Correlation between expected values and actual observed values

We have $\hat{Y}_i = Y_i$ في $\hat{Y} = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$

$$r_{yy} = \frac{S_{yy}}{\sqrt{S_{yy}S_{\hat{y}\hat{y}}}} = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{S_{yy}(\sum(\hat{y}_i - \bar{\hat{y}})^2)}}$$

$$= \frac{\sum(y_i - \bar{y})(\bar{y} - \hat{\beta}_1(x_i - \bar{x}) - \bar{y})}{\sqrt{S_{yy}\sum(\bar{y} - \hat{\beta}_1(x_i - \bar{x}) - \bar{y})^2}} = \frac{\hat{\beta}_1\sum(y_i - \bar{y})(x_i - \bar{x})}{[S_{yy}\sum\hat{\beta}_1^2(x_i - \bar{x})^2]^{\frac{1}{2}}}$$

$$= \frac{\hat{\beta}_1\sum(y_i - \bar{y})(x_i - \bar{x})}{[S_{yy}\hat{\beta}_1^2\sum(x_i - \bar{x})^2]^{\frac{1}{2}}} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{S_{yy}S_{xx}}}] = r_{xy} \Longrightarrow r_{y\hat{y}} = r_{xy}$$

$\therefore R = r_{y\hat{y}} = |r_{xy}|$ Multiple correlation coefficient

This relationship occurs in simple linear regression, and is called multiple because $\hat{y}$ comes from several values of *x*, but in reality it is a simple relationship, i.e.:
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_1 x_{i2} + \ \ldots$

As for the sign of $r_{y\hat{y}}$, it is:

$$r_{y\hat{y}} = \frac{\hat{\beta}_1 S_{xy}}{\sqrt{(S_{yy})(\hat{\beta}_1^2 S_{xx})}} = \frac{\frac{S_{xy}}{S_{xx}}S_{xy}}{\sqrt{(S_{yy})(\hat{\beta}_1^2 S_{xx})}} = \frac{\frac{S^2{}_{xy}}{S_{xx}}}{\sqrt{(S_{yy})(\hat{\beta}_1^2 S_{xx})}} = \frac{\text{positive}}{\sqrt{\text{positive}}} = \text{positive}$$

$\therefore R = r_{y\hat{y}} = |r_{xy}| \implies 0 \le r_{y\hat{y}} \le 1$

Now you can use the correlation coefficient $r_{y\hat{y}}$ that we explained previously to obtain the coefficient of determination as follows:

$$r_{y\hat{y}} = \frac{S_{y\hat{y}}}{\sqrt{S_{yy}S_{y\hat{y}}}} \Longrightarrow r^2{}_{y\hat{y}} = \frac{S^2{}_{y\hat{y}}}{S_{yy}S_{y\hat{y}}}$$

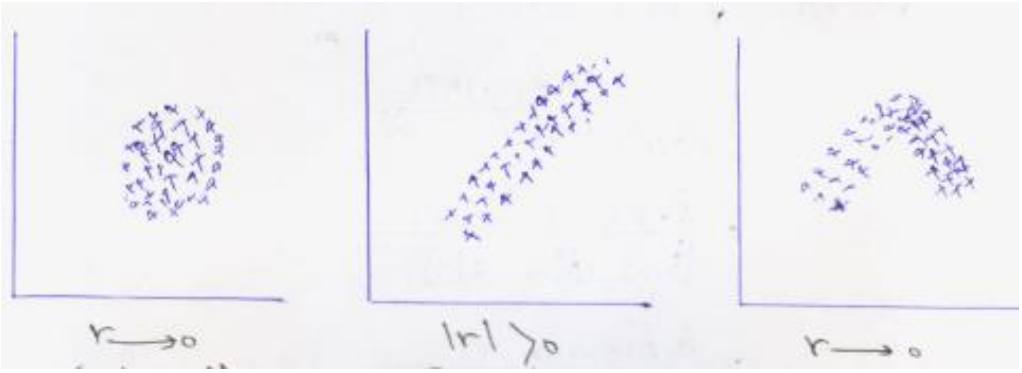
$$S^2{}_{y\hat{y}} = [\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2$$

$$= [\sum(y_i - \bar{y})(\bar{y} + \hat{\beta}_1(X_i - \bar{X}) - \bar{y})]^2$$

$$= [\hat{\beta}_1\sum(y_i - \bar{y})(X_i - \bar{X})]^2 = [\hat{\beta}_1 S_{xy}]^2 = [SS\ Due\ to\ Regression]^2$$

