5.2.2 Inference about the slope parameter

Inference concerning the slope parameter can either be done using the F-test or the T-test as follows:

1. F – test
   To test the hypothesis

$$H_0: \beta = 0 \quad against$$
$$H_0: \beta \neq 0$$

we partition the total variation in y as follows:

$$Total\ variation = explained\ variation + unexplained\ variation$$

or

$$SST = SSR + SSE$$

where

$$SST = Total\ sum\ of\ squares$$
$$= \sum_{i=1}^{n} (y_i - \bar{y})^2$$
$$= \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$
$$= S_{yy}$$

$$SSR = Regression\ sum\ of\ squares$$
$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
$$= \hat{\beta} S_{xy}$$

$$SSE = Error\ sum\ of\ squares$$
$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= SST - SSR$$
$$= S_{yy} - \hat{\beta} S_{xy}$$

Prepare the ANOVA table:

Source of variation (Source)	Sum of squares (SS)	Degrees of freedom (df)	Mean squares (MS)	F^*
Regression	SSR	1	$MSR = SSR$	$\frac{MSR}{MSE}$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

We reject H_0 if $F^* > f_\alpha(1, n-2)$

Note:

(a) $MSR = SSR$ since $MSR = \frac{SSR}{1}$.

(b) MSE is used to estimate σ^2.

(c) We can easily verify that $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}S_{xy}$ i.e.

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{n}\left(\hat{\alpha} + \hat{\beta}x_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i - \bar{y}\right)^2 \\ &= \sum_{i=1}^{n}\left(-\hat{\beta}\bar{x} + \hat{\beta}x_i\right)^2 = \sum_{i=1}^{n}\hat{\beta}^2(x_i - \bar{x})^2 \\ &= \hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}^2 S_{xx} = \hat{\beta}\frac{S_{xy}}{S_{xx}}S_{xx} = \hat{\beta}S_{xy} \\ &= \hat{\beta}S_{xy}\end{aligned}$$

2. T – test

To test the hypothesis

$H_0: \beta = \beta_0 \quad against$
$H_a: \beta < \beta_0 \quad or$
$H_1: \beta \neq \beta_0 \quad or$
$H_A: \beta > \beta_0$

we use the test statistic

$$T = \frac{\hat{\beta} - \beta_0}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

which is a t – distribution with $n-2$ degrees of freedom, where $\hat{\sigma}^2 = MSE$.

We reject

$H_a: \beta < \beta_0 \quad if \quad t^* < -t_{\alpha,n-2}$
$H_1: \beta \neq \beta_0 \quad if \quad |t^*| > t_{\frac{\alpha}{2},n-2}$
$H_A: \beta > \beta_0 \quad if \quad t^* > t_{\alpha,n-2}$

Note: Note that the T – test is more general as specific values of β_0 can be tested while the F – test is only used to test $\beta = 0$.

3. Confidence interval

A $100(1-\alpha)\%$ confidence interval for β in the regression line $y = \alpha + \beta x$ is given by

$$\hat{\beta} \pm t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

4. Coefficient of determination
   The ratio

$$R^2 = \frac{explained\ variation}{total\ variation} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

   called the coefficient of determination gives the proportion of variation in y values explained by the linear relationship with x (the regression line). The closer the points are to the line, the greater the value of R^2.

5.3 **Correlation analysis**

A regression is only useful if the x and y values show some degree of linear relationship (i.e. a cluster of points about a nonhorizontal line).

The degree of linear relationship between the x and y values is measured by the linear correlation coefficient r given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where

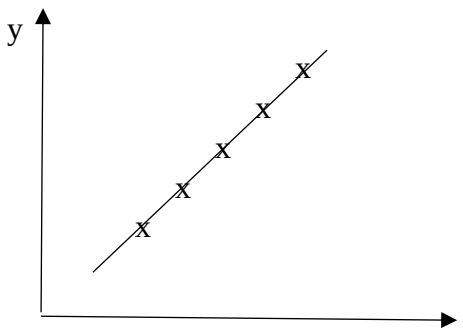
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

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

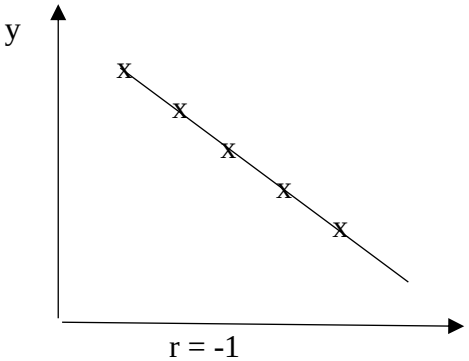
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

5.3.1 Properties of r

1. The value of r is always between -1 and 1 i.e. $-1 \leq r \leq 1$.
2. If $r = 1$, then all the points lie exactly on a line with a positive slope. Similarly $r = -1$ if all the points lie on a line with a negative slope.

