Introduction to Statistics and Data Science using *eStat*

Chapter 12 Correlation and Regression Analysis

12.2 Simple Linear Regression Analysis

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

12.2 Simple Linear Regression Analysis

- Regression analysis is a statistical method
  - establishes a mathematical model of relationships between variables,
  - estimates model using measured values of the variables,
  - uses estimated model to describe the relationship between variables, or to apply it to the analysis such as forecasting.
- Mathematical model ⇨ regression equation
- Variable affected by other variables is called a dependent variable.
  ⇨ response variable
- Variables that affect dependent variable are called independent variables.
  ⇨ explanatory variable

12.2 Simple Linear Regression Analysis

- **Population Regression Model** $Y_i = \alpha + \beta X_i + \epsilon_i$, $i = 1, 2, \ldots, n$

  Estimated Regression Equation $\hat{Y}_i = a + b X_i$

  Residuals $e_i = Y_i - \hat{Y}_i$

- **Method of Least Squares Method**

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

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

- **Least Square Estimator of α and β**

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

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

12.2 Simple Linear Regression Analysis

[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 amount of sales when you have spent on advertising by 10.

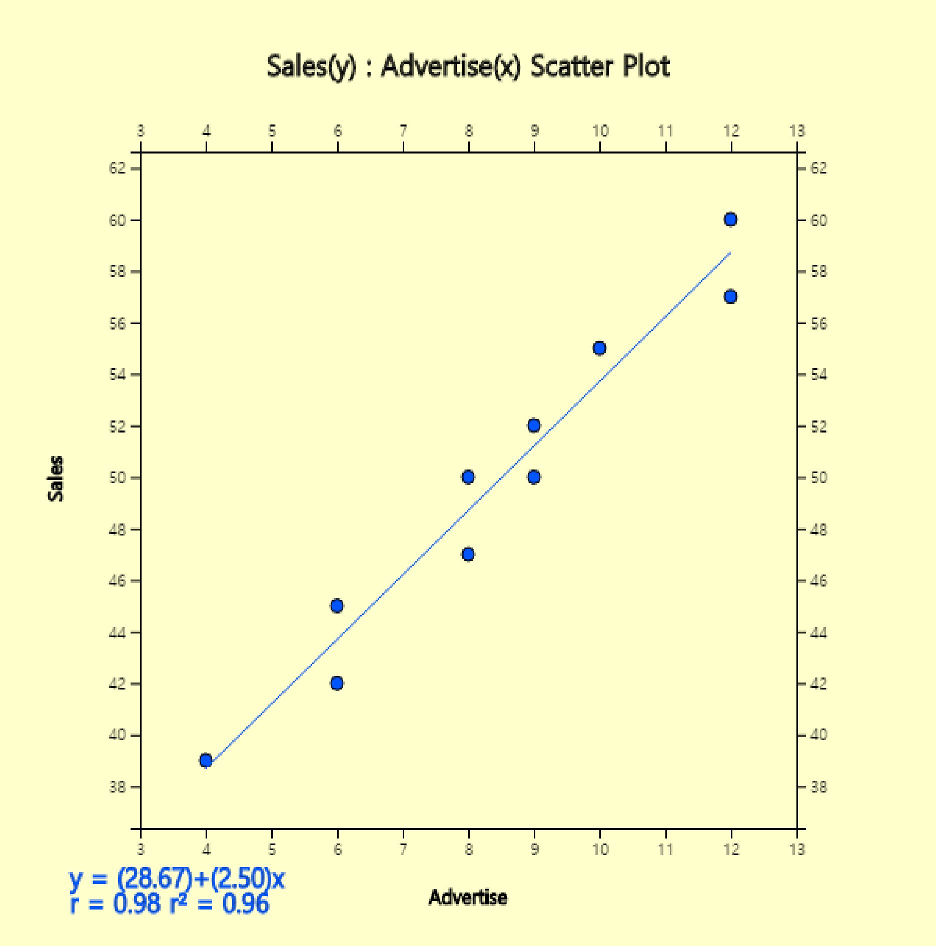
<Answer>

$$b = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{151.2}{60.4} = 2.503$$

