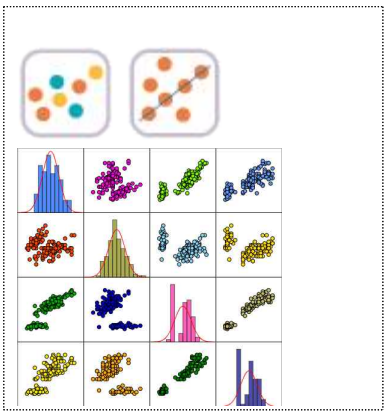
12

Correlation and Regression Analysis



SECTIONS

12.1 Correlation Analysis

12.2 Simple Linear Regression Analysis
- 12.2.1 Simple Linear Regression Model
- 12.2.2 Estimation of Regression Coefficient
- 12.2.3 Goodness of Fit for Regression Line
- 12.2.4 Analysis of Variance for Regression
- 12.2.5 Inference for Regression
- 12.2.6 Residual Analysis

12.3 Multiple Linear Regression Analysis
- 12.3.1 Multiple Linear Regression Model
- 12.3.2 Estimation of Regression Coefficient
- 12.3.3 Goodness of Fit for Regression and Analysis of Variance
- 12.3.4 Inference for Multiple Linear Regression

CHAPTER OBJECTIVES

From Chapter 7 to Chapter 10, we discussed the estimation and the testing hypothesis of parameters such as population mean and variance for single variable.

This chapter describes a correlation analysis for two or more variables.

If variables are related with each other, then a regression analysis is described to see how this association can be used. Simple linear regression analysis and multiple regression analysis are discussed.

12.1 Correlation Analysis

- The easiest way to observe the relation of two variables is to draw a scatter plot with one variable as X axis and the other as Y axis. If two variables are related, data will gather together with a certain pattern, and if not related, data will be scattered around. The correlation analysis is a method of analyzing the degree of linear relationship between two variables. It is to investigate how linearly the other variable increases or decreases as one variable increases.

Example 12.1.1

Based on the survey of advertising costs and sales for 10 companies that make the same product, we obtained the following data as in Table 12.1.1. Using『eStat』, draw a scatter plot for this data and investigate the relation of the two variables.

Table 12.1.1 Advertising costs and sales (unit: 1 million USD)

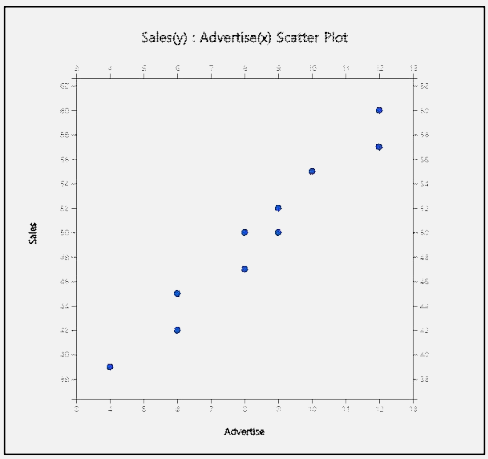
Company	1	2	3	4	5	6	7	8	9	10
Advertise (X)	4	6	6	8	8	9	9	10	12	12
Sales (Y)	39	42	45	47	50	50	52	55	57	60

Ex ⇨ eBook ⇨ EX120101_SalesByAdvertise.csv.

Answer

- Using『eStat』, enter data as shown in <Figure 12.1.1>. If you select the Sales as 'Y Var' and the Advertise 'by X Var' in the variable selection box that appears when you click the scatter plot icon on the main menu, the scatter plot will appear as shown in <Figure 12.1.2>. As we can expect, the scatter plot show that the more investments in advertising, the more sales increase, and not only that, the form of increase is linear.

<Figure 12.1.1> Data input in『eStat』



<Figure 12.1.2> Scatter plot of sales by advertise

- The relation between two variables can be roughly investigated using a scatter plot like this. However, a measure of the extent of the relation can be used together to provide a more accurate and objective view of the relation between two variables. As a measure of the relation between two variables, there is a **covariance**. The population covariance of the two variables X and Y is denoted as $Cov(X, Y)$. When the random samples of two variables are given as

(X_1, Y_1), $\cdots$, (X_n, Y_n), the estimate of the population covariance using samples, which is called the sample covariance, s_{XY}, is defined as follows:

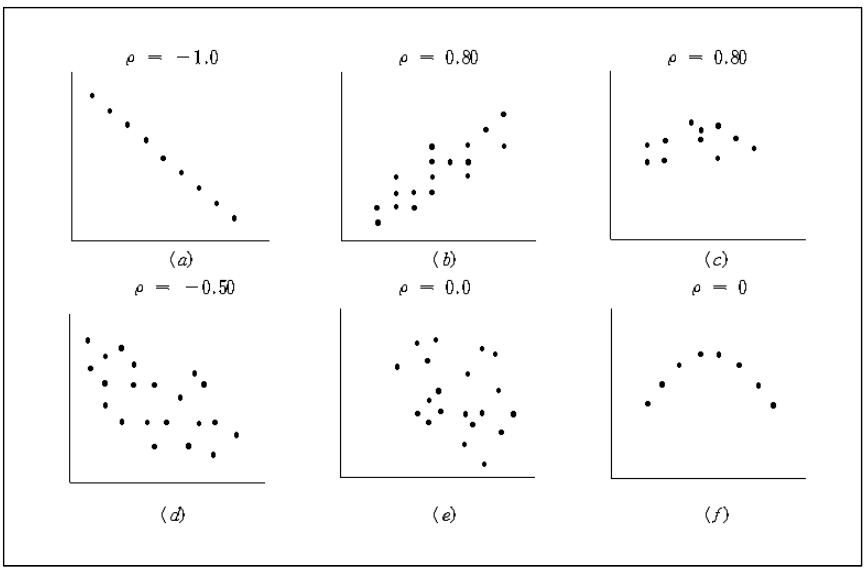
$$\begin{aligned} s_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}) \\ &= \frac{1}{n-1}(\sum_{i=1}^{n}X_iY_i-n\overline{X}\overline{Y}) \end{aligned}$$

In the above equation, $\overline{X}$ and $\overline{Y}$ represent the sample means of X and Y respectively.

- In order to understand the meaning of covariance, consider a case that Y increases if X increases. If the value of X is larger than $\overline{X}$ and the value of Y is larger than $\overline{Y}$, then $(X-\overline{X})(Y-\overline{Y})$ always has a positive value. Also, if the value of X is smaller than $\overline{X}$ and the value of Y is smaller than $\overline{Y}$, then $(X-\overline{X})(Y-\overline{Y})$ has a positive value. Therefore, their mean value which is the covariance tends to be positive. Conversely, if the value of the covariance is negative, the value of the other variable decreases as the value of one variable increases. Hence, by calculating covariance, we can see the relation between two variables: positive correlation (i.e., increasing the value of one variable will increase the value of the other) or negative correlation (i.e., decreasing the value of the other).
- Covariance itself is a good measure, but, since the covariance depends on the unit of X and Y, it makes difficult to interpret the covariance according to the size of the value and inconvenient to compare with other data. Standardized covariance which divides the covariance by the standard deviation of X and Y, σ_X and σ_Y, to obtain a measurement unrelated to the type of variable or specific unit, is called the population correlation coefficient and denoted as ρ.

Population Correlation Coefficient: $\rho = \dfrac{Cov(X,Y)}{\sigma_X\,\sigma_Y}$

- <Figure 12.1.3> shows different scatter plots and its values of the correlation coefficient.

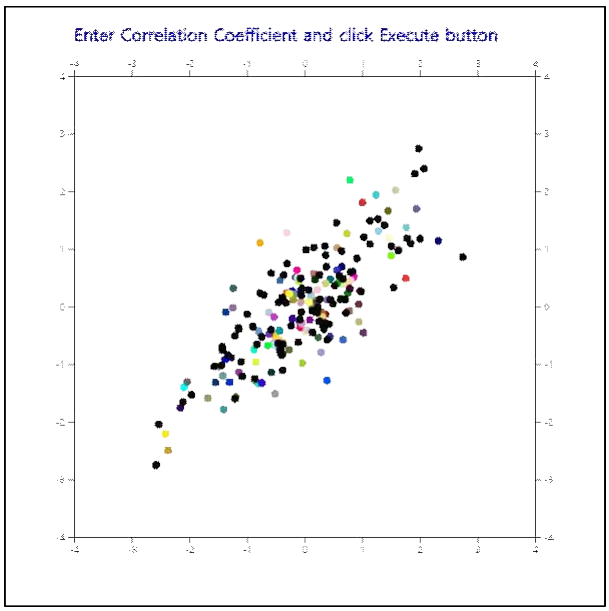


<Figure 12.1.3> Different scatter plots and their correlation coefficients.

- The correlation coefficient ρ is interpreted as follows:

  1) ρ has a value between -1 and +1. A ρ value closer to +1 indicates a strong positive linear relation and a ρ value closer to -1 indicates a strong negative linear relation. Linear relationship weakens as the value of ρ is close to 0.
  2) If all the corresponding values of X and Y are located on a straight line, the value of ρ has either +1 (if the slope of the straight line is positive) or -1 (if the slope of the straight line is negative).
  3) The correlation coefficient ρ is only a measure of linear relationship between two variables. Therefore, in the case of $\rho = 0$, there is no linear relationship between the two variables, but there may be a different relationship. (see the scatter plot (f) in <Figure 12.1.3>)

- 『eStatU』 provides a simulation of scatter plot shapes for different correlations as in <Figure 12.1.4>.



<Figure 12.1.4> Simulation of correlation coefficient at 『eStatU』

- An estimate of the population correlation coefficient using samples of two variables is called the **sample correlation coefficient** and denoted as r. The formula for the sample correlation coefficient r can be obtained by replacing each parameter with the estimates in the formula for the population correlation coefficient.

$$r = \frac{s_{XY}}{s_X \, s_Y}$$

where s_{XY} is the sample covariance and s_X, s_Y are the sample standard deviations of X and Y as follows:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

$$s_X = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$s_Y = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Therefore, the formula r can be written as follows:

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

Example 12.1.2

Find the sample covariance and correlation coefficient for the advertising costs and sales of [Example 12.1.1].

Answer

- To calculate the sample covariance and correlation coefficient, it is convenient to make the following table. This table can also be used for calculations in regression analysis.

