Chapter 19

SIMPLE LINEAR REGRESSION AND CORRELATION ANALYSIS

Let X and Y be two random variables with joint probability density function $f(x,y)$. Then the conditional density of Y given that $X = x$ is

$$f(y/x) = \frac{f(x,y)}{g(x)}$$

where

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$$

is the marginal density of X. The conditional mean of Y

$$E\left(Y|X=x\right) = \int_{-\infty}^{\infty} y f(y/x)\, dy$$

is called the regression equation of Y on X.

Example 19.1. Let X and Y be two random variables with the joint probability density function

$$f(x,y) = \begin{cases} xe^{-x(1+y)} & \text{if } x > 0,\ y > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find the regression equation of Y on X and then sketch the regression curve.

Answer: The marginal density of X is given by

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} x e^{-x(1+y)}\, dy \\ &= \int_{-\infty}^{\infty} x e^{-x}\, e^{-xy}\, dy \\ &= x e^{-x} \int_{-\infty}^{\infty} e^{-xy}\, dy \\ &= x e^{-x} \left[-\frac{1}{x}\, e^{-xy} \right]_0^{\infty} \\ &= e^{-x}. \end{aligned}$$

The conditional density of Y given $X = x$ is

$$f(y/x) = \frac{f(x,y)}{g(x)} = \frac{x e^{-x(1+y)}}{e^{-x}} = x e^{-xy}, \qquad y > 0.$$

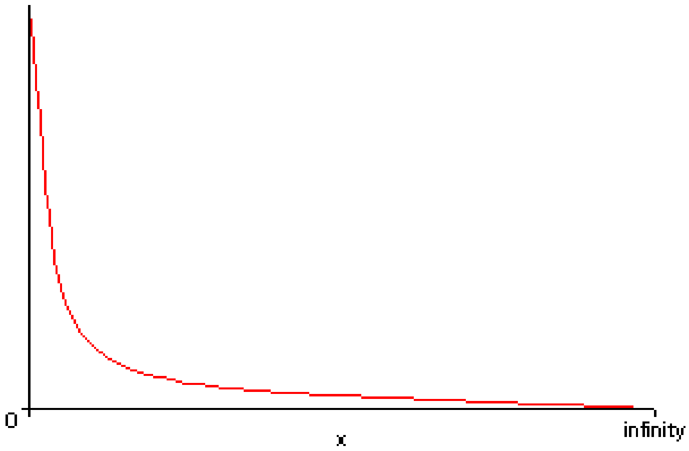
The conditional mean of Y given $X = x$ is given by

$$E(Y/x) = \int_{-\infty}^{\infty} y f(y/x)\, dy = \int_{-\infty}^{\infty} y\, x\, e^{-xy}\, dy = \frac{1}{x}.$$

Thus the regression equation of Y on X is

$$E(Y/x) = \frac{1}{x}, \quad x > 0.$$

The graph of this equation of Y on X is shown below.



Graph of the regression equation E(Y/x) = 1/ x

From this example it is clear that the conditional mean $E(Y/x)$ is a function of x. If this function is of the form $\alpha + \beta x$, then the corresponding regression equation is called a linear regression equation; otherwise it is called a nonlinear regression equation. The term linear regression refers to a specification that is linear in the parameters. Thus $E(Y/x) = \alpha + \beta x^2$ is also a linear regression equation. The regression equation $E(Y/x) = \alpha x^\beta$ is an example of a nonlinear regression equation.

The main purpose of regression analysis is to predict Y_i from the knowledge of x_i using the relationship like

$$E(Y_i/x_i) = \alpha + \beta x_i.$$

The Y_i is called the response or dependent variable where as x_i is called the predictor or independent variable. The term regression has an interesting history, dating back to Francis Galton (1822-1911). Galton studied the heights of fathers and sons, in which he observed a regression (a "turning back") from the heights of sons to the heights of their fathers. That is tall fathers tend to have tall sons and short fathers tend to have short sons. However, he also found that very tall fathers tend to have shorter sons and very short fathers tend to have taller sons. Galton called this phenomenon regression towards the mean.

In regression analysis, that is when investigating the relationship between a predictor and response variable, there are two steps to the analysis. The first step is totally data oriented. This step is always performed. The second step is the statistical one, in which we draw conclusions about the (population) regression equation $E(Y_i/x_i)$. Normally the regression equation contains several parameters. There are two well known methods for finding the estimates of the parameters of the regression equation. These two methods are: (1) The least square method and (2) the normal regression method.

19.1. The Least Squares Method

Let $\{(x_i, y_i) \,|\, i = 1, 2, ..., n\}$ be a set of data. Assume that

$$E(Y_i/x_i) = \alpha + \beta x_i, \tag{1}$$

that is

$$y_i = \alpha + \beta x_i, \qquad i = 1, 2, ..., n.$$

Then the sum of the squares of the error is given by

$$\mathcal{E}(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 . \tag{2}$$

The least squares estimates of α and β are defined to be those values which minimize $\mathcal{E}(\alpha, \beta)$. That is,

$$\left(\widehat{\alpha}, \widehat{\beta}\right) = \arg \min_{(\alpha,\beta)} \mathcal{E}(\alpha, \beta).$$

This least squares method is due to Adrien M. Legendre (1752-1833). Note that the least squares method also works even if the regression equation is nonlinear (that is, not of the form (1)).

Next, we give several examples to illustrate the method of least squares.

Example 19.2. Given the five pairs of points (x, y) shown in table below

x	4	0	-2	3	1
y	5	0	0	6	3

what is the line of the form $y = x + b$ best fits the data by method of least squares?

Answer: Suppose the best fit line is $y = x + b$. Then for each x_i, $x_i + b$ is the estimated value of y_i. The difference between y_i and the estimated value of y_i is the error or the residual corresponding to the i^{th} measurement. That is, the error corresponding to the i^{th} measurement is given by

$$\epsilon_i = y_i - x_i - b.$$

Hence the sum of the squares of the errors is

$$\begin{aligned} \mathcal{E}(b) &= \sum_{i=1}^{5} \epsilon_i^2 \\ &= \sum_{i=1}^{5} (y_i - x_i - b)^2 . \end{aligned}$$

Differentiating $\mathcal{E}(b)$ with respect to b, we get

$$\frac{d}{db}\mathcal{E}(b) = 2 \sum_{i=1}^{5} (y_i - x_i - b)\,(-1).$$

Setting $\frac{d}{db}\mathcal{E}(b)$ equal to 0, we get

$$\sum_{i=1}^{5} (y_i - x_i - b) = 0$$

which is

$$5b = \sum_{i=1}^{5} y_i - \sum_{i=1}^{5} x_i.$$

Using the data, we see that

$$5b = 14 - 6$$

which yields $b = \frac{8}{5}$. Hence the best fitted line is

$$y = x + \frac{8}{5}.$$

Example 19.3. Suppose the line $y = bx + 1$ is fit by the method of least squares to the 3 data points

x	1	2	4
y	2	2	0

What is the value of the constant b?

Answer: The error corresponding to the i^{th} measurement is given by

$$\epsilon_i = y_i - bx_i - 1.$$

Hence the sum of the squares of the errors is

$$\begin{aligned} \mathcal{E}(b) &= \sum_{i=1}^{3} \epsilon_i^2 \\ &= \sum_{i=1}^{3} (y_i - bx_i - 1)^2 . \end{aligned}$$

Differentiating $\mathcal{E}(b)$ with respect to b, we get

$$\frac{d}{db}\mathcal{E}(b) = 2 \sum_{i=1}^{3} (y_i - bx_i - 1)\,(-x_i).$$

Setting $\frac{d}{db}\mathcal{E}(b)$ equal to 0, we get

$$\sum_{i=1}^{3} (y_i - b x_i - 1)\, x_i = 0$$

which in turn yields

$$b = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i}{\displaystyle\sum_{i=1}^{n} x_i^2}$$

Using the given data we see that

$$b = \frac{6-7}{21} = -\frac{1}{21},$$

and the best fitted line is

$$y = -\frac{1}{21} x + 1.$$

Example 19.4. Observations $y_1, y_2, ..., y_n$ are assumed to come from a model with

$$E(Y_i / x_i) = \theta + 2 \ln x_i$$

where θ is an unknown parameter and $x_1, x_2, ..., x_n$ are given constants. What is the least square estimate of the parameter θ?

