Introduction to Statistics and Data Science using *eStat*

Chapter 12 Correlation and Regression Analysis

12.1 Correlation Analysis

Jung Jin Lee
Professor of Soongsil University, Korea
Visiting Professor of ADA University, Azerbaijan

12.1 Correlation Analysis

12.2 Simple Linear Regression Analysis

12.3 Multiple Linear Regression Analysis

12.1 Correlation Analysis

[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.

- Using 『eStat』, draw a scatter plot for this data and investigate the relation of the two variables.

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

12.1 Correlation Analysis

<Answer of Example 12.1.1>

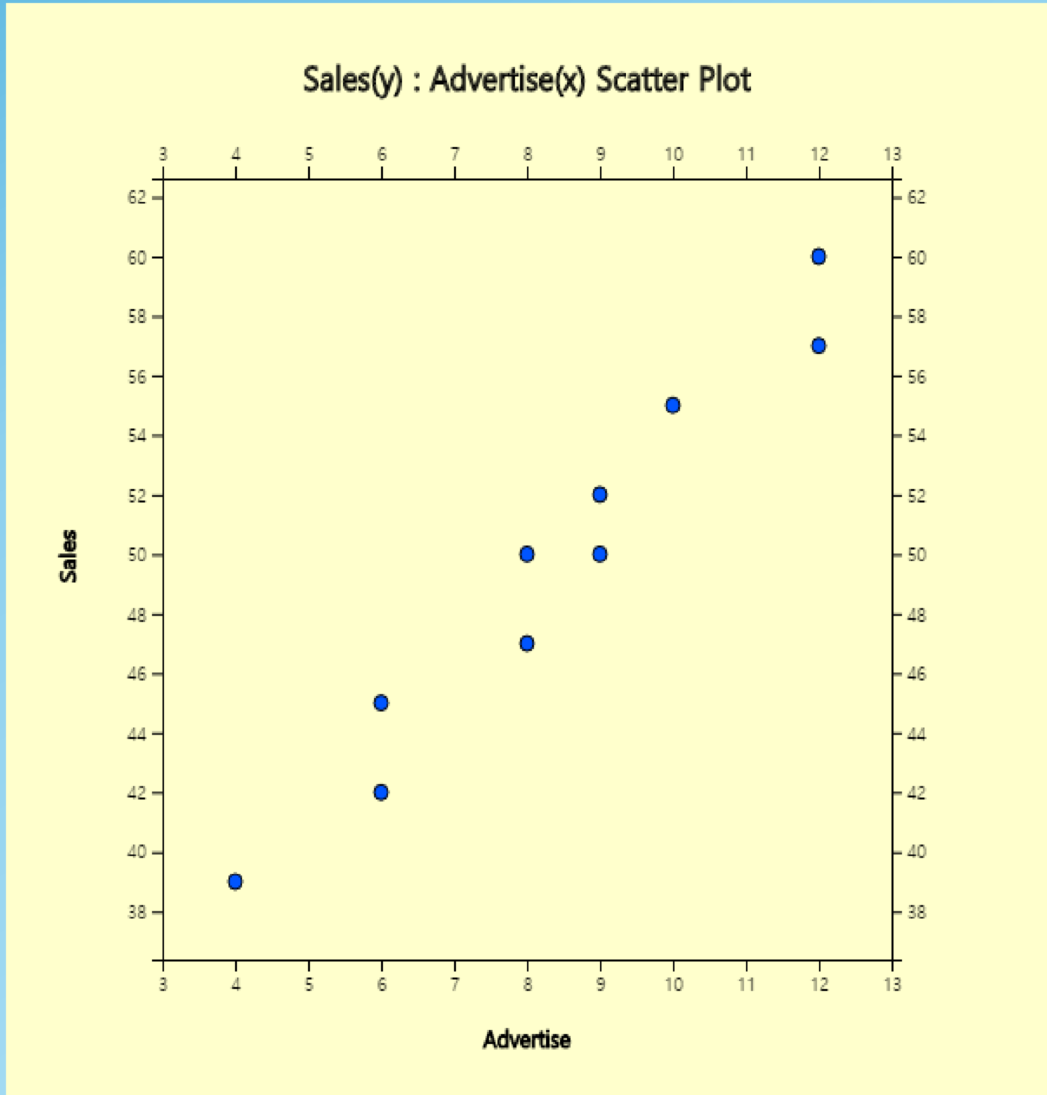
File: EX120101_SalesByAdvertise.csv

Y Var: 2: Sales by X Var: 1: Advertise

(Selected data: Raw Data) (Multiple Selection)

SelectedVar: V2 by V1,

	Advertise	Sales	V3	V4	V5
1	4	39			
2	6	42			
3	6	45			
4	8	47			
5	8	50			
6	9	50			
7	9	52			
8	10	55			
9	12	57			
10	12	60			
11					



