Tutorial Business Analytics

Tutorial 3 - Solution

Exercise 3.1

The following table displays the per capita gross national product ($X\ in\ 1000\$$) and the percentage of literate people among the population (Y).

Country	X	Y
Nepal	0.5	5
Uganda	0.6	28
Thailand	1.0	68
South Korea	1.4	77
Peru	1.8	48
Lebanon	3.6	48
Ireland	5.7	98
France	6.4	96
New Zealand	13.0	99

Note: $\sum x_i = 34$, $\sum {x_i}^2 = 262.22$, $\sum y_i = 567$, $\sum x_i y_i = 2914.3$, $\bar{x} = 3.78,\ \bar{y} = 63$

a) Calculate the coefficients $\hat{\beta}_o$ and $\hat{\beta}_1$ of the simple linear regression model using the ordinary least squares. Find the regression line using the formulas below:

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\frac{1}{n-1}\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$$

$$= \frac{\frac{1}{n}\sum_i^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_i^n x_i^2 - \bar{x}^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

b) Interpret the coefficients calculated in exercise a).

c) Test the zero hypothesis $H_0 : \beta_1 \leq 0$ with significance level $\alpha = 0.05$. Use the following t-test with $RSS = 4411.4$ and $\sum_{i=1}^n (x_i - \bar{x})^2 = 133.77$:

$$t_0 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{RSS}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \frac{1}{n-2}}$$

d) Now the above linear regression model will be used to estimate the percentage of literates among a country with known gross national product. Which problems might occur? Briefly explain your concerns using an example.

e) Repeat c) using R (Exercise 3.1_R-Script.R).

Solution

a) The formulas for the coefficients are:

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_i^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_i^n x_i^2 - \bar{x}^2}$$

$$= \frac{\frac{1}{9}(2914.3) - (3.78)(63)}{\frac{1}{9}(262.22) - (3.78)(3.78)} = 5.77$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$= 63 - (5.77)(3.78) = 41.2$$

⇒ regression line: $\hat{y}_i = 41.2 + 5.77 x_i$

b) $\hat{\beta}_0$: a country with a capita gross national product of 0 has 41.2 percent of literate people among the population
$\hat{\beta}_1$: with each increase of \$1000, the percentage of literate people among the population increases by 5.77.

c) Use the "test manual" from tutorial 2 to solve the exercise.

1.) Not needed, because only one test is used in regression analysis.

2.) $H_0: \beta_1 \leq 0$ vs. $H_1: \beta_1 > 0$

3.) simple t-test for coefficient:

$$t_0 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{RSS}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \frac{1}{n-2}} = 2.17$$

with $RSS = 4411.4$, $\sum_{i=1}^n (x_i - \bar{x})^2 = 133.77$

$t_0 = 2.66$

4.) $\alpha = 0.05$

5.) $df = n - 2 = 7, \quad t^c_{7;0.95} = 1.895$

6.) $t_0 > t^c \quad \Rightarrow$ reject H_0. $\hat{\beta}_1$ is statistically significant.

d) A prediction for countries with per capita gross national product outside the sample range is problematic. For example, any country with a gross national product smaller than Nepal or greater than New Zealand may have illogical predicted percentage of literate people among population (< 0% or >100%).

e) Note: R always outputs a two-tail p-value for t-test when you read the summary of the model, i.e., $H_0: \beta_1 = 0$.

Exercise 3.2

t	Demand
0	**28.20**
1	**37.65**
2	**47.28**
3	**59.76**
4	**73.44**
5	**86.19**
6	**100.31**
7	**112.58**
8	**121.63**
9	
10	

a) For the time series above, calculate the forecasted demand value for $t = 10$ using the simple linear regression and the formula below:
$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \cdot t$

b) Calculate the RMSE and explain its meaning.

c) For the time series above, calculate the forecasted demand value for $t = 10$, assuming a biannual seasonal component of the following form: Starting from the first period $t = 0$, suppose after every second period a new year begins. Make use of the formula below:
$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \cdot t + \hat{\beta}_2 \cdot Q_1$

d) Does the data reflect biannual data?

Note: You can use R to solve this exercise (exercise 3.2_R-Script.R).

Solution

a) Forecast

$\hat{\beta}_0 = 25.3822, \quad \hat{\beta}_1 = 12.1833$

$\Rightarrow \qquad \hat{y}_{10} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 10 = 25.3822 + 12.1833 \cdot 10 = 147.2152$

b) RMSE

$\text{RMSE} = \sqrt{MSE}$

$\text{MSE} = \frac{RSS}{T}$

$RSS = \sum_{t=1}^{n} \hat{e}_t^2$

$\hat{e}_t = y_t - \hat{y}_t$

$\hat{e}_1 = 28.20 - 25.38222 = 2.81777$

$\hat{e}_2 = 37.65 - 37.56556 = 0.08444$

$(\ldots)$

$RSS = 27.72076$

$MSE = 3.08$

$RMSE = 1.755$

RMSE is the average deviation of the prediction from the actual values of the data. It is useful for comparing different machine learning models for numerical prediction.

c) Forecast

$\hat{\beta}_0 = 25.3117, \ \hat{\beta}_1 = 12.1833, \ \hat{\beta}_2 = 0.1270$

$\hat{y}_{10} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 10 + 1 \cdot \hat{\beta}_2 = 25.3117 + 12.1833 \cdot 10 + 0.1270 = 147.2717$

d) No, because the seasonal variable is not statistically significant.