Answer: The sum of the squares of errors is

$$\mathcal{E}(\theta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \theta - 2 \ln x_i)^2 .$$

Differentiating $\mathcal{E}(\theta)$ with respect to θ, we get

$$\frac{d}{d\theta}\mathcal{E}(\theta) = 2 \sum_{i=1}^{n} (y_i - \theta - 2 \ln x_i)\,(-1).$$

Setting $\frac{d}{d\theta}\mathcal{E}(\theta)$ equal to 0, we get

$$\sum_{i=1}^{n} (y_i - \theta - 2 \ln x_i) = 0$$

which is

$$\theta = \frac{1}{n} \left(\sum_{i=1}^{n} y_i - 2 \sum_{i=1}^{n} \ln x_i \right).$$

Hence the least squares estimate of θ is $\widehat{\theta} = \overline{y} - \frac{2}{n}\sum_{i=1}^{n} \ln x_i$.

Example 19.5. Given the three pairs of points (x, y) shown below:

x	4	1	2
y	2	1	0

What is the curve of the form $y = x^\beta$ best fits the data by method of least squares?

Answer: The sum of the squares of the errors is given by

$$\begin{aligned} \mathcal{E}(\beta) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - x_i^\beta \right)^2. \end{aligned}$$

Differentiating $\mathcal{E}(\beta)$ with respect to β, we get

$$\frac{d}{d\beta}\mathcal{E}(\beta) = 2 \sum_{i=1}^{n} \left(y_i - x_i^\beta \right) \left(- x_i^\beta \, \ln x_i \right)$$

Setting this derivative $\frac{d}{d\beta}\mathcal{E}(\beta)$ to 0, we get

$$\sum_{i=1}^{n} y_i x_i^\beta \, \ln x_i = \sum_{i=1}^{n} x_i^\beta \, x_i^\beta \, \ln x_i.$$

Using the given data we obtain

$$(2)\, 4^\beta \ln 4 = 4^{2\beta} \ln 4 + 2^{2\beta} \ln 2$$

which simplifies to

$$4 = (2)\, 4^\beta + 1$$

or

$$4^\beta = \frac{3}{2}.$$

Taking the natural logarithm of both sides of the above expression, we get

$$\beta = \frac{\ln 3 - \ln 2}{\ln 4} = 0.2925$$

Thus the least squares best fit model is $y = x^{0.2925}$.

Example 19.6. Observations $y_1, y_2, ..., y_n$ are assumed to come from a model with $E(Y_i/x_i) = \alpha + \beta x_i$, where α and β are unknown parameters, and $x_1, x_2, ..., x_n$ are given constants. What are the least squares estimate of the parameters α and β?

Answer: The sum of the squares of the errors is given by

$$\begin{aligned}\mathcal{E}(\alpha, \beta) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - -\alpha - \beta x_i)^2 .\end{aligned}$$

Differentiating $\mathcal{E}(\alpha, \beta)$ with respect to α and β respectively, we get

$$\frac{\partial}{\partial \alpha} \mathcal{E}(\alpha, \beta) = 2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) (-1)$$

and

$$\frac{\partial}{\partial \beta} \mathcal{E}(\alpha, \beta) = 2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) (-x_i).$$

Setting these partial derivatives $\frac{\partial}{\partial \alpha} \mathcal{E}(\alpha, \beta)$ and $\frac{\partial}{\partial \beta} \mathcal{E}(\alpha, \beta)$ to 0, we get

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \tag{3}$$

and

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) \, x_i = 0. \tag{4}$$

From (3), we obtain

$$\sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i$$

which is

$$\overline{y} = \alpha + \beta \, \overline{x}. \tag{5}$$

Similarly, from (4), we have

$$\sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2$$

which can be rewritten as follows

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+n\overline{x}\,\overline{y}=n\,\alpha\,\overline{x}+\beta\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})+n\beta\,\overline{x}^2 \tag{6}$$

Defining

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

we see that (6) reduces to

$$S_{xy}+n\overline{x}\,\overline{y}=\alpha\,n\,\overline{x}+\beta\,\left[S_{xx}+n\overline{x}^2\right] \tag{7}$$

Substituting (5) into (7), we have

$$S_{xy}+n\overline{x}\,\overline{y}=\left[\overline{y}-\beta\,\overline{x}\right]n\,\overline{x}+\beta\,\left[S_{xx}+n\overline{x}^2\right].$$

Simplifying the last equation, we get

$$S_{xy}=\beta\,S_{xx}$$

which is

$$\beta=\frac{S_{xy}}{S_{xx}}. \tag{8}$$

In view of (8) and (5), we get

$$\alpha=\overline{y}-\frac{S_{xy}}{S_{xx}}\,\overline{x}. \tag{9}$$

Thus the least squares estimates of α and β are

$$\widehat{\alpha}=\overline{y}-\frac{S_{xy}}{S_{xx}}\,\overline{x} \quad \text{and} \quad \widehat{\beta}=\frac{S_{xy}}{S_{xx}},$$

respectively.

We need some notations. The random variable Y given $X=x$ will be denoted by Y_x. Note that this is the variable appears in the model $E(Y/x)=\alpha+\beta x$. When one chooses in succession values $x_1,x_2,...,x_n$ for x, a sequence $Y_{x_1},Y_{x_2},...,Y_{x_n}$ of random variable is obtained. For the sake of convenience, we denote the random variables $Y_{x_1},Y_{x_2},...,Y_{x_n}$ simply as $Y_1,Y_2,...,Y_n$. To do some statistical analysis, we make following three assumptions:

(1) $E(Y_x)=\alpha+\beta\,x$ so that $\mu_i=E(Y_i)=\alpha+\beta\,x_i$;

(2) $Y_1, Y_2, ..., Y_n$ are independent;

(3) Each of the random variables $Y_1, Y_2, ..., Y_n$ has the same variance σ^2.

Theorem 19.1. Under the above three assumptions, the least squares estimators $\widehat{\alpha}$ and $\widehat{\beta}$ of a linear model $E(Y/x) = \alpha + \beta\, x$ are unbiased.

Proof: From the previous example, we know that the least squares estimators of α and β are

$$\widehat{\alpha} = \overline{Y} - \frac{S_{xY}}{S_{xx}}\; \overline{X} \quad \text{and} \quad \widehat{\beta} = \frac{S_{xY}}{S_{xx}},$$

where

$$S_{xY} := \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}).$$

First, we show $\widehat{\beta}$ is unbiased. Consider

$$\begin{aligned}
E\left(\widehat{\beta}\right) &= E\left(\frac{S_{xY}}{S_{xx}}\right) = \frac{1}{S_{xx}}\; E\left(S_{xY}\right) \\
&= \frac{1}{S_{xx}}\; E\left(\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})\right) \\
&= \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; E\left(Y_i - \overline{Y}\right) \\
&= \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; E\left(Y_i\right) - \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; E\left(\overline{Y}\right) \\
&= \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; E\left(Y_i\right) - \frac{1}{S_{xx}}\; E\left(\overline{Y}\right)\; \sum_{i=1}^{n} (x_i - \overline{x}) \\
&= \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; E\left(Y_i\right) = \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; \left(\alpha + \beta x_i\right) \\
&= \alpha\; \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x}) + \beta\; \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; x_i \\
&= \beta\; \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; x_i \\
&= \beta\; \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; x_i - \beta\; \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; \overline{x} \\
&= \beta\; \frac{1}{S_{xx}}\; \sum_{i=1}^{n} (x_i - \overline{x})\; (x_i - \overline{x}) \\
&= \beta\; \frac{1}{S_{xx}}\; S_{xx} = \beta.
\end{aligned}$$

Thus the estimator $\widehat{\beta}$ is unbiased estimator of the parameter β.