$$a = \overline{Y} - b\,\overline{X} = 49.7 - 2.503 \times 8.4 = 28.672$$

- Forecasting

  $28.671 + 2.503 \times 10 = 53.705$



12.2 Simple Linear Regression Analysis

12.2.3 Goodness of Fit for Regression Line

- Residual standard error s is a measure of the extent to which observations are scattered around the estimated line.

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

The residual standard error s is defined as the square root of s^2 .

- SST = SSE + SSR

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad df \quad n-1$$

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad df \quad n-2$$

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad df \quad 1$$

$$R^2 = \frac{SSR}{SST}$$

[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>

$$\hat{Y}_i = 28.672 + 2.503\,X_i$$

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

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

	X_i	Y_i	$\hat{Y}_i$	SST $\sum(Y_i - \bar{Y})^2$	SSR $\sum(\hat{Y}_i - \bar{Y})^2$	SSE $\sum(Y_i - \hat{Y}_i)^2$
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				

12.2 Simple Linear Regression Analysis

<Answer of Example 12.2.2>

Regression Analysis				
Regression	$y =$	28.672 +	2.503 x	
Correlation Coefficient	$r = 0.978$	$H_0: \rho = 0$ $H_1: \rho \neq 0$	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			

12.2 Simple Linear Regression Analysis

[Example 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			

12.2 Simple Linear Regression Analysis

❑ **Inference for the parameter β**

- **Point estimate:** $b = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sim N\left(\beta, \dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$
- **Standard error of estimate b:** $SE(b) = \dfrac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- **Confidence interval of β:** $b \pm t_{n-2;\,\alpha/2} \times SE(b)$
- **Testing hypothesis:**

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

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

  1) $H_1 : \beta < \beta_0$ Reject H_0 if $t <$ - $t_{n-2;\,\alpha}$
  2) $H_1 : \beta > \beta_0$ Reject H_0 if $t >$ $t_{n-2;\,\alpha}$
  3) $H_1 : \beta \neq \beta_0$ Reject H_0 if $|\,t\,| >$ $t_{n-2;\,\alpha/2}$

$y = \alpha + \beta x$

12.2 Simple Linear Regression Analysis

❑ Inference for the parameter α

- Point estimate: $a = \bar{Y} - b\bar{X} \sim N(\alpha, (\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})\sigma^2)$
- Standard error of estimate a: $SE(a) = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- Confidence interval of β: $a \pm t_{n-2;\,\alpha/2} \times SE(a)$
- Testing hypothesis:

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

  Test statistic: $t = \frac{a - \alpha_0}{SE(a)}$

  1) $H_1 : \alpha < \alpha_0$ Reject H_0 if t < - $t_{n-2;\,\alpha}$
  2) $H_1 : \alpha > \alpha_0$ Reject H_0 if t > $t_{n-2;\,\alpha}$
  3) $H_1 : \alpha \neq \alpha_0$ Reject H_0 if $|\,t\,|$ > $t_{n-2;\,\alpha/2}$

12.2 Simple Linear Regression Analysis

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

- Point estimate: $\hat{Y}_0 = a + bX_0$
- Standard error of estimate $\hat{Y}_0$: $SE(\hat{Y}_0) = s\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$
- Confidence interval of $\mu_{Y|x}$: $\hat{Y}_0 \pm t_{n-2;\,\alpha/2} \times SE(\hat{Y}_0)$

12.2 Simple Linear Regression Analysis

[Example 12.2.4]

1) Inference for β

- $b = 2.50333$

$$SE(b) = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{1.484}{60.4} = 0.1908$$

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

$$2.5033 \pm 3.833 \times 0.1908 \quad \Leftrightarrow \quad (1.7720, 3.2346)$$

- Test statistic for $H_0: \beta = 0 \quad H_1: \beta \neq 0$

Reject H_0 if $|t| > t_{n-2;\, \alpha/2}$

$$t = \frac{b - \beta_0}{SE(b)} = \frac{2.5033 - 0}{0.1908} = 13.22$$

Since $t_{8;\, 0.025} = 3.833$, H_0 is rejected.