Correlation Coefficient

It is a measure of the strength of the linear relationship between two variables that the correlation coefficient considers to be independent. We can observe the types of relationships between variables through the following graphic forms:



No Relation (r approaches zero) | Strong Linear Relation | Non-Linear Relation

$$r = \frac{S_{xy}}{\sqrt{S_{yy}S_{xx}}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}\ ]\ =\ = \sqrt{\frac{S_{xy}}{S_{xx}}}.\frac{S_{xy}}{S_{yy}} = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}}$$

Relation Between r and $\hat{\beta}_1$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}.\frac{\sqrt{S_{xx}}}{\sqrt{S_{xx}}} = \frac{S_{xy}}{S_{xx}}.\sqrt{\frac{S_{xx}}{S_{yy}}} = \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}} \qquad \dots 1$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.\frac{\sqrt{S_{yy}}}{\sqrt{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.\sqrt{\frac{S_{yy}}{S_{xx}}} = r.\sqrt{\frac{S_{yy}}{S_{xx}}} \qquad \dots 2$$

Or we find 2 in 1 and using these relations the ANOVA table becomes as follows:

S.O.V	d.f.	SS
$R(X_1)$	1	$SSR(X_1) = \hat{\beta}_1^2 S_{xx} = (r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}})^2 S_{xx} = r^2\frac{S_{yy}}{S_{xx}}S_{xx} = r^2 S_{yy}$
Error	n-2	$SSe = S_{yy} - \hat{\beta}_1^2 S_{xx} = S_{yy} - r^2 S_{yy} = (1-r^2)S_{yy}$
Total	n-1	$SST = S_{yy}$

Example: If the correlation coefficient between the dependent and independent variable is r=0.5 and the total sum of squares = 100, test the importance of the model in explaining the changes that occurred in y if the model was found in a study of 12 observations (n=12)

Solution:-

$H_0 : \beta_0 = 0$
$H_1 : \beta_1 \neq 0$

S.O.V	d.f.	SS	MS	M
$\mathbf{R(X_1)}$	1	$SSR = r^2 S_{yy} = (0.5)^2 100 = 25$	25	$\text{Cal. F} = \frac{25}{75} = 3.33$
Error	10	$\text{SSe} = (1 - r^2)\text{S} = (1 - 0.25)(100) = 75$	7.5	
Total	11	$SS_T = 100$		

$$tab.F_{(0.5,1.10)} = 4.96$$

Accept the null hypothesis, i.e. there is no significance of the model through x in explaining the changes occurring in y.

Test of Lack of fit

In the problem of regression there are two types of errors, as is clear from the figure.

There are two reasons for this error:

1. The values are scattered around the regression line
2. The model is not a good fit to the data

The error here is due to the scattering of values around the regression line only.

If the error resulting from the model not fitting is significant (influential), we say that the model does not fit the data, where: -

Pure Error:- The pure or unadulterated error.

Lack of Fit:- The error of lack of fit.

The lack of fit test is conducted to test the fit of the linear model to the data or explain whether the relationship between X and Y is linear or not.

$\mathbf{H_0}$**:-** There is no lack of fit, meaning that the linear model fits the data.
$\mathbf{H_1}$**:-** There is a lack of fit, meaning that the linear model does not fit the data.

Note:- If the hypothesis $\mathbf{H_0}$ is accepted and we test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ and accept H_0, this means that the relationship is indeed linear, but x is not the variable that explains the changes in y, meaning that the whole model is useless in explaining the changes in y.

