WEEK 13

11. Simple Linear Regression

Let X and Y be two random variables with join probability (or probability density) function $f(x,y)$ and the marginal probability (or probability density) functions $f_X(x)$ and $f_Y(y)$ respectively. From the joint probability function we can find the conditional probability (or probability density) function of Y given $X = x$ as

$$f_{Y|X=x}(y \mid x) = \frac{f(x,y)}{f_X(x)} \text{ , } f_X(x) > 0$$

we can also calculate the conditional expectation of Y given $X = x$ as

$$E(Y \mid X = x) = \begin{cases} \sum_y y\, P(Y = y \mid X = x) & , \quad \textit{discrete case} \\ \int_y y\, f_{Y|X=x}(y \mid x)dy & , \quad \textit{continuous case.} \end{cases}$$

Obviously, this conditional expectation is a function of x , that is $E(Y \mid X = x) = h(x)$. This conditional expectation is known as the regression of the random variable Y on X . If the function h is a linear function of x then the regression is called a linear regression, otherwise it is a non-linear regression of Y on x . In this class we will consider the case a linear regression. That is,

$$E(Y \mid X = x) = h(x) = \alpha + \beta x .$$

Moreover, let $(Y, X_1, X_2, ..., X_p)$ be the random variables with joint probability (or probability density) function $f_{Y,X_1,X_2,...,X_p}(y, x_1, x_2, ..., x_p)$. In a similar way, we can find the conditional probability (or probability density) function of Y given $X_1 = x_1, X_2 = x_2, ..., X_p = x_p$ as

$$f_{Y|X_1=x_1, X_2=x_2,..., X_p=x_p}(y \mid x_1, x_2, ..., x_p) = \frac{f_{Y,X_1,X_2,...,X_p}(y, x_1, x_2, ..., x_p)}{f_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p)}$$

where $f_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p) > 0$. And in a similar way, we can calculate the conditional expectation of Y given $X_1 = x_1, X_2 = x_2, ..., X_p = x_p$ as

$$E(Y \mid X_1 = x_1, X_2 = x_2, ..., X_p = x_p) = h(x_1, x_2, ..., x_p) .$$

This conditional expectation is known as multiple regression of Y on $X_1, X_2, ..., X_p$. As it is obviously seen, this conditional expectation is a function of $x_1, x_2, ..., x_p$, namely, $h(x_1, x_2, ..., x_p)$. If $h(x_1, x_2, ..., x_p)$ is a linear function of x's then it is a multiple linear regression of Y on the variables $x_1, x_2, ..., x_p$ namely, if $h(x_1, x_2, ..., x_p) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p$ then it is multiple linear regression of Y on the variables $x_1, x_2, ..., x_p$, otherwise it is a non-linear regression. For the case $p = 1$ the linear regression is named as "simple" linear regression. In this class, we are going to investigate the simple linear regression.

The main goal in the linear regression is to estimate the function h. In the real life, we fixed the value of x and measure the value of Y. As it is clear, it is possible that we can observe different values for Y at the same value of x's. That is, we consider a function

$$Y_i = \alpha + \beta x_i + e_i \ , \ i = 1, 2, ..., n \qquad (1)$$

and we say that it is a regression equation if

- e_i's are independent and identically distributed random variables such that $E(e_i) = 0$ and $Var(e_i) = \sigma^2$
- x_i's are fixed in the sense that they are not random.

Moreover, in order to make statistical inference we also assume the normality of e_i's. That is, in order to say that the equation in (1)

- $e_i \sim i.i.d\ N(0, \sigma^2)$
- x_i's are fixed in the sense that they are not random.

Note that if $e_i \sim i.i.d\ N(0, \sigma^2)$ then

$$E(Y_i) = E(\alpha + \beta x_i + e_i) = \alpha + \beta x_i + E(e_i) = \alpha + \beta x_i$$

and

$$Var(Y_i) = Var(\alpha + \beta x_i + e_i) = Var(e_i) = \sigma^2$$

and therefore, Y_i's are independent (not identically) normally distributed random variables. That is,

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2).$$

Consider the simple linear regression equation given in (1)

$$Y_i = \alpha + \beta x_i + e_i \ , \ i = 1, 2, ..., n.$$

Here,

- Y_i's are the dependent variables (dependency means that they are function of x's). Actually they are independent random variables.
- x_i's are independent variables (or explonatory variables which are not random)
- e_i's are the error terms (they are independent normally distributed random variables such that $E(e_i)=0$, $Var(e_i)=\sigma^2$, or simply $e_i \sim i.i.d\ N(0,\sigma^2)$)
- α and β are the parameters to be estimated. Actually there is another parameter to be estimated which is the variance of the error term, σ^2.

$$\underset{dependent\ \text{var.}}{Y_i} = \underbrace{\alpha + \beta}_{parameters\ _i} \underset{\exp lonatory\ \text{var.}}{x_i} + \underset{error\ term}{e_i}, \quad i=1,2,...,n$$

Consider two variables X and Y. If there is a functional relationship (e.g. $Y=f(X)$) between these variables it is a deterministic relationship (shown in the following figure (a)). For example, if there is a relationship like $Y=2X+3$, we observe $Y=5$ for $X=1$ and $Y=3$ for $X=0$. However, in the reality, it is possibe that we can observe $Y=4.8$ for $X=1$ and $Y=3.2$ for $X=0$. Moreover, if we repart the experiment at the same conditions, it is possible that we can observe $Y=5.1$ for $X=1$ and $Y=2.7$ for $X=0$. That is there is a stochastic relationship between these two random variable (see the following figure (b) below). .

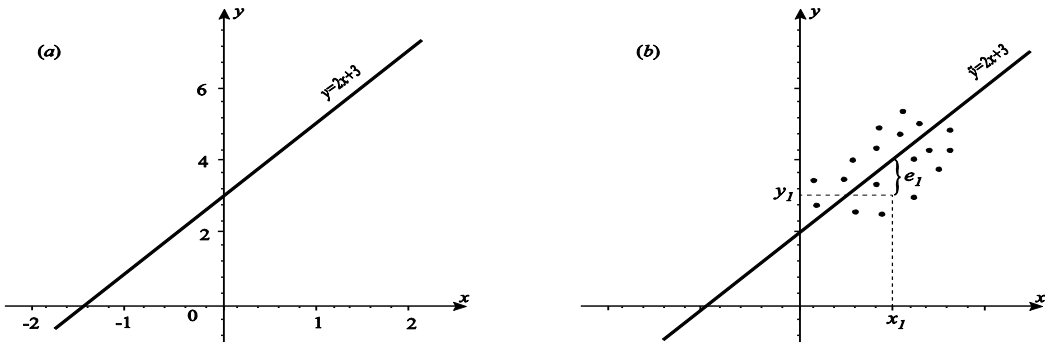


Figure. Graph of $y=2x+3$ line

We estimate the parameters α and β by minimizing error sum of squares. Note that from the regression equation the error term can be written as $e_i = Y_i - \alpha - \beta x_i$ and therefore error sum of squares can be written as

$$Q(\alpha,\beta)=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)^2 . \qquad (2)$$

In order to minimize this sum of squares, we take the first derivatives and equate to zero. The solutions (say $\hat{\alpha}$ and $\hat{\beta}$) are either a minimum or a maximum. To make sure that they are minimum we need to look at the second derivatives. If the second derivatives at these solutions are positive then they are minimum (we are not going to look at the second derivatives here and assume that they are minimum). The derivatives are;

$$\left.\begin{aligned}\frac{\partial Q(\alpha,\beta)}{\partial \alpha} &= -2\sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)\\ \frac{\partial Q(\alpha,\beta)}{\partial \beta} &= -2\sum_{i=1}^{n}x_i(Y_i-\alpha-\beta x_i)\end{aligned}\right\} = \begin{bmatrix}0\\0\end{bmatrix}$$

and from these equations we have the following equations:

$$-2\sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)=0 \quad \Rightarrow \quad \sum_{i=1}^{n}Y_i = \alpha n + \beta\sum_{i=1}^{n}x_i$$

$$-2\sum_{i=1}^{n}x_i(Y_i-\alpha-\beta x_i)=0 \Rightarrow \quad \sum_{i=1}^{n}x_iY_i = \alpha\sum_{i=1}^{n}x_i + \beta\sum_{i=1}^{n}x_i^2.$$