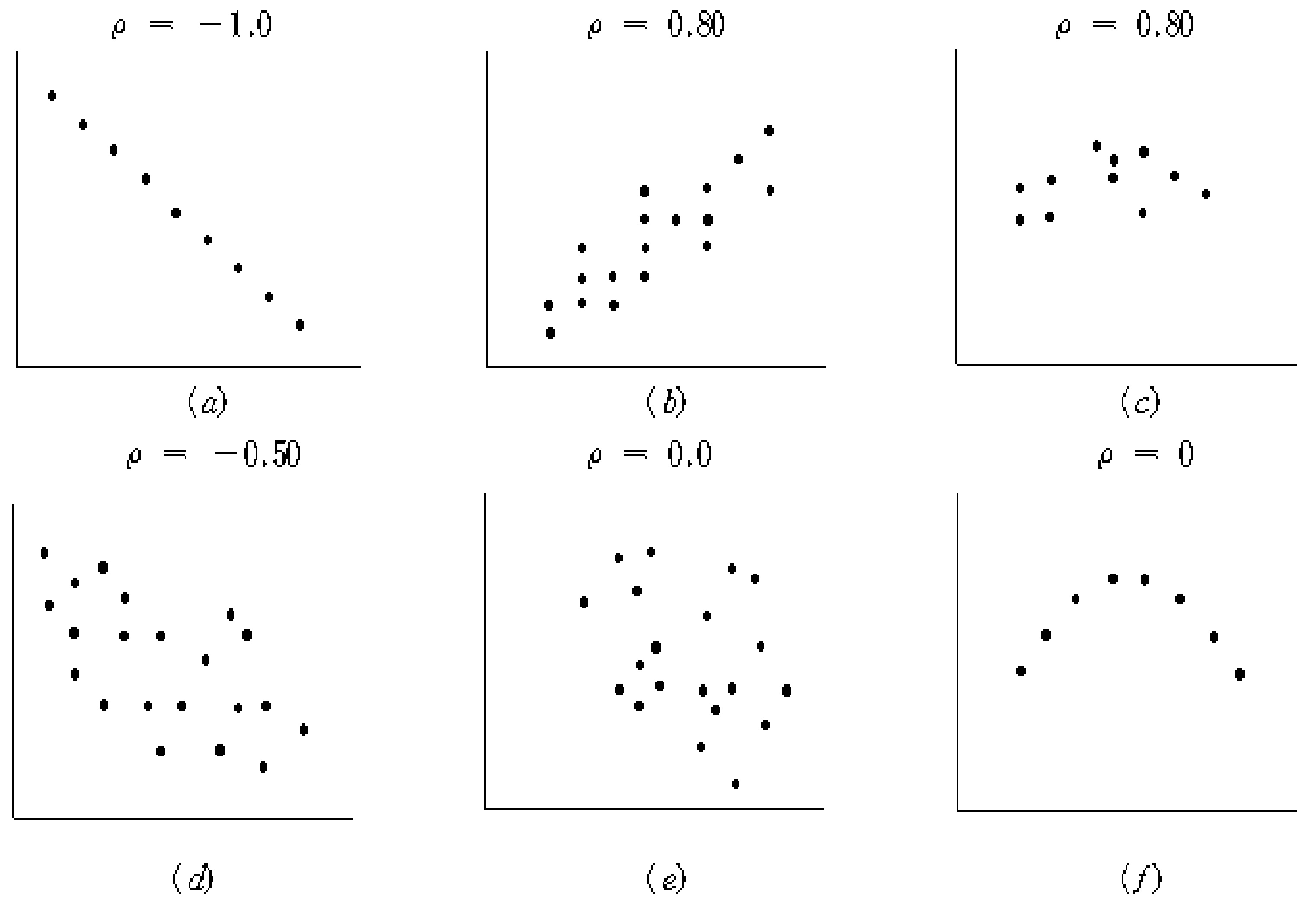
12.1 Correlation Analysis

- **Random Sample** $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$
  from a population with (μ_X, μ_Y) and (σ_X^2, σ_Y^2)

- **Population Covariance** $\sigma_{XY} = Cov(X, Y) = E(X_i - \mu_X)(Y_i - \mu_Y)$
- **Sample Covariance** $S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$

- **Population Correlation** $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
- **Sample Correlation** $r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$