Table 12.1.2 A table for calculating the covariance

	X	Y	X^2	Y^2	XY
1	4	39	16	1521	156
2	6	42	36	1764	252
3	6	45	36	2025	270
4	8	47	64	2209	376
5	8	50	64	2500	400
6	9	50	81	2500	450
7	9	52	81	2704	468
8	10	55	100	3025	550
9	12	57	144	3249	684
10	12	60	144	3600	720
Sum	84	497	766	25097	4326
Mean	8.4	49.7			

- Terms which are necessary to calculate the covariance and correlation coefficient are as follows:

$$SXX = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = 766 - 10 \times 8.4^2 = 60.4$$

$$SYY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = 25097 - 10 \times 49.7^2 = 396.1$$

$$SXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} = 4326 - 10 \times 8.4 \times 49.7 = 151.2$$

- SXX, SYY, SXY represent the sum of squares of X, the sum of squares of Y, the sum of squares of XY. Hence, the covariance and correlation coefficient are as follows:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{151.2}{10-1} = 16.8$$

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{151.2}{\sqrt{60.4 \times 396.1}} = 0.978$$

This value of the correlation coefficient is consistent with the scatter plot which shows a strong positive correlation of the two variables.

- Sample correlation coefficient r can be used for testing hypothesis of the population correlation coefficient. The main interest in testing hypothesis of ρ is $H_0 : \rho = 0$ which tests the existence of linear correlation. This test can be done using t distribution as follows:

Testing the population correlation coefficient ρ:

Null Hypothesis: $H_0 : \rho = 0$

Test Statistic: $t_0 = \sqrt{n-2}\,\dfrac{r}{\sqrt{1-r^2}}$

t_0 follows t distribution with $(n-2)$ degrees of freedom

Rejection Region of H_0:

1) $H_1 : \rho < 0$: Reject if $t_0 < -t_{n-2;\,\alpha}$
2) $H_1 : \rho > 0$: Reject if $t_0 > t_{n-2;\,\alpha}$
3) $H_1 : \rho \neq 0$: Reject if $|t_0| > t_{n-2;\,\alpha/2}$

Example 12.1.3	In the Example 12.1.2, test the hypothesis that the population correlation coefficient between advertising cost and the sales amount is zero at the significance level of 0.05. (Since the sample correlation coefficient is 0.978 which is close to 1, this test will not be required in practice.)
Answer	◆ The value of the test statistic t is as follows: $$t = \sqrt{10-2}\,\frac{0.978}{\sqrt{1-0.978^2}} = 13.26$$ Since it is greater than $t_{8;0.025}$ = 2.306, $H_0 : \rho = 0$ should be rejected. ◆ With the selected variables of 『eStat』 as <Figure 12.1.1>, click the regression icon on the main menu, then the scatter plot with a regression line will appear. Clicking the [Correlation and Regression] button below this graph will show the output as <Figure 12.1.5> in the Log Area with the result of the regression analysis. The values of this result are slightly different from the textbook, which is the error associated with the number of digits below the decimal point. The same conclusion is obtained that the p-value for the correlation test is 0.0001, less than the significance level of 0.05 and, therefore, the null hypothesis is rejected.

Regression Analysis				
Regression	y = 28.672 + 2.503 x			
Correlation Coefficient	r = 0.978	H_0: ρ = 0 H_1: ρ ≠ 0	t value = 13.117	p value < 0.0001
Coefficient of Determination	r^2 = 0.956			
Standard Error	s = 1.483			

<Figure 12.1.5> Testing hypothesis of correlation using 『eStat』

[Practice 12.1.1]

A professor of statistics argues that a student's final test score can be predicted from his/her midterm. Ten students were randomly selected and their mid-term and final exam scores are as follows:

id	Mid-term X	Final Y
1	92	87
2	65	71
3	75	75
4	83	84
5	95	93
6	87	82
7	96	98
8	53	42
9	77	82
10	68	60

Ex ⇨ eBook ⇨ PR120101_MidtermFinal.csv.

1) Draw a scatter plot of this data with the mid-term score on X axis and final score on Y axis. What do you think is the relationship between mid-term and final scores?
2) Find the sample correlation coefficient and test the hypothesis that the population correlation coefficient is zero with the significance level of 0.05.

- If there are more than three variables in the analysis, the relationship can be viewed using the scatter plots for each combination of two variables and the sample correlation coefficients can be obtained. However, to make it easier to see the relationship between the variables, the correlations between the variables can be arranged in a matrix format which is called a correlation matrix. 『eStat』 shows the result of a correlation matrix and the significance test for those values. The result of the test shows the t value and p-value.

Example 12.1.4

Draw a scatter plot matrix and correlation coefficient matrix using four variables of the iris data saved in the following location of 『eStat』.

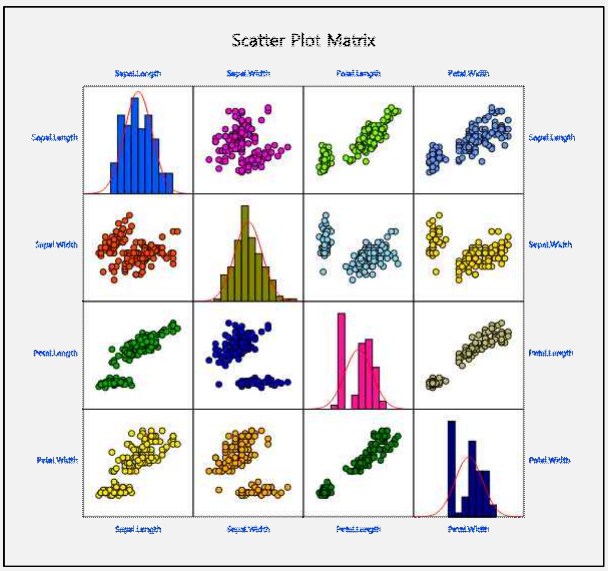
Ex ⇨ eBook ⇨ EX120104_Iris.csv

The variables are Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width. Test the hypothesis whether the correlation coefficients are equal to zero.

Answer

- From 『eStat』, load the data and click the 'Regression' icon. When the variable selection box appears, select the four variables of Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width, then the scatter plot matrix will be shown as <Figure 12.1.6>.
- It is observed that the Sepal.Length and the Petal.Length, and the Petal.Length and the Petal.Width are related.

Example 12.1.4 Answer (continued)



<Figure 12.1.6> Scatter plot matrix using 『eStat』

- When selecting [Regression Analysis] button from the options below the graph, the basic statistics and correlation coefficient matrix such as <Figure 12.1.7> appear in the Log Area with the test result. It can be seen that all correlations are significant except the correlation coefficient between the Sepal.Length and Sepal.Width.

Descriptive Statistics						
Variable	Variable Name	Observation	Mean	Std Dev	std err	95% Confidence Interval
Variable 1	Sepal.Length	150	5.843	0.828	0.068	(5.710, 5.977)
Variable 2	Sepal.Width	150	3.057	0.436	0.036	(2.987, 3.128)
Variable 3	Petal.Length	150	3.758	1.765	0.144	(3.473, 4.043)
Variable 4	Petal.Width	150	1.199	0.762	0.062	(1.076, 1.322)
Missing Observations	0					

Correlation Matrix					
Correlation Analysis H_0: ρ=0 ρ≠0 t-value p-value	Variable Name	Variable 1	Variable 2	Variable 3	Variable 4
Variable 1	Sepal.Length	1	-0.118 t-value = -1.440 p-value 0.1519	0.872 t-value = 21.646 p-value < 0.0001	0.818 t-value = 17.296 p-value < 0.0001
Variable 2	Sepal.Width	-0.118 t-value = -1.440 p-value 0.1519	1	-0.428 t-value = -5.768 p-value < 0.0001	-0.366 t-value = -4.786 p-value < 0.0001
Variable 3	Petal.Length	0.872 t-value = 21.646 p-value < 0.0001	-0.428 t-value = -5.768 p-value < 0.0001	1	0.963 t-value = 43.387 p-value < 0.0001
Variable 4	Petal.Width	0.818 t-value = 17.296 p-value < 0.0001	-0.366 t-value = -4.786 p-value < 0.0001	0.963 t-value = 43.387 p-value < 0.0001	1

<Figure 12.1.7> Descriptive statistics and correlation matrix using 『eStat』

[Practice 12.1.2]

A health scientist randomly selected 20 people to determine the effects of smoking and obesity on their physical strength and examined the average daily smoking rate (x_1, number/day), the ratio of weight by height (x_2, kg/m), and the time to exercise with a certain intensity (y, in hours). Draw a scatterplot matrix and test whether there is a correlation among smoking, obesity and exercising time with a certain intensity.

smoking rate x_1	ratio of weight by height x_2	time to exercise y
24	53	11
0	47	22
25	50	7
0	52	26
5	40	22
18	44	15
20	46	9
0	45	23
15	56	15
6	40	24
0	45	27
15	47	14
18	41	13
5	38	21
10	51	20
0	43	24
12	38	15
0	36	24
15	43	12
12	45	16

Ex ⇨ eBook ⇨ PR120102_SmokingObesityExercis.csv.

12.2 Simple Linear Regression Analysis

- **Regression analysis** is a statistical method that first establishes a reasonable mathematical model of relationships between variables, estimates the model using measured values of the variables, and then uses the estimated model to describe the relationship between the variables, or to apply it to the analysis such as forecasting. For example, a mathematical model of the relationship between sales (Y) and advertising costs (X) would not only explain the relationship between sales and advertising costs, but would also be able to predict the amount of sales that a given investment.

Definition

Regression analysis is a statistical method that first establishes a reasonable mathematical model of relationships between variables, estimates the model using measured values of the variables, and then uses the estimated model to describe the relationship between the variables, or to apply it to the analysis such as forecasting.

- As such, the regression analysis is intended to investigate and predict the degree of relation between variables and the shape of the relation. In regression analysis, a mathematical model of the relation between variables is called a regression equation, and the variable affected by other related variables is called a **dependent variable**. The dependent variable is the variable we would like to describe which is usually observed in response to other variables, so it is also called a **response variable**. In addition, variables that affect the dependent variable are called **independent variables**. The independent variable is also referred to as

the **explanatory variable**, because it is used to describe the dependent variable. In the previous example, if the objective is to analyse the change in sales amounts resulting from increases and decreases in advertising costs, the sales amount is a dependent variable and the advertising cost is an independent variable.