That is, we have the following equations (called NORMAL EQUATIONS):

$$\boxed{\begin{aligned}\sum_{i=1}^{n}Y_i &= \alpha n + \beta\sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}x_iY_i &= \alpha\sum_{i=1}^{n}x_i + \beta\sum_{i=1}^{n}x_i^2\end{aligned}}. \qquad (3)$$

Note that these solutions are obtained by minimizing error sum of squares. And therefore these estimators are also known as "*ordinary least square"* (OLS) estimators. Moreover, the estimator $\hat{\alpha}$ is the "*intercept term"* and $\hat{\beta}$ is *the "slope"* of the regression equation. The solutions (say $\hat{\alpha}$ and $\hat{\beta}$) to these equations are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_n)(Y_i-\bar{Y}_n)}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n$$

where

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n}Y_i,\ \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n}x_i,\ S_{xx} = \sum_{i=1}^{n}(x_i-\bar{x}_n)^2 \text{ and } S_{xy} = \sum_{i=1}^{n}(x_i-\bar{x}_n)(Y_i-\bar{Y}_n)$$

and since $\sum_{i=1}^{n}(x_i-\bar{x}_n)=0$ the sum in the denominator in $\hat{\beta}$ can be written as

$$\sum_{i=1}^{n}(x_i-\bar{x}_n)(Y_i-\bar{Y}_n) = \sum_{i=1}^{n}(x_i-\bar{x}_n)Y_i\,.$$

and therefore the estimator of β can be written as

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)Y_i}{S_{xx}} = \sum_{i=1}^{n}\left(\frac{(x_i - \bar{x}_n)}{S_{xx}}\right)Y_i = \sum_{i=1}^{n} w_i Y_i .$$

That is, $\hat{\beta}$ is a linear combination of Y_i's. Similarly, the estimator of α is also linear combination of Y_i's because

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{x}_n \sum_{i=1}^{n} w_i Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}_n w_i\right)Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}_n(x_i - \bar{x}_n)}{S_{xx}}\right)Y_i = \sum_{i=1}^{n} s_i Y_i .$$

That is, the OLS estimators of α and β are linear combinations of Y_i's. In other words, they are linear in Y_i's.

In a summary,

$$\hat{\beta} = \sum_{i=1}^{n} w_i Y_i \qquad w_i = \frac{(x_i - \bar{x}_n)}{\sum_{j=1}^{n}(x_j - \bar{x}_n)^2} = \frac{(x_i - \bar{x}_n)}{S_{xx}}$$

$$\hat{\alpha} = \sum_{i=1}^{n} s_i Y_i \qquad s_i = \frac{1}{n} - \frac{\bar{x}_n(x_i - \bar{x}_n)}{\sum_{j=1}^{n}(x_j - \bar{x}_n)^2} = \frac{1}{n} - \frac{\bar{x}_n(x_i - \bar{x}_n)}{S_{xx}} .$$

Notes:

a) Notes on the linear estimator of $\hat{\beta} = \sum_{i=1}^{n} w_i Y_i$

- $\sum_{i=1}^{n} w_i = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x}_n) = 0$
- $\sum_{i=1}^{n} w_i^2 = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{S_{xx}}{S_{xx}^2} = \frac{1}{S_{xx}} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$
- $$\sum_{i=1}^{n} x_i w_i = \frac{1}{S_{xx}}\sum_{i=1}^{n} x_i(x_i - \bar{x}_n) = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x}_n + \bar{x}_n)(x_i - \bar{x}_n)$$
$$= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + \frac{\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x}_n) = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{S_{xx}}{S_{xx}} = 1$$

AND

b)) Notes on the linear estimator of $\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n$

- $\sum_{i=1}^{n} s_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}_n(x_i - \bar{x}_n)}{S_{xx}}\right) = 1 - \frac{\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x}_n) = 1$

$$\sum_{i=1}^{n} s_i^2 = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}_n(x_i-\bar{x}_n)}{S_{xx}}\right)^2 = \sum_{i=1}^{n}\left(\frac{1}{n^2} - \frac{\bar{x}_n^2(x_i-\bar{x}_n)^2}{S_{xx}^2} - \frac{2\bar{x}_n(x_i-\bar{x}_n)}{S_{xx}}\right)$$

- $$= \sum_{i=1}^{n}\frac{1}{n^2} + \frac{\bar{x}_n^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 - \frac{2\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}_n) = \sum_{i=1}^{n}\frac{1}{n^2} + \frac{\bar{x}_n^2 S_{xx}}{S_{xx}^2} = \frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}}$$

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

$$\sum_{i=1}^{n} x_i s_i = \sum_{i=1}^{n} x_i\left(\frac{1}{n} - \frac{\bar{x}_n(x_i-\bar{x}_n)}{S_{xx}}\right) = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}x_i(x_i-\bar{x}_n)$$

- $$= \bar{x}_n - \frac{\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}_n+\bar{x}_n)(x_i-\bar{x}_n) = \bar{x}_n - \frac{\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 + \frac{\bar{x}_n}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}_n)$$

$$= \bar{x}_n - \frac{\bar{x}_n S_{xx}}{S_{xx}} = \bar{x}_n - \bar{x}_n = 0.$$

<u>Statistical Properties of OLS Estimators</u>:

a) The OLS estimator of β

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_n)(Y_i-\bar{Y}_n)}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_n)(Y_i-\bar{Y}_n)}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2} = \sum_{i=1}^{n} w_i Y_i$$

- $\hat{\beta}$ is an unbiased estimator of β because

$$E(\hat{\beta}) = \sum_{i=1}^{n} w_i E(Y_i) = \sum_{i=1}^{n} w_i(\alpha+\beta x_i) = \alpha\sum_{i=1}^{n} w_i + \beta\sum_{i=1}^{n} x_i w_i = 0 + \beta(1) = \beta\,.$$

- $Var(\hat{\beta}) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2} = \dfrac{\sigma^2}{S_{xx}}$ because

$$Var(\hat{\beta}) = Var\left(\sum_{i=1}^{n} w_i Y_i\right) = \sum_{i=1}^{n} w_i^2 Var(Y_i) = \sigma^2\sum_{i=1}^{n} w_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2} = \frac{\sigma^2}{S_{xx}}.$$

- If e_i's are normally distributed random variables ($e_i \sim i.i.d\ N(0,\sigma^2)$) then Y_i's are independent and normally distributed random variables and therefore any linear combinations or independent normally distributed random variables is also normally distributed. Therefore, $\hat{\beta} \sim N(\beta, \sigma^2 / S_{xx})$.

b) The OLS estimator of α

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}_n(x_i - \bar{x}_n)}{S_{xx}}\right)Y_i = \sum_{i=1}^{n} s_i Y_i$$

- $\hat{\alpha}$ is an unbiased estimator of α because

  $E(\hat{\alpha}) = \sum_{i=1}^{n} s_i E(Y_i) = \sum_{i=1}^{n} s_i(\alpha + \beta x_i) = \alpha\sum_{i=1}^{n} s_i + \beta\sum_{i=1}^{n} x_i s_i = \alpha + \beta(0) = \alpha$.

- $Var(\hat{\alpha}) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}}\right]$ because

  $Var(\hat{\alpha}) = \sum_{i=1}^{n} s_i^2 Var(Y_i) = \sigma^2\sum_{i=1}^{n} s_i^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}}\right]$.