Next, we show that $\widehat{\alpha}$ is also an unbiased estimator of α. Consider

$$\begin{aligned}
E\left(\widehat{\alpha}\right) &= E\left(\overline{Y} - \frac{S_{xY}}{S_{xx}}\,\overline{x}\right) = E\left(\overline{Y}\right) - \overline{x}\;E\left(\frac{S_{xY}}{S_{xx}}\right) \\
&= E\left(\overline{Y}\right) - \overline{x}\,E\left(\widehat{\beta}\right) = E\left(\overline{Y}\right) - \overline{x}\;\beta \\
&= \frac{1}{n}\;\left(\sum_{i=1}^{n} E\left(Y_i\right)\right) - \overline{x}\;\beta \\
&= \frac{1}{n}\;\left(\sum_{i=1}^{n} E\left(\alpha + \beta x_i\right)\right) - \overline{x}\;\beta \\
&= \frac{1}{n}\;\left(n\alpha + \beta \sum_{i=1}^{n} x_i\right) - \overline{x}\;\beta \\
&= \alpha + \beta\,\overline{x} - \overline{x}\;\beta = \alpha
\end{aligned}$$

This proves that $\widehat{\alpha}$ is an unbiased estimator of α and the proof of the theorem is now complete.

19.2. The Normal Regression Analysis

In a regression analysis, we assume that the x_i's are constants while y_i's are values of the random variables Y_i's. A regression analysis is called a normal regression analysis if the conditional density of Y_i given $X_i = x_i$ is of the form

$$f(y_i/x_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right)^2},$$

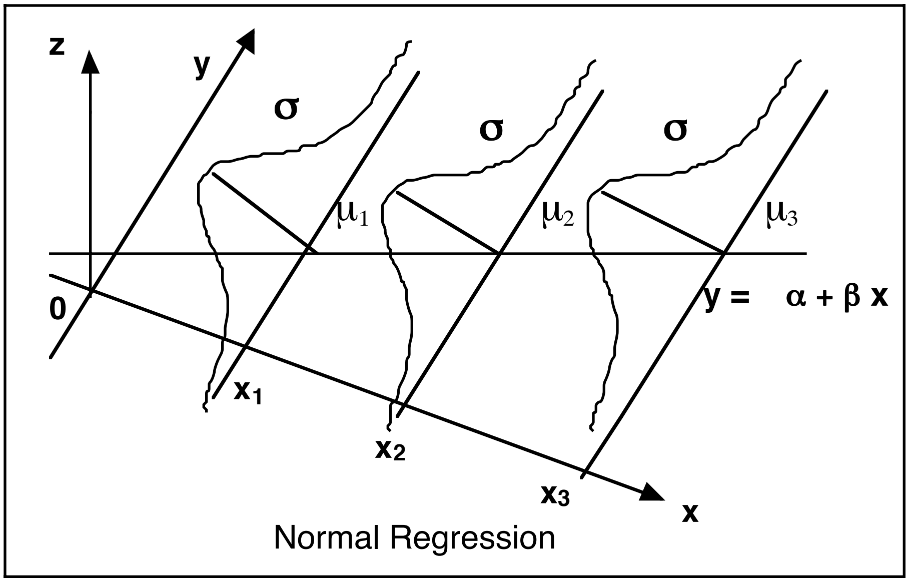
where σ^2 denotes the variance, and α and β are the regression coefficients. That is $Y|_{x_i} \sim N(\alpha + \beta x, \sigma^2)$. If there is no danger of confusion, then we will write Y_i for $Y|_{x_i}$. The figure on the next page shows the regression model of Y with equal variances, and with means falling on the straight line $\mu_y = \alpha + \beta\, x$.

Normal regression analysis concerns with the estimation of σ, α, and β. We use maximum likelihood method to estimate these parameters. The maximum likelihood function of the sample is given by

$$L(\sigma, \alpha, \beta) = \prod_{i=1}^{n} f(y_i/x_i)$$

and

$$\begin{aligned}\ln L(\sigma,\alpha,\beta) &= \sum_{i=1}^{n} \ln f(y_i/x_i)\\ &= -n\ln\sigma - \frac{n}{2}\,\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\alpha-\beta\,x_i\right)^2.\end{aligned}$$



Taking the partial derivatives of $\ln L(\sigma,\alpha,\beta)$ with respect to α, β and σ respectively, we get

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ln L(\sigma,\alpha,\beta) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\alpha-\beta\,x_i\right)\\ \frac{\partial}{\partial\beta}\ln L(\sigma,\alpha,\beta) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\alpha-\beta\,x_i\right)x_i\\ \frac{\partial}{\partial\sigma}\ln L(\sigma,\alpha,\beta) &= -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}\left(y_i-\alpha-\beta\,x_i\right)^2.\end{aligned}$$

Equating each of these partial derivatives to zero and solving the system of three equations, we obtain the maximum likelihood estimator of β, α, σ as

$$\widehat{\beta} = \frac{S_{xY}}{S_{xx}}, \qquad \widehat{\alpha} = \overline{Y} - \frac{S_{xY}}{S_{xx}}\,\overline{x}, \quad \text{and} \quad \widehat{\sigma} = \sqrt{\frac{1}{n}\left[S_{YY} - \frac{S_{xY}}{S_{xx}}\,S_{xY}\right]},$$

where

$$S_{xY} = \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y}\right).$$

Theorem 19.2. In the normal regression analysis, the likelihood estimators $\widehat{\beta}$ and $\widehat{\alpha}$ are unbiased estimators of β and α, respectively.

Proof: Recall that

$$\begin{aligned}
\widehat{\beta} &= \frac{S_{xY}}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) Y_i,
\end{aligned}$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$. Thus $\widehat{\beta}$ is a linear combination of Y_i's. Since $Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$, we see that $\widehat{\beta}$ is also a normal random variable.

First we show $\widehat{\beta}$ is an unbiased estimator of β. Since

$$\begin{aligned}
E\left(\widehat{\beta}\right) &= E\left(\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) Y_i\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) E\left(Y_i\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) \left(\alpha + \beta x_i\right) = \beta,
\end{aligned}$$

the maximum likelihood estimator of β is unbiased.

Next, we show that $\widehat{\alpha}$ is also an unbiased estimator of α. Consider

$$\begin{aligned}
E\left(\widehat{\alpha}\right) &= E\left(\overline{Y} - \frac{S_{xY}}{S_{xx}}\, \overline{x}\right) = E\left(\overline{Y}\right) - \overline{x}\; E\left(\frac{S_{xY}}{S_{xx}}\right) \\
&= E\left(\overline{Y}\right) - \overline{x}\, E\left(\widehat{\beta}\right) = E\left(\overline{Y}\right) - \overline{x}\; \beta \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} E\left(Y_i\right)\right) - \overline{x}\; \beta \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} E\left(\alpha + \beta x_i\right)\right) - \overline{x}\; \beta \\
&= \frac{1}{n}\left(n\alpha + \beta \sum_{i=1}^{n} x_i\right) - \overline{x}\; \beta \\
&= \alpha + \beta\, \overline{x} - \overline{x}\; \beta = \alpha.
\end{aligned}$$

This proves that $\widehat{\alpha}$ is an unbiased estimator of α and the proof of the theorem is now complete.

Theorem 19.3. In normal regression analysis, the distributions of the estimators $\widehat{\beta}$ and $\widehat{\alpha}$ are given by

$$\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right) \qquad \text{and} \qquad \widehat{\alpha} \sim N\left(\alpha,\ \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right)$$

where

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

Proof: Since

$$\begin{aligned} \widehat{\beta} &= \frac{S_{xY}}{S_{xx}} \\ &= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) \\ &= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) Y_i, \end{aligned}$$

the $\widehat{\beta}$ is a linear combination of Y_i's. As $Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$, we see that $\widehat{\beta}$ is also a normal random variable. By Theorem 19.2, $\widehat{\beta}$ is an unbiased estimator of β.

The variance of $\widehat{\beta}$ is given by

$$\begin{aligned} Var\left(\widehat{\beta}\right) &= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 Var\left(Y_i/x_i\right) \\ &= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 \sigma^2 \\ &= \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 \, \sigma^2 \\ &= \frac{\sigma^2}{S_{xx}}. \end{aligned}$$

Hence $\widehat{\beta}$ is a normal random variable with mean (or expected value) β and variance $\frac{\sigma^2}{S_{xx}}$. That is $\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)$.