- If the number of independent variables included in the regression equation is one, it is called a simple linear regression. If the number of independent variables are two or more, it is called a multiple linear regression.

12.2.1 Simple Linear Regression Model

- Simple linear regression analysis has only one independent variable and the regression equation is shown as follows:

$$Y = f(X, \alpha, \beta) = \alpha + \beta X$$

In other words, the regression equation is represented by the linear equation of the independent variable, and α and β are unknown parameters which represent the intercept and slope respectively. The α and β are called the regression coefficients. The above equation represents an unknown linear relationship between Y and X in population and is therefore, referred to as the population regression equation.

- In order to estimate the regression coefficients α and β, observations of the dependent and independent variable are required, i.e., samples. In general, all of these observations are not located in a line. This is because, even if the Y and X have an exact linear relation, there may be a measurement error in the observations, or there may not be an exact linear relationship between Y and X. Therefore, the regression formula can be written by considering these errors together as follows:

$$Y_i = \alpha + \beta X_i + \epsilon_i \ , \qquad i = 1, 2, \cdots, n$$

where i is the subscript representing the ith observation, and ϵ_i is the random variable indicating an error with a mean of zero and a variance σ^2 which is independent of each other. The error ϵ_i indicates that the observation Y_i is how far away from the population regression equation. The above equation includes unknown population parameters α, β and σ^2 and is therefore, referred to as a population regression model.

- If a and b are the estimated regression coefficients using samples, the fitted regression equation can be written as follows. It is referred to as the sample regression equation.

$$\widehat{Y}_i = a + bX_i$$

In this expression, $\widehat{Y}_i$ represents the estimated value of Y at $X = X_i$ as predicted by the appropriate regression equation. These predicted values can not match the actual observed values of Y, and differences between these two values are called residuals and denoted as e_i

Residuals: $e_i = Y_i - \widehat{Y}_i \ , \qquad i = 1, 2, \cdots, n$

- The regression analysis makes some assumptions about the unobservable error ϵ_i. Since the residuals e_i calculated using the sample values have similar characteristics as ϵ_i, they are used to investigate the validity of these assumptions. (Refer to Section 12.2.6 for residual analysis.)

12.2.2 Estimation of Regression Coefficient

- When sample data, $(X_1, Y_1), \cdots, (X_n, Y_n)$, are given, a straight line representing it can be drawn in many ways. Since one of the main objectives of regression analysis is prediction, we would like to use the estimated regression line that would make the residuals smallest that the error occurs when predicting the value of Y. However, it is not possible to minimize the value of the residuals at all points, and it should be chosen to make the residuals 'totally' smaller. The most widely used of these methods is the method which minimizes the total sum of squared residuals, that is called the method of least squares regression.

Definition

Method of Least Squares Regression

A method of estimating regression coefficients so that the total sum of the squared errors occurring in each observation is minimized. i.e.,

Find α and β which minimize $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$

- To obtain the values of α and β by the least squares method, the sum of squares above should be differentiated partially with respect to α and β, and equate them zero respectively. If the solution of α and β of these equations is a and b, the equations can be written as follows:

$$a \cdot n + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$
$$a\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

- The above expression is called a **normal equation**. The solution a and b of this normal equation is called the least squares estimator of α and β and is given as follows:

Definition

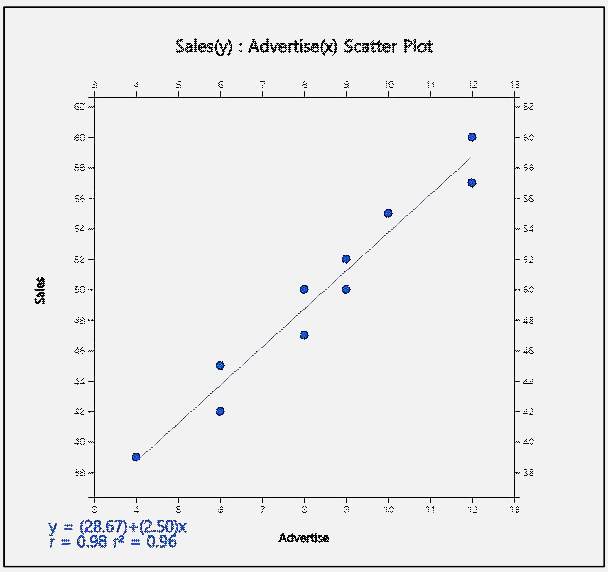
Least Squares Estimator of α and β

$$b = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$a = \overline{Y} - b\overline{X}$$

If we divide both the numerator and the denominator of b by $n-1$, b can be written as $b = s_{XY}/s_X^2$. Since the correlation coefficient is $r = \frac{s_{XY}}{s_X s_Y}$ and, therefore, $s_{XY} = r s_X s_Y$, the slope b can also be calculated by using the correlation coefficient as follows:

$$b = \frac{s_{XY}}{s_X^2} = \frac{r s_X s_Y}{s_X^2} = r\frac{s_Y}{s_X}$$

Example 12.2.1	In [Example 12.1.1], find the least squares estimate of the slope and intercept if the sales amount is a dependent variable and the advertising cost is an independent variable. Predict the amount of sales when you have spent on advertising by 10.
Answer	◆ In [Example 12.1.1], the calculation required to obtain the intercept and slope has already been made. The intercept and slope using this are as follows: $$b = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{151.2}{60.4} = 2.503$$ $$a = \overline{Y} - b\overline{X} = 49.7 - 2.503 \times 8.4 = 28.672$$ Therefore, the fitted regression line is $\widehat{Y}_i = 28.672 + 2.503X_i$ ◆ <Figure 12.2.1> shows the fitted regression line on the original data. The meaning of slope value, 2.5033, is that, if advertising cost increases by one (i.e., one million), sales increases by about 2.5 million.  <Figure 12.2.1> Simple linear regression using 『eStat』 ◆ Prediction of the sales amount of a company with an advertising cost of 10 can be obtained by using the fitted sample regression line as follows: $$28.672 + (2.503)(10) = 53.702$$ In other words, sales of 53.705 million are expected. That is not to say that all companies with advertising costs of 10 million USD have sales of 53.705 million USD, but that the average amount of their sales is about that. Therefore, there may be some differences in individual companies.

[Practice 12.2.1]	Using the data of [Practice 12.1.1] for the mid-term and final exam score, find the least squares estimate of the slope and intercept if the final exam score is a dependent variable and the mid-term score is an independent variable. Predict the final exam score when you have a mid-term score of 80.

12.2.3 Goodness of Fit for Regression Line

- After estimating the regression line, it should be investigated how valid the regression line is. Since the objective of a regression analysis is to describe a dependent variable as a function of an independent variable, it is necessary to find out how much the explanation is. A residual standard error and a coefficient of determination are used for such validation studies.
- Residual standard error s is a measure of the extent to which observations are scattered around the estimated line. First, you can define the sample variance of residuals as follows:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

The **residual standard error** s is defined as the square root of s^2. The s^2 is an estimate of σ^2 which is the extent that the observations Y are spread around the population regression line. A small value of s or s^2 indicates that the observations are close to the estimated regression line, which in turn the regression line represents well the relationship between the two variables.

- However, it is not clear how small the residual standard error s is, although the smaller value is the better. In addition, the size of the value of s depends on the unit of Y. To eliminate this shortcoming, a relative measure called the coefficient of determination is defined. The **coefficient of determination** is the ratio of the variation described by the regression line over the total variation of observation Y_i, so that it is a relative measure that can be used regardless of the type and unit of the variable.
- As in the analysis of variance in Chapter 9, the following partitions of the sum of squares and degrees of freedom are formed in the regression analysis:

Partitions of the sum of squares and degrees of freedom

Sum of squares: $SST = SSE + SSR$

Degrees of freedom: $(n-1) = (n-2) + 1$

- Description of the above three sums of squares is as follows:

Total Sum of Squares $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$:

The total sum of squares indicating the total variation in observed values of Y is called the total sum of squares (SST). This SST has the degree of freedom, $(n-1)$, and if SST is divided by the degrees of freedom, it becomes the sample variance of Y_i.

Error Sum of Squares $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$:

The error sum of squares (SSE) of the residuals represents the unexplained variation of the total variation of the Y. Since the calculation of this sum of squares requires the estimation of two parameters α and β, SSE has the degree of freedom $(n-2)$. This is the reason why, in the calculation of the sample variance of residuals s^2, it was divided by $(n-2)$.

Regression Sum of Squares $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$:

The regression sum of squares (SSR) indicates the variation explained by the

regression line among the total variation of Y. This sum of squares has the degree of freedom of 1.

- If the estimated regression equation fully explains the variation in all samples (i.e., if all observations are on the sample regression line), the unexplained variation SSE will be zero. Thus, if the portion of SSE is small among the total sum of squares SST, or if the portion of SSR is large, the estimated regression model is more suitable. Therefore, the ratio of SSR to the total variation SST, called the coefficient of determination, is defined as a measure of the suitability of the regression line as follows:

$$R^2 = \frac{Explained \quad Variation}{Total \quad Variation} = \frac{SSR}{SST}$$

The value of the coefficient of determination is always between 0 and 1 and the closer the value is to 1, the more concentrated the samples are around the regression line, which means that the estimated regression line explains the observations well.

Example 12.2.2

Calculate the value of the residual standard error and the coefficient of determination in the data on advertising costs and sales.

Answer

- To obtain the residual standard error and the coefficient of determination, it is convenient to make the following Table 12.2.1. Here, the estimated value $\widehat{Y}_i$ of the sales from each value of X_i uses the fitted regression line.

$$\widehat{Y}_i = 28.672 + 2.503X_i$$

Table 12.2.1 Useful calculations for the residual standard error and coefficient of determination

	X	Y	$\widehat{Y}_i$	SST $(Y_i - \overline{Y})^2$	SSR $(\widehat{Y}_i - \overline{Y})^2$	SSE $(Y_i - \widehat{Y}_i)^2$
1	4	39	38.639	114.49	122.346	0.130
2	6	42	43.645	59.29	36.663	2.706
3	6	45	43.645	22.09	36.663	1.836
4	8	47	48.651	7.29	1.100	2.726
5	8	50	48.651	0.09	1.100	1.820
6	9	50	51.154	0.09	2.114	1.332
7	9	52	51.154	5.29	2.114	0.716
8	10	55	53.657	28.09	15.658	1.804
9	12	57	58.663	53.29	80.335	2.766
10	12	60	58.663	106.09	80.335	1.788
Sum	84	497	496.522	396.1	378.429	17.622
Average	8.4	49.7				

- In Table 12.2.1, SST = 396.1, SSR = 378.429, SSE = 17.622. Here, the relationship of SST = SSE + SSR does not exactly match because of the error in the number of digits calculation. The sample variance of residuals is as follows:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{17.622}{(10-2)} = 2.203$$

Example 12.2.2 Answer (continued)

Hence, the residual standard error is $s = 1.484$. The coefficient of determination is as follows:

$$R^2 = \frac{SSR}{SST} = \frac{378.429}{396.1} = 0.956$$

This means that 95.6% of the total variation in the observed 10 sales amounts can be explained by the simple linear regression model using a variable of advertising costs, so this regression line is quite useful.