- If e_i's are normally distributed random variables ($e_i \sim i.i.d\ N(0,\sigma^2)$) then Y_i's are independent and normally distributed random variables and therefore any linear combinations or independent normally distributed random variables is also normally distributed. Therefore, $\hat{\alpha} \sim N(\alpha, Var(\hat{\alpha}))$.

Note (IMPORTANT) : As we have shown above, the OLS estimators of α and β are unbiased and linear combinations of Y_i's. It is possible that we can find many linear and unbiased estimators. However, these OLS estimators have the smallest variance among all linear and unbiased estimators of α and β. That is, the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are the **B**est **L**inear **U**nbiased **E**stimators (BLUE) of α and β.

The value of OLS estimator of β can also be written as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2}.$$

After we calculate the values of the OLS estimaor, we write the "*fitted regression line"* as

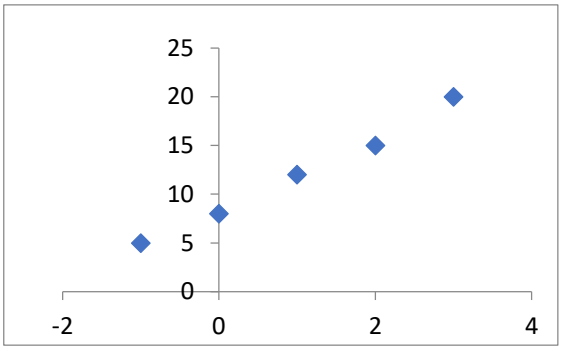
$$\boxed{\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i\ ,\ i = 1,2,...,n} \qquad (4)$$

and the "*residuals*" are calculated as $\hat{e}_i = y_i - \hat{y}_i\ ,\ i = 1,2,...,n$.

Example: Assume that the following data is appropriate for a simple linear regression. That is, we have only 5 observations.

$X = x$	–1	0	1	2	3
$Y = y$	5	8	12	15	20

The scatter plot and some of calculated values are given below.



$n = 5$, $\bar{x}_n = 1$, $\bar{y}_n = 12$

$\sum_{i=1}^{5} x_i = 5$, $\sum_{i=1}^{5} x_i^2 = 15$, $\sum_{i=1}^{5} y_i = 60$,

$\sum_{i=1}^{5} x_i y_i = 97$, $\sum_{i=1}^{5} y_i^2 = 858$

Now, we can calculate the values of the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ as

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2} = \frac{97 - 5(1)(12)}{15 - 5(1)^2} = \frac{37}{10} = 3.7$$

and

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\,\bar{x}_n = 12 - (3.7)\,(1) = 8.3$$

and thus the fitted regression line is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}\,x_i = 8.3 + 3.7\,x_i \ , \ i = 1, 2, ..., n$$

and the residuals are $\hat{e}_i = y_i - \hat{y}_i$, $i = 1, 2, ..., n$. The calculated fitted values (or predicted values) and the residuals are calculated and given below.

$\hat{y}_1 = 8.3 + 3.7x_1 = \hat{y}_1 = 8.3 + 3.7(-1) = 4.6$ $\qquad$ $\hat{e}_1 = y_1 - \hat{y}_1 = 5 - 4.6 = 0.4$

$\hat{y}_2 = 8.3 + 3.7x_2 = \hat{y}_1 = 8.3 + 3.7(0) = 8.3$ $\qquad$ $\hat{e}_2 = y_2 - \hat{y}_2 = 8 - 8.3 = -0.3$

$\hat{y}_3 = 8.3 + 3.7x_3 = \hat{y}_1 = 8.3 + 3.7(1) = 12.0$ $\qquad$ $\hat{e}_3 = y_3 - \hat{y}_3 = 12 - 12 = 0.0$

$\hat{y}_4 = 8.3 + 3.7x_4 = \hat{y}_1 = 8.3 + 3.7(2) = 15.7$ $\qquad$ $\hat{e}_4 = y_4 - \hat{y}_4 = 17 - 17.7 = -0.7$

$\hat{y}_5 = 8.3 + 3.7x_5 = \hat{y}_1 = 8.3 + 3.7(3) = 19.4$ $\qquad$ $\hat{e}_5 = y_5 - \hat{y}_5 = 20 - 19.4 = 0.6$

Predicted values and the residuals

We put all the observed, predicted and the residuals in the following table:

i	x	x^2	y	y^2	xy	$\hat{y}$	$\hat{y}^2$	$\hat{e}$	$\hat{e}^2$	$x\hat{e}$	$\hat{e}\,\hat{y}$
1	-1	1	5	25	-5	4.6	21.16	0.4	0.16	-0.4	1.84

2	0	0	8	64	0	8.3	68.89	-0.3	0.09	0	-2.49
3	1	1	12	144	12	12	144.0	0	0	0	0
4	2	4	15	225	30	15.7	246.49	-0.7	0.49	-1.4	-10.99
5	3	9	20	400	60	19.4	376.36	0.6	0.36	1.8	11.64
Σ	**5**	**15**	**60**	**858**	**97**	**60**	**856.9**	**0.0**	**1.1**	**0.0**	**0.0**

When we check the table, we have the folloing interesting results:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \ , \qquad \sum_{i=1}^{n} \hat{e}_i = 0 \ , \qquad \sum_{i=1}^{n} x_i \hat{e}_i = 0 \, , \quad \sum_{i=1}^{n} \hat{y}_i \hat{e}_i = 0 \, , \quad \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \hat{e}_i^2 \ .$$

Actually, for any regression equation including an intercept term, the above equalities are always valid. That's why we use the residual plots in order to check the assumptions. However, the the regression equatio ndoes not include and intercept term ($\alpha = 0$) the above equalities may not satisfy (e.g. the sum of the residual may not be zero, etc.) . In the following, we show that the above equalities are actually true in general.

Consider the simple linear regression line given in (1) and the normal equations given in (3). The simple linear regression equation

$Y_i = \alpha + \beta x_i + e_i \ , i = 1, 2, ..., n$

and the normal equations

$$\sum_{i=1}^{n} Y_i = \hat{\alpha} n \ + \hat{\beta} \ \sum_{i=1}^{n} x_i \quad , \quad \sum_{i=1}^{n} x_i Y_i = \hat{\alpha} \ \sum_{i=1}^{n} x_i \ + \hat{\beta} \ \sum_{i=1}^{n} x_i^2 \ .$$

- $\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = \sum_{i=1}^{n} y_i - \left[n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} x_i \right] = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0$
- $0 = \sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i \ \Rightarrow \ \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$
- $$\sum_{i=1}^{n} x_i \hat{e}_i = \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = \sum_{i=1}^{n} (x_i y_i - \hat{\alpha} x_i - \hat{\beta} x_i^2)$$
  $$= \sum_{i=1}^{n} x_i y_i - \left[\hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2 \right] = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i = 0$$
- $\sum_{i=1}^{n} \hat{y}_i \hat{e}_i = \sum_{i=1}^{n} \hat{e}_i (\hat{\alpha} + \hat{\beta} x_i) = \hat{\alpha} \sum_{i=1}^{n} \hat{e}_i + \hat{\beta} \sum_{i=1}^{n} x_i \hat{e}_i = 0 + 0 = 0 \, .$

ANOVA Table: From the above example, we have also observe that

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \hat{e}_i^2 \text{ and } \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i .$$

The second equality also implies that $\bar{y}_n = \hat{\bar{y}}_n$. Using these equalities, we define the folloing sum of sqaures

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2 \quad , \quad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2 = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}_n^2 \text{ and } SSE = \sum_{i=1}^{n} \hat{e}_i^2 .$$

Here, SST stands for "*total sum of squares*", SSR stands for "*regression sum of squares*" and SSE for "*error sum of squares*". Now, we are ready to construct the ANOVA table. In the following,

$d.f$ is the "*number of degrees of freedom*", SoV *is "source of variation"*, MS is the "*mean squares"* and F is *the "value of* F *statistic"*. The values MS 's are calculated by the corresponding SS 's divided by its degrees of freedoms and the value of F statistic is the ration of MSR and MSE (mean square regression devided by mean square errors)

SoV	$d.f$	SS	MS	F
Regression	1	SSR	MSR	F
Error	$n-2$	SSE	MSE	
Total	$n-1$	SST		

Example: Consider the previous example. We have calculated from the table above ($n = 5$) as

$$\sum_{i=1}^{n} y_i = 60 \ , \quad \sum_{i=1}^{n} y_i^2 = 858 \ , \ \sum_{i=1}^{n} \hat{y}_i^2 = 856.9 \text{ and } \sum_{i=1}^{n} e_i^2 = 1.1 .$$

Therefore,

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2 = 858 - 5(12)^2 = 858 - 5(144) = 138$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2 = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}_n^2 = 856.9 - 5(12)^2 = 856.9 - 5(144) = 136.9 \text{ and } \sum_{i=1}^{n} e_i^2 = 1.1$$

.

