$\beta_1' - t_{\alpha/2,\, n-2}$ s.e. $(\beta_1') \leq \beta_1 \leq \beta_1' + t_{\alpha/2,\, n-2}$ s.e. (β_1')

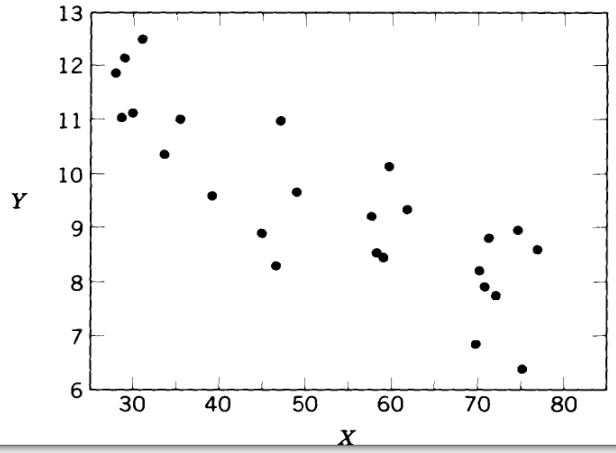
s.e. $(\beta_1') = \sqrt{\frac{MSR}{S_{xx}}}$ s.e. $(\beta_0') = \sqrt{MSR\,(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$

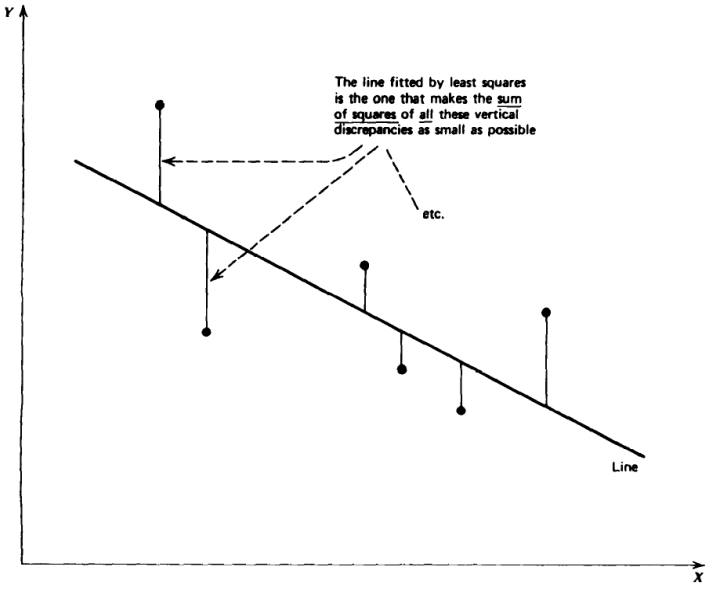
and RSS/(n-2) = MSR is an unbiased estimator of σ^2

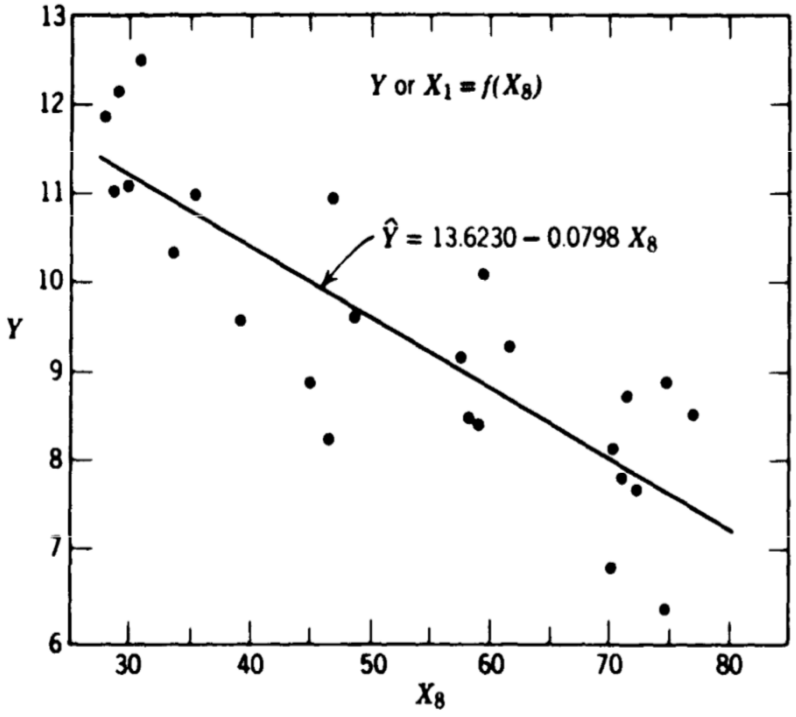
Example:

Observation Number	Variable Number	
	1 (Y)	8 (X)
1	10.98	35.3
2	11.13	29.7
3	12.51	30.8
4	8.40	58.8
5	9.27	61.4
6	8.73	71.3
7	6.36	74.4
8	8.50	76.7
9	7.82	70.7
10	9.14	57.5
11	8.24	46.4
12	12.19	28.9
13	11.88	28.1
14	9.57	39.1
15	10.94	46.8
16	9.58	48.5
17	10.09	59.3
18	8.11	70.0
19	6.83	70.0
20	8.88	74.5
21	7.68	72.1
22	8.47	58.1
23	8.86	44.6
24	10.36	33.4
25	11.08	28.6

FITTING A STRAIGHT LINE BY LEAST SQUARES




Y
The line fitted by least squares is the one that makes the sum of squares of all these vertical discrepancies as small as possible
etc.
Line
X


Y or $X_1 = f(X_8)$
$\hat{Y} = 13.6230 - 0.0798\ X_8$
Y
X_8
13
12
11
10
9
8
7
6
30
40
50
60
70
80

A. A study was made on the effect of temperature on the yield of a chemical process. The following data (in coded form) were collected:

X	Y
−5	1
−4	5
−3	4
−2	7
−1	10
0	8
1	9
2	13
3	14
4	13
5	18

1. Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least squares estimates of β_0 and β_1? What is the prediction equation?
2. Construct the analysis of variance table and test the hypothesis $H_0: \beta_1 = 0$ with an α risk of 0.05.
3. What are the confidence limits ($\alpha = 0.05$) for β_1?
4. What are the confidence limits ($\alpha = 0.05$) for the true mean value of Y when $X = 3$?
5. What are the confidence limits ($\alpha = 0.05$) for the difference between the true mean value of Y when $X_1 = 3$ and the true mean value of Y when $X_2 = -2$?
6. Are there any indications that a better model should be tried?

Detection and Treatment of missing values

1. Deleting the observations

If you have large number of observations in your dataset, then try deleting (or not to include missing values while model building, for example by setting

na.action = na.omit) those observations (rows) that contain missing values. Make sure after deleting the observations, you have:

1. Have sufficent data points, so the model doesn't lose power.
2. Not to introduce bias (meaning, disproportionate or non-representation of classes).

2. Deleting the variable

If a paricular variable is having more missing values that rest of the variables in the dataset, and, if by removing that one variable you can save many observations, then you are better off without that variable unless it is a really important predictor that makes a lot of business sense. It is a matter of deciding between the importance of the variable and losing out on a number of observations.

3. Imputation with mean / median / mode

Replacing the missing values with the mean / median / mode is a crude way of treating missing values. Depending on the context, like if the variation is low or if the variable has low leverage over the response, such a rough approximation is acceptable and could possibly give satisfactory results.

Detection and Treatment of Outliers

An outlier is an extreme observation; one that is considerably different from the majority of the data. Residuals that are considerably larger in absolute value than the others, say three or four standard deviations from the mean, indicate outliers. Depending on their location in x space, outliers can have moderate to severe effects on the regression model.

- Residual plots against $\hat{y}_1$ and the normal probability plot, boxplot are helpful in identifying outliers.