12.1 Correlation Analysis



12.1 Correlation Analysis

- Characteristics of ρ

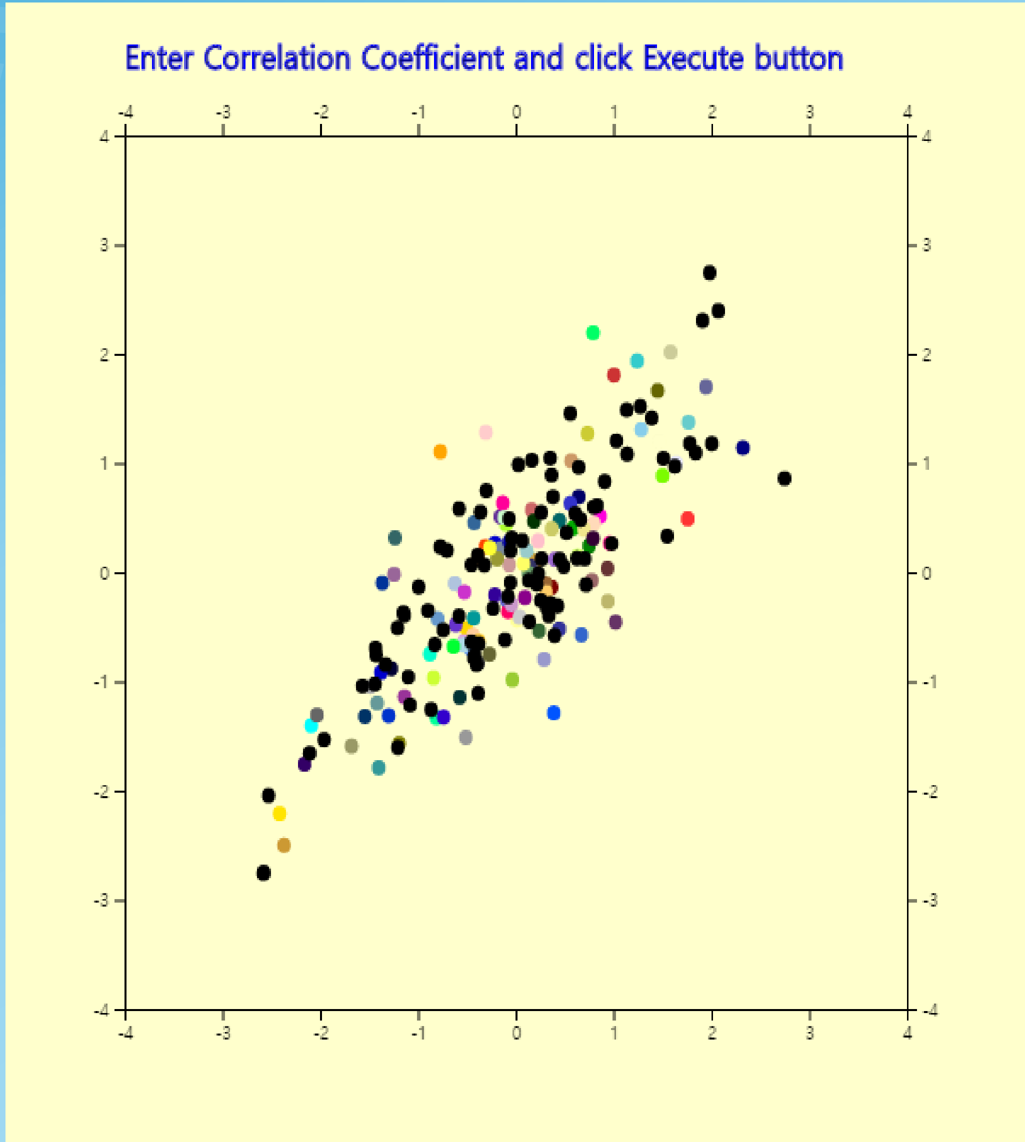
1) ρ has a value between -1 and +1.
   - closer to +1 ⇨ strong positive linear relation
   - closer to -1 ⇨ strong negative linear relation.
   - closer to 0 ⇨ weak linear relation

2) If all values of X and Y are located on a straight line, ρ is either +1 or -1.

3) ρ is only a measure of linear relationship between two variables.
   - if ρ =0, there is no linear relationship between the two variables, but there may be a different relationship

12.1 Correlation Analysis

- **Simulation of correlation coefficient**



12.1 Correlation Analysis

[Example 12.1.2] Find the sample covariance and correlation coefficient for the advertising costs and sales of [Example 12.1.1].

<Answer>

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

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

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

id	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			

12.1 Correlation Analysis

<Answer of Example 12.1.2>

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

$$r = \frac{s_{XY}}{s_X\, s_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{SXY}{\sqrt{SXX\ \ SXY}} = \frac{151.2}{\sqrt{60.4\ \times 396.1}} = 0.978$$

12.1 Correlation Analysis

❑ Testing the population correlation coefficient ρ

Null Hypothesis: $H_0 : \rho = 0$

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

Rejection Region of H_0 :

1) $H_1 : \rho < 0$ Reject H_0 if $t_0 < -t_{n-2;\ \alpha}$
2) $H_1 : \rho > 0$ Reject H_0 if $t_0 > t_{n-2;\ \alpha}$
3) $H_1 : \rho \neq 0$ Reject H_0 if $|\,t_0\,| > t_{n-2;\ \alpha/2}$

12.1 Correlation Analysis

[Example 12.1.3] In 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.