But if we accept H_0: there is no deficiency in the fit and we accept the hypothesis $H_1 : \beta_1 \neq 0$, meaning that the model is actually linear and that x is important in explaining the changes that occur in y, meaning that the model is ideal for explaining the changes that occur in y.

If you accept H_1, there is a lack of fit, then you accept $H_0 : \beta_1 = 0$, then the correct model is non-linear, i.e. the model is basically rejected.
If you accept H_1 (the correct model is non-linear), then you accept $H_1 : \beta_1 \neq 0$, then if we reject H_1, it is not acceptable to test $\beta_1 \neq 0$, because, the model is not linear, i.e. there is no need for it, i.e. the test is useless.
So, the lack of fit test is done before testing β_1
Note: The lack of fit test is performed if there is a real repetition of the data, where
The real repetition: repetition of a specific case.
The unreal repetition: re-reading a specific value, for example, a 17-year-old person whose height is 110 cm was measured, and his age is 50 years and his height appeared to be 180, so this is an unreal repetition.
Suppose we have:

X_i	Y_i
1.5	2
1.6	2.1
1.5	1.9
1.3	1.5
1.5	2.1
1.3	1.8
1.9	2.2
2	3
2	3.1

Repeated X_i values	Y_i values corresponding to repeated X_i values		
$X_1 = 1.5$	2	1.9	2.1
$X_1 = 1.3$	1.5	1.8	
$X_1 = 2$	3	3.1	

Example: The following data represents the sales volume (y) of a certain product (in thousands of dinars) and the amount spent on advertising it on television (x) in hundreds of dinars for a foreign company.

$\boldsymbol{X_i} = 35 \quad 25 \quad 40 \quad 35 \quad 64 \quad 25 \quad 50 \quad 67 \quad 69 \quad 50 \quad 70 \quad 50$

$\boldsymbol{Y_i} = 112 \quad 125 \quad 128 \quad 115 \quad 162 \quad 130 \quad 142 \quad 158 \quad 175 \quad 140 \quad 170 \quad 145$

Test whether the linear model fits the data? Solution: - In order to clarify the frequency in the X values, we will arrange these values in ascending order based on the X values as follows: -

Test whether the linear model fits the data?

Solution: - In order to clarify the frequency in the X values, we will arrange these values in ascending order based on the X values as follows: -

$\boldsymbol{X_i} = 25 \quad 25 \quad 35 \quad 35 \quad 40 \quad 50 \quad 50 \quad 50 \quad 64 \quad 67 \quad 69 \quad 70$

$\boldsymbol{Y_i} = 125 \quad 130 \quad 112 \quad 115 \quad 128 \quad 142 \quad 140 \quad 145 \quad 162 \quad 158 \quad 175 \quad 170$

Now we make a table to calculate the means and sum of squares of the net error.

X_{i1}	Y_{ij}			$\bar{Y}_i$	$SS_{p.e.}$	d.f.
25	125	130		127.5	12.5	**1**
35	112	115		113.5	4.5	**1**
40	128			---	---	**---**
50	142	140	145	142.33	12.67	**2**
64	162			---	---	**----**
67	158			---	---	**---**
69	175			---	---	**---**
70	170			---	---	**---**
SUM					29.67	**4**

Thus, ANOVA table will be as follows:

H_0: There is no deficiency in the fit.
H_1: There is a deficiency in the fit.

S.O.V	d.f.	SS	MS	F
$R(X_1)$	1	$SSR = \hat{\beta}_1 S_{xy} = 3963.65$	396.65	$\text{Cal. F} = 47.4$
Error	n-2=10	$SSe = SST - SSR(X_1) = 836.02$	83.602	
L.O.F.	$(n-2) - [\sum (n-1)]$ 10-4=6	806.35	134.39	$\text{Cal. F} = 18.12$
p.e.	4	29.67	7.4175	
Total	n-1=11	$SS_T = 100$		

By comparing the calculated F value with its table value, we notice:

$$\text{Cal. F} = 18.12 > tab.F(0.05,6,4) = 6.16$$

The null hypothesis is rejected and the alternative hypothesis is accepted, i.e. the linear model is not suitable for the data, although:

$$\text{Cal. F} = 47.4 > tab.F(0.05,1,10) = 4.96$$

That means that we must look for another non-linear model that fits the data.