- Examining **scaled residuals**, such as the studentized and R student residuals, is an excellent way to identify potential outliers.

- Outliers should be carefully investigated to see if a reason for their unusual behaviour can be found.

- Sometimes outliers are "bad" values, occurring as a result of unusual but explainable events. Examples include faulty measurement or analysis, incorrect recording of data, and failure of a measuring instrument. If this is the case, then the outlier should be corrected (if possible) or deleted from the data set.

- However, we emphasize that there should be strong nonstatistical evidence that the outlier is a bad value before it is discarded.

- Sometimes we find that the outlier is an unusual but perfectly plausible observation. Deleting these points to "improve the fit of the equation" can be dangerous, as it can give the user a false sense of precision in estimation or prediction.

- An outlier that has an unusual x value but does not affect the estimates of the regression coefficients is called **leverage point.**

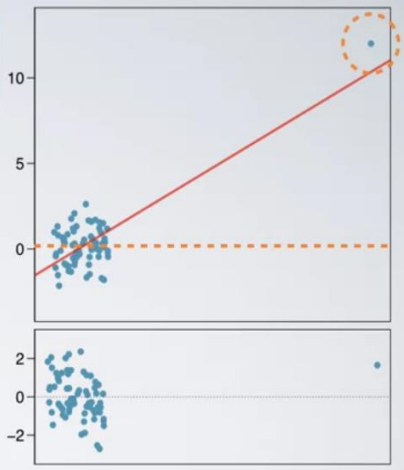
- An outlier that has an unusual x value and affects the estimates of the regression coefficients is called **influential point.**

types of outliers

- outliers are points that fall away from the cloud of points
- outliers that fall horizontally away from the center of the cloud but don't influence the slope of the regression line are called leverage points
- outliers that actually influence the slope of the regression line are called influential points
  - usually high leverage points
  - to determine if a point is influential, visualize the regression line with and without the point, and ask: *Does the slope of the line change considerably?*

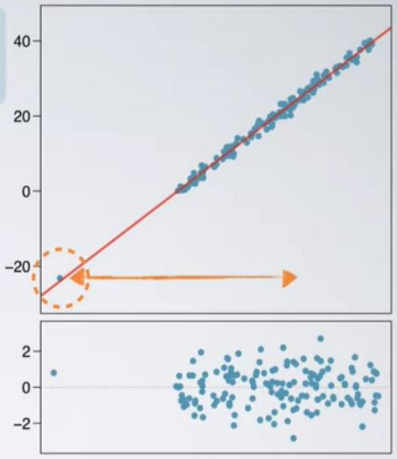
How does the outlier influence the least squares line?

Without the outlier there is no relationship between x and y.



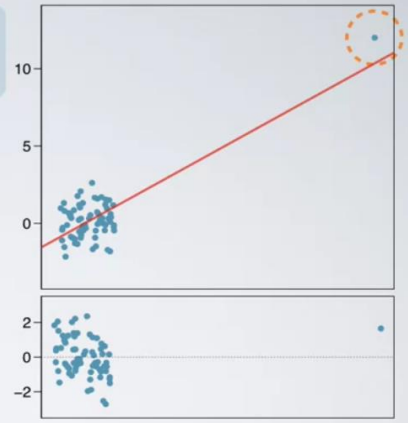
What type of outlier is this?