- Click the [Correlation and Regression] button in the option below the graph of <Figure 12.2.1> to show the coefficient of determinations and estimation errors shown in <Figure 12.2.2>.

Regression Analysis				
Regression	y = 28.672 + 2.503 x			
Correlation Coefficient	r = 0.978	H_0: ρ = 0 H_1: ρ ≠ 0	t value = 13.117	p value < 0.0001
Coefficient of Determination	r^2 = 0.956			
Standard Error	s = 1.483			
Variable	Variable Name	Observation	Mean	Std Dev
Independent Variable x	Advertise	10	8.400	2.591
Dependent Variable y	Sales	10	49.700	6.634
Missing Observations	0			

<Figure 12.2.2> Correlation and descriptive statistics

[Practice 12.2.2]

Using the data of [Practice 12.1.1] for the mid-term and final exam scores, calculate the value of the residual standard error and coefficient of determination.

12.2.4 Analysis of Variance for Regression

- If we divide three sums of squares obtained in the above example by its degree of freedom, each one becomes a kind of variance. For example, if you divide the SST by $(n-1)$ degrees of freedom, then it becomes the sample variance of the observed values $Y_1, Y_2, \cdots, Y_n$. If you divide the SSE by $(n-2)$ degrees of freedom, it becomes s^2 which is an estimate of the variance of error σ^2. For this reason, addressing the problems associated with the regression using the partition of the sum of squares is called the ANOVA of regression. Information required for ANOVA, such as calculated sum of squares and degrees of freedom, can be compiled in the ANOVA table as shown in Table 12.2.2.

Table 12.2.2 Analysis of variance table for simple linear regression

Factor	Sum of squares	Degrees of freedom	Mean squares	F value
Regression	SSR	1	MSR=SSR/1	Fo=MSR/MSE
Error	SSE	$n-2$	MSE=SSE/$(n-2)$	
Total	SST	$n-1$		

- The sum of squares divided by its degrees of freedom is referred to as mean squares, and Table 12.2.2 defines the regression mean squares (MSR) and error mean squares (MSE) respectively. As the expression indicates, MSE is the same statistic as s^2 which is the estimate of σ^2.
- The F value given in the last column are used for testing hypothesis $H_0 : \beta = 0$, $H_1 : \beta \neq 0$. If β is not 0, the F value can be expected to be large, because the assumed regression line is valid and the variation of Y is explained in large part by the regression line. Therefore, we can reversely decide that β is not zero if the calculated F ratio is large enough. If the assumptions about the error terms mentioned in the population regression model are valid and if the error terms follows a normal distribution, the distribution of F value, when the null hypothesis is true follows F distribution with 1 and $(n-2)$ degrees of freedom. Therefore, if $F_0 > F_{1,n-2;\alpha}$, then we can reject $H_0 : \beta = 0$.

F Test for simple linear regression:

Hypothesis: $H_0 : \beta = 0$, $H_1 : \beta \neq 0$

Decision rule: If $F_0 = \dfrac{MSR}{MSE} > F_{1,\ n-2;\alpha}$, then reject H_0

(In 『eStat』, the p-value for this test is calculated and the decision can be made using this p-value. That is, if the p-value is less than the significance level, the null hypothesis H_0 is rejected.)

Example 12.2.3

Prepare an ANOVA table for the example of advertising cost and test it using the 5% significance level.

Answer

- Using the sum of squares calculated in [Example 12.2.2], the ANOVA table is prepared as follows:

Factor	Sum of Squares	Degrees of freedom	Mean squares	F value
Regression	378.42	1	MSR=378.42/1=378.42	F_0=378.42/2.2=172.0
Error	17.62	10-2	MSE =17.62/8 =2.20	
Total	396.04	10-1		

- Since the calculated F value of 172.0 is much greater than $F_{1,8;0.05} = 5.32$, we reject the null hypothesis $H_0 : \beta = 0$ with the significance level $\alpha = 0.05$.
- Click the [Correlation and Regression] button in the options window below the graph <Figure 12.2.1> to show the result of the ANOVA as shown in <Figure 12.2.3>.

[ANOVA]					
Factor	Sum of Squares	deg of freedom	Mean Squares	F value	p value
Regression	378.501	1	378.501	172.052	< 0.0001
Error	17.599	8	2.200		
Total	396.100	9			

<Figure 12.2.3> Regression Analysis of Variance using 『eStat』

[Practice 12.2.3]

Using the data in [Practice 12.1.1] for the mid-term and final exam scores, prepare an ANOVA table and test it using the 5% significance level.

12.2.5 Inference for Regression

- One assumption of the error term ϵ in the population regression model is that it follows a normal distribution with the mean of zero and variance of σ^2. Under this assumption the regression coefficients and other parameters can be estimated and tested. Note that, under the assumption above, the regression model $Y = \alpha + \beta X + \epsilon$ follows a normal distribution with the mean $\alpha + \beta X$ and variance σ^2

1) Inference for the parameter β

The parameter β, which is the slope of the regression line, indicates the existence and extent of a linear relationship between the dependent and the independent variables. The inference for β can be summarized as follows. Especially, the test for hypotheses $H_0 : \beta = 0$ is used whether the independent variable describes the dependent variable significantly. The F test for the hypothesis $H_0 : \beta = 0$ described in the ANOVA of regression is theoretically the same as in the test below. 『eStat』 calculates the p- value under the null hypothesis. If this p-value is less than the significance level, the null hypothesis is rejected and the regression line is said to be significant.

Inference for the parameter β

Point estimate: $b = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$, $b \sim N(\beta, \dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2})$

Standard error of estimate b: $SE(b) = \dfrac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$

Confidence interval of β: $b \pm t_{n-2;\alpha/2} \cdot SE(b)$

Testing hypothesis:

Null hypothesis: $H_0 : \beta = \beta_0$

Test statistic:: $t = \dfrac{b - \beta_0}{SE(b)}$

H_0 rejection region: if $H_1 : \beta < \beta_0$, then $t < -t_{n-2;\alpha}$

if $H_1 : \beta > \beta_0$, then $t > t_{n-2;\alpha}$

if $H_1 : \beta \neq \beta_0$, then $|t| > t_{n-2;\alpha/2}$

2) Inference for the parameter α

The inference for the parameter α, which is the intercept of the regression line, can be summarized as below. The parameter α is not much of interest in most of the analysis, because it represents the average value of the response variable when an independent variable is 0.

Inference for the parameter α

Point estimate: $a = \overline{Y} - b\overline{X}$, $\quad a \sim N(\alpha, (\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}) \cdot \sigma^2)$

Standard error of estimate a: $SE(a) = s \cdot \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$

Confidence interval α: $a \pm t_{n-2;\alpha/2} \cdot SE(a)$

Testing hypothesis:

Null hypothesis: $H_0 : \alpha = \alpha_0$

Test Statistic:: $t = \frac{a - \alpha_o}{SE(a)}$

H_0 rejection region: if $H_1 : \alpha < \alpha_0$, then $t < -t_{n-2;\alpha}$

if $H_1 : \alpha > \alpha_0$, $t > t_{n-2;\alpha}$

if $H_1 : \alpha \neq \alpha_0$, $|t| > t_{n-2;\alpha/2}$

3) Inference for the average of Y

At any point in $X = X_0$, the dependent variable Y has an average value $\mu_{Y|x} = \alpha + \beta X_0$. Estimation of $\mu_{Y|x}$ is also considered as an important parameter, because it means predicting the mean value of Y.

Inference for the average value $\mu_{Y|x} = \alpha + \beta X_0$

Point estimate: $\hat{Y}_0 = a + bX_0$

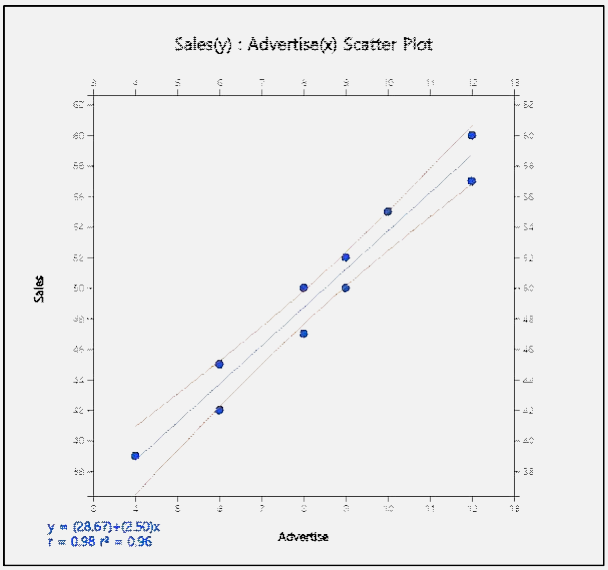
Standard error of estimate $\hat{Y}_0$: $SE(\hat{Y}_0) = s \cdot \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$

Confidence interval of $\mu_{Y|x}$: $\hat{Y}_0 \pm t_{n-2;\alpha/2} \cdot SE(\hat{Y}_0)$

The confidence interval formula of the mean value $\mu_{Y|x}$ depends on the value of the X given the standard error of the estimate, so the width of the confidence interval depends on the value of the given X. As the formula for the standard error shows, this width is the narrowest at a time $X = \overline{X}$, and if X is the farther away from $\overline{X}$, the wider it becomes. If we calculate the confidence interval for the mean value of Y at each point of X, and then if we connect the upper and lower limits to each other, we have a confidence band of the regression line on the above and below the sample regression line.

