Tutorial Business Analytics

Tutorial 3

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 3.2

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 3.3

Given the data in Exercise_3.csv (Y- response variable), test the Gauss-Markov assumptions using R. If any of the properties is violated, make the necessary corrections until you have a BLUE estimator.

a) Test if there is a linear dependency between predictors.

b) Check the linearity assumption and fit a model to the data.

c) Check if the homoscedasticity assumption holds.

d) Check for autocorrelation.

e) Test the exogeneity assumption. Briefly explain the results.

<u>Note:</u> You can use appropriate R functions from R packages like car and lmtest to solve this exercise.