MATH 4752/6752 – Mathematical Statistics II

Regression and Correlation

Xiaojing Ye
Department of Mathematics & Statistics
Georgia State University

In statistical inference, we are often interested in predicting the value of a variable based on observation of one (or multiple) other variables, which is called **bivariate regression** (or **multiple regression**).

In bivariate regression, we want to obtain the **regression equation** of Y on X defined as the conditional expectation of Y given $X = x$:

$$\mu_{Y|x} = \mathbb{E}[Y|X = x] = \int y f_{Y|X}(y|x)\, dy$$

For discrete random variables, we replace integral with sum.

The regression equation of X on Y and regression equation of Y on multiple variables $X_1, \ldots, X_k$ can be defined similarly.

Example. Given the two random variables X and Y that have the joint density

$$f(x,y) = \begin{cases} x \cdot e^{-x(1+y)} & \text{for } x > 0 \text{ and } y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find the regression equation of Y on X.

Solution. We first compute the marginal pdf of X:

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy = \begin{cases} e^{-x} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

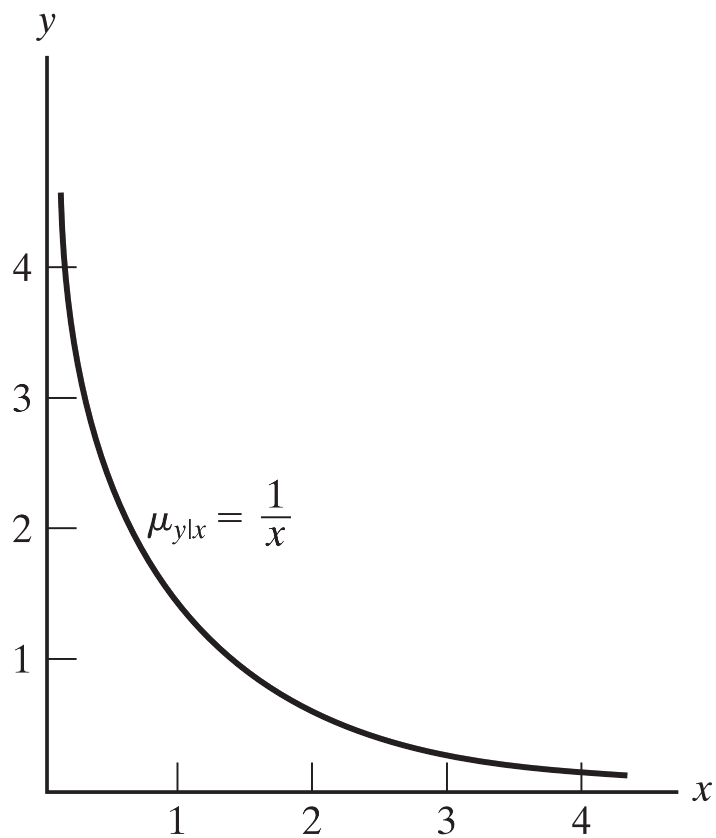
and hence the conditional pdf of Y given $X = x$ is

$$w(y \mid x) = \frac{f(x,y)}{g(x)} = \frac{x \cdot e^{-x(1+y)}}{e^{-x}} = x \cdot e^{-xy}$$

for $y > 0$ and $w(y|x) = 0$ elsewhere. Notice that $(Y|X = x) \sim$ Exponential$(1/x)$. Hence

$$\mu_{Y|x} = \mathbb{E}[Y|X = x] = \int_0^{\infty} y \cdot x \cdot e^{-xy}\, dy = \frac{1}{x}.$$

Here is the plot of the regression equation $\mu_{Y|x} = \frac{1}{x}$ for $x > 0$:



Example. If X and Y have the multinomial distribution

$$f(x,y) = \binom{n}{x, y, n-x-y} \cdot \theta_1^x \theta_2^y (1-\theta_1-\theta_2)^{n-x-y}$$

for $x, y = 0, 1, \dots, n$ with $x + y \leq n$, find the regression equation of Y on X.

Solution. The marginal pmf of X is given by

$$\begin{aligned} g(x) &= \sum_{y=0}^{n-x} \binom{n}{x, y, n-x-y} \cdot \theta_1^x \theta_2^y (1-\theta_1-\theta_2)^{n-x-y} \\ &= \binom{n}{x} \theta_1^x \sum_{y=0}^{n-x} \binom{n-x}{y} \theta_2^y (1-\theta_1-\theta_2)^{n-x-y} \\ &= \binom{n}{x} \theta_1^x (1-\theta_1)^{n-x} \end{aligned}$$

for $x = 0, 1, \dots, n$, which means that X follows Binomial (n, θ_1) distribution.

Solution (cont). Therefore we obtain the condition pmf of Y given $X = x$:

$$w(y \mid x) = \frac{f(x,y)}{g(x)} = \frac{\binom{n-x}{y} \theta_2^y (1-\theta_1-\theta_2)^{n-x-y}}{(1-\theta_1)^{n-x}}$$
$$= \binom{n-x}{y} \left(\frac{\theta_2}{1-\theta_1}\right)^y \left(\frac{1-\theta_1-\theta_2}{1-\theta_1}\right)^{n-x-y}$$

for $y = 0, 1, \dots, n-x$.