Now determine the distribution of $\widehat{\alpha}$. Since each $Y_i \sim N(\alpha + \beta x_i,\ \sigma^2)$, the distribution of $\overline{Y}$ is given by

$$\overline{Y} \sim N\left(\alpha + \beta\overline{x},\ \frac{\sigma^2}{n}\right).$$

Since

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

the distribution of $\overline{x}\,\widehat{\beta}$ is given by

$$\overline{x}\,\widehat{\beta} \sim N\left(\overline{x}\,\beta,\ \overline{x}^2\,\frac{\sigma^2}{S_{xx}}\right).$$

Since $\widehat{\alpha} = \overline{Y} - \overline{x}\,\widehat{\beta}$ and $\overline{Y}$ and $\overline{x}\,\widehat{\beta}$ being two normal random variables, $\widehat{\alpha}$ is also a normal random variable with mean equal to $\alpha + \beta\,\overline{x} - \beta\,\overline{x} = \alpha$ and variance variance equal to $\frac{\sigma^2}{n} + \frac{\overline{x}^2\,\sigma^2}{S_{xx}}$. That is

$$\widehat{\alpha} \sim N\left(\alpha,\ \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right)$$

and the proof of the theorem is now complete.

It should be noted that in the proof of the last theorem, we have assumed the fact that $\overline{Y}$ and $\overline{x}\,\widehat{\beta}$ are statistically independent.

In the next theorem, we give an unbiased estimator of the variance σ^2. For this we need the distribution of the statistic U given by

$$U = \frac{n\,\widehat{\sigma}^2}{\sigma^2}.$$

It can be shown (we will omit the proof, for a proof see Graybill (1961)) that the distribution of the statistic

$$U = \frac{n\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Theorem 19.4. An unbiased estimator S^2 of σ^2 is given by

$$S^2 = \frac{n\,\widehat{\sigma}^2}{n-2},$$

where $\widehat{\sigma} = \sqrt{\frac{1}{n}\,\left[S_{YY} - \frac{S_{xY}}{S_{xx}}\,S_{xY}\right]}$.

Proof: Since

$$\begin{aligned} E(S^2) &= E\left(\frac{n\,\widehat{\sigma}^2}{n-2}\right) \\ &= \frac{\sigma^2}{n-2}\,E\left(\frac{n\,\widehat{\sigma}^2}{\sigma^2}\right) \\ &= \frac{\sigma^2}{n-2}\,E(\chi^2(n-2)) \\ &= \frac{\sigma^2}{n-2}\,(n-2) = \sigma^2. \end{aligned}$$

The proof of the theorem is now complete.

Note that the estimator S^2 can be written as $S^2 = \frac{SSE}{n-2}$, where

$$SSE = S_{YY} = \widehat{\beta}\, S_{xY} = \sum_{i=1}^{2} [y_i - \widehat{\alpha} - \widehat{\beta}\, x_i]$$

the estimator S^2 is unbiased estimator of σ^2. The proof of the theorem is now complete.

In the next theorem we give the distribution of two statistics that can be used for testing hypothesis and constructing confidence interval for the regression parameters α and β.

Theorem 19.5. The statistics

$$Q_\beta = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n}}$$

and

$$Q_\alpha = \frac{\widehat{\alpha} - \alpha}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n\, (\overline{x})^2 + S_{xx}}}$$

have both a t-distribution with $n-2$ degrees of freedom.

Proof: From Theorem 19.3, we know that

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

Hence by standardizing, we get

$$Z = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1).$$

Further, we know that the likelihood estimator of σ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\left[S_{YY} - \frac{S_{xY}}{S_{xx}}\, S_{xY}\right]}$$

and the distribution of the statistic $U = \frac{n\,\widehat{\sigma}^2}{\sigma^2}$ is chi-square with $n-2$ degrees of freedom.

Since $Z = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$ and $U = \frac{n\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$, by Theorem 14.6, the statistic $\frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2)$. Hence

$$Q_\beta = \frac{\widehat{\beta}-\beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{n}} = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,S_{xx}}}} = \frac{\frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,\sigma^2}}} \sim t(n-2).$$

Similarly, it can be shown that

$$Q_\alpha = \frac{\widehat{\alpha}-\alpha}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{n\,(\overline{x})^2 + S_{xx}}} \sim t(n-2).$$

This completes the proof of the theorem.

In the normal regression model, if $\beta = 0$, then $E(Y_x) = \alpha$. This implies that $E(Y_x)$ does not depend on x. Therefore if $\beta \neq 0$, then $E(Y_x)$ is dependent on x. Thus the null hypothesis $H_o : \beta = 0$ should be tested against $H_a : \beta \neq 0$. To devise a test we need the distribution of $\widehat{\beta}$. Theorem 19.3 says that $\widehat{\beta}$ is normally distributed with mean β and variance $\frac{\sigma^2}{S_{xx}}$. Therefore, we have

$$Z = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1).$$

In practice the variance $Var(Y_i/x_i)$ which is σ^2 is usually unknown. Hence the above statistic Z is not very useful. However, using the statistic Q_β, we can devise a hypothesis test to test the hypothesis $H_o : \beta = \beta_o$ against $H_a : \beta \neq \beta_o$ at a significance level γ. For this one has to evaluate the quantity

$$|t| = \left| \frac{\widehat{\beta}-\beta}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,S_{xx}}}} \right|$$

$$= \left| \frac{\widehat{\beta}-\beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{n}} \right|$$

and compare it to quantile $t_{\gamma/2}(n-2)$. The hypothesis test, at significance level γ, is then "Reject $H_o : \beta = \beta_o$ if $|t| > t_{\gamma/2}(n-2)$".

The statistic

$$Q_\beta = \frac{\widehat{\beta}-\beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{n}}$$

is a pivotal quantity for the parameter β since the distribution of this quantity Q_β is a t-distribution with $n-2$ degrees of freedom. Thus it can be used for the construction of a $(1-\gamma)100\%$ confidence interval for the parameter β as follows:

$$\begin{aligned}
&1-\gamma \\
&= P\left(-t_{\frac{\gamma}{2}}(n-2) \leq \frac{\widehat{\beta}-\beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)S_{xx}}{n}} \leq t_{\frac{\gamma}{2}}(n-2) \right) \\
&= P\left(\widehat{\beta} - t_{\frac{\gamma}{2}}(n-2)\widehat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} \leq \beta \leq \widehat{\beta} + t_{\frac{\gamma}{2}}(n-2)\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}} \right).
\end{aligned}$$

Hence, the $(1-\gamma)\%$ confidence interval for β is given by

$$\left[\widehat{\beta} - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\,,\ \ \widehat{\beta} + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}} \right].$$

In a similar manner one can devise hypothesis test for α and construct confidence interval for α using the statistic Q_α. We leave these to the reader.

Now we give two examples to illustrate how to find the normal regression line and related things.

Example 19.7. Let the following data on the number of hours, x which ten persons studied for a French test and their scores, y on the test is shown below:

x	4	9	10	14	4	7	12	22	1	17
y	31	58	65	73	37	44	60	91	21	84

Find the normal regression line that approximates the regression of test scores on the number of hours studied. Further test the hypothesis $H_o : \beta = 3$ versus $H_a : \beta \neq 3$ at the significance level 0.02.

Answer: From the above data, we have

$$\sum_{i=1}^{10} x_i = 100, \qquad \sum_{i=1}^{10} x_i^2 = 1376$$

$$\sum_{i=1}^{10} y_i = 564, \qquad \sum_{i=1}^{10} y_i^2 =$$

$$\sum_{i=1}^{10} x_i y_i = 6945$$

$$S_{xx} = 376, \qquad S_{xy} = 1305, \qquad S_{yy} = 4752.4.$$

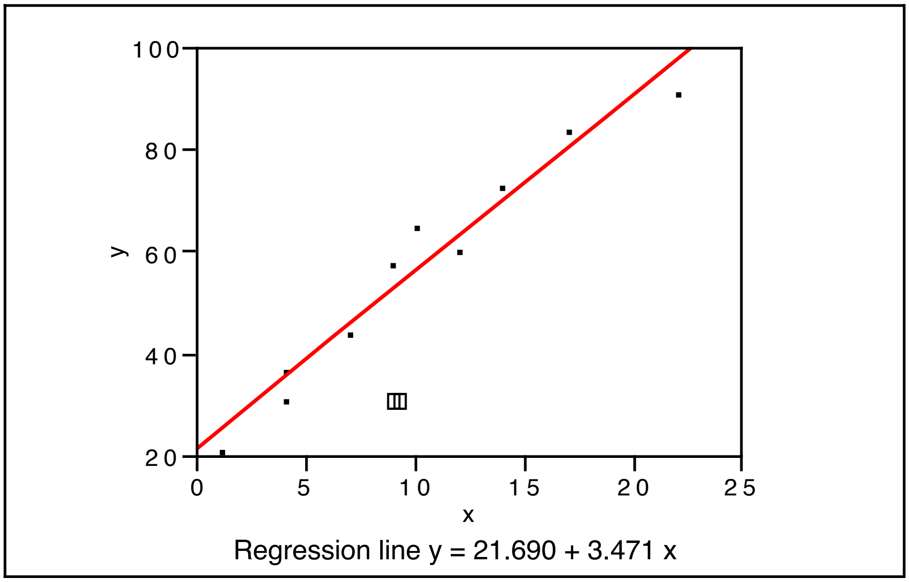
Hence

$$\widehat{\beta} = \frac{s_{xy}}{s_{xx}} = 3.471 \quad \text{and} \quad \widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = 21.690.$$

Thus the normal regression line is

$$y = 21.690 + 3.471x.$$

This regression line is shown below.