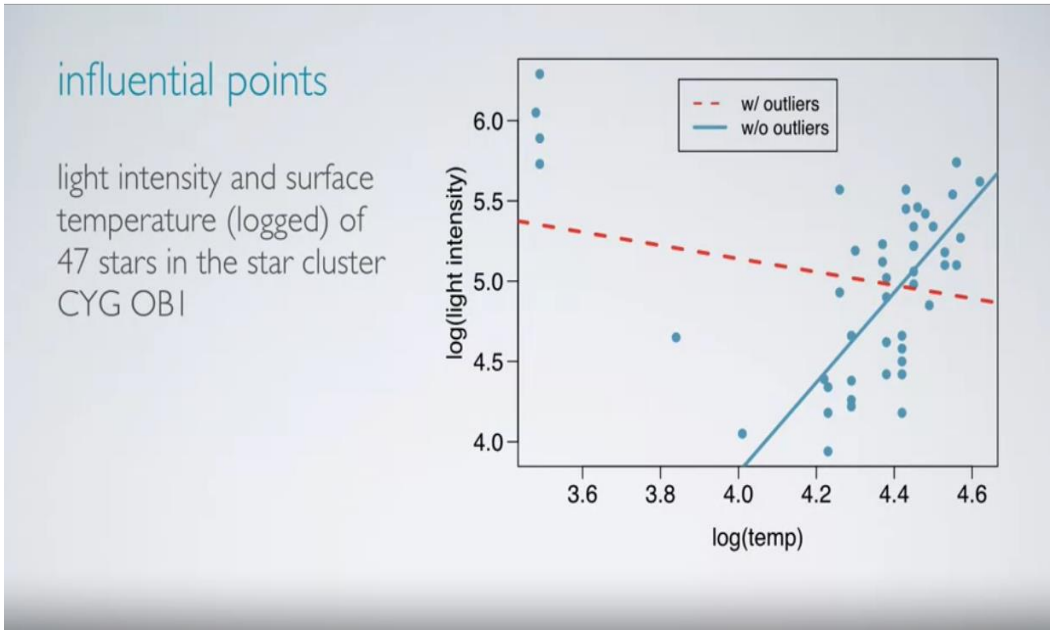
leverage point



What type of outlier is this?

influential point





Treatment of Outliers:

- Sometimes outliers are “bad” values, occurring as a result of unusual but explainable events. Examples include faulty measurement or analysis, incorrect recording of data, and failure of a measuring instrument. If this is the case, then the outlier should be **corrected (if possible) or deleted** from the data set.

- **Transforming variables** can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values.

- **Imputing:** We can use mean, median, mode imputation method.

- **Treat Separately:** If there are significant number of outliers we should treat them separately in the regression model. One of the approach is to treat two groups as two separate groups and build individual models for two different groups and then combine the outputs.

Note:

Various statistical tests have been proposed for detecting and rejecting outliers. For example Barnett and Lewis [1994]. Stefansky [1971, 1972] has proposed an approximate test for identifying outliers based on the maximum normed residual that is particularly easy to apply.

While these tests may be useful for identifying outliers, they should not be interpreted to imply that the points so discovered should be automatically rejected. As we have noted, these points may be important clues containing valuable information

In most practical problems, especially those involving historical data, the analyst has a rather large pool of possible **candidate regressors**, of which only a few are likely to be important. Finding an appropriate subset of regressors for the model is often called the **variable selection problem.**

Weighted Least Squares:

The assumptions were: The model errors have mean zero and constant variance and are uncorrelated. Here we focus on methods and procedures for building regression models when some of the above assumptions are violated.

The method of weighted least squares is an useful method in building regression models in situations where some of the underlying assumptions are violated.

We will illustrate how weighted least squares can be used when the equal-variance assumption is not appropriate.

In this method of estimation, the deviation between the observed and expected values of yi is multiplied by a weight w_i chosen inversely proportional to the variance of yi.

Since each weight is inversely proportional to the error variance, it reflects the information in that observation. So, an observation with small error variance has a large weight since it contains relatively more information than an observation with large error variance (small weight). These standard deviations reflect the information in the response *Y* values (remember these are averages) and so in estimating a regression model we should downweight the obervations with a large standard deviation and upweight the observations with a small standard deviation. In other words we should use weighted least squares with weights equal to 1/*SD*2.