Example 12.2.4	Let's make inferences about each parameter with the result of a regression analysis of the previous data for the sales amount and advertising costs. Use 『eStat』 to check the test result and confidence band.
Answer	1) Inference for β The point estimate of β is $b = 2.5033$ and the standard error of b is as follows: $$SE(b) = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} = \frac{1.484}{\sqrt{60.4}} = 0.1908$$ ◆ Hence, the 95% confidence interval of β using $t_{8\,;\,0.025} = 3.833$ is as follows: $2.5033 \pm (3.833)(0.1908)$ 2.5033 ± 0.7313 i.e. the interval (1.7720, 3.2346). ◆ The test statistic for the hypothesis $H_0 : \beta = 0$, $H_1 : \beta \neq 0$ is as follows: $$t = \frac{2.5033 - 0}{0.1908} = 13.12$$ Since $t_{8\,;\,0.025} = 3.833$, the null hypothesis $H_0 : \beta = 0$ is rejected with the significance level of $\alpha = 0.05$. This result of two sided test can be obtained from the confidence interval. Since 95% confidence interval (1.7720, 3.2346) do not include 0, the null hypothesis $H_0 : \beta = 0$ can be rejected. 2) Inference for α The point estimate of α is $a = 29.672$ and its standard error is as follows: $$SE(a) = s \cdot \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} = 1.484 \cdot \sqrt{\frac{1}{10} + \frac{8.4^2}{60.4}} = 1.670$$ Since the value of t statistic is $29.672/1.67 = 17.1657$ and $t_{8\,;\,0.025} = 3.833$, the null hypothesis $H_0 : \alpha = 0$ is also rejected with the significance level $\alpha = 0.05$. 3) Inference for the average value of Y In 『eStat』, the standard error of $\hat{Y}$, which is the estimate of $\mu_{Y\|x}$, is calculated at each point of X. For example, the point estimate of $\hat{Y}$ at $X = 8$ is $\hat{Y} = 49.699$ and its standard error is 0.475. Hence, the 95% confidence interval of $\mu_{Y\|x}$ is as follows: $49.699 \pm (3.833)(0.475)$ 49.699 ± 1.821 i.e., the inteval is (46.878, 50.520). We can calculate the confidence interval for other value of X in a similar way as follows: At $X = 4$, $39.685 \pm (3.833)(0.962) \Rightarrow (33.998, 43.372)$ At $X = 6$, $43.692 \pm (3.833)(0.656) \Rightarrow (41.178, 46.206)$ At $X = 9$, $51.202 \pm (3.833)(0.483) \Rightarrow (49.351, 53.053)$ At $X = 12$, $59.712 \pm (3.833)(0.832) \Rightarrow (56.063, 61.361)$ As we discussed, the confidence interval becomes wider as X is far from $\overline{X}$. ◆ If you select the [Confidence Band] button from the options below the regression graph of <Figure 12.2.1>, you can see the confidence band graph on the scatter plot together with regression line as <Figure 12.2.4>. If you click the [Correlation and Regression] button, the inference result of each parameter will appear in the Log Area as shown in <Figure 12.2.5>.

Example 12.2.4 Answer (continued)



<Figure 12.2.4> Confidence band using 『eStat』

Parameter	Estimated Value	std err	t value	p value
Intercept	28.672	1.670	17.166	< 0.0001
Slope	2.503	0.191	13.117	< 0.0001

<Figure 12.2.5> Testing hypothesis of regression coefficients

[Practice 12.2.4] Using the data in [Practice 12.1.1] for the mid-term and final exam scores, make inferences about each parameter using 『eStat』 and draw the confidence band.

12.2.6 Residual Analysis

- The inference for each regression parameter in the previous section is all based on some assumptions about the error term ϵ included in the population regression model. Therefore, the satisfaction of these assumptions is an important precondition for making a valid inference. However, because the error term is unobservable, the residuals as estimate of the error term are used to investigate the validity of these assumptions which are referred to as a **residual analysis**.
- First, let's look at the assumptions in the regression model.

Assumptions in regression model

$A1$: The assumed model $Y = \alpha + \beta X + \epsilon$ is correct.

$A2$: The expectation of error terms ϵ_i is 0.

$A3$: (Homoscedasticity) The variance of ϵ_i is σ^2 which is the same for all X.

$A4$: (Independence) Error terms ϵ_i are independent.

$A5$: (Normality) Error terms ϵ_i's are normally distributed.

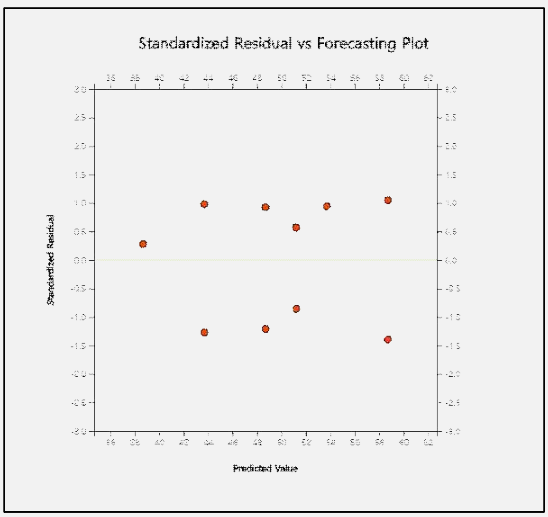
- Review the references for the meaning of these assumptions. The validity of these assumptions is generally investigated using scatter plots of the residuals. The

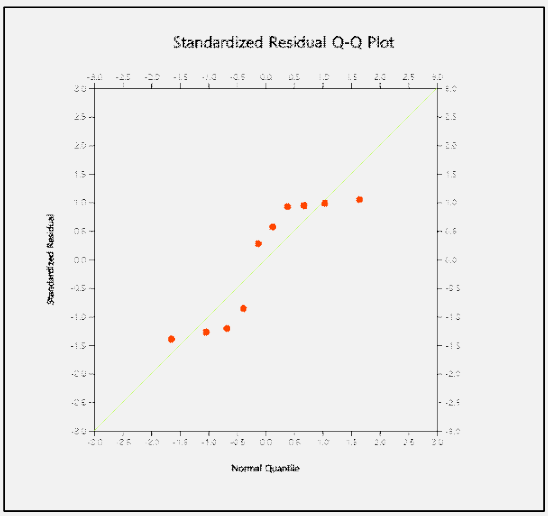
following scatter plots used primarily for each assumption:

1) Residuals versus predicted values (i.e., e_i vs $\widehat{Y}_i$) : $A3$
2) Residuals versus independent variables (i.e., e_i vs X_i) : $A1$
3) Residuals versus observations (i.e., e_i vs i) : $A2$, $A4$

In the above scatter plots, if the residuals show no particular trend around zero, and appear randomly, then each assumption is valid.

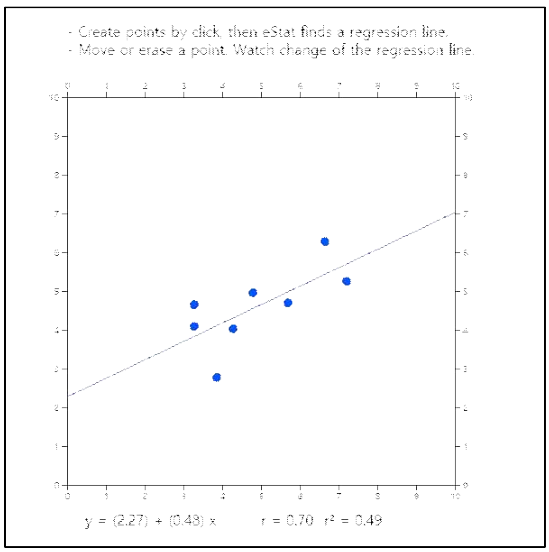
- The assumption that the error term ϵ follows a normal distribution can be investigated by drawing a histogram of the residuals in case of a large amount of data to see if the distribution is similar to the shape of the normal distribution. Another method is to use the quantile-quantile (Q-Q) scatter plot of the residuals. In general, if the Q-Q scatter plot of the residuals forms a straight line, it can be considered as a normal distribution.
- Since residuals are also dependent on the unit of the dependent variable, standardized values of the residuals are used for consistent analysis of the residuals, which are called standardized residuals. Both the scatter plots of the residuals described above and the Q-Q scatter plot are created using the standardized residuals. In particular, if the value of the standardized residuals is outside the ± 2, an anomaly value or an outlier value can be suspected.

Example 12.2.5	Draw a scatter plot of residuals and a Q-Q scatter plot for the advertising cost example.
Answer	◆ When you click the [Residual Plot] button from the options below the regression graph of <Figure 12.2.1>, the scatter plot of the standardized residuals and predicted values are appeared as shown in <Figure 12.2.6>. If you click [Residual Q-Q Plot] button, <Figure 12.2.7> is appeared. Although the scatter plot of the residuals has no significant pattern, the Q-Q plot deviates much from the straight line and so, the normality of the error term is somewhat questionable. In such cases, the values of the response variable need to be re-analyzed by taking logarithmic or square root transformation.  <Figure 12.2.6> Residual plot

Example 12.2.5 Answer (continued)	 <Figure 12.2.7> Residual Q-Q Plot

[Practice 12.2.4]	Using the data in [Practice 12.1.1] for the mid-term and final exam scores, draw a scatter plot of the residuals and a Q-Q scatter plot.

- In 『eStatU』, it is possible to do experiments on how much a regression line is affected by an extreme point (<Figure 12.2.8>). A point can be created by clicking the mouse on the screen in the link below. If you create multiple dots, you can see how much the regression line changes each time. You can observe how sensitive the correlation coefficient and the coefficient of determination are as you move a point along with the mouse.



<Figure 12.2.8> Simulation experiment of regression analysis at 『eStatU』

12.3 Multiple Linear Regression Analysis

- For actual applications of the regression analysis, the multiple regression models with two or more independent variables are more frequently used than the simple linear regression with one independent variable. This is because it is rare for a dependent variable to be sufficiently explained by a single independent variable, and in most cases, a dependent variable has a relationship with several independent variables. For example, it may be expected that sales will be significantly affected by advertising costs, which are examples of simple linear regression, but will also be affected by product quality ratings, the number and size of stores sold. The statistical model used to identify the relationship between one dependent variable and several independent variables is called a multiple linear regression analysis. However, the simple linear regression and multiple linear regression analysis differ only in the number of independent variables involved, and there is no difference in the method of analysis.