12.2 Simple Linear Regression Analysis

[Example 12.2.4]

2) Inference for α

- $a = 29.672$

$$SE(a) = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = 1.484\sqrt{\frac{1}{10} + \frac{8.4^2}{60.4}} = 1.670$$

- Test statistic for $H_0: \alpha = 0 \quad H_1: \alpha \neq 0$

Reject H_0 if $|t| > t_{n-2;\ \alpha/2}$

$$t = t = \frac{a - \alpha_0}{SE(a)} = \frac{29.672 - 0}{1.670} = 17.1657$$

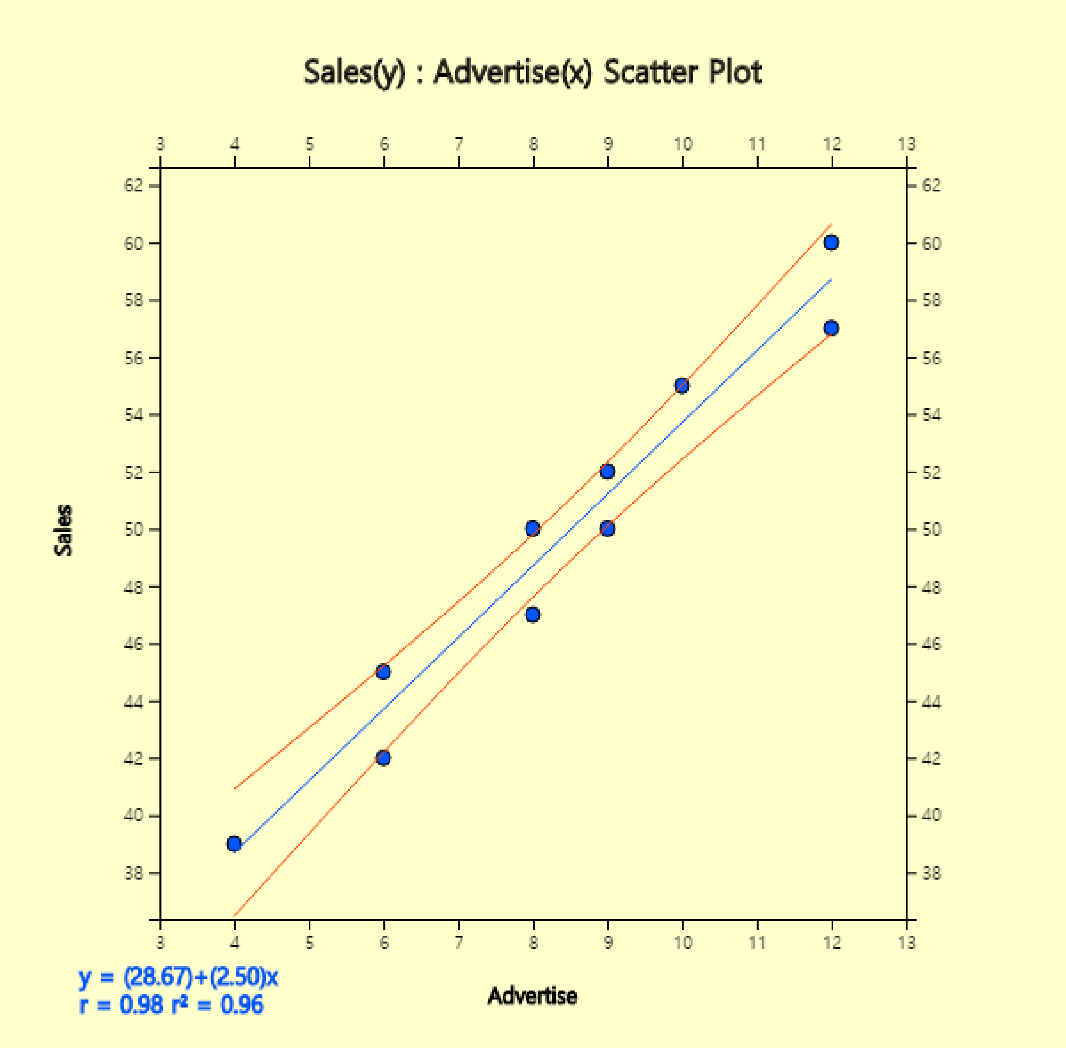
Since $t_{8;\ 0.025} = 3.833$, H_0 is rejected.