Therefore we know $(Y|X = x) \sim$Binomial($n-x, \frac{\theta_2}{1-\theta_1}$), and hence the regression equation of Y on X is

$$\mu_{Y|x} = \mathbb{E}[Y|X = x] = (n-x) \cdot \frac{\theta_2}{1-\theta_1} = \frac{(n-x)\theta_2}{1-\theta_1}.$$

Example. If the joint density of X_1, X_2, X_3 is given by

$$f(x_1, x_2, x_3) = \begin{cases} (x_1 + x_2)\, e^{-x_3} & \text{for } 0 < x_1 < 1, 0 < x_2 < 1, x_3 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find the regression equation of X_2 on X_1 and X_3.

Solution. The joint density of X_1 and X_3 is given by

$$m(x_1, x_3) = \begin{cases} \left(x_1 + \frac{1}{2}\right) e^{-x_3} & \text{for } 0 < x_1 < 1, x_3 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Therefore

$$\mu_{X_2|x_1,x_3} = \int_{-\infty}^{\infty} x_2 \cdot \frac{f(x_1, x_2, x_3)}{m(x_1, x_3)} dx_2 = \int_0^1 \frac{x_2 (x_1 + x_2)}{\left(x_1 + \frac{1}{2}\right)} dx_2$$

$$= \frac{x_1 + \frac{2}{3}}{2x_1 + 1}.$$

An important class of regression equations is linear (affine) in x:

$$\mu_{Y|x} = \alpha + \beta x$$

for some constants α and β, which are called **regression coefficients**.

Remarks. Linear regression equations are important because:

- They lend themselves readily to further mathematical treatment;
- They often provide good approximations to otherwise complicated regression equations;
- In the case of the bivariate normal distribution, the regression equations are, in fact, linear.

Theorem. If the regression of Y on X is linear, then

$$\mu_{Y|x} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

where

$$\mathbb{E}[X] = \mu_1, \quad \mathbb{E}[Y] = \mu_2, \quad \text{var}[X] = \sigma_1^2, \quad \text{var}[Y] = \sigma_2^2.$$

and

$$\text{cov}(X, Y) = \sigma_{12}, \quad \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}.$$

Proof. Since $\mu_{Y|x} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$ for some α, β, it follows that

$$\int y \cdot w(y|x)dy = \alpha + \beta x \qquad (*)$$

Proof (cont). Multiplying both sides of $(*)$ by $g(x)$ and integrating on x yield

$$\mu_2 = \iint y \cdot w(y \mid x) g(x) dy dx = \alpha \int g(x) dx + \beta \int x \cdot g(x) dx = \alpha + \beta \mu_1.$$

Multiplying both sides of $(*)$ by $x \cdot g(x)$ and integrating on x yield

$$\begin{aligned} \mathbb{E}[XY] &= \iint xy \cdot w(y \mid x) g(x) dy dx \\ &= \alpha \int x \cdot g(x) dx + \beta \int x^2 \cdot g(x) dx \\ &= \alpha \mu_1 + \beta \, \mathbb{E}[X^2]. \end{aligned}$$

Recall that

$$\mathbb{E}[XY] = \sigma_{12} + \mu_1 \mu_2, \qquad \mathbb{E}[X^2] = \sigma_1^2 + \mu_1^2.$$

Then solving the equations above for α and β yields

$$\begin{aligned} \alpha &= \mu_2 - \frac{\sigma_{12}}{\sigma_1^2} \cdot \mu_1 = \mu_2 - \rho \frac{\sigma_2}{\sigma_1} \cdot \mu_1 \\ \beta &= \frac{\sigma_{12}}{\sigma_1^2} = \rho \frac{\sigma_2}{\sigma_1} \end{aligned}$$

We have discussed the problem of regression only in connection with random variables having **known** joint distributions.

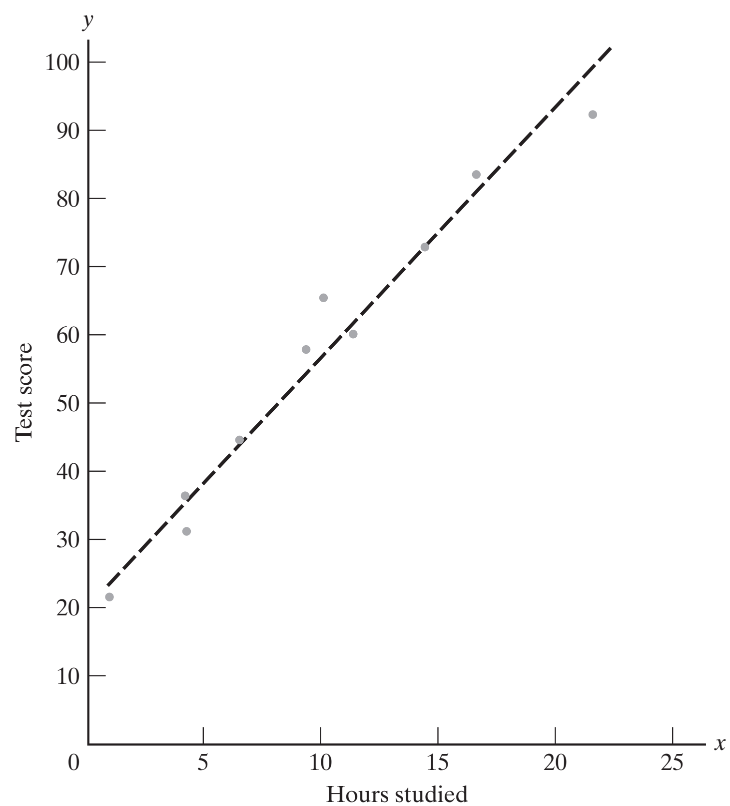
In practice, there are many problems where a set of paired data gives the indication that the regression is linear, where we do not know the joint distribution but want to estimate the regression coefficients α and β.