12.3.1 Multiple Linear Regression Model

- In the multiple linear regression model, it is assumed that the dependent variable Y and k number of independent variables have the following relational formulas:

$$Y_i = \beta_o + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$$

This means that the dependent variable is represented by the linear function of independent variables and a random variable that represents the error term as in the simple linear regression model. The assumption of the error terms is the same as the assumption in the simple linear regression. In the above equation, β_0 is the intercept of Y axis and β_i is the slope of the Y axis and X_i which indicates the effect of X_i to Y when other independent variables are fixed.

Example 12.3.1

When logging trees in forest areas, it is necessary to investigate the amount of timber in those areas. Since it is difficult to measure the volume of a tree directly, we can think of ways to estimate the volume using the diameter and height of a tree that is relatively easy to measure. The data in Table 12.3.1 are the values for measuring diameter, height and volume after sampling of 15 trees in a region. (The diameter was measured at a point 1.5 meters above the ground.) Draw a scatter plot matrix of this data and consider a regression model for this problem.

Table 12.3.1 Diameter, height and volume of tree

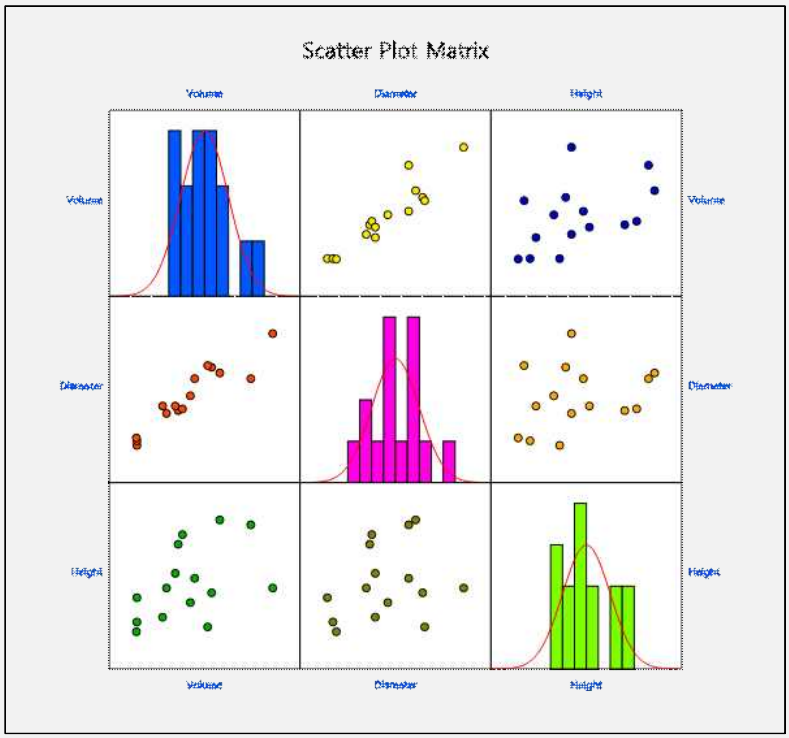
Diameter(cm)	Height(m)	Volume(m^3)
21.0	21.33	0.291
21.8	19.81	0.291
22.3	19.20	0.288
26.6	21.94	0.464
27.1	24.68	0.532
27.4	25.29	0.557
27.9	20.11	0.441
27.9	22.86	0.515
29.7	21.03	0.603
32.7	22.55	0.628
32.7	25.90	0.956
33.7	26.21	0.775
34.7	21.64	0.727
35.0	19.50	0.704
40.6	21.94	1.084

Ex ⇨ eBook ⇨ EX120301_TreeVolume.csv.

Example 12.3.1 Answer

- Load the data saved at the following location of 『eStat』.
  Ex ⇨ eBook ⇨ EX120301_TreeVolume.csv

- In the variable selection box which appears by selecting the regression icon, select 'Y variable' by volume and select 'by X variable' as the diameter and height to display a scatter plot matrix as shown in <Figure 12.3.1>. It can be observed that there is a high correlation between volume and diameter, and that volume and height, and diameter and height are also somewhat related.



<Figure 12.3.1> Scatterplot matrix

Correlation Matrix				
Correlation Analysis H_0: ρ=0 ρ≠0 t-value p-value	Variable Name	Variable 1	Variable 2	Variable 3
Variable 1	Volume	1	0.934 t-value = 9.456 p-value < 0.0001	0.464 t-value = 1.889 p-value 0.0814
Variable 2	Diameter	0.934 t-value = 9.456 p-value < 0.0001	1	0.263 t-value = 0.984 p-value 0.3431
Variable 3	Height	0.464 t-value = 1.889 p-value 0.0814	0.263 t-value = 0.984 p-value 0.3431	1

<Figure 12.3.2> Correlation matrix

Example 12.3.1 Answer (continued)

- Since the volume is to be estimated using the diameter and height of the tree, the volume is the dependent variable Y, and the diameter and height are independent variables X_1, X_2 respectively, and the following regression model can be considered.

$$Y_i = \beta_o + \beta_i X_{i1} + \beta_2 X_{i2} + \epsilon_i\,, \quad i = 1, 2, \cdots, 15$$

[Practice 12.3.1]

A health scientist randomly selected 20 people to determine the effect of smoking and obesity on their physical strength and examined the average daily smoking rate (x_1, number/day), the ratio of weight by height (x_2, kg/m), and the time to continue to exercise with a certain intensity (y, in hours). Draw a scatter plot matrix of this data and consider a regression model for this problem.

smoking rate x_1	ratio of weight by height x_2	time to continue to exercise y
24	53	11
0	47	22
25	50	7
0	52	26
5	40	22
18	44	15
20	46	9
0	45	23
15	56	15
6	40	24
0	45	27
15	47	14
18	41	13
5	38	21
10	51	20
0	43	24
12	38	15
0	36	24
15	43	12
12	45	16

Ex ⇨ eBook ⇨ PR120301_SmokingObesityExercis.csv.

- In general, matrix and vectors are used to facilitate expression of formula and calculation of expressions. For example, if there are k number of independent variables, the population multiple regression model at the observation point $i = 1, 2, \cdots, n$ is presented in a simple manner as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

Here $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, ϵ are defined as follows:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ & & \cdots & & \\ & & \cdots & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}$$

12.3.2 Estimation of Regression Coefficient

- In a multiple regression analysis, it is necessary to estimate the $(k+1)$ number of regression coefficients $\beta_0, \beta_1, \cdots, \beta_k$ using samples. In this case, the least squares method which minimizes the sum of the squared errors is also used. We find $\boldsymbol{\beta}$ which minimizes the following sum of the error squares as follows:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

As in the simple linear regression, the above error sum of squares is differentiated with respect to $\boldsymbol{\beta}$ and then, equate to zero which is called a normal equation. The solution of the equation, denoted as $\boldsymbol{b}$ which is called the least squares estimate of $\boldsymbol{\beta}$, should satisfy the following normal equation.

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$$

Therefore, if there exists an inverse matrix of $\boldsymbol{X}'\boldsymbol{X}$, the least squares estimator of $\boldsymbol{\beta}$, $\boldsymbol{b}$, is as follows:

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

(Note: Statistical packages uses a different formula, because the above formula causes large amount of computing error)

- If the estimated regression coefficients are $\boldsymbol{b} = (b_0, b_1, \cdots, b_k)'$, the estimate of the response variable Y is as follows:

$$\widehat{Y}_i = b_0 + b_1 X_{i1} + \cdots + b_k X_{ik}$$

The residuals are as follows:

$$\begin{aligned} e_i &= Y_i - \widehat{Y}_i \\ &= Y_i - (b_0 + b_1 X_{i1} + \cdots + b_k X_{ik}) \end{aligned}$$

By using a vector notation, the residual vector $\boldsymbol{e}$ can be defined as follows:

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}$$

12.3.3 Goodness of Fit for Regression and Analysis of Variance

- In order to investigate the validity of the estimated regression line in the multiple regression analysis, the standardized residual error and coefficient of determination are also used. In the simple linear regression analysis, the computational formula for these measures was given as a function of the residuals, i.e., observed value of Y and its predicted value, so there is nothing to do with the number of independent variables. Therefore, the same formula can be used in the multiple linear regression and there is only a difference in the value of the degrees of freedom that each sum of squares has.
- In the multiple linear regression analysis, the standard error of residuals is defined as follows:

$$s = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}$$

- The difference from the simple linear regression is that the degrees of freedom for residuals is $(n-k-1)$, because the k number of regression coefficients must be estimated in order to calculate residuals. As in simple linear regression, s^2 is a statistic such as the residual mean squares(MSE). The coefficient of determination is given in $R^2 = SSR/SST$ and its interpretation is as shown in the simple linear regression.
- The sum of squares is defined by the same formula as in the simple linear regression, and can be divided with corresponding degrees of freedom as follows and the table of the analysis of variance is shown in Table 12.3.2.

Sum of squares: $SST = SSE + SSR$
Degrees of freedom: $n-1 = (n-k-1) + k$

Table 12.3.2 Analysis of variance table for multiple linear regression analysis

Source	Sum of Squares	Degrees of Freedom	Mean Squares	F value
Regression	SSR	k	MSR=SSR/k	F_0=MSR/MSE
Error	SSE	$n-k-1$	MSE=SSE/$(n-k-1)$	
Total	SST	$n-1$		

- The F value in the above ANOVA table is used to test the significance of the regression equation, where the null hypothesis is that all independent variables are not linearly related to the dependent variables.

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
H_1 : At least one of k number of β_is is not equal to 0

- Since F_0 follows F distribution with k and $(n-k-1)$ degrees of freedom under the null hypothesis, we can reject H_0 at the significance level α if $F_0 > F_{k,n-k-1;\alpha}$. Each β_i can also be tested which is described in the following sections. (Also, 『eStat』 calculates the p-value for this test, so use this p-value to test. That is, if the p-value is less than the significance level, the null hypothesis is rejected.)