Thus the ANOVA table is constructed as follows.

SoV	$d.f$	SS	MS	F

Regression	1	136.9	136.9	373.36
Error	3	1.1	0.3667	
Total	4	138.0		

From the ANOVA table, the value of MSE is an unbiased estimate of σ^2 ($E(MSE) = \sigma^2$) and the ratio

$$R^2 = \frac{SSR}{SST} \tag{5}$$

is the percentage of the variability explained by the model. That is, the percentage of the variability in the dependent variable Y explained by the explanatory variable x. The larger value of R^2 indicates that a better model (x's explain Y's better). In the above example, the OLS estimate of σ^2 is $\hat{\sigma}^2 = MSE = 0.3667$ and

$$R^2 = \frac{MSR}{MSE} = \frac{136.9}{138} \cong 0.992$$

which means that more than 99% of all variability in Y's are explained by the model (or the explanatory variable x).

Example: Assume that the folloowing data set is appropriate for a simple linear regression equation.

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
16	**26.6**	15	**25.4**	21	**31.7**	20	**30.8**	25	**35.7**	19	**29.0**	18	**28.5**
18	**28.7**	23	**33.1**	17	**27.8**	22	**32.5**	20	**30.5**	14	**24.0**	22	**32.5**
17	**27.1**	18	**28.3**	25	**35.4**	21	**31.1**	15	**25.4**	19	**29.3**	12	**23.0**
17	**27.9**	21	**31.6**	19	**29.0**	18	**28.9**	21	**31.2**	23	**33.7**	15	**28.5**
20	**30.2**	19	**30.0**	23	**33.7**	21	**31.1**	21	**31.4**	20	**30.5**	16	**28.7**
16	**26.7**	19	**29.9**	20	**30.6**	19	**29.3**	18	**28.9**	22	**32.2**	13	**24.0**
16	**26.2**	21	**31.5**	15	**25.6**	20	**30.3**	20	**30.8**	13	**23.2**	15	**28.0**
18	**28.3**	27	**37.3**	18	**28.3**	21	**31.1**	19	**29.2**	17	**27.4**	16	**32.2**

For this data set, we assume that a simple linear regression equation is appropriate. That is, we have

$Y_i = \alpha + \beta x_i + e_i$, $i = 1,2,...,56$.

In order to calculate the values of the OLS estimators, some of the calculated values are given below:

$$\sum_{i=1}^{56} x_i = 1054, \qquad \sum_{i=1}^{56} x_i^2 = 20380, \qquad \sum_{i=1}^{56} (x_i - \bar{x}_n)^2 = 542.214286, \ \bar{x}_n = 18.824286$$

$$\sum_{i=1}^{56} y_i = 1653.8, \qquad \bar{y}_n = 29.5321429, \qquad \sum_{i=1}^{56} x_i y_i = 31618.6.$$

Therefore (by considering the roundig errors) we have

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n \bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2} = \frac{31618.6 - 56(18.82143)(29.53214)}{542.21} = \frac{491.722}{542.214} \cong 0.90688$$

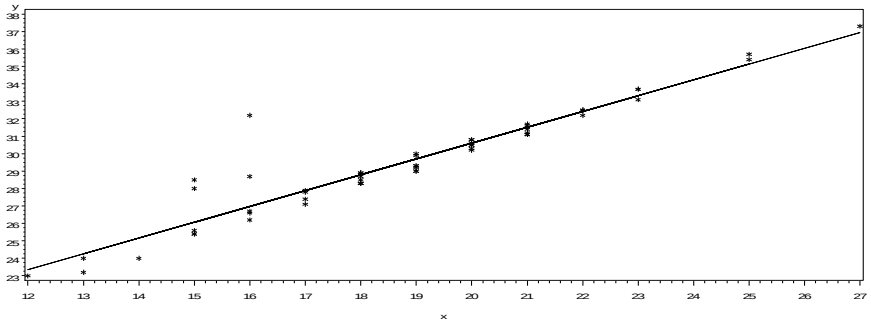
and

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n = 29.532149 - (0.90688)(18.8214286) \cong 12.46343$$

and therefore the fitted regression line is,

$$\hat{Y}_i = 12.46 + 0.907 x_i \ , i = 1,2,...,56 .$$

The plot of the observed values (with "stars") and the fitted regression line (the straight line) is given in the following figure.



The plot of observed values and the fitted regression line

Some other statistical values have been calculated bu running the following SAS codes. You can also verify the same results in any statistical package program (or in excel).

```
data a; input x y@@; cards;
16 26.6 15 25.4 21 31.7 20 30.8 25 35.7 19 29.0 18 28.5
18 28.7 23 33.1 17 27.8 22 32.5 20 30.5 14 24.0 22 32.5
17 27.1 18 28.3 25 35.4 21 31.1 15 25.4 19 29.3 12 23.0
17 27.9 21 31.6 19 29.0 18 28.9 21 31.2 23 33.7 15 28.5
```

```
20 30.2 19 30.0 23 33.7 21 31.1 21 31.4 20 30.5 16 28.7
16 26.7 19 29.9 20 30.6 19 29.3 18 28.9 22 32.2 13 24.0
16 26.2 21 31.5 15 25.6 20 30.3 20 30.8 13 23.2 15 28.0
18 28.3 27 37.3 18 28.3 21 31.1 19 29.2 17 27.4 16 32.2
;
proc reg; model y=x;
output out=out predicted=yhat residual=resid;
proc print data=out; var x y yhat resid;
run;
```

SAS Codes to analyze the data

The output of the above SAS codes is given below. When we investigate the output, we observe that almost 90% of all variability in the dependent variable is explained by the model (x's). The parameter estimates are almost the same as we calculated above (rounding error). later, we are going to use the same data to make some statistical inferences about the parameters (hypothesis testing and confidence intervals). In the following, we also printed out the predicted values and the residuals obtained from the model.