A typical method is called the **method of least squares**.

Consider the following data on the number of hours that 10 persons studied for a French test and their scores on the test:

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

From the plot of the data below, we get the impression that a straight line provides a reasonably good fit:



Although the points do not all fall exactly on a straight line, the overall pattern suggests that the average test score for a given number of hours studied may well be related to the number of hours studied in a linear pattern.

Suppose we are given a set of paired data

$$\{(x_i, y_i) : i = 1, \dots, n\}.$$

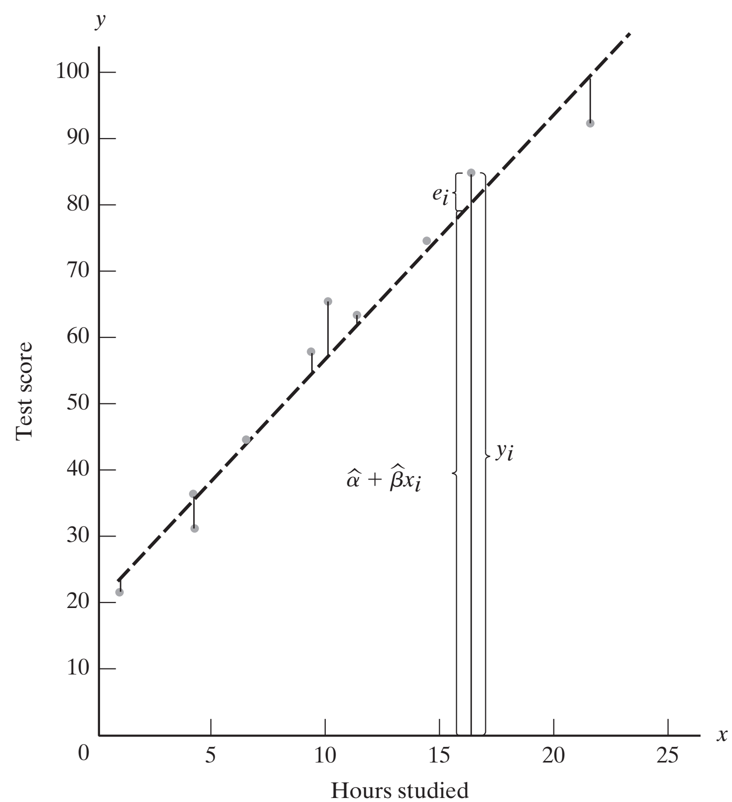
Then the **least squares estimates** of the regression coefficients $(\widehat{\alpha}, \widehat{\beta})$ in bivariate linear regression are the minimizer of

$$q(\alpha, \beta) = \sum_{i=1}^{n} \Big(y_i - (\alpha + \beta x_i)\Big)^2.$$

In other words,

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{arg\,min}_{\alpha,\beta} q(\alpha, \beta).$$

Notice that $q(\alpha, \beta)$ is the sum of squared errors, i.e., $\sum_{i=1}^{n} e_i^2$ where e_i is the discrepancy between y_i and $\alpha + \beta x_i$:



So the least squares estimates $(\widehat{\alpha}, \widehat{\beta})$ are the interception and slope combination that yield smallest sum of squared errors.

To find the minimizer $(\widehat{\alpha}, \widehat{\beta})$ of $q(\alpha, \beta)$, we take partial derivatives of q with respect to α and β, setting them to 0, and solving for α and β:

$$\frac{\partial q}{\partial \widehat{\alpha}} = \sum_{i=1}^{n} (-2) \left[y_i - \left(\widehat{\alpha} + \widehat{\beta} x_i \right) \right] = 0$$
$$\frac{\partial q}{\partial \widehat{\beta}} = \sum_{i=1}^{n} (-2) x_i \left[y_i - \left(\widehat{\alpha} + \widehat{\beta} x_i \right) \right] = 0$$

These two equations can be written as a system of **normal equations** of $(\widehat{\alpha}, \widehat{\beta})$:

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

Notice that the system above is a system of linear equations of $(\widehat{\alpha}, \widehat{\beta})$. Solving this system yields the solution

$$\widehat{\alpha} = \frac{\sum_{i=1}^n y_i - \widehat{\beta} \cdot \sum_{i=1}^n x_i}{n}$$

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

It is customary to use the following notations:

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

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

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

Then we can simplify the expressions of $\widehat{\alpha}$ and $\widehat{\beta}$ as

$$\widehat{\alpha} = \bar{y} - \widehat{\beta} \cdot \bar{x}, \qquad \widehat{\beta} = \frac{S_{xy}}{S_{xx}}$$

Example. Consider the data in the following table.
(a) find the equation of the least squares line that approximates the regression of the test scores on the number of hours studied;
(b) predict the average test score of a person who studied 14 hours for test.