12.3.4 Inference for Multiple Linear Regression

- Parameters that are of interest in multiple linear regression, as in the simple linear regression, are the expected value of Y and each regression coefficients $\beta_0, \beta_1, \cdots, \beta_k$. The inference of these parameters $\beta_0, \beta_1, \cdots, \beta_k$ is made possible by obtaining a probability distribution of the point estimates b_i. Under the assumption that the error terms ϵ_i are independent and all have a distribution of $N(0,\sigma^2)$, it can be shown that the distribution of b_i is as follows:

$$b_i \sim \mathrm{N}(\beta_i,\ c_{ii} \cdot \sigma^2), \qquad i = 0, 1, \cdots, k$$

The above c_{ii} is the ith diagonal element of the $(k+1)\times(k+1)$ matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. In addition, using an estimate s^2 instead of a parameter σ^2, you can make inferences about each regression coefficient using the t distribution.

Inference on regression coefficient β_i

Point estimate: b_i

Standard error of point estimate: $SE(b_i) = \sqrt{c_{ii}} \cdot s$

Confidence interval of β_i : $b_i \pm t_{n-k-1\,;\,\alpha/2} \cdot SE(b_i)$

Testing hypothesis:

Null hypothesis: $H_0 : \beta_i = \beta_{i0}$

Test Statistic:: $t = \dfrac{b_i - \beta_{i0}}{SE(b_i)}$

H_0 rejection region:

if $H_1 : \beta_i < \beta_{i0},\ t < -t_{n-k-1\,;\,\alpha}$

if $H_1 : \beta_i > \beta_{i0},\ t > t_{n-k-1\,;\,\alpha}$

if $H_1 : \beta_i \neq \beta_{i0},\ |t| > t_{n-k-1\,;\,\alpha/2}$

(Since 『eStat』 calculates the p-value under the null hypothesis $H_0 : \beta_i = 0$, p-value is used for testing hypothesis.)

- Residual analysis of the multiple linear regression is the same as in the simple linear regression.

Example 12.3.2	For the tree data of [Example 12.3.1], obtain the least squares estimate of each coefficient of the proposed regression equation using 『eStat』 and apply the analysis of variance, test for goodness of fit and test for regression coefficients.
Answer	◆ In the options window below the scatter plot matrix in <Figure 12.3.1>, click [Regression Analysis] button. Then you can find the estimated regression line, ANOVA table as shown in <Figure 12.3.3> in the Log Area. The estimated regression equation is as follows: $\widehat{Y}_i = -1.024 + 0.037X_1 + 0.024X_2$ In the above equation, 0.037 represents the increase of the volume of the tree when the diameter (X_1) increases 1(cm). ◆ The p-value calculated from the ANOVA table in <Figure 12.3.3> at F value of 73.12 is less than 0.0001, so you can reject the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ at the significance level $\alpha = 0.05$. The coefficient of determination, $R^2 = 0.924$, implies that 92.4% of the total variances of the dependent variable are explained by the regression line. Based on the above two results, we can conclude that the diameter and height of the tree are quite useful in estimating the volume.

Example 12.3.2 Answer (continued)

Regression Analysis					
Regression y =	(-1.024) + (0.037) X_1 + (0.024) X_2				
Multiple Correlation Coeff	0.961	Coefficient of Determination	0.924	Standard Error	0.069

Parameter	Estimated Value	std err	t value	p value	95% Confidence Interval
β_0	-1.024	0.188	-5.458	0.0001	(-1.358 ,-0.689)
β_1 Diameter	0.037	0.003	10.590	< 0.0001	(0.031 ,0.043)
β_2 Height	0.024	0.008	2.844	0.0148	(0.009 ,0.038)

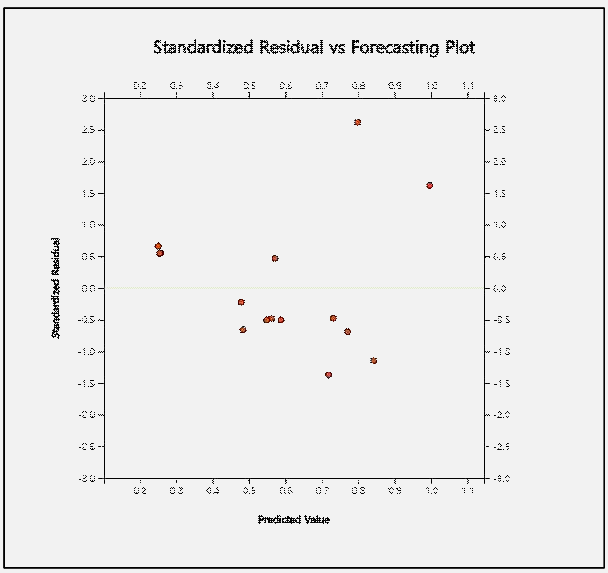
[ANOVA]					
Factor	Sum of Squares	deg of freedom	Mean Squares	F value	p value
Regression	0.7058	2	0.3529	73.1191	< 0.0001
Error	0.0579	12	0.0048		
Total	0.7638	14			

<Figure 12.3.3> Result of Multiple Linear Regression

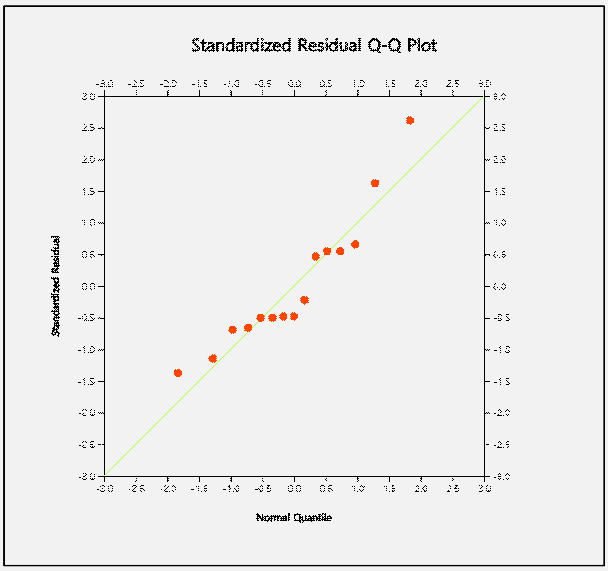
- Since $SE(b_1) = 0.003$, $SE(b_2) = 0.008$ and $t_{12\,;\,0.025} = 2.179$ from the result in <Figure 12.3.3>, the 95% confidence intervals for each regression coefficients can be calculated as follows. The difference between this result and the Figure 12.3.3 due to the error in the calculation below the decimal point.

  95% confidence interval for β_1: $0.037 \pm (2.179)(0.003) \;=>\; (0.029,\ 0.045)$
  95% confidence interval for β_2: $0.024 \pm (2.179)(0.008) \;=>\; (0.006,\ 0.042)$

- In the hypothesis test of $H_0 : \beta_i = 0,\ H_1 : \beta_i \neq 0,\ i = 1,2$, each p-value is less than the significance level of 0.05, so you can reject each null hypothesis.
- The scatter plot of the standardized residuals is shown in <Figure 12.3.4> and the Q-Q scatter plot is shown in <Figure 12.3.5>. There is no particular pattern in the scatter plot of the standardized residuals, but there is one outlier value, and the Q-Q scatter plot shows that the assumption of normality is somewhat satisfactory.



<Figure 12.3.4> Residual analysis of multiple linear regression

Example 12.3.2 Answer (continued)	 <Figure 12.3.5> Q-Q plot of multiple linear regression

[Practice 12.3.2]	Apply a multiple regression model by using『eStat』on the regression model of [Practice 12.3.1]. Obtain the least squares estimate of each coefficient of the proposed regression equation and apply the analysis of variance, test for goodness of fit and test for regression coefficients.

Exercise

12.1 A survey was conducted on the level of education(X, the period after graduating a high school, unit: year) for 10 businessmen and annual income (Y, unit: one thousand USD) after graduating from the high school.

id	Education Period (X)	Annual Income (Y)
1	4	50
2	2	37
3	0	35
4	3	45
5	4	57
6	4	49
7	5	60
8	5	47
9	2	39
10	2	50

1) Draw a scatter plot of data and interpret.
2) Calculate the sample correlation coefficient.
3) Apply the regression analysis with annual income as the dependent variable and the level of education as the independent variable.

12.2 The following data shows studying time for a week (X) and the grade (Y) of six students.

Studying time X (unit: hour)	Grade Y
15	2.0
28	2.7
13	1.3
20	1.9
4	0.9
10	1.7

1) Find a regression line and 95% confidence interval for β (it is a further grade score that is expected to be raised when a student studies one more hour a week.)
2) Calculate a 99% confidence interval in the average score of a student who studies an average of 12 hours a week.
3) Test for hypothesis $H_0 : \beta = 0.10, \quad H_1 : \beta < 0.10$ (significance level = 0.01).

12.3 A professor of statistics argues that a student's final test score can be predicted from his/her midterm. Five students were randomly selected and their mid-term and final exam scores are as follows:

id	Mid-term X	Final Y
1	92	87
2	65	71
3	75	75
4	83	84
5	95	93

1) Draw a scatter plot of this data with mid-term score on X axis and final score on Y axis. What do you think is the relationship between mid-term and final scores?

2) Find the regression line and analyse the result.

12.4 An economist argues that there is a clear relationship between coffee and sugar prices. 'When people buy coffee, they will also buy sugar. Isn't it natural that the higher the demand, the higher the price?' We collected the following sample data to test his theory.

Year	Coffee Price	Sugar Price
1985	0.68	0.245
1986	1.21	0.126
1987	1.92	0.092
1988	1.81	0.086
1989	1.55	0.101
1990	1.87	0.223
1991	1.56	0.212

1) Prepare a scatter plot with the coffee price on X axis and sugar price on Y axis. Is this data true to this economist's theory?
2) Test this economist's theory by using a regression analysis.

12.5 A rope manufacturer thinks that the strength of the rope is proportional to the nylon content of the rope. Ten ropes are randomly selected and their data are as follows:

% Nylon X	Strength (psi) Y
0	260
10	360
20	490
20	510
30	600
30	600
40	680
50	820
60	910
70	990

1) Draw a scatter plot with the % Nylon on X axis and strength on Y axis. Find a regression line using the least squares method. Draw this estimated regression line on the scatter plot.
2) Estimate the strength of a rope in case of 33% nylon.
3) Estimate the strength of a rope in case of 66% nylon.
4) The strength of two ropes in case of 20% nylon on the data are different. How can you explain this variation in a regression model?
5) Estimate the strength of a rope in case of 0% nylon. Why is this estimate different from the observed value of 260?
6) Obtain a 95% confidence interval for the strength of the 0% nylon rope.
7) If the observed strength of the 0% nylon rope was outside the confidence interval in 6), how would you interpret this result?