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

if x= 8, $\hat{Y}_0 = 49.699$, ⇨ $49.699 \pm 3.833 \times 0.475$

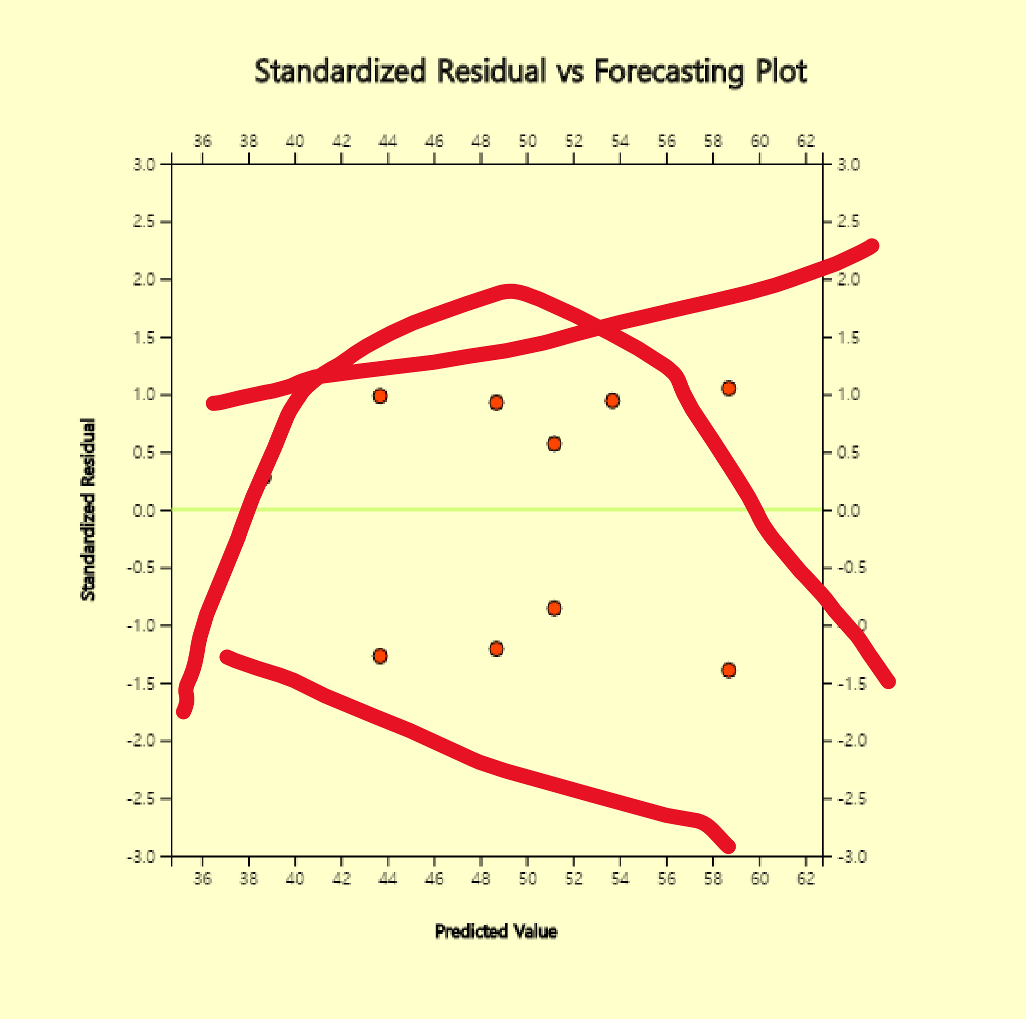