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

Solution. (a) We have $n = 10$ and compute

$$\sum_{i=1}^{n} x_i = 100, \quad \sum_{i=1}^{n} x_i^2 = 1,376, \quad \sum_{i=1}^{n} y_i = 564, \quad \sum_{i=1}^{n} x_i y_i = 6,945.$$

From these we obtain

$$S_{xx} = 1,376 - \frac{1}{10}(100)^2 = 376, \quad S_{xy} = 6,945 - \frac{1}{10}(100)(564) = 1,305$$

Therefore

$$\widehat{\beta} = \frac{1,305}{376} = 3.471, \quad \widehat{\alpha} = \frac{564}{10} - 3.471 \cdot \frac{100}{10} = 21.69.$$

So the equation of the least squares line is $\widehat{y} = 21.69 + 3.471x$.

(b) Substituting $x = 14$ into the equation obtained in part (a), we get

$$\widehat{y} = 21.69 + 3.471 \cdot 14 = 70.284 \approx 70.$$

Given a set of paired data $\{x_i, y_i) : i = 1, \ldots, n\}$, there are two ways to interpret the data:

- **Regression analysis**: we analyze by treating x_i's as constants and y_i's as values of corresponding independent random variables Y_i.

- **Correlation analysis**: we look upon the (x_i, y_i) as values of the independent random vectors (X_i, Y_i).

We first consider regression analysis, in particular, **normal regression analysis**, where the conditional density of Y_i is given by:

$$w(y_i \mid x_i) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left[\frac{y_i-(\alpha+\beta x_i)}{\sigma}\right]^2} \qquad -\infty < y_i < \infty$$

and α, β, and σ are the same for each i.

We will be interested in the following questions:

- Point and interval estimations $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\sigma}$ of α, β, and σ.
- Hypothesis testings involving $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\sigma}$.
- Prediction using $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$ for new x.

Suppose we use maximum likelihood estimates of α, β, and σ, then we first form the log-likelihood function:

$$\ell(\alpha,\beta,\sigma) = \ln \prod_{i=1}^{n} w(y_i|x_i) = -n\ln\sigma - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\frac{y_i-(\alpha+\beta x_i)}{\sigma}\right]^2$$

Taking partial derivatives of ℓ with respect to α, β, σ and setting them to 0:

$$\frac{\partial \ell}{\partial \alpha} = \frac{1}{\sigma^2}\cdot\sum_{i=1}^{n}\left[y_i-(\alpha+\beta x_i)\right] = 0$$

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2}\cdot\sum_{i=1}^{n}x_i\left[y_i-(\alpha+\beta x_i)\right] = 0$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\cdot\sum_{i=1}^{n}\left[y_i-(\alpha+\beta x_i)\right]^2 = 0$$

Solving for α, β, σ yields $\widehat{\alpha} = \bar{y} - \widehat{\beta}\cdot\bar{x}$ and $\widehat{\beta} = \frac{S_{xy}}{S_{xx}}$ as before, and

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}\left[y_i-(\alpha+\beta x_i)\right]^2} = \sqrt{\frac{1}{n}(S_{yy}-\widehat{\beta}\cdot S_{xy})}$$

Let $\widehat{A}$, $\widehat{B}$, $\widehat{\Sigma}$ denote the corresponding maximum likelihood estimators obtained above. Then

$$\widehat{B} = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right)}{S_{xx}} = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right) Y_i$$

which is a linear combination of Y_i's. Therefore $\widehat{B}$ also follows normal distribution, and

$$\mathbb{E}[\widehat{B}] = \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{S_{xx}}\right] \cdot E\left(Y_i \mid x_i\right) = \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{S_{xx}}\right] (\alpha + \beta x_i) = \beta$$

and

$$\text{var}[\widehat{B}] = \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{S_{xx}}\right]^2 \cdot \text{var}\left(Y_i \mid x_i\right) = \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{S_{xx}}\right]^2 \cdot \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

Theorem. For normal population,

$$\hat{B} \sim N(\beta, \frac{\sigma^2}{S_{xx}}) \quad \text{and} \quad \frac{n\hat{\Sigma}^2}{\sigma^2} \sim \chi^2_{n-2},$$

and they are independent.

The theorem above implies that

$$T = \frac{\frac{\hat{B}-\beta}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{n\hat{\Sigma}^2}{\sigma^2}/(n-2)}} = \frac{\hat{B}-\beta}{\hat{\Sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}}$$

Example. With reference to the data in the table in Section 3 pertaining to the amount of time that 10 persons studied for a certain test and the scores that they obtained, test the null hypothesis $\beta = 3$ against the alternative hypothesis $\beta > 3$ at the 0.01 level of significance.

Solution. We proceed with the four steps:

- **Step 1.** Set up the test

$$H_0 : \ \beta = 3 \quad \text{vs} \quad H_1 : \ \beta > 3$$

  with level of significance $\alpha = 0.01$.

- **Step 2.** Decide to use test statistic $T = \frac{\widehat{B} - \beta}{\widehat{\Sigma}} \sqrt{\frac{(n-2)S_{xx}}{n}}$ and reject if

$$t = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)S_{xx}}{n}} > t_{\alpha,n-1} = t_{0.01,8} = 2.896.$$

- **Step 3.** Based on the data table, we obtain

$$\sum_{i=1}^{n} y_i^2 = 36,562,$$

$$S_{yy} = 36,562 - \frac{564^2}{10} = 4,752.4$$

$$\hat{\sigma} = \sqrt{\frac{1}{10}(4,752.4 - 3.471 \cdot 1,305)} = 4.720$$

$$t = \frac{3.471 - 3}{4.720}\sqrt{\frac{8 \cdot 376}{10}} = 1.73.$$

- **Step 4.** Since $t = 1.73 < 2.896$, we cannot reject H_0.