Now we can make a scatter plot of the data as follows:

$X_i = 2 \quad 5 \quad 4 \quad 3 \quad 6 \quad 9 \quad 5 \quad 3 \quad 3 \Rightarrow \sum X_i = 40 \ , \overline{X} = 4.44$

$Y_i = 14 \quad 18 \quad 15 \quad 15 \quad 16 \quad 20 \quad 17 \quad 14 \quad 15 \Rightarrow \sum Y_i = 144 \ , \overline{Y} = 16$

$S_{xx} = 36.22 \ , \sum X^2 = 214 \ , S_{yy} = 32 \ , \quad \sum Y^2 = 2336$

$\sum X_i Y_i = 671 \ , S_{xy} = 31$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{31}{36.22} = 0.856$$

$SSR(X_1) = \hat{\beta}_1 S_{xy} = 0.856(31) = 26.536$

$SST = S_{yy} = 32$

$SSe = 32 - 26.536 = 5.464 \ , \text{d. f. (e)} = 9 - 2 = 7$

$$SS(\text{p. e.}) = (15-14.67)^2 + (14-14.67)^2 + (15-14.67)^2 = 0.1089 + 0.4489 + 0.1089 = 0.6667$$

$\text{d. f.}_{(1)} = 2$

$SS(\text{p. e.}_{(2)}) = (18-17.5)^2 + (17-17.5)^2 = 0.25 + 0.25 = 0.5 \ , \text{d. f.}_{(\text{p.e.})} = 1$

$\text{SS(p. e.)} = 0.6667 + 0.5 = 1.1667 \quad , \text{d. f.}_{(\text{p.e.})} = 3$

$\text{SS(L. o. f.)} = \text{d. f.} - \text{d. f.}_{(\text{p.e.})} = 7 - 3 = 4$

$$\text{MS(L. o. f.)} = \frac{4.2973}{4} = 1.074325$$

$$\text{MS(p. e.)} = \frac{1.1667}{3} = 0.3889$$

$$\text{tab. F} = \frac{\text{MS(L. o. f.)}}{\text{MS(p. e.)}} = \frac{1.074325}{0.3889} = 2.7625$$

$\text{tab. F}_{(0.05,4,3)} = 9.12$

$\text{cal. F} < tab.\text{F}$

$2.7623 < 9.12$

Yes, SSR can be found from $S_{y\hat{y}}$ as follows:

$\text{SSR}(X_1) = \ S_{y\hat{y}}$

$$\text{and we have : } - S_{y\hat{y}} = \sum (\hat{y} - \bar{\hat{y}})^2 = \sum (\bar{y} + \hat{\beta}_1(x_i - \bar{x}) - \bar{y})^2$$

$$= \sum \left(\hat{\beta}_1(x_i - \bar{x})\right)^2 = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2$$

$$= \hat{\beta}_1^2 S_{xx} = \text{SS due to Regression}$$

$\therefore S_{y\hat{y}} = S_{\hat{y}\hat{y}} = SS$ due to Regression

$$r_{y\hat{y}} = \frac{\text{SS due to Regression}}{SS_{Total}[\text{SS due to Regression}]}$$

$$\frac{[\text{SS due to Regr.}]^2}{SS_{Total}[\text{SS due to Regr.}]} = R^2$$

$\therefore R^2 = r^2{}_{y\hat{y}} = r_{xy}^2$

$R = r_{y\hat{y}}$

} In simple linear regression only

$\therefore 0 \leq R^2 = r^2{}_{y\hat{y}} = r_{xy}^2 \leq 1$

Coefficient of non determination = $(1 - R^2)$

Or it is the part that is not explained by the total sum of squares

Or it is the part of the changes in Y that are not explained by the model that explains the relationship between x and Y