12.2 Simple Linear Regression Analysis

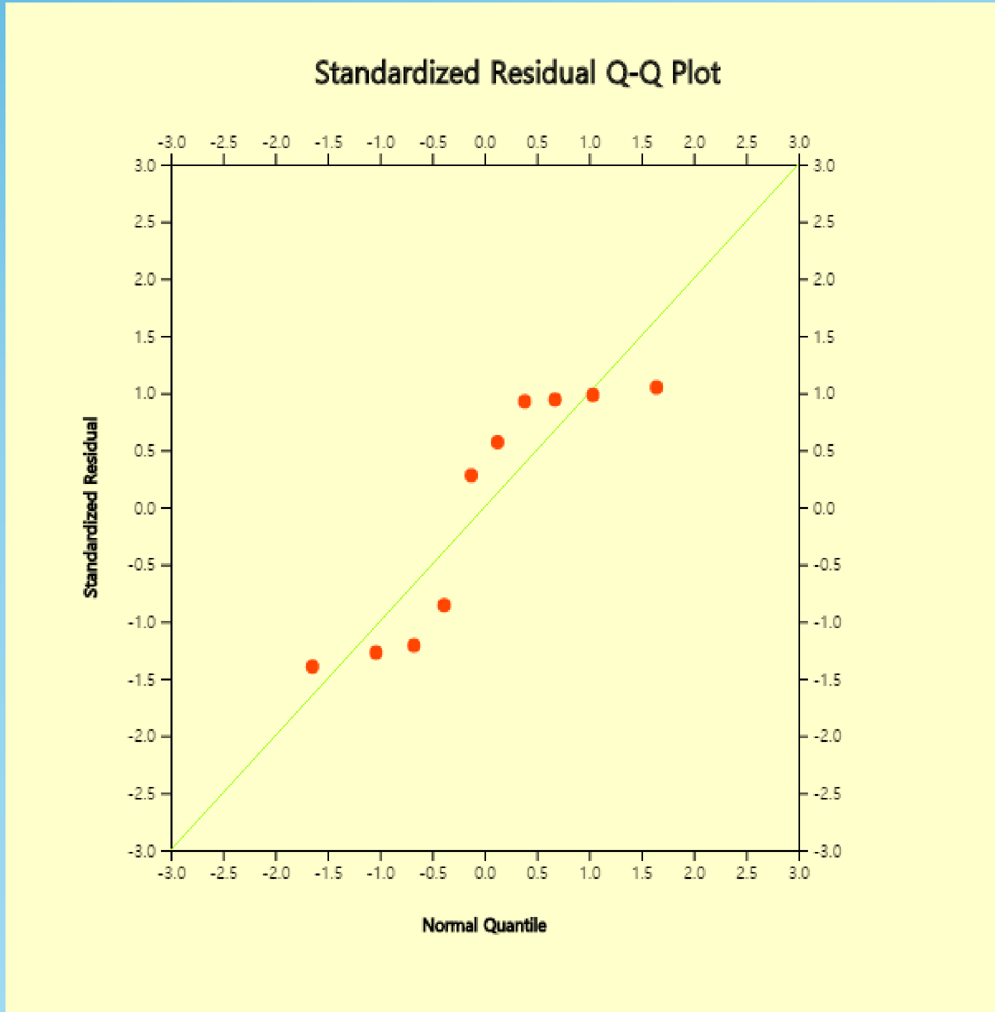
<Answer of Example 12.2.4>



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