Obs	x	y	yhat	resid
1	16	26.6	26.9735	-0.37346
2	15	25.4	26.0666	-0.66658
3	21	31.7	31.5078	0.19216
4	20	30.8	30.6010	0.19904
5	25	35.7	35.1353	0.56466
6	19	29.0	29.6941	-0.69409
7	18	28.5	28.7872	-0.28721
8	18	28.7	28.7872	-0.08721
9	23	33.1	33.3216	-0.22159
10	17	27.8	27.8803	-0.08033

11	22	32.5	32.4147	0.08529
12	20	30.5	30.6010	-0.10096
13	14	24.0	25.1597	-1.15970
14	22	32.5	32.4147	0.08529
15	17	27.1	27.8803	-0.78033
16	18	28.3	28.7872	-0.48721
17	25	35.4	35.1353	0.26466
18	21	31.1	31.5078	-0.40784
19	15	25.4	26.0666	-0.66658
20	19	29.3	29.6941	-0.39409
21	12	23.0	23.3459	-0.34595
22	17	27.9	27.8803	0.01967
23	21	31.6	31.5078	0.09216
24	19	29.0	29.6941	-0.69409
25	18	28.9	28.7872	0.11279
26	21	31.2	31.5078	-0.30784
27	23	33.7	33.3216	0.37841
28	15	28.5	26.0666	2.43342
29	20	30.2	30.6010	-0.40096
30	19	30.0	29.6941	0.30591
31	23	33.7	33.3216	0.37841
32	21	31.1	31.5078	-0.40784
33	21	31.4	31.5078	-0.10784
34	20	30.5	30.6010	-0.10096
35	16	28.7	26.9735	1.72654
36	16	26.7	26.9735	-0.27346
37	19	29.9	29.6941	0.20591

38	20	30.6	30.6010	-0.00096
39	19	29.3	29.6941	-0.39409
40	18	28.9	28.7872	0.11279
41	22	32.2	32.4147	-0.21471
42	13	24.0	24.2528	-0.25283
43	16	26.2	26.9735	-0.77346
44	21	31.5	31.5078	-0.00784
45	15	25.6	26.0666	-0.46658
46	20	30.3	30.6010	-0.30096
47	20	30.8	30.6010	0.19904
48	13	23.2	24.2528	-1.05283
49	15	28.0	26.0666	1.93342
50	18	28.3	28.7872	-0.48721
51	27	37.3	36.9491	0.35090
52	18	28.3	28.7872	-0.48721
53	21	31.1	31.5078	-0.40784
54	19	29.2	29.6941	-0.49409
55	17	27.4	27.8803	-0.48033
56	16	32.2	26.9735	5.22654

The data, observed and predicted values and residuals

the parameter estimates and the results of regression analysis of dependent variable Y on the explanatory (independent variable) variable x is given below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	445.93064	445.93064	**484.01**	**<.0001**

Error 54 49.75150 0.92132

Corrected Total 55 495.68214

****

Root MSE 0.95986 R-Square **0.8996**

Dependent Mean 29.53214 Adj R-Sq **0.8978**

Coeff Var 3.25021

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	**12.46343**	0.78637	15.85	<.0001
x	1	**0.90688**	0.04122	22.00	<.0001

Estimation results and the ANOVA table

Statistical Inference on the Regression Parameters:

Consider the following simple linear regression equation given in (1)

$Y_i = \alpha + \beta x_i + e_i$, $i = 1, 2, ..., n$.

Here, α and β are the parameters to be estimated. We estimated these parameters (OLS) as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{\alpha} = \bar{Y}_n - \hat{\beta}\,\bar{x}_n$$

and we calculeted their means and the variances as

$$E(\hat{\beta}) = \beta, \quad E(\hat{\alpha}) = \alpha, \; Var(\hat{\beta}) = \frac{\sigma^2}{S_{xx}} \text{ and } Var(\hat{\alpha}) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}}\right].$$

When we assume the normality of the error terms, the OLS estimators $\hat{\alpha}$ ând $\hat{\beta}$ are also normally distributed random variables (because they are linear combinations of independent normally distributed random variables), namely

$\hat{\beta} \sim N(\beta, Var(\hat{\beta}))$ and $\hat{\alpha} \sim N(\alpha, Var(\hat{\alpha}))$.

These distributional properties imply that

$$\frac{\hat{\beta}-\beta}{\sqrt{Var(\hat{\beta})}} \sim N(0,1), \quad \frac{\hat{\alpha}-\alpha}{\sqrt{Var(\hat{\alpha})}} \sim N(0,1)$$

and when the variance of the error term is unknown

$$\frac{\hat{\beta}-\beta}{s(\hat{\beta})} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{\alpha}-\alpha}{s(\hat{\alpha})} \sim t_{n-2}$$

where

$$s^2(\hat{\alpha}) = MSE\left(\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right) \text{ and } s^2(\hat{\beta}) = \frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}.$$

Hypothesis testing for β :

We are going to consider the following hypotheses. For any one of the hypothesis we calculate the same value of t statistic as

$$t_h = \frac{\hat{\beta}-\beta_0}{s(\hat{\beta})}.$$

The hypotheses and their rejection rules are summarized below.

	Hypothesis	Rejection rule.
a	$H_0 : \beta = \beta_0$ against $H_a : \beta > \beta_0$	Reject H_0 if $t_h > t_{n-2}(\alpha)$, $P(t_{n-2} > t_{n-2}(\alpha)) = \alpha$
b	$H_0 : \beta = \beta_0$ against $H_a : \beta < \beta_0$	Reject H_0 if $t_h < -t_{n-2}(\alpha)$, $P(t_{n-2} > t_{n-2}(\alpha)) = \alpha$
c	$H_0 : \beta = \beta_0$ against $H_a : \beta \neq \beta_0$	Reject H_0 if $\lvert t_h \rvert > t_{n-2}(\alpha/2)$, $P(t_{n-2} > t_{n-2}(\alpha)) = \alpha$

Hypothesis testing for α :

We are going to consider the following hypotheses. For any one of the hypothesis we calculate the same value of t statistic as

$$t_h = \frac{\hat{\alpha}-\alpha_0}{s(\hat{\alpha})}.$$

The hypotheses and their rejection rules are summarized below.

	Hypothesis	Rejection rule.
a	$H_0 : \alpha = \alpha_0$ against $H_a : \alpha > \alpha_0$	Reject H_0 if $t_h > t_{n-2}(\alpha)$, $P(t_{n-2} > t_{n-2}(\alpha)) = \alpha$
b	$H_0 : \alpha = \alpha_0$ against $H_a : \alpha < \alpha_0$	Reject H_0 if $t_h < -t_{n-2}(\alpha)$, $P(t_{n-2} > t_{n-2}(\alpha)) = \alpha$
c	$H_0 : \alpha = \alpha_0$ against $H_a : \alpha \neq \alpha_0$	Reject H_0 if $\lvert t_h \rvert > t_{n-2}(\alpha/2)$, $P(t_{n-2} > t_{n-2}(\alpha)) = \alpha$

Example: Consider the above data set. Assume that the following data is appropriate for a simple linear regression. That is, we have only 5 observations.

$X = x$	–1	0	1	2	3
$Y = y$	5	8	12	15	20

We have calculated the values of the OLS estimators and some related values of the statistics as

$\hat{\alpha} = 8.3$, $s(\hat{\alpha}) = 0.33166$, $\hat{\beta} = 3.7$ and $s(\hat{\beta}) = 0.19149$.

The SAS codes and the results of the analysis are summarized below.

```
data a; input x y@@; cards;
-1 5 0 8 1 12 2 15 3 20
;
proc reg; model y=x; run;
```

```
**********************************************************************

                Analysis of Variance

                   Sum of      Mean

Source          DF     Squares    Square    F Value    Pr > F

Model            1   136.90000   136.90000   373.36   0.0003

Error            3     1.10000     0.36667
```

Corrected Total	4	138.00000			
**********					
Root MSE		0.60553	R-Square	0.9920	
Dependent Mean		12.00000	Adj R-Sq	0.9894	
Coeff Var		5.04608			
**********					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.30000	0.33166	25.03	0.0001
x	1	3.70000	0.19149	19.32	0.0003

SAS Codes and the results of the analysis

Tests for slope β

a) Suppose we want to test the null hypothesis $H_0 : \beta = 3$ against $H_a : \beta > 3$ at 5% level. The value of the test statistic and the critical value are found as

$$t_h = \frac{\hat{\beta} - 3}{s(\hat{\beta})} = \frac{3.7 - 3}{0.19149} \cong 3.655 \text{ and } P(t_3 > t_3(0.05)) = 0.05 \Rightarrow t_3(0.05) = 2.353$$

and since $t_h = 3.655 > 2.353 = t_3(0.05))$ we reject the null hypothesis at 5% level.