12.6 A health scientist randomly selected 20 people to determine the effects of smoking and obesity on their physical strength and examined the average daily smoking rate (x_1, number/day), the ratio of weight by height (x_2, kg/m), and the time to continue to exercise with a certain intensity (y, in hours). Test whether smoking and obesity can affect your exercising time with a certain intensity. Apply a multiple regression model by using 『eStat』.

smoking rate x_1	ratio of weight by height x_2	time to continue to exercise y
24	53	11
0	47	22
25	50	7
0	52	26
5	40	22
18	44	15
20	46	9
0	45	23
15	56	15
6	40	24
0	45	27
15	47	14
18	41	13
5	38	21
10	51	20
0	43	24
12	38	15
0	36	24
15	43	12
12	45	16

12.7 The price of old watches in an antique auction is said to be determined by the year of making the watch and the number of bidders. In order to see if this is true, the 32 recently auctioned alarm clocks were examined for the elapsed period (in years) after manufacture, the number of bidders and the auction price (in 1,000USD) as follows. Test the hypothesis that the auction price of the alarm clock increases with the increase in the number of bidders using the multiple linear regression model. (significance level: 0.05)

Elapsed Period x_1	Number of bidders x_2	Auction Price y
127	13	1235
115	12	1080
127	7	845
150	9	1522
156	6	1047
182	11	1979
156	12	1822
132	10	1253
137	9	1297
113	9	946
137	15	1713
117	11	1024
137	8	1147
153	6	1092
117	13	1152
126	10	1336
170	14	2131
182	8	1550
162	11	1884
184	10	2041
143	6	854
159	9	1483
108	14	1055
175	8	1545
108	6	729
179	9	1792
111	15	1175
187	8	1593
111	7	785
115	7	744
194	5	1356
168	7	1262

Multiple Choice Exercise

12.1 The variables X and Y have a strong relationship with a quadratic equation ($y = x^2$) as shown in the following table. What is their sample correlation coefficient?

X	⋯	-3	-2	-1	0	1	2	3	⋯
Y	⋯	9	4	1	0	1	4	9	⋯

① 1 ② 0
③ -1 ④ $\frac{1}{2}$

12.2 Which is a wrong description of the correlation coefficient?

① $-1 < r < 1$ ② if $r = -1$, perfect negative correlation
③ if $r = 0$, no linear correlation ④ if r < 0, negative correlation

12.3 Which is a right description of the correlation coefficient?

① if $r > 1$, there is strong positive correlation between x and y.
② if $|r|$ closes to 0, there exist a weak linear correlation between x and y.
③ If r is negative, then y is increasing when x increases.
④ If r is near -1, there exist a weak linear correlation between x and y.

12.4 If the sample correlation coefficient between x_i and y_i $(i = 1, 2, \cdots, n)$ is r, what is the sample correlation coefficient between $10x_i + 2$ and $5y_i + 3$?

① r ② $2r$
③ $5r$ + 3 ④ $10r$ + 2

12.5 If the sample correlation coefficient between x and y is r, what is the sample correlation coefficient between $2x$ and $3y+1$?

① r ② $2r$
③ $3r$ ④ $3r$ + 1

12.6 When not all points on a scatter plot tend to be linear, what is the sample correlation coefficient r close to:

① $r \geq 1$ ② $r \leq -1$
③ $|r|$ is close to 1 ④ $|r|$ is close to 0

12.7 Find the sample correlation coefficient between x and y of the following data.

x	10	20	30	40
y	2	4	6	8

① 1 ② 0.3 ③ 0.4 ④ 0.5

12.8 If the correlation coefficient of two variables x, y is 0, what is the right description?

① There is no linear relationship between two variables x, y.
② There is a linear relationship between two variables x, y.
③ Two variables x, y has a strong relationship.
④ Two variables x, y has a strong linear relationship.

12.9 Which one of the following descriptions on the sample correlation coefficient r is not right?

① r is a random variable.
② $-1 \le r \le 1$
③ r is a measure of linear relationship between two variables.
④ Distribution of r is a normal distribution.

12.10 Find the sample correlation coefficient between x and y of the following data.

x	1	2	3	4	5
y	5	4	3	2	1

① -1 ② $-\frac{1}{2}$
③ 0 ④ $\frac{1}{2}$

12.11 Find the sample correlation coefficient r between x and y of the following data?

x	1	2	3	4	5	6
y	-1	1	3	5	7	9

① -0.5 ② 0 ③ 0.5 ④ 1

12.12 If X and Y are independent, what is the sample correlation coefficient r?

① 1 ② $\frac{1}{2}$
③ 0 ④ $-\frac{1}{2}$

12.13 Which one of the followings is right for description of the sample correlation coefficient r?

① $0 \le r \le 1$ ② $-1 \le r \le 0$
③ $-1 \le r \le 1$ ④ $-\infty < r < \infty$

12.14 Which one of the followings is right for description of the sample correlation coefficient r between X and Y?

① if r = -1, the value of X is directly proportional to the value of Y.
② if r = 1, the value of X is directly proportional to the value of Y.
③ if r = 0, the value of X is inversely proportional to the value of Y.
④ if r = -1, the value of X is not related with the value of Y.

12.15 Which one of the followings is not right for description of the sample correlation coefficient r between X and Y?

① $-1 \le r \le 1$
② Distribution of r is a normal distribution.
③ r is a random variable.
④ The formula to calculate r is $r = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$.

12.16 If two variables X and Y have a strong quadratic relation, what is the sample correlation coefficient r?

① $r \fallingdotseq 1$ ② $r \fallingdotseq -1$
③ $r \fallingdotseq 0$ ④ know information on r.

12.17 Which one of the followings has positive correlation?

① height of mountain and pressure
② weight and height
③ monthly income and Engel' coefficient
④ amount of production and price

12.18 If all points lie on a straight line in a scatter plot, what is the characteristic of the correlation coefficient?

① perfect correlation ② strong correlation
③ weak correlation ④ no correlation

12.19 If the sample correlation coefficient is $r = -1$, what is the characteristic of the correlation coefficient?

① inverse correlation ② positive correlation
③ weak correlation ④ usual correlation

12.20 Find the sample correlation coefficient between x and y of the following data.

x	1	2	3	4
y	1	4	3	6

① 0.29 ② 0.53 ③ 0.87 ④ 0.98

12.21 Find the sample covariance between x and y of the following data.

x	y
1	5
2	5
3	5
4	5

① 1
② 0
③ 0.5
④ -1

12.22 Find the sample covariance between x and y of the following data.

x	1	2	3	4	5
y	17	15	13	11	9

① 0 ② 1
③ $-\frac{1}{2}$ ④ -1

12.23 Find the sample covariance between x and y of the following data.

x	1	2	3	4	5
y	6	8	10	12	14

① 3 ② 4
③ 10 ④ 20

12.24 Find the regression line between x and y using the following data.

x	1	2	3	4	5
y	1	4	7	10	13

① $y - 7 = 3(x - 3)$ ② $y - 7 = 2(x - 3)$
③ $y - 3 = 3(x - 7)$ ④ $y - 3 = 2(x - 7)$

12.25 If the standard deviations of the X and Y variables are 4.06 and 2.65 respectively, the covariance is 10.50, what is the sample correlation coefficient r?

① 10.759 ② 0.532
③ 1.025 ④ 0.976

12.26 If we know the sample correlation coefficient r and the standard deviations of X and Y, s_x and s_y respectively, what is the regression line equation?

① $y = \bar{y} + \frac{s_y}{s_x} r(x - \bar{x})$ ② $y = \bar{x} + \frac{s_y}{s_x} r(y - \bar{y})$
③ $y = \bar{y} + \frac{s_x}{s_y} r(x - \bar{x})$ ④ $y = \bar{x} + \frac{s_x}{s_y} r(y - \bar{y})$

12.27 If the sample correlation coefficient of two random variables x and y is $r = 1/2$, the sample means are $\bar{x} = 10$, $\bar{y} = 14$, and the sample standard deviations are $s_x = 2$, $s_y = 3$, what is the regression line of y on x?

① $y = \frac{3}{4}x + \frac{13}{2}$ ② $y = \frac{3}{4}x - \frac{13}{4}$
③ $y = \frac{3}{x} - 1$ ④ $y = \frac{3}{4}x + 1$

12.28 Find the regression coefficient b of the regression line $Y = a + bX$ using the following data.

	sample mean	sample standard deviation	correlation coefficient
X	40	4	0.75
Y	30	3	

① 0.56
② 0.07
③ 1.00
④ 1.53

12.29 Which one of the following statements is true about the regression line of two variables X and Y, the regression line of Y on X and the regression line of X on Y?

① The two regression lines are always consistent.
② The two regression lines are always parallel.
③ The two regression lines meet at one point $(\overline{X}, \overline{Y})$ and do not match.
④ The two regression lines are always perpendicular.

12.30 Find the regression coefficient b of the regression line $Y = a + bX$ using the following data.

	sample mean	sample standard deviation	correlation coefficient
x	12	3	r = 0.6
y	13	4	

① 0.6 ② 0.7 ③ 0.8 ④ 0.9

12.31 Which one is a wrong explanation about the regression coefficient b and the sample correlation coefficient r?

① If $b = 0,\ \ r = 0$ (no correlation)
② If $b > 0,\ \ r > 0$ (positive correlation)
③ If $b = 1,\ \ r = 1$ (perfect correlation)
④ If $b < 0,\ \ r < 0$ (negative correlation)

12.32 If a regression line is $Y = 4 + 0.4X$ and the sample standard deviations of X and Y are 4, 2 respectively, what is the value of the sample correlation coefficient r?

① 1 ② 0.8 ③ 0.5 ④ 0.4

(Answers)
12.1 ②, 12.2 ①, 12.3 ②, 12.4 ①, 12.5 ①, 12.6 ④, 12.7 ①, 12.8 ①, 12.9 ④, 12.10 ①, 12.11 ④, 12.12 ③, 12.13 ③, 12.14 ②, 12.15 ②, 12.16 ③, 12.17 ②, 12.18 ①, 12.19 ①, 12.20 ③, 12.21 ②, 12.22 ④, 12.23 ②, 12.24 ①, 12.25 ④, 12.26 ①, 12.27 ①, 12.28 ①, 12.29 ③, 12.30 ③), 12.31 ③, 12.32 ②