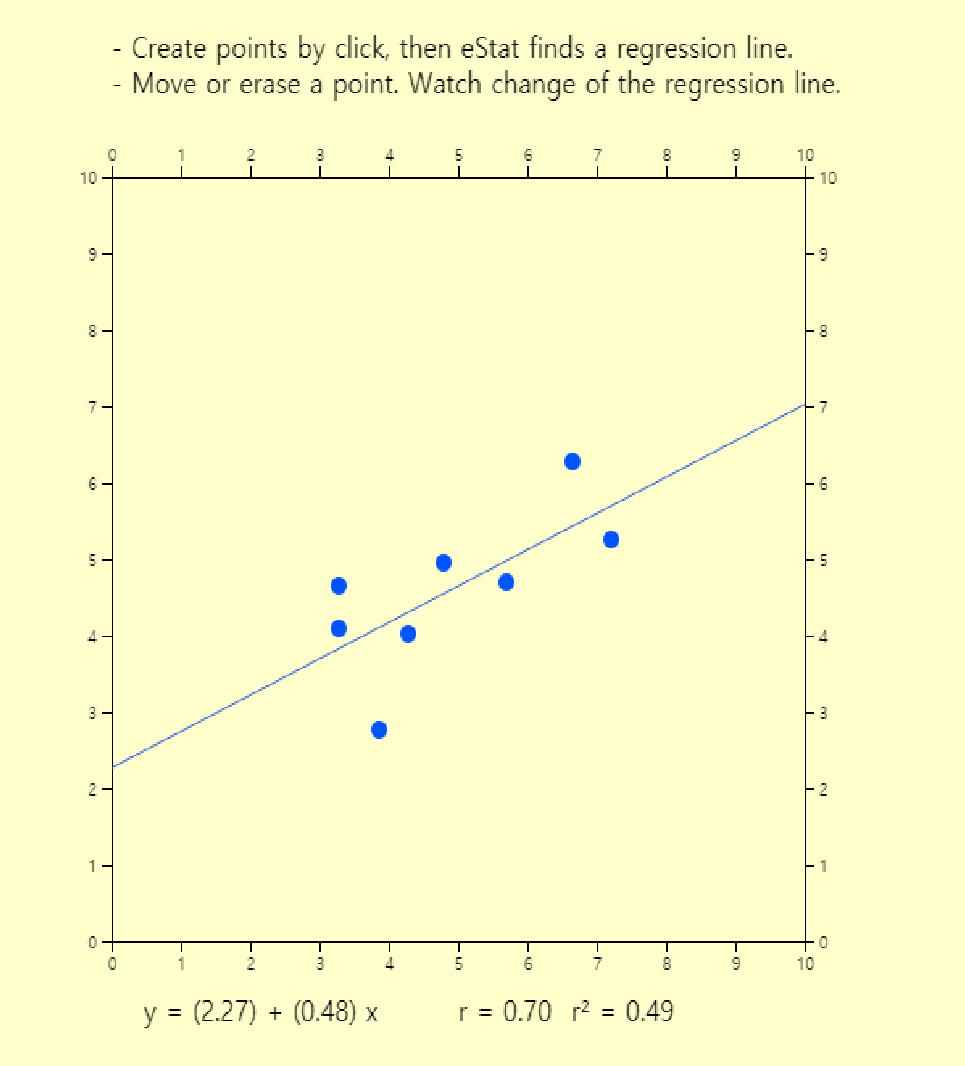
[Example 12.2.5] Residual Analysis





12.2 Simple Linear Regression Analysis

- **Simulation of Regression Analysis in eStatU**



Thank you