Regression line y = 21.690 + 3.471 x

Now we test the hypothesis $H_o : \beta = 3$ against $H_a : \beta \neq 3$ at 0.02 level of significance. From the data, the maximum likelihood estimate of σ is

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\,S_{xy}\right]} \\ &= \sqrt{\frac{1}{n}\left[S_{yy} - \widehat{\beta}\,S_{xy}\right]} \\ &= \sqrt{\frac{1}{10}\,[4752.4 - (3.471)(1305)]} \\ &= 4.720\end{aligned}$$

and

$$|t| = \left| \frac{3.471 - 3}{4.720} \sqrt{\frac{(8)\,(376)}{10}} \right| = 1.73.$$

Hence

$$1.73 = |t| < t_{0.01}(8) = 2.896.$$

Thus we do not reject the null hypothesis that $H_o : \beta = 3$ at the significance level 0.02.

This means that we can not conclude that on the average an extra hour of study will increase the score by more than 3 points.

Example 19.8. The frequency of chirping of a cricket is thought to be related to temperature. This suggests the possibility that temperature can be estimated from the chirp frequency. Let the following data on the number chirps per second, x by the striped ground cricket and the temperature, y in Fahrenheit is shown below:

x	20	16	20	18	17	16	15	17	15	16
y	89	72	93	84	81	75	70	82	69	83

Find the normal regression line that approximates the regression of temperature on the number chirps per second by the striped ground cricket. Further test the hypothesis $H_o : \beta = 4$ versus $H_a : \beta \neq 4$ at the significance level 0.1.

Answer: From the above data, we have

$$\sum_{i=1}^{10} x_i = 170, \qquad \sum_{i=1}^{10} x_i^2 = 2920$$

$$\sum_{i=1}^{10} y_i = 789, \qquad \sum_{i=1}^{10} y_i^2 = 64270$$

$$\sum_{i=1}^{10} x_i y_i = 13688$$

$$S_{xx} = 376, \qquad S_{xy} = 1305, \qquad S_{yy} = 4752.4.$$

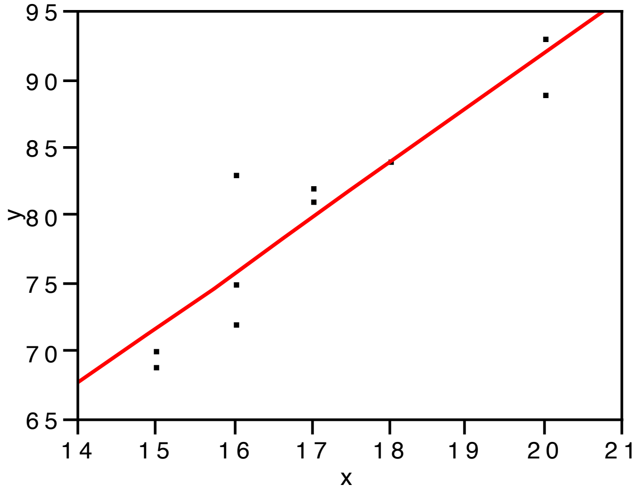
Hence

$$\widehat{\beta} = \frac{s_{xy}}{s_{xx}} = 4.067 \quad \text{and} \quad \widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = 9.761.$$

Thus the normal regression line is

$$y = 9.761 + 4.067x.$$

This regression line is shown below.



Regression line y = 9.761 + 4.067x

Now we test the hypothesis $H_o : \beta = 4$ against $H_a : \beta \neq 4$ at 0.1 level of significance. From the data, the maximum likelihood estimate of σ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\, S_{xy}\right]}$$

$$= \sqrt{\frac{1}{n}\left[S_{yy} - \widehat{\beta}\, S_{xy}\right]}$$

$$= \sqrt{\frac{1}{10}\,[589 - (4.067)(122)]}$$

$$= 3.047$$

and

$$|t| = \left| \frac{4.067 - 4}{3.047} \sqrt{\frac{(8)\,(30)}{10}} \right| = 0.528.$$

Hence

$$0.528 = |t| < t_{0.05}(8) = 1.860.$$

Thus we do not reject the null hypothesis that $H_o : \beta = 4$ at a significance level 0.1.

Let $\mu_x = \alpha + \beta\, x$ and write $\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}\, x$ for an arbitrary but fixed x. Then $\widehat{Y}_x$ is an estimator of μ_x. The following theorem gives various properties of this estimator.

Theorem 19.6. Let x be an arbitrary but fixed real number. Then
(i) $\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$,
(ii) $\widehat{Y}_x$ is an unbiased estimator of μ_x, and
(iii) $Var\left(\widehat{Y}_x\right) = \left\{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right\}\sigma^2$.

Proof: First we show that $\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$. Since

$$\begin{aligned}
\widehat{Y}_x &= \widehat{\alpha} + \widehat{\beta}\, x \\
&= \overline{Y} - \widehat{\beta}\overline{x} + \widehat{\beta}\, x \\
&= \overline{Y} + \widehat{\beta}\,(x - \overline{x}) \\
&= \overline{Y} + \sum_{k=1}^{n} \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}}\, Y_k \\
&= \sum_{k=1}^{n} \frac{Y_k}{n} + \sum_{k=1}^{n} \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}}\, Y_k \\
&= \sum_{k=1}^{n} \left(\frac{1}{n} + \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}} \right) Y_k
\end{aligned}$$

$\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$.

Next, we show that $\widehat{Y}_x$ is an unbiased estimator of μ_x. Since

$$\begin{aligned}
E\left(\widehat{Y}_x\right) &= E\left(\widehat{\alpha} + \widehat{\beta}\, x\right) \\
&= E\left(\widehat{\alpha}\right) + E\left(\widehat{\beta}\, x\right) \\
&= \alpha + \beta\, x \\
&= \mu_x
\end{aligned}$$

$\widehat{Y}_x$ is an unbiased estimator of μ_x.

Finally, we calculate the variance of $\widehat{Y}_x$ using Theorem 19.3. The variance of $\widehat{Y}_x$ is given by

$$\begin{aligned}
Var\left(\widehat{Y}_x\right) &= Var\left(\widehat{\alpha} + \widehat{\beta}\, x\right) \\
&= Var\left(\widehat{\alpha}\right) + Var\left(\widehat{\beta}\, x\right) + 2Cov\left(\widehat{\alpha},\, \widehat{\beta}\, x\right) \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) + x^2\, \frac{\sigma^2}{S_{xx}} + 2\, x\, Cov\left(\widehat{\alpha},\, \widehat{\beta}\right) \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) - 2\, x\, \frac{\overline{x}\, \sigma^2}{S_{xx}} \\
&= \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right) \sigma^2.
\end{aligned}$$

In this computation we have used the fact that

$$Cov\left(\widehat{\alpha},\, \widehat{\beta}\right) = -\frac{\overline{x}\, \sigma^2}{S_{xx}}$$

whose proof is left to the reader as an exercise. The proof of the theorem is now complete.