For the case of simple linear regression, the weighted least squares function is

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

The resulting least-squares normal equations are:

$$\widehat{\beta_0} \sum_{i=1}^{n} w_i + \widehat{\beta_1} \sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i y_i$$

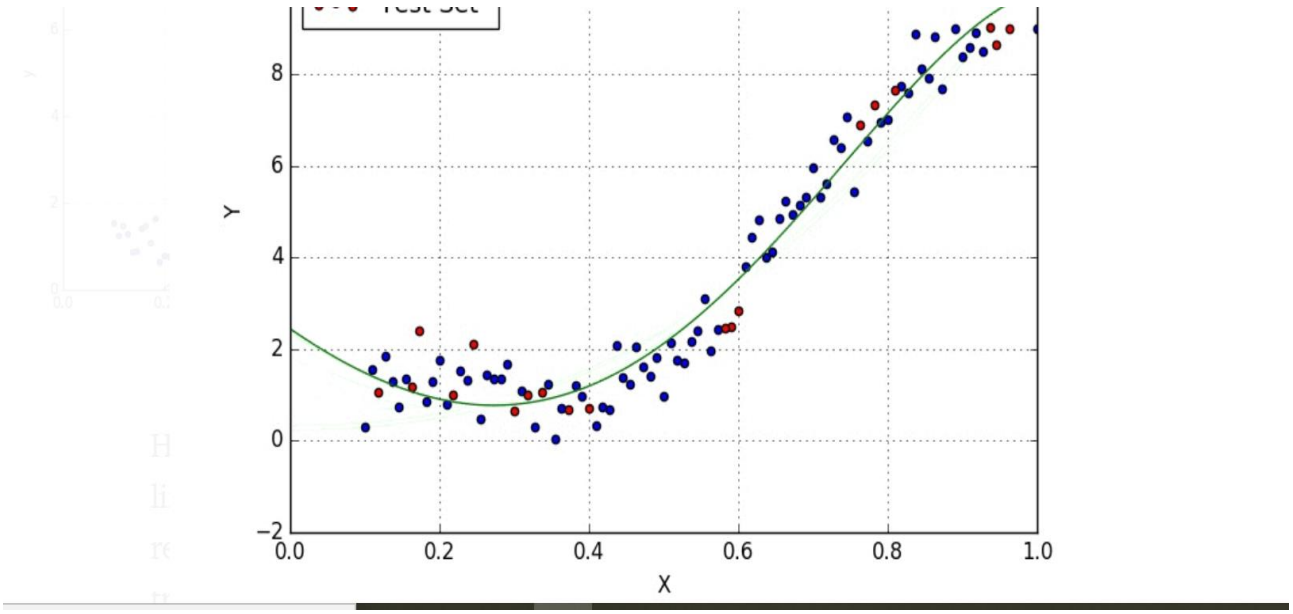
$$\widehat{\beta_0} \sum_{i=1}^{n} w_i x_i + \widehat{\beta_1} \sum_{i=1}^{n} w_i x_i^2 = \sum_{i=1}^{n} w_i y_i \; x_i$$

Solving the above equations we will get weighted least-squares estimates of β_0 and β_1.

Polynomial Linear Regression

In the last section, we saw two variables in your data set were correlated but what happens if we know that our data is correlated, but the relationship doesn't look linear? So hence depending on what the data looks like, we can do a polynomial regression on the data to fit a polynomial equation to it.

Hence If we try to use a simple linear regression in the below graph then the linear regression line won't fit very well. It is very difficult to fit a linear regression line in the below graph with a low value of error. Hence we can try to use the polynomial regression to fit a polynomial line so that we can achieve a minimum error



The equation of the polynomial regression for the above graph data would be

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

The general equation of a polynomial regression is:

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots .. \beta_k x^k + \epsilon$.....Eq *

It is important to keep the order of the model as low as possible. In an extreme case it is always possible to pass a polynomial of order $n - 1$ through n points so that a polynomial of sufficiently high

degree can always be found that provides a “good” fit to the data.
If we set $x_i = x^i$, j = 1, 2,.. ., k, then

Eq. * becomes a multiple linear regression model in the k regressors $x_1, x_2, \ldots . x_k$ Thus, a polynomial model of order k may be fitted using the techniques of Multiple regression model.