<Answer>

$$t_0 = \sqrt{n-2}\ \frac{r}{\sqrt{1-r^2}} = \sqrt{10-2}\ \frac{0.978}{\sqrt{1-0.978^2}} = 13.26$$

$$t_{10-2;0.025} = 2.306$$

Hence $H_0 : \rho = 0$ is rejected

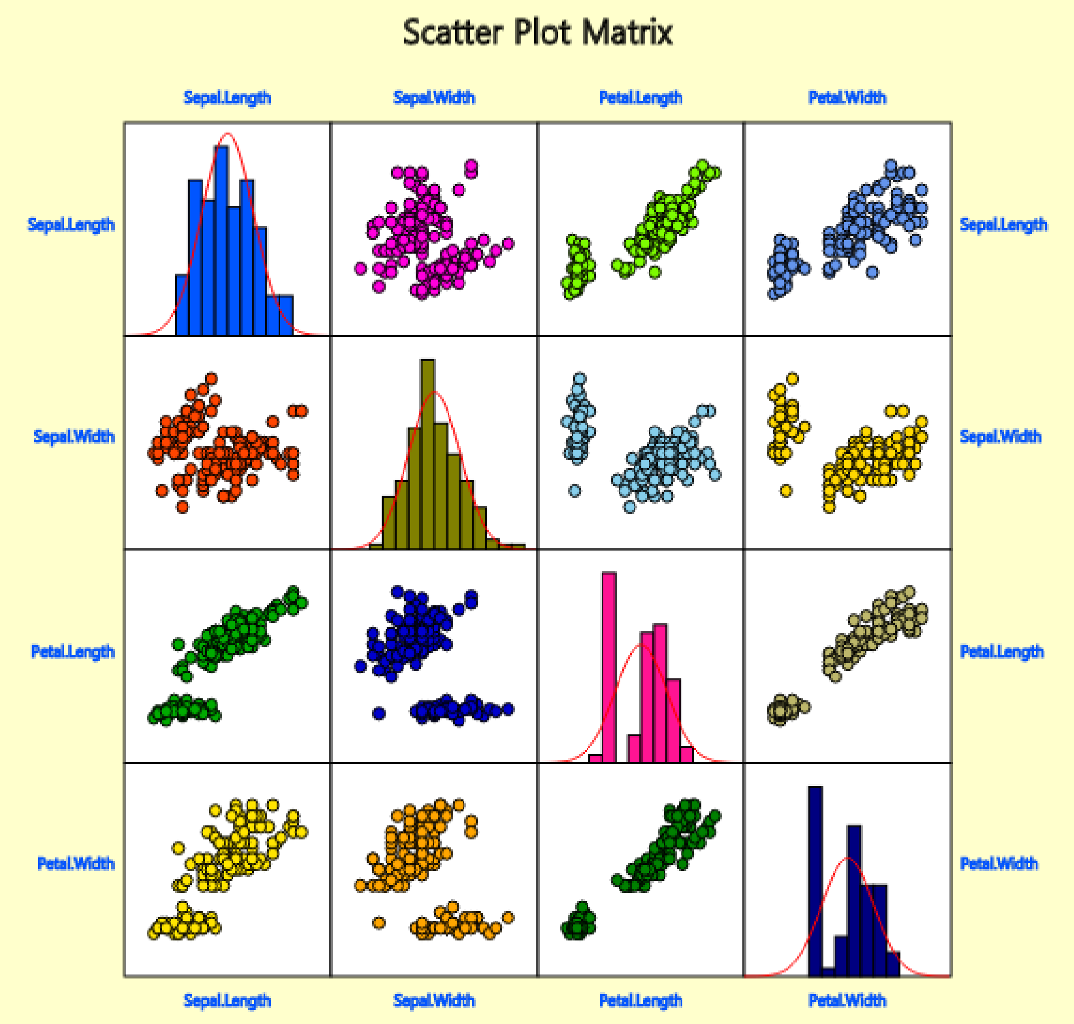
Regression Analysis				
Regression	y =	28.672 +	2.503 x	
Correlation Coefficient	r = 0.978	H_0: ρ = 0 H_1: ρ ≠ 0	t value = 13.117	p value < 0.0001
Coefficient of Determination	r^2 = 0.956			
Standard Error	s = 1.483			

12.1 Correlation Analysis

[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.**



12.1 Correlation Analysis

<Answer of Example 12.1.4>

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

Thank you