By Theorem 19.3, we see that

$$\widehat{\beta} \sim N\left(\beta,\; \frac{\sigma^2}{S_{xx}}\right) \qquad \text{and} \qquad \widehat{\alpha} \sim N\left(\alpha,\; \frac{\sigma^2}{n} + \frac{\overline{x}^2 \sigma^2}{S_{xx}}\right).$$

Since $\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}\, x$, the random variable $\widehat{Y}_x$ is also a normal random variable with mean μ_x and variance

$$Var\left(\widehat{Y}_x\right) = \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right) \sigma^2.$$

Hence standardizing $\widehat{Y}_x$, we have

$$\frac{\widehat{Y}_x - \mu_x}{\sqrt{Var\left(\widehat{Y}_x\right)}} \sim N(0,1).$$

If σ^2 is known, then one can take the statistic $Q = \frac{\widehat{Y}_x - \mu_x}{\sqrt{Var\left(\widehat{Y}_x\right)}}$ as a pivotal quantity to construct a confidence interval for μ_x. The $(1-\gamma)100\%$ confidence interval for μ_x when σ^2 is known is given by

$$\left[\widehat{Y}_x - z_{\frac{\gamma}{2}} \sqrt{Var(\widehat{Y}_x)},\quad \widehat{Y}_x + z_{\frac{\gamma}{2}} \sqrt{Var(\widehat{Y}_x)}\;\right].$$

Example 19.9. Let the following data on the number chirps per second, x by the striped ground cricket and the temperature, y in Fahrenheit is shown below:

x	20	16	20	18	17	16	15	17	15	16
y	89	72	93	84	81	75	70	82	69	83

What is the 95% confidence interval for β? What is the 95% confidence interval for μ_x when $x = 14$ and $\sigma = 3.047$?

Answer: From Example 19.8, we have

$$n = 10, \quad \widehat{\beta} = 4.067, \quad \widehat{\sigma} = 3.047 \quad \text{and} \quad S_{xx} = 376.$$

The $(1-\gamma)\%$ confidence interval for β is given by

$$\left[\widehat{\beta} - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\,, \;\; \widehat{\beta} + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}} \right].$$

Therefore the 90% confidence interval for β is

$$\left[4.067 - t_{0.025}(8)\,(3.047)\sqrt{\frac{10}{(8)\,(376)}}\,, \; 4.067 + t_{0.025}(8)\,(3.047)\sqrt{\frac{10}{(8)\,(376)}} \right]$$

which is

$$[\,4.067 - t_{0.025}(8)\,(0.1755)\,, \quad 4.067 + t_{0.025}(8)\,(0.1755)]\,.$$

Since from the t-table, we have $t_{0.025}(8) = 2.306$, the 90% confidence interval for β becomes

$$[\,4.067 - (2.306)\,(0.1755)\,, \quad 4.067 + (2.306)\,(0.1755)]$$

which is [3.6623, 4.4717].

If variance σ^2 is not known, then we can use the fact that the statistic $U = \frac{n\,\widehat{\sigma}^2}{\sigma^2}$ is chi-squares with $n-2$ degrees of freedom to obtain a pivotal quantity for μ_x. This can be done as follows:

$$\begin{aligned} Q &= \frac{\widehat{Y}_x - \mu_x}{\widehat{\sigma}}\sqrt{\frac{(n-2)\,S_{xx}}{S_{xx} + n\,(x-\overline{x})^2}} \\ &= \frac{\frac{\widehat{Y}_x - \mu_x}{\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2}}}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,\sigma^2}}} \sim t(n-2). \end{aligned}$$

Using this pivotal quantity one can construct a $(1-\gamma)100\%$ confidence interval for mean μ as

$$\left[\widehat{Y}_x - t_{\frac{\gamma}{2}}(n-2) \sqrt{\frac{S_{xx} + n(x-\overline{x})^2}{(n-2)\, S_{xx}}}, \;\; \widehat{Y}_x + t_{\frac{\gamma}{2}}(n-2) \sqrt{\frac{S_{xx} + n(x-\overline{x})^2}{(n-2)\, S_{xx}}} \;\right].$$

Next we determine the 90% confidence interval for μ_x when $x = 14$ and $\sigma = 3.047$. The $(1-\gamma)100\%$ confidence interval for μ_x when σ^2 is known is given by

$$\left[\widehat{Y}_x - z_{\frac{\gamma}{2}} \sqrt{Var(\widehat{Y}_x)}, \;\; \widehat{Y}_x + z_{\frac{\gamma}{2}} \sqrt{Var(\widehat{Y}_x)} \;\right].$$

From the data, we have

$$\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}\, x = 9.761 + (4.067)\,(14) = 66.699$$

and

$$Var\left(\widehat{Y}_x\right) = \left(\frac{1}{10} + \frac{(14-17)^2}{376} \right) \sigma^2 = (0.124)\,(3.047)^2 = 1.1512.$$

The 90% confidence interval for μ_x is given by

$$\left[66.699 - z_{0.025}\, \sqrt{1.1512}, \; 66.699 + z_{0.025}\, \sqrt{1.1512} \;\right]$$

and since $z_{0.025} = 1.96$ (from the normal table), we have

$$[66.699 - (1.96)\,(1.073), \; 66.699 + (1.96)\,(1.073)]$$

which is $[64.596, \; 68.802]$.

We now consider the predictions made by the normal regression equation $\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta} x$. The quantity $\widehat{Y}_x$ gives an estimate of $\mu_x = \alpha + \beta x$. Each time we compute a regression line from a random sample we are observing one possible linear equation in a population consisting all possible linear equations. Further, the actual value of Y_x that will be observed for given value of x is normal with mean $\alpha + \beta x$ and variance σ^2. So the actual observed value will be different from μ_x. Thus, the predicted value for $\widehat{Y}_x$ will be in error from two different sources, namely (1) $\widehat{\alpha}$ and $\widehat{\beta}$ are randomly distributed about α and β, and (2) Y_x is randomly distributed about μ_x.

Let y_x denote the actual value of Y_x that will be observed for the value x and consider the random variable

$$\mathcal{D} = Y_x - \widehat{\alpha} - \widehat{\beta}\, x.$$

Since $\mathcal{D}$ is a linear combination of normal random variables, $\mathcal{D}$ is also a normal random variable.

The mean of $\mathcal{D}$ is given by

$$\begin{aligned} E(\mathcal{D}) &= E(Y_x) - E(\widehat{\alpha}) - x\, E(\widehat{\beta}) \\ &= \alpha + \beta\, x - \alpha - x\, \beta \\ &= 0. \end{aligned}$$

The variance of $\mathcal{D}$ is given by

$$\begin{aligned} Var(\mathcal{D}) &= Var(Y_x - \widehat{\alpha} - \widehat{\beta}\, x) \\ &= Var(Y_x) + Var(\widehat{\alpha}) + x^2\, Var(\widehat{\beta}) + 2\, x\, Cov(\widehat{\alpha}, \widehat{\beta}) \\ &= \sigma^2 + \frac{\sigma^2}{n} + \frac{\overline{x}^2\, \sigma^2}{S_{xx}} + x^2\, \frac{\sigma^2}{S_{xx}} - 2\, x\, \frac{\overline{x}}{S_{xx}} \\ &= \sigma^2 + \frac{\sigma^2}{n} + \frac{(x - \overline{x})^2\, \sigma^2}{S_{xx}} \\ &= \frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2. \end{aligned}$$

Therefore

$$\mathcal{D} \sim N\left(0,\ \frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2\right).$$

We standardize $\mathcal{D}$ to get

$$Z = \frac{\mathcal{D} - 0}{\sqrt{\frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2}} \sim N(0, 1).$$

Since in practice the variance of Y_x which is σ^2 is unknown, we can not use Z to construct a confidence interval for a predicted value y_x.