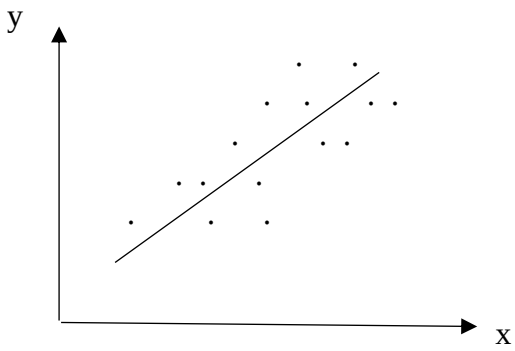


r = 1
perfect positive correlation

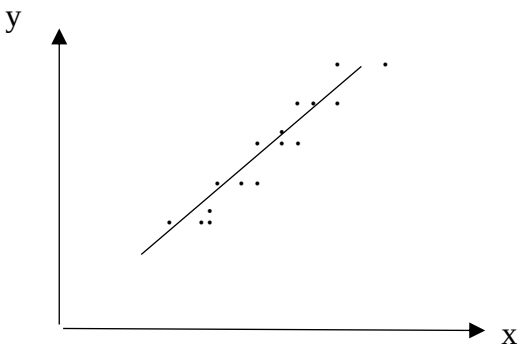


r = -1
perfect negative correlation

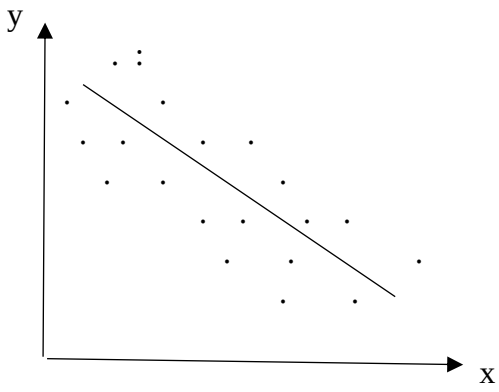
3. Suppose $r \neq 0$. If r is positive, the line has a positive slope. If r is negative, the line has a negative slope. The closer r is to 1 or -1, the closer the points tend to cluster about the line. e.g.



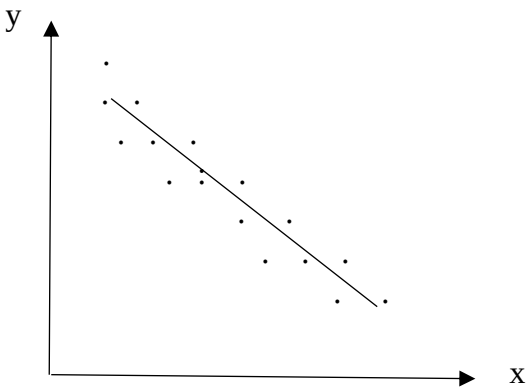
positive correlation r = 0.5



positive correlation r = 0.9

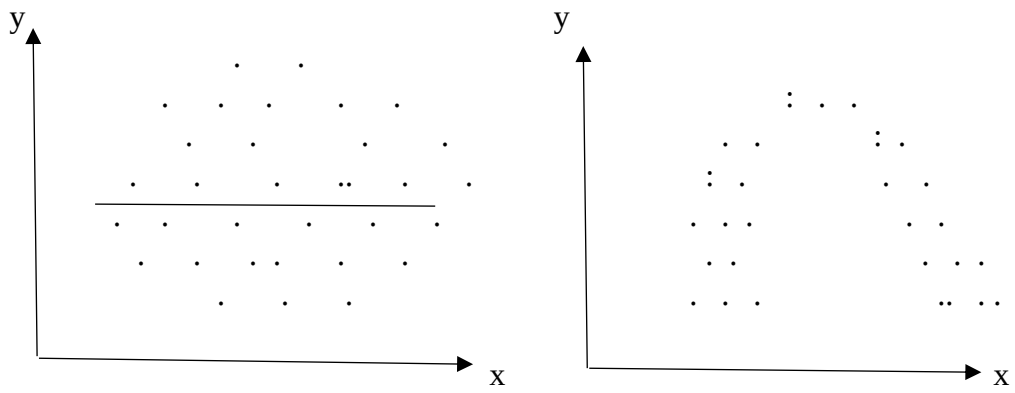


negative correlation r = -0.6



negative correlation r = -0.9

4. If $r = 0$, then there is no linear correlation between the x and y values.
   e.g.



r = 0 r = 0

5. $r = \sqrt{R^2}$ where $R^2 = coefficient\ of\ determination.$