The derivations above also implies the interval estimation of β: we know

$$\mathsf{P}\left(-t_{\alpha/2,n-2} < \frac{\widehat{\mathsf{B}} - \beta}{\widehat{\Sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

which implies that

$$\widehat{\beta} - t_{\alpha/2,n-2} \cdot \widehat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} < \beta < \widehat{\beta} + t_{\alpha/2,n-2} \cdot \widehat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}$$

is a $(1-\alpha)\cdot$100% confidence interval for β.

Example. With reference to the data in the table in Section 3 pertaining to the amount of time that 10 persons studied for a certain test and the scores that they obtained, construct a 95% confidence interval for β.

Solution. We have $\alpha/2 = 0.025$ and find that $t_{0.025,8} = 2.306$. Then the 95% confidence interval of β is

$$3.471-(2.306)(4.720)\sqrt{\frac{10}{8(376)}} < \beta < 3.471+(2.306)(4.720)\sqrt{\frac{10}{8(376)}}$$

which is

$$2.84 < \beta < 4.10.$$

Now we consider correlation analysis for normal data pairs $\{(x_i, y_i) : i = 1, \dots, n\}$. Suppose they are samples from the bivariate normal distribution

$$N\left((\mu_1, \mu_2), \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

To obtain maximum likelihood estimates of $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$, we first write the likelihood function

$$L(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \prod_{i=1}^{n} f(x_i, y_i; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$$

or the log-likelihood function

$$\ell(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \sum_{i=1}^{n} \ln f(x_i, y_i; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho),$$

where $f(x, y; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ is the pdf of the bivariate normal distribution above.

To obtain maximum likelihood estimates, we take partial derivatives of ℓ with respect to $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$, set to 0, and solve for $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ to obtain

$$\begin{aligned}
\widehat{\mu}_1 &= \bar{x}, \\
\widehat{\mu}_2 &= \bar{y} \\
\widehat{\sigma}_1 &= \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \\
\widehat{\sigma}_2 &= \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}} \\
\widehat{\rho} &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}
\end{aligned}$$

The sample correlation coefficient $\widehat{\rho}$, as the maximum likelihood estimate of ρ, is often denoted by r, and the corresponding maximum estimator is denoted by R.

Recall that for bivariate normal distribution, there is

$$\sigma^2_{Y|x} = \text{var}[Y|X = x] = \sigma_2^2(1 - \rho^2)$$

Notice that, if $\rho = 1$, then $\sigma^2_{Y|x} = 0$ and there is a perfect linear relation between X and Y (so one determines the other and vice versa).

Similarly, if $\widehat{\rho} = 1$, then the data pairs $\{(x_i, y_i) : 1 \le i \le n\}$ lie on a straight line.

Example. Suppose that we want to determine on the basis of the following data whether there is a relationship between the time, in minutes, it takes a secretary to complete a certain form in the morning and in the late afternoon:

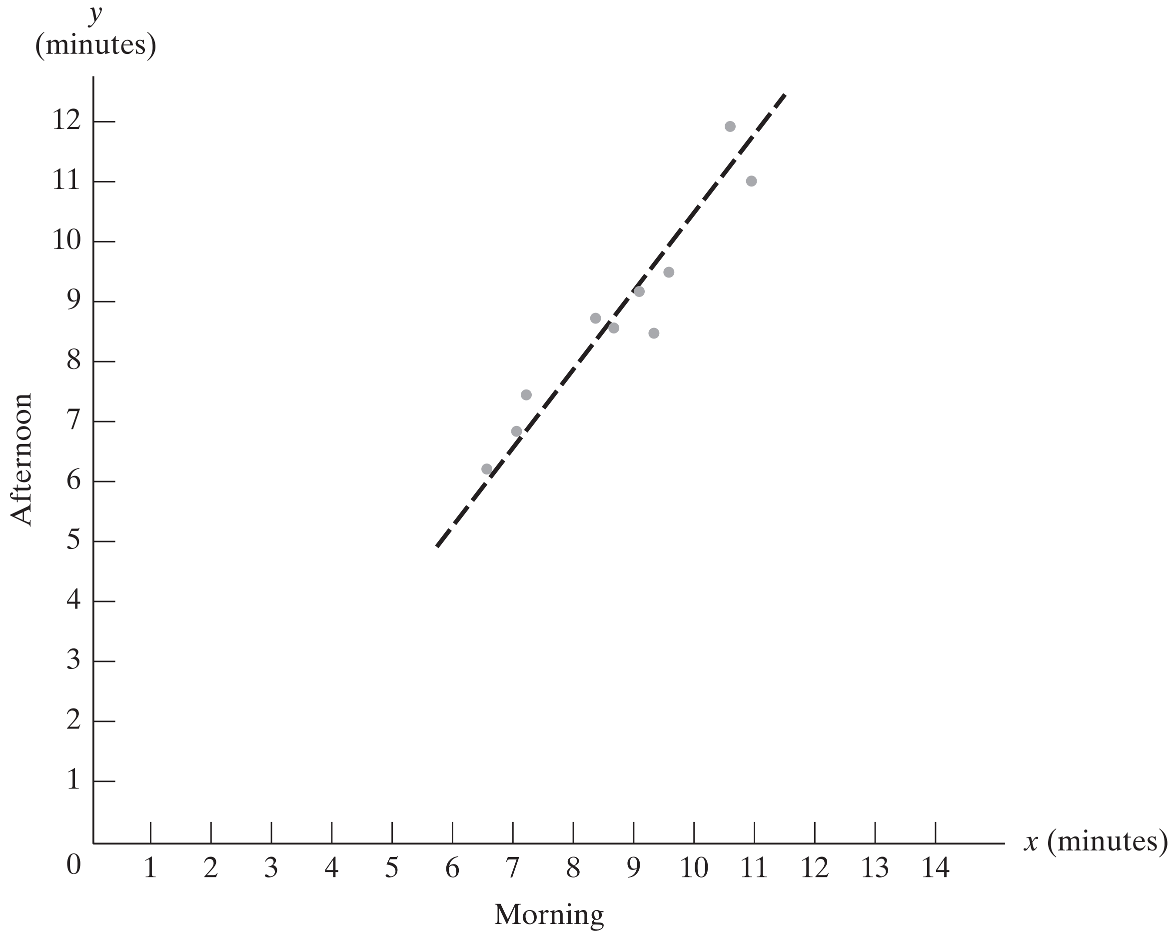
Morning x	Afternoon y
8.2	8.7
9.6	9.6
7.0	6.9
9.4	8.5
10.9	11.3
7.1	7.6
9.0	9.2
6.6	6.3
8.4	8.4
10.5	12.3