We know that $U = \frac{n\, \widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)$. By Theorem 14.6, the statistic

$\frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2)$. Hence

$$
\begin{aligned}
Q &= \frac{y_x - \widehat{\alpha} - \widehat{\beta}\,x}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{(n+1)\,S_{xx}+n}} \\
&= \frac{\frac{y_x - \widehat{\alpha} - \widehat{\beta}\,x}{\sqrt{\frac{(n+1)\,S_{xx}+n}{n\,S_{xx}}\,\sigma^2}}}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,\sigma^2}}} \\
&= \frac{\frac{\mathcal{D}-0}{\sqrt{Var(\mathcal{D})}}}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,\sigma^2}}} \\
&= \frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2).
\end{aligned}
$$

The statistic Q is a pivotal quantity for the predicted value y_x and one can use it to construct a $(1-\gamma)100\%$ confidence interval for y_x. The $(1-\gamma)100\%$ confidence interval, $[a, b]$, for y_x is given by

$$
\begin{aligned}
1-\gamma &= P\left(-t_{\frac{\gamma}{2}}(n-2) \leq Q \leq t_{\frac{\gamma}{2}}(n-2)\right) \\
&= P(a \leq y_x \leq b),
\end{aligned}
$$

where

$$
a = \widehat{\alpha} + \widehat{\beta}\,x - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\,\sqrt{\frac{(n+1)\,S_{xx}+n}{(n-2)\,S_{xx}}}
$$

and

$$
b = \widehat{\alpha} + \widehat{\beta}\,x + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\,\sqrt{\frac{(n+1)\,S_{xx}+n}{(n-2)\,S_{xx}}}.
$$

This confidence interval for y_x is usually known as the *prediction interval* for predicted value y_x based on the given x. The prediction interval represents an interval that has a probability equal to $1-\gamma$ of containing not a parameter but a future value y_x of the random variable Y_x. In many instances the prediction interval is more relevant to a scientist or engineer than the confidence interval on the mean μ_x.

Example 19.10. Let the following data on the number chirps per second, x by the striped ground cricket and the temperature, y in Fahrenheit is shown below:

x	20	16	20	18	17	16	15	17	15	16
y	89	72	93	84	81	75	70	82	69	83

What is the 95% prediction interval for y_x when $x = 14$?

Answer: From Example 19.8, we have

$$n = 10, \quad \widehat{\beta} = 4.067, \quad \widehat{\alpha} = 9.761, \quad \widehat{\sigma} = 3.047 \quad \text{and} \quad S_{xx} = 376.$$

Thus the normal regression line is

$$y_x = 9.761 + 4.067x.$$

Since $x = 14$, the corresponding predicted value y_x is given by

$$y_x = 9.761 + (4.067)\,(14) = 66.699.$$

Therefore

$$\begin{aligned} a &= \widehat{\alpha} + \widehat{\beta}\,x - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\,\sqrt{\frac{(n+1)\,S_{xx} + n}{(n-2)\,S_{xx}}} \\ &= 66.699 - t_{0.025}(8)\,(3.047)\,\sqrt{\frac{(11)\,(376) + 10}{(8)\,(376)}} \\ &= 66.699 - (2.306)\,(3.047)\,(1.1740) \\ &= 58.4501. \end{aligned}$$

Similarly

$$\begin{aligned} b &= \widehat{\alpha} + \widehat{\beta}\,x + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\,\sqrt{\frac{(n+1)\,S_{xx} + n}{(n-2)\,S_{xx}}} \\ &= 66.699 + t_{0.025}(8)\,(3.047)\,\sqrt{\frac{(11)\,(376) + 10}{(8)\,(376)}} \\ &= 66.699 + (2.306)\,(3.047)\,(1.1740) \\ &= 74.9479. \end{aligned}$$

Hence the 95% prediction interval for y_x when $x = 14$ is [58.4501, 74.9479].

19.3. The Correlation Analysis

In the first two sections of this chapter, we examine the regression problem and have done an in-depth study of the least squares and the normal regression analysis. In the regression analysis, we assumed that the values of X are not random variables, but are fixed. However, the values of Y_x for a given value of x are randomly distributed about $E(Y_x) = \mu_x = \alpha + \beta x$. Further, letting ε to be a random variable with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$, one can model the so called regression problem by

$$Y_x = \alpha + \beta\, x + \varepsilon.$$

In this section, we examine the correlation problem. Unlike the regression problem, here both X and Y are random variables and the correlation problem can be modeled by

$$E(Y) = \alpha + \beta\, E(X).$$

From an experimental point of view this means that we are observing random vector (X, Y) drawn from some bivariate population.

Recall that if (X, Y) is a bivariate random variable then the correlation coefficient ρ is defined as

$$\rho = \frac{E\left((X - \mu_X)\,(Y - \mu_Y)\right)}{\sqrt{E\left((X - \mu_X)^2\right)\; E\left((Y - \mu_Y)^2\right)}}$$

where μ_X and μ_Y are the mean of the random variables X and Y, respectively.

Definition 19.1. If $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is a random sample from a bivariate population, then the sample correlation coefficient is defined as

$$R = \frac{\displaystyle\sum_{i=1}^{n} (X_i - \overline{X})\,(Y_i - \overline{Y})}{\sqrt{\displaystyle\sum_{i=1}^{n} (X_i - \overline{X})^2}\;\sqrt{\displaystyle\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}.$$

The corresponding quantity computed from data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ will be denoted by r and it is an estimate of the correlation coefficient ρ.

Now we give a geometrical interpretation of the sample correlation coefficient based on a paired data set $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. We can associate this data set with two vectors $\vec{x} = (x_1, x_2, ..., x_n)$ and $\vec{y} = (y_1, y_2, ..., y_n)$ in $\mathbb{R}^n$. Let $\mathcal{L}$ be the subset $\{\lambda\, \vec{e}\,|\, \lambda \in \mathbb{R}\}$ of $\mathbb{R}^n$, where $\vec{e} = (1, 1, ..., 1) \in \mathbb{R}^n$. Consider the linear space V given by $\mathbb{R}^n$ modulo $\mathcal{L}$, that is $V = \mathbb{R}^n / \mathcal{L}$. The linear space V is illustrated in a figure on next page when $n = 2$.

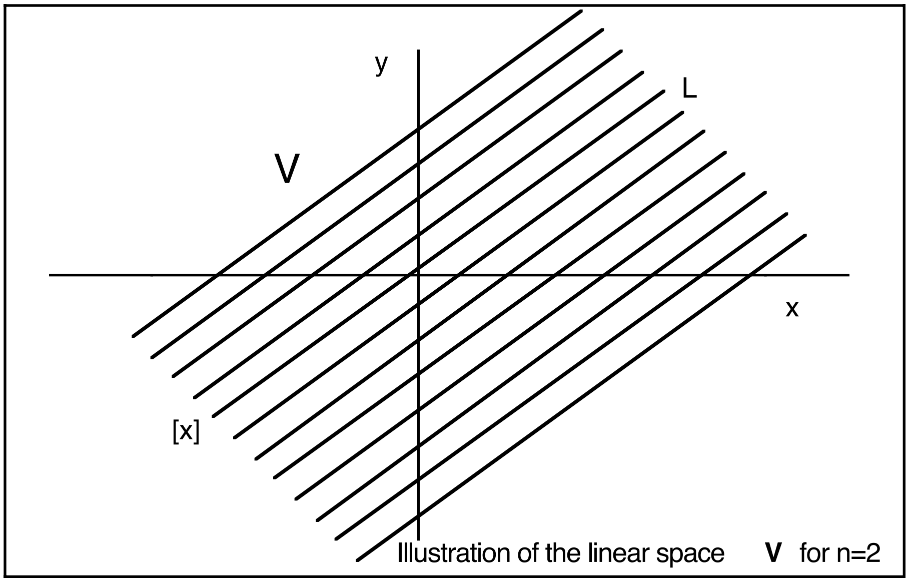


Illustration of the linear space **V** for n=2

We denote the equivalence class associated with the vector $\vec{x}$ by $[\vec{x}]$. In the linear space V it can be shown that the points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ are collinear if and only if the the vectors $[\vec{x}]$ and $[\vec{y}]$ in V are proportional.

We define an inner product on this linear space V by

$$\langle [\vec{x}], [\vec{y}] \rangle = \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}).$$

Then the angle θ between the vectors $[\vec{x}]$ and $[\vec{y}]$ is given by

$$\cos(\theta) = \frac{\langle [\vec{x}], [\vec{y}] \rangle}{\sqrt{\langle [\vec{x}], [\vec{x}] \rangle}\,\sqrt{\langle [\vec{y}], [\vec{y}] \rangle}}$$

which is

$$\cos(\theta) = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}\,\sqrt{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y})^2}} = r.$$

Thus the sample correlation coefficient r can be interpreted geometrically as the cosine of the angle between the vectors $[\vec{x}]$ and $[\vec{y}]$. From this view point the following theorem is obvious.