Power:Tthe power of the test is always the probability of rejecting the null hypothesis. Therefore, the power function

$$Power = P(reject\, H_0) = P_\beta\left(\frac{\hat{\beta} - 3}{s(\hat{\beta})} > 2.353\right) = P_\beta\left(\frac{\hat{\beta} - \beta + \beta - 3}{s(\hat{\beta})} > 2.353\right)$$

$$= P\left(t_3 > 2.353 - \frac{\beta - 3}{s(\hat{\beta})}\right)$$

and therefore the ampirical power (at the calculated value $\beta = 3.7$) is

$$Power = P\left(t_3 > 2.353 - \frac{3.7 - 3}{0.19149}\right) = P\left(t_3 > -1.3025\right) \cong 0.858$$

b) Now, we want to test the null hypothesis of $H_0 : \beta = 4$ against the alternative $H_a : \beta < 4$ at 5% level. We are going to use the same critical value ($t_3(0.05) = 2.353$) and the value of the test statistic is

$$t_h = \frac{\hat{\beta} - 4}{s(\hat{\beta})} = \frac{3.7 - 4}{0.19149} \cong -1.567 .$$

As a conclution, we are going to reject $H_0 : \beta = 4$ if $t_h < -t_3(0.05))$ but since

$$t_h = -1.567 > -2.353 = -t_3(0.05))$$

we fail to reject the null hypothesis $H_0 : \beta = 4$ at 5% level.

c) Finally we want to test the null hypothesis of $H_0 : \beta = 0$ against the alternative $H_a : \beta \neq 0$ at 5% level. Now, the value of the test statistic

$$t_h = \frac{\hat{\beta}}{s(\hat{\beta})} = \frac{3.7}{0.19149} \cong 19.32$$

and the critical value is $P(t_3 > t_3(0.025)) = 0.025 \Rightarrow t_3(0.025) = 3.182$ ($\alpha = 0.05$ we look at for $\alpha / 2$) .

As a conclusion we are going to reject the null if $|t_h| > t_{n-2}(\alpha / 2)$ and since

$$|t_h| = 19.32 > 3.182 = t_{n-2}(\alpha / 2)$$

we reject the null hypothesis at 5% level.

Tests for intercept α

a) Let us try to test the null hypothesis of $H_0 : \alpha = 8$ against $H_a : \alpha \neq 8$ at 5% level. the value of the test statistics is

$$t_h = \frac{\hat{\alpha} - 8}{s(\hat{\alpha})} = \frac{8.3 - 8}{0.33166} \cong 0.9045$$

and the critical value (for 2-sided case) is $P(t_3 > t_3(0.025)) = 0.025 \Rightarrow t_3(0.025) = 3.182$. Since,

$$|t_h| = 0.9045 < 3.182 = t_{n-2}(\alpha / 2)$$

we fail to reject the null hypothesis at 5% level.

b) If we wanted to test $H_0 : \alpha = 7$ against $H_a : \alpha > 7$ at 5% level, the value of the test statistic is

$$t_h = \frac{\hat{\alpha} - 7}{s(\hat{\alpha})} = \frac{8.3 - 7}{0.33166} \cong 3.92$$

and we reject the null hypothesis if $t_h > t_{n-2}(\alpha)$. For $\alpha = 0.05$ the critical value $t_3(0.05) = 2.353$. Since, $t_h = 3.92 > 2.353 = t_3(0.05)$ we reject $H_0 : \alpha = 7$ at 5% level against the alternative of $H_a : \alpha > 7$.

Power: Tthe power of the test is the probability of rejecting the null hypothesis. Therefore, the power function

$$Power = P(reject\, H_0) = P_\alpha\left(\frac{\hat{\alpha} - 7}{s(\hat{\alpha})} > 2.353\right) = P_\alpha\left(\frac{\hat{\alpha} - \alpha + \alpha - 7}{s(\hat{\alpha})} > 2.353\right)$$
$$= P\left(t_3 > 2.353 - \frac{\alpha - 7}{s(\hat{\alpha})}\right).$$

Thus the ampirical power (at the calculated value $\alpha = 8.3$) is

$$Power = P\left(t_3 > 2.353 - \frac{8.3 - 7}{0.33166}\right) = P\left(t_3 > -1.5667\right) \cong 0.892.$$

Confidence Intervals:

Confidence intervals for the parameters are very similar to the confidence intervals for the normal means which we have already studied. The confidence intervals for α and β are

$(1-\alpha)100\%$ confidence interval for α is $\hat{\alpha} \pm s(\hat{\alpha}) t_{n-2}(\alpha/2)$

and

$(1-\alpha)100\%$ confidence interval for β is $\hat{\beta} \pm s(\hat{\beta}) t_{n-2}(\alpha/2)$.

Example: We consider the data set given in the previous example and try to write 95% confidence intervals for the parameters α and β. In the above discussion, we have calculated

$$\hat{\alpha} = 8.3,\ s(\hat{\alpha}) = 0.33166\ ,\ \hat{\beta} = 3.7 \text{ and } s(\hat{\beta}) = 0.19149$$

If we want to write a 95% confidence interval $\alpha = 0.05$ and from $P(t_3 > t_3(0.025)) = 0.025$ we get the critical value from the t table as $t_3(0.025) = 3.182$. Now we are ready to write 95% confidence intervals for the regression parameters as:

- 95% confidence interval for α is $\hat{\alpha} \pm s(\hat{\alpha}) t_{n-2}(\alpha/2)$

  $8.3 \pm (0.33166)(3.182)$ or $(7.245, 9.355)$

- 95% confidence interval for β is $\hat{\beta} \pm s(\hat{\beta}) t_{n-2}(\alpha/2)$

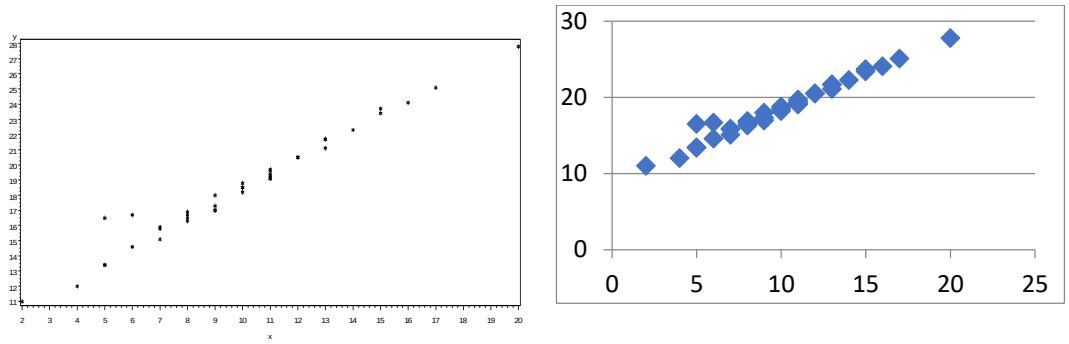
  $3.7 \pm (0.19149)(3.182)$ or $(3.091, 4.309)$.

Example: In the above example we had only 5 observations and therefore some of the statistical inferences may not be significant. To summarize the regression concept we consider the following data set (now we have $n = 40$ observations).