Compute and interpret the sample correlation coefficient.

Solution. From the data we get $n = 10$, $\sum_{i=1}^{n} x = 86.7$, $\sum_{i=1}^{n} x_i^2 = 771.35$, $\sum_{i=1}^{n} y_i = 88.8$, $\sum_{i=1}^{n} y_i^2 = 819.34$, and $\sum_{i=1}^{n} x_i y_i = 792.92$, then

$$
\begin{aligned}
S_{xx} &= 771.35 - \frac{1}{10}(86.7)^2 = 19.661 \\
S_{yy} &= 819.34 - \frac{1}{10}(88.8)^2 = 30.796 \\
S_{xy} &= 792.92 - \frac{1}{10}(86.7)(88.8) = 23.024 \\
r &= \frac{23.024}{\sqrt{(19.661)(30.796)}} = 0.936
\end{aligned}
$$

The **scattergram** of data and the fitted line is given by



The distribution of the maximum likelihood estimator R is complicated. However, there is approximately

$$\frac{1}{2} \cdot \ln \frac{1+R}{1-R} \in N\left(\frac{1}{2} \cdot \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right).$$

Therefore, we know

$$z = \frac{\frac{1}{2} \cdot \ln \frac{1+r}{1-r} - \frac{1}{2} \cdot \ln \frac{1+\rho}{1-\rho}}{\frac{1}{\sqrt{n-3}}} = \frac{\sqrt{n-3}}{2} \cdot \ln \frac{(1+r)(1-\rho)}{(1-r)(1+\rho)}$$

is approximately $N(0, 1)$. We conduct hypothesis test or construct confidence interval based on this approximation.

Example. Suppose that we want to determine on the basis of the following data whether there is a relationship between the time, in minutes, it takes a secretary to complete a certain form in the morning and in the late afternoon:

Morning x	Afternoon y
8.2	8.7
9.6	9.6
7.0	6.9
9.4	8.5
10.9	11.3
7.1	7.6
9.0	9.2
6.6	6.3
8.4	8.4
10.5	12.3

Test the null hypothesis $\rho = 0$ against the alternative hypothesis $\rho \neq 0$ at the 0.01 level of significance.

Solution. We proceed with the four steps:

- **Step 1.** Set up the test

$$H_0: \ \rho = 0 \quad \text{vs} \quad H_1: \ \rho \neq 0$$

  with level of significance $\alpha = 0.01$.

- **Step 2.** Decide to use test statistic $Z = \frac{\sqrt{n-3}}{2} \cdot \ln \frac{1+R}{1-R}$ and reject if

$$|z| = \left| \frac{\sqrt{n-3}}{2} \cdot \ln \frac{1+r}{1-r} \right| > z_{\alpha/2} = z_{0.005} = 2.575.$$

- **Step 3.** Based on the data table, we obtain $r = 0.936$ and thus

$$z = \frac{\sqrt{10}}{2} \cdot \ln \frac{1+0.936}{1-0.936} = 4.5$$

- **Step 4.** Since $z = 4.5 > 2.575$, we reject H_0.

We can extend the bivariate linear regression to multiple linear regression:

$$\mu_{Y|x_1,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

In this case, given data $\{(x_{i1}, \ldots, x_{ik}, y_i : i = 1, \ldots, n\}$, we consider least squares estimates $\widehat{\beta}_0, \ldots, \widehat{\beta}_k$ to minimize the sum of squared errors:

$$q(\beta_0, \ldots, \beta_k) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)\Big)^2$$

To obtain the minimizer, we take partial derivatives of q with respect to β_j for $j = 0, 1, \ldots, k$, set to 0:

$$\frac{\partial q}{\partial \widehat{\beta}_0} = \sum_{i=1}^{n} (-2) \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik} \right) \right] = 0$$

$$\frac{\partial q}{\partial \widehat{\beta}_1} = \sum_{i=1}^{n} (-2) x_{i1} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik} \right) \right] = 0$$

$$\frac{\partial q}{\partial \widehat{\beta}_2} = \sum_{i=1}^{n} (-2) x_{i2} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik} \right) \right] = 0$$

$$\cdots$$

$$\frac{\partial q}{\partial \widehat{\beta}_k} = \sum_{i=1}^{n} (-2) x_{ik} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik} \right) \right] = 0$$