Theorem 19.7. The sample correlation coefficient r satisfies the inequality

$$-1 \leq r \leq 1.$$

The sample correlation coefficient $r = \pm 1$ if and only if the set of points $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ for $n \geq 3$ are collinear.

To do some statistical analysis, we assume that the paired data is a random sample of size n from a bivariate normal population $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then the conditional distribution of the random variable Y given $X = x$ is normal, that is

$$Y|_x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),\, \sigma_2^2(1 - \rho^2)\right).$$

This can be viewed as a normal regression model $E(Y|_x) = \alpha + \beta\, x$ where $\alpha = \mu - \rho\frac{\sigma_2}{\sigma_1}\mu_1$, $\beta = \rho\frac{\sigma_2}{\sigma_1}$, and $Var(Y|_x) = \sigma_2^2(1 - \rho^2)$.

Since $\beta = \rho\frac{\sigma_2}{\sigma_1}$, if $\rho = 0$, then $\beta = 0$. Hence the null hypothesis $H_o : \rho = 0$ is equivalent to $H_o : \beta = 0$. In the previous section, we devised a hypothesis test for testing $H_o : \beta = \beta_o$ against $H_a : \beta \neq \beta_o$. This hypothesis test, at significance level γ, is "Reject $H_o : \beta = \beta_o$ if $|t| \geq t_{\frac{\gamma}{2}}(n-2)$", where

$$t = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}}\sqrt{\frac{(n-2)\, S_{xx}}{n}}.$$

If $\beta = 0$, then we have

$$t = \frac{\widehat{\beta}}{\widehat{\sigma}}\sqrt{\frac{(n-2)\, S_{xx}}{n}}. \tag{10}$$

Now we express t in term of the sample correlation coefficient r. Recall that

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}, \tag{11}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\, S_{xy}\right], \tag{12}$$

and

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}. \tag{13}$$

Now using (11), (12), and (13), we compute

$$\begin{aligned} t &= \frac{\widehat{\beta}}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n}} \\ &= \frac{S_{xy}}{S_{xx}} \frac{\sqrt{n}}{\sqrt{\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\, S_{xy}\right]}} \sqrt{\frac{(n-2)\, S_{xx}}{n}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} \frac{1}{\sqrt{\left[1 - \frac{S_{xy}}{S_{xx}} \frac{S_{xy}}{S_{yy}}\right]}} \sqrt{n-2} \\ &= \sqrt{n-2}\, \frac{r}{\sqrt{1-r^2}}. \end{aligned}$$

Hence to test the null hypothesis $H_o : \rho = 0$ against $H_a : \rho \neq 0$, at significance level γ, is "Reject $H_o : \rho = 0$ if $|t| \geq t_{\frac{\gamma}{2}}(n-2)$", where $t = \sqrt{n-2}\, \frac{r}{1-r^2}$.

This above test does not extend to test other values of ρ except $\rho = 0$. However, tests for the nonzero values of ρ can be achieved by the following result.

Theorem 19.8. Let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be a random sample from a bivariate normal population $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. If

$$V = \frac{1}{2} \ln \left(\frac{1+R}{1-R} \right) \quad \text{and} \quad m = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right),$$

then

$$Z = \sqrt{n-3}\, (V - m) \to N(0,1) \quad \text{as } n \to \infty.$$

This theorem says that the statistic V is approximately normal with mean m and variance $\frac{1}{n-3}$ when n is large. This statistic can be used to devise a hypothesis test for the nonzero values of ρ. Hence to test the null hypothesis $H_o : \rho = \rho_o$ against $H_a : \rho \neq \rho_o$, at significance level γ, is "Reject $H_o : \rho = \rho_o$ if $|z| \geq z_{\frac{\gamma}{2}}$", where $z = \sqrt{n-3}\, (V - m_o)$ and $m_o = \frac{1}{2} \ln \left(\frac{1+\rho_o}{1-\rho_o} \right)$.

Example 19.11. The following data were obtained in a study of the relationship between the weight and chest size of infants at birth:

x, weight in kg	2.76	2.17	5.53	4.31	2.30	3.70
y, chest size in cm	29.5	26.3	36.6	27.8	28.3	28.6

Determine the sample correlation coefficient r and then test the null hypothesis $H_o : \rho = 0$ against the alternative hypothesis $H_a : \rho \neq 0$ at a significance level 0.01.

Answer: From the above data we find that

$$\overline{x} = 3.46 \qquad \text{and} \qquad \overline{y} = 29.51.$$

Next, we compute S_{xx}, S_{yy} and S_{xy} using a tabular representation.

$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})(y - \overline{y})$	$(x - \overline{x})^2$	$(y - \overline{y})^2$
-0.70	-0.01	0.007	0.490	0.000
-1.29	-3.21	4.141	1.664	10.304
2.07	7.09	14.676	4.285	50.268
0.85	-1.71	-1.453	0.722	2.924
-1.16	-1.21	1.404	1.346	1.464
0.24	-0.91	-0.218	0.058	0.828
		$S_{xy} = 18.557$	$S_{xx} = 8.565$	$S_{yy} = 65.788$

Hence, the correlation coefficient r is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} = \frac{18.557}{\sqrt{(8.565)\,(65.788)}} = 0.782.$$

The computed t value is give by

$$t = \sqrt{n-2}\, \frac{r}{\sqrt{1-r^2}} = \sqrt{(6-2)}\, \frac{0.782}{\sqrt{1-(0.782)^2}} = 2.509.$$

From the t-table we have $t_{0.005}(4) = 4.604$. Since

$$2.509 = |t| \not\geq t_{0.005}(4) = 4.604$$

we do not reject the null hypothesis $H_o : \rho = 0$.

19.4. Review Exercises

1. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimators of $\widehat{\beta}$ and $\widehat{\sigma}^2$ of β and σ^2.

2. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then show that the maximum likelihood estimator of $\widehat{\beta}$ is normally distributed. What are the mean and variance of $\widehat{\beta}$?

3. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find an unbiased estimator $\widehat{\sigma}^2$ of σ^2 and then find a constant k such that $k\,\widehat{\sigma}^2 \sim \chi^2(2n)$.

4. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find a pivotal quantity for β and using this pivotal quantity construct a $(1-\gamma)100\%$ confidence interval for β.

5. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find a pivotal quantity for σ^2 and using this pivotal quantity construct a $(1-\gamma)100\%$ confidence interval for σ^2.

6. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim EXP(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimator of $\widehat{\beta}$ of β.

7. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim EXP(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the least squares estimator of $\widehat{\beta}$ of β.

8. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim POI(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimator of $\widehat{\beta}$ of β.

9. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim POI(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the least squares estimator of $\widehat{\beta}$ of β.

10. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim POI(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, show that the least squares estimator and the maximum likelihood estimator of β are both unbiased estimator of β.

11. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim POI(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, the find the variances of both the least squares estimator and the maximum likelihood estimator of β.

12. Given the five pairs of points (x, y) shown below:

x	10	20	30	40	50
y	50.071	0.078	0.112	0.120	0.131

What is the curve of the form $y = a + bx + cx^2$ best fits the data by method of least squares?

13. Given the five pairs of points (x, y) shown below:

x	4	7	9	10	11
y	10	16	22	20	25

What is the curve of the form $y = a + b\,x$ best fits the data by method of least squares?

14. The following data were obtained from the grades of six students selected at random:

Mathematics Grade, x	72	94	82	74	65	85
English Grade, y	76	86	65	89	80	92

Find the sample correlation coefficient r and then test the null hypothesis $H_o : \rho = 0$ against the alternative hypothesis $H_a : \rho \neq 0$ at a significance level 0.01.

15. Given a set of data $\{(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)\}$ what is the least square estimate of α if $y = \alpha$ is fitted to this data set.

16. Given a set of data points $\{(2, 3), (4, 6), (5, 7)\}$ what is the curve of the form $y = \alpha + \beta\, x^2$ best fits the data by method of least squares?

17. Given a data set $\{(1, 1), (2, 1), (2, 3), (3, 2), (4, 3)\}$ and $Y_x \sim N(\alpha + \beta\, x,\ \sigma^2)$, find the point estimate of σ^2 and then construct a 90% confidence interval for σ.

18. For the data set $\{(1, 1), (2, 1), (2, 3), (3, 2), (4, 3)\}$ determine the correlation coefficient r. Test the null hypothesis $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ at a significance level 0.01.