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
6	14.6	9	17.0	12	20.5	8	16.3	2	11.0	11	19.2	13	21.7	11	19.3
5	13.4	8	16.5	10	18.5	15	23.4	7	15.9	13	21.7	11	19.1	14	22.3
11	19.7	8	16.7	4	12.0	11	19.1	11	19.6	5	16.5	11	19.4	16	24.1
10	18.8	13	21.1	12	20.5	5	13.4	9	17.0	10	18.2	10	18.5	17	25.1
15	23.7	7	15.8	7	15.1	9	17.3	8	16.9	9	18.0	6	16.7	20	27.8

A scatter plot of Y values against x's are given below figure. As it is seen from the scatter plot, there seems to be a linear relationship between Y's and x's. Therefore, it is reasonable to consider a simple linear regression equation between these two variables as

$$Y_i = \beta_0 + \beta_1 x_i + e_i \ , \ i = 1, 2, ..., 40 .$$



Scatter plot of Y values against x's

Based on the data, we calculate

$$\sum_{i=1}^{40} x_i = 399 \ , \qquad \bar{x}_n = 9.975 \ , \qquad \sum_{i=1}^{40} (x_i - \bar{x}_n)^2 = 546.975 ,$$

$$\sum_{i=1}^{40} y_i = 741.4 , \qquad \bar{y}_n = 18.535 , \qquad \sum_{i=1}^{40} (y_i - \bar{y}_n)^2 = 486.991 , \qquad \sum_{i=1}^{40} x_i y_i = 7903.7 .$$

Using these values, we calculate the values of the OLS estimators as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{40} x_i y_i - n \bar{x}_n \bar{y}_n}{\sum_{i=1}^{40} (x_i - \bar{x}_n)^2} = \frac{7903.7 - 40(9.975)(18.535)}{546.975} = 0.929174093 \cong 0.92917$$

and

$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n = 18.535 - 0.929174093(9.975) = 9.266488422 \cong 9.26649$

and the fitted regression line is

$$\boxed{\hat{Y}_i = 9.26649 + 0.92917\, x_i \ , \ i = 1.2,...,40.}$$

In order to construct the ANOVA table, from the fitted regression equation, we calculate the predicted values and the residual. The sum of squares are calculated as,

$SST = \sum_{i=1}^{40}(y_i - \bar{y}_n)^2 = 486.991$, $SSR = \sum_{i=1}^{40}(\hat{y}_i - \bar{y}_n)^2 = 472.2388$ and $SSE = \sum_{i=1}^{40}\hat{e}_i^2 = 14.7522$

and thus the ANOVA table constructed as foloows.

SoV	*d.f*	*SS*	*MS*	*F*
Regression	1	472.2388	472.2388	1216.43
Error	38	14.7522	0.38822	
Total	39	486.9910		

From the ANOVA table the value of $R^2 = SSR / SST = 472.2388 / 486.991 = 0.9697$ which means that more than 96% of all variability in Y is explained by the model (or x's). The other results of the regression analysis incluging the predicted values and the residuall are calculated by using the SAS codes given below.

```
data a; input x y@@; cards;
 6 14.6  9 17.0 12 20.5  8 16.3  2 11.0 11 19.2 13 21.7 11 19.3
 5 13.4  8 16.5 10 18.5 15 23.4  7 15.9 13 21.7 11 19.1 14 22.3
11 19.7  8 16.7  4 12.0 11 19.1 11 19.6  5 16.5 11 19.4 16 24.1
10 18.8 13 21.1 12 20.5  5 13.4  9 17.0 10 18.2 10 18.5 17 25.1
15 23.7  7 15.8  7 15.1  9 17.3  8 16.9  9 18.0  6 16.7 20 27.8
;
proc reg; model y=x;
output out=out predicted=yhat residual=ehat;
proc print data=out; var x y yhat ehat;
run;
```

****

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	472.23880	472.23880	1216.43	<.0001
Error	38	14.75220	0.38822		
Corrected Total	39	486.99100			

****

Root MSE	0.62307	R-Square	0.9697
Dependent Mean	18.53500	Adj R-Sq	0.9689
Coeff Var	3.36158		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.26649	0.28342	32.70	<.0001
x	1	0.92917	0.02664	34.88	<.0001

SAS Codes, ANOVA table and Parameter estimates

Some statistical Inferences about the parameters:

From the results of SAS codes in the above we have

$$\hat{\beta}_1 = 0.92917,\ \hat{\beta}_0 = 9.26649,\ s(\hat{\beta}_1) = 0.02664 \text{ and } s(\hat{\beta}_0) = 0.28342$$

and from the t table,

$$t_{38}(0.05) = 1.6859 \text{ and } t_{38}(0.025) = 2.0244$$

where $P(t_{38} > t_{38}(0.05)) = 0.05$ and $P(t_{38} > t_{38}(0.025)) = 0.025$.

If we want to test the null hypothesis $H_0 : \beta_1 = 0.9$ against $H_a : \beta_1 > 0.9$ at 5% level the value of the test statistic is

$$t_h = \frac{\hat{\beta}_1 - 0.9}{s(\hat{\beta}_1)} = \frac{0.92917 - 0.9}{0.02664} \cong 1.095 .$$

As a conclusion we reject the null hypothesis if $t_h > t_{38}(0.05)$ and since

$$t_h = 1.095 < 1.6859 = t_{38}(0.05)$$

we fail to reject the null hypothesis $H_0 : \beta_1 = 0.9$ at 5% level.

We can also write confidence interval for the slope parameter β_1. A 95% confidence interval for the slope (with rounding) is

$\hat{\beta}_1 \pm s(\hat{\beta}_1) t_{38}(0.025)$ or $0.929 \pm (0.02664)(2.0244)$ or 0.929 ± 0.054 or $(0.875, 0.983)$.

Now, let us test the null hypothesis of $H_0 : \beta_0 = 8.5$ against $H_a : \beta_0 > 8.5$ at 5% level the value of the test statistic is

$$t_h = \frac{\hat{\beta}_0 - 8.5}{s(\hat{\beta}_0)} = \frac{9.26 - 8.5}{0.28342} \cong 2.68 .$$

As a conclusion we reject the null hypothesis if $t_h > t_{38}(0.05)$ and since

$$t_h = 2.68 > 1.6859 = t_{38}(0.05)$$

we reject the null hypothesis $H_0 : \beta_0 = 8.5$ at 5% level.

We can also write confidence interval for the slope parameter β_0. A 95% confidence interval for the intercept term (with rounding) is

$\hat{\beta}_0 \pm s(\hat{\beta}_0) t_{38}(0.025)$ or $9.26 \pm (0.28342)(2.0244)$ or 9.26 ± 0.574 or $(8.686, 9.834)$.

Assumptions:

In order to say that the model

$$Y_i = \beta_0 + \beta_1 x_i + e_i \ , \ i = 1, 2, ..., 40$$

is a regression model and to do any statistical inference about the parameters, we assumed that the error terms will be independent and identically distributed random variables. We usually use the residuals to check these assumptions. As we remember, we the residuals are orthogolal to the explonatory observations and to the predicted values (sum of the products is zero). Moreover, the sum of the residuals is also zero. That is, any pattern in these plots indicates a violotion of the assumtions. That is, we need to look at plot of residuals versus x's, residuals versus $\hat{y}$'s and residuals versus observations . If we see any kind of pattern we need to make some transformation to verify the assumtion. Moreover, in order to do any statistical inferences about the parameters we need the normality assumption of the error terms. There are many ways

to check the normality assumptions (Kolmogorov-Simirnov, Shapiro-Wilk, Cramer-von Mises, Anderson-Darling tests, and some plots like PP-plot, QQ plot and normal probability plot, Box-Cox plot, etc.) but here we are going to look at the normal probability plot. If we observe a linearity in the normal probability plot, we can conclude that the error terms are normally distributed. The residuals and the predicted values are obtained from the fitted regression line and given in the following table.