This yields the system of $k+1$ normal equations of the least squares estimates $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k$:

$$\sum_{i=1}^{n} y = \widehat{\beta}_0 \cdot n + \widehat{\beta}_1 \cdot \sum_{i=1}^{n} x_1 + \widehat{\beta}_2 \cdot \sum_{i=1}^{n} x_2 + \cdots + \widehat{\beta}_k \cdot \sum_{i=1}^{n} x_k$$

$$\sum_{i=1}^{n} x_1 y = \widehat{\beta}_0 \cdot \sum_{i=1}^{n} x_1 + \widehat{\beta}_1 \cdot \sum_{i=1}^{n} x_1^2 + \widehat{\beta}_2 \cdot \sum_{i=1}^{n} x_1 x_2 + \cdots + \widehat{\beta}_k \cdot \sum_{i=1}^{n} x_1 x_k$$

$$\sum_{i=1}^{n} x_2 y = \widehat{\beta}_0 \cdot \sum_{i=1}^{n} x_2 + \widehat{\beta}_1 \cdot \sum_{i=1}^{n} x_2 x_1 + \widehat{\beta}_2 \cdot \sum_{i=1}^{n} x_2^2 + \cdots + \widehat{\beta}_k \cdot \sum_{i=1}^{n} x_2 x_k$$

$$\cdots$$

$$\sum_{i=1}^{n} x_k y = \widehat{\beta}_0 \cdot \sum_{i=1}^{n} x_k + \widehat{\beta}_1 \cdot \sum_{i=1}^{n} x_k x_1 + \widehat{\beta}_2 \cdot \sum_{i=1}^{n} x_k x_2 + \cdots + \widehat{\beta}_k \cdot \sum_{i=1}^{n} x_k^2$$

Solving this system yields the least squares estimates $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k$.

Example. The following data show the number of bedrooms, the number of baths, and the prices at which a random sample of eight one-family houses sold in a certain large housing development:

Number of bedrooms x_1	Number of baths x_2	Price (dollars) y
3	2	292,000
2	1	264,600
4	3	317,500
2	1	265,500
3	2	302,000
2	2	275,500
5	3	333,000
4	2	307,500

Use the method of least squares to fit a linear equation of sale price on the numbers of bedrooms and baths. Predict the price of a three-bedroom with two baths house.

Solution. We compute that

$$\sum_{i=1}^{n} x_{i1}y_i = 7,558,200, \quad \sum_{i=1}^{n} x_{i2}y_i = 4,835,600$$

and $n = 8$, $\sum_{i=1}^{n} x_{i1} = 25$, $\sum_{i=1}^{n} x_{i2} = 16$,

$$\sum_{i=1}^{n} y_i = 2,357,600, \quad \sum_{i=1}^{n} x_{i1}^2 = 87, \quad \sum_{i=1}^{n} x_{i1}x_{i2} = 55, \quad \sum_{i=1}^{n} x_{i2}^2 = 36$$

Then we obtain the normal equations:

$$\begin{aligned} 2,357,600 &= 8\widehat{\beta}_0 + 25\widehat{\beta}_1 + 16\widehat{\beta}_2 \\ 7,558,200 &= 25\widehat{\beta}_0 + 87\widehat{\beta}_1 + 55\widehat{\beta}_2 \\ 4,835,600 &= 16\widehat{\beta}_0 + 55\widehat{\beta}_1 + 36\widehat{\beta}_2 \end{aligned}$$

solving which yields:

$$\widehat{\beta}_1 = 224,929, \quad \widehat{\beta}_2 = 15,314, \quad \widehat{\beta}_3 = 10,957.$$

Therefore the linear regression equation is $\widehat{y} = 224,929 + 15,314x_1 + 10,957x_2$. For $x_1 = 3$ and $x_2 = 2$, we obtain $\widehat{y} = 292,785$.

Multiple linear regression computation can be written in matrix notations. Let us denote

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Then the least squares estimate of $\mathbf{B}$ is given by

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where $\mathbf{X}'$ is the transpose of $\mathbf{X}$.

To see this, we notice that $q(\mathbf{B}) = \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2$. Set its gradient $\nabla q(\mathbf{B})$ to $\mathbf{0}$, that is

$$\nabla q(\mathbf{B}) = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{B}) = \mathbf{0}$$

which reduces to the normal equation of $\mathbf{B}$. Solving for $\mathbf{B}$ yields the estimate $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Example. The following data show the number of bedrooms, the number of baths, and the prices at which a random sample of eight one-family houses sold in a certain large housing development:

Number of bedrooms x_1	Number of baths x_2	Price (dollars) y
3	2	292,000
2	1	264,600
4	3	317,500
2	1	265,500
3	2	302,000
2	2	275,500
5	3	333,000
4	2	307,500

Determine the least squares estimates of the multiple regression coefficients using the matrix notations.

Solution. Following the matrix notation, we can compute

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 8 & 25 & 16 \\ 25 & 87 & 55 \\ 16 & 55 & 36 \end{pmatrix}$$

Hence we can compute its inverse:

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \frac{1}{84} \cdot \begin{pmatrix} 107 & -20 & -17 \\ -20 & 32 & -40 \\ -17 & -40 & 71 \end{pmatrix}$$

Moreover, we have

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 2,357,600 \\ 7,558,200 \\ 4,835,600 \end{pmatrix}$$

Solution (cont). Finally, we have

$$\hat{\mathrm{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

$$= \frac{1}{84} \cdot \begin{pmatrix} 07 & -20 & -17 \\ -20 & 32 & -40 \\ -17 & -40 & 71 \end{pmatrix} \begin{pmatrix} 2,357,600 \\ 7,558,200 \\ 4,835,600 \end{pmatrix}$$

$$= \frac{1}{84} \cdot \begin{pmatrix} 18,894,000 \\ 1,286,400 \\ 920,400 \end{pmatrix}$$

$$= \begin{pmatrix} 224,929 \\ 15,314 \\ 10,957 \end{pmatrix}$$

Recall that the maximum likelihood estimate of the standard deviation is given by

$$\widehat{\sigma} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} [y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}]^2}$$

This maximum likelihood estimator can also be written in matrix notation:

$$\widehat{\sigma} = \sqrt{\frac{\mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y}}{n}}.$$

Example. The following data show the number of bedrooms, the number of baths, and the prices at which a random sample of eight one-family houses sold in a certain large housing development:

Number of bedrooms x_1	Number of baths x_2	Price (dollars) y
3	2	292,000
2	1	264,600
4	3	317,500
2	1	265,500
3	2	302,000
2	2	275,500
5	3	333,000
4	2	307,500

Use this data to determine the value of $\widehat{\sigma}$.

Solution. We first compute that

$$\begin{aligned}\mathbf{Y}'\mathbf{Y} &= (292,000)^2 + (264,600)^2 + \ldots + (307,500)^2 \\ &= 699,123,160,0001\end{aligned}$$

Then we can compute

$$\begin{aligned}\mathbf{B}'\mathbf{X}'\mathbf{Y} &= \frac{1}{84} \cdot (18,894,000 \quad 286,400 \quad 920,400) \begin{pmatrix} 637,000 \\ 7,558,200 \\ 4,835,600 \end{pmatrix} \\ &= 699,024,394,285\end{aligned}$$

Using the formula of $\hat{\sigma} = \sqrt{\frac{\mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y}}{n}}$, we obtain

$$\hat{\sigma} = \sqrt{\frac{699,123,160,000 - 699,024,394,285}{8}} = 3,514$$

Remark. Note that the maximum likelihood estimator corresponding to $\widehat{\sigma}$ is not unbiased. The unbiased estimator of σ^2 is given by

$$S_e^2 = \frac{\mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y}}{n - k - 1}$$

Therefore, we would get

$$s_e = \sqrt{\frac{699,123,160,000 - 699,024,394,285}{8 - 2 - 1}} = 4,444$$

for this estimator, which is different from $\widehat{\sigma} = 3,514$ above.

Theorem. For multivariate normal distributions, there are

$$\widehat{B}_i \sim N\left(\beta_i, c_{ii}\sigma^2\right), \quad \text{and} \quad \frac{n\widehat{\Sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}$$

where c_{ij} is the (i,j)th entry of $(\mathbf{X}'\mathbf{X})^{-1}$. Moreover, $\widehat{B}_i$ and $\frac{n\widehat{\Sigma}^2}{\sigma^2}$ are independent.

The theorem above provides a means for hypothesis testing and interval estimation involving $\widehat{\beta}_i$'s. Specifically,

$$T = \frac{\widehat{B}_i - \beta_i}{\widehat{\Sigma} \cdot \sqrt{\frac{n|c_{ii}|}{n-k-1}}} \sim t_{n-k-1}$$

for $i = 0, 1, \ldots, k$.

Example. The following data show the number of bedrooms, the number of baths, and the prices at which a random sample of eight one-family houses sold in a certain large housing development:

Number of bedrooms x_1	Number of baths x_2	Price (dollars) y
3	2	292,000
2	1	264,600
4	3	317,500
2	1	265,500
3	2	302,000
2	2	275,500
5	3	333,000
4	2	307,500

Test the null hypothesis $\beta_1 = 9,500$ against the alter- native hypothesis $\beta_1 > 9,500$ at the 0.05 level of significance.

Solution. We proceed with the four steps:

- **Step 1.** Set up the test

$$H_0 : \beta_1 = 9,500 \quad \text{vs} \quad H_1 : \beta_1 > 9,500.$$

  with level of significance $\alpha = 0.05$.

- **Step 2.** Decide to use test statistic $T = \frac{\widehat{B}_1 - \beta_i}{\widehat{\Sigma} \cdot \sqrt{\frac{n|c_{11}|}{n-k-1}}}$ and reject if

$$t = \frac{\widehat{\beta}_1 - \beta_i}{\widehat{\sigma} \cdot \sqrt{\frac{n|c_{11}|}{n-k-1}}} > t_{\alpha,n-k-1} = t_{0.05,5}.$$

- **Step 3.** Based on the data table, we obtain $n = 8$, $k = 2$, $\widehat{\beta}_1 = 15,314$, $c_{11} = \frac{32}{84}$, and $\widehat{\sigma} = 3,546$ and thus

$$t = \frac{15,314 - 9,500}{3,514\sqrt{\frac{8\cdot\left|\frac{32}{84}\right|}{5}}} = \frac{5,814}{2,743} = 2.12$$

- **Step 4.** Since $t = 2.12 > 2.015$, we reject H_0.