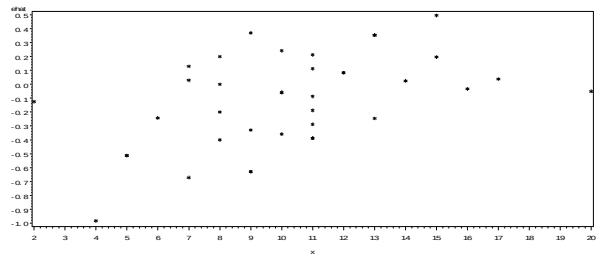
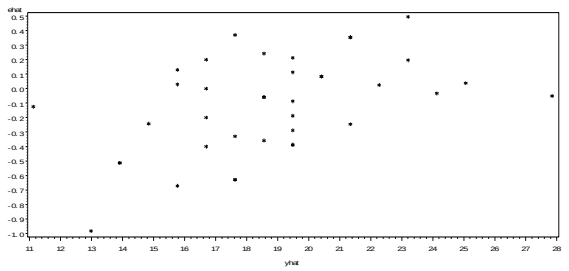
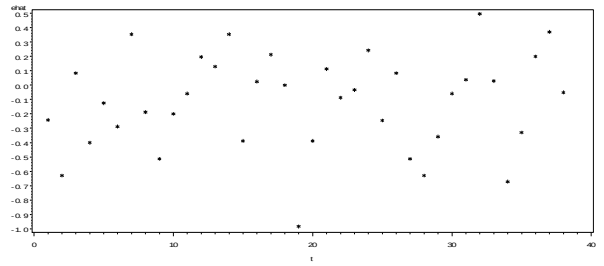
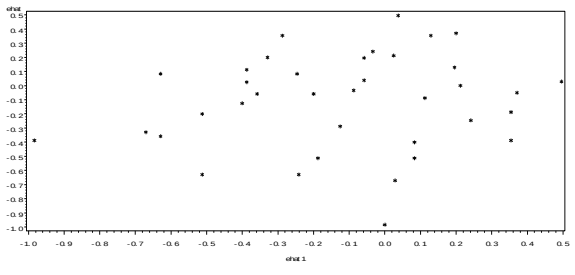
Obs	x	y	yhat	ehat
1	6	14.6	14.8415	-0.24153
2	9	17.0	17.6291	-0.62906
3	12	20.5	20.4166	0.08342
4	8	16.3	16.6999	-0.39988
5	2	11.0	11.1248	-0.12484
6	11	19.2	19.4874	-0.28740
7	13	21.7	21.3458	0.35425
8	11	19.3	19.4874	-0.18740
9	5	13.4	13.9124	-0.51236
10	8	16.5	16.6999	-0.19988
11	10	18.5	18.5582	-0.05823
12	15	23.4	23.2041	0.19590
13	7	15.9	15.7707	0.12929
14	13	21.7	21.3458	0.35425
15	11	19.1	19.4874	-0.38740
16	14	22.3	22.2749	0.02507
17	11	19.7	19.4874	0.21260
18	8	16.7	16.6999	0.00012
19	4	12.0	12.9832	-0.98318
20	11	19.1	19.4874	-0.38740

21	11	19.6	19.4874	0.11260
22	5	16.5	13.9124	2.58764
23	11	19.4	19.4874	-0.08740
24	16	24.1	24.1333	-0.03327
25	10	18.8	18.5582	0.24177
26	13	21.1	21.3458	-0.24575
27	12	20.5	20.4166	0.08342
28	5	13.4	13.9124	-0.51236
29	9	17.0	17.6291	-0.62906
30	10	18.2	18.5582	-0.35823
31	10	18.5	18.5582	-0.05823
32	17	25.1	25.0624	0.03755
33	15	23.7	23.2041	0.49590
34	7	15.8	15.7707	0.02929
35	7	15.1	15.7707	-0.67071
36	9	17.3	17.6291	-0.32906
37	8	16.9	16.6999	0.20012
38	9	18.0	17.6291	0.37094
39	6	16.7	14.8415	1.85847
40	20	27.8	27.8500	-0.04997

Predicted values (yhat) and residuals (ehat)

In the following residual plots indicate that the assumptions of the regression equation are valid. As we see from the plot of residuals against the x's, $\hat{y}_i$ s and i's no pattern (linear, parabolic, etc.) appears. Moreover, in the last plot ($\hat{e}_i$ versus $\hat{e}_{i-1}$) again no pattern appears. This plot indicates that there is no first order autocorrecation in the error term. We may also check some other order of autocorrelation (plotting $\hat{e}_i$ against for example $\hat{e}_{i-2}$ for the second

order of autocorrelation). There is also test for checking autocorrelations (Durbin-Watson test) but we are not going to investigate here.

Residuals versus x's ($\hat{e}_i$ vs. x_i)	Residuals versus predicted values $\hat{e}_i$ vs. $\hat{y}_i$)
	
Residuals versus observations ($\hat{e}_i$ vs. i)	Plot of $\hat{e}_i$ versus $\hat{e}_{i-1}$
	

Checking for Normality:

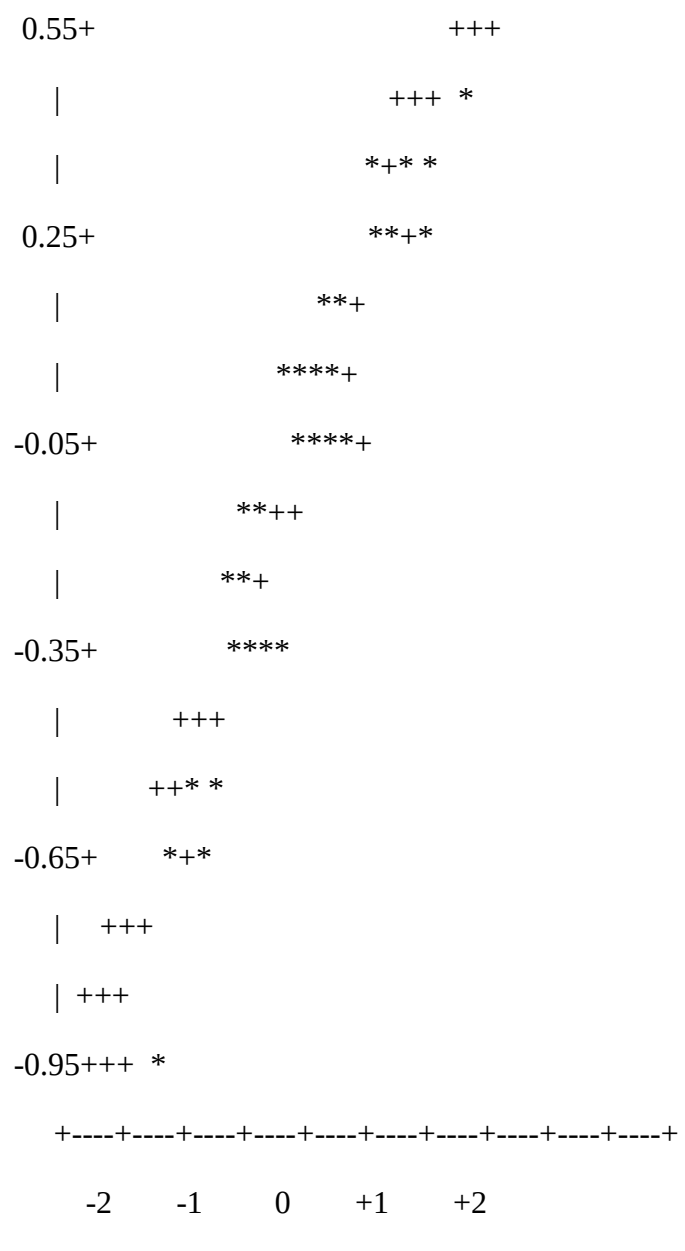
One of the main assumption in any statistical inference is the normality of the data. As it is mentioned above, there are many ways to look at the normality. In the following, we are going to look at the normal probability plot of the residuals.

Tests for Normality

Test	--Statistic---	-----p Value------
Shapiro-Wilk	W 0.980542	Pr < W 0.7367
Kolmogorov-Smirnov	D 0.096539	Pr > D >0.1500
Cramer-von Mises	W-Sq 0.037226	Pr > W-Sq >0.2500
Anderson-Darling	A-Sq 0.23493	Pr > A-Sq >0.2500

****

Normal Probability Plot



Normality tests and the Normal Probability Plot

If we observe a linearity in the normal probability plot, we can assume the normality of the error term. In the following, the normal probability plot and some results of the normality test have been run and given in the above table. A linearity is obvious in the normal probability plot and therefore it is reasonable to assume the normality of the error terms in the regression equation.