Chapter VI

Correlation
and
Simple Linear Regression Analysis

Exercises

Part 1:

1.

The following data were collected on the height (inches) and weight (pounds) of women swimmers:

Height	68	64	62	65	66
Weight	132	108	102	115	128

1. Develop a scatter diagram for these data with height as the independent variable.
2. What does the scatter diagram indicate about the relationship between the two variables?
3. Try to approximate the relationship between height and weight by drawing a straight line through the data.
4. Develop the estimated regression equation by computing the values of b_0 and b_1.
5. If a swimmer's height is 63 inches, what would you estimate her weight to be?

2.

The following data are the monthly salaries y and the grade point averages x for students who obtained a bachelor's degree in business administration with a major in information systems:

GPA	**Monthly Salary [$]**
2.6	3300
3.4	3600
3.6	4000
3.2	3500
3.5	3900
2.9	3600

The estimated regression equation for these data is

$$y^* = 1790.5 + 581.1x$$

1. Compute *SST, SSR, und SSE.*
2. Compute the coefficient of determination. Comment on the goodness of fit.
3. What is the value of the sample correlation coefficient?
4. Does the t test indicate a significant relationship between grade point average and monthly salary? What is your conclusion? Use $\alpha = 0.05$.
5. Test for a significant relationship using the F test. What is your conclusion? Use $\alpha = 0.05$.
6. Show the *ANOVA* table.
7. Develop a 95% confidence interval for the mean starting salary for all students with a 3.0 GPA.
8. Develop a 95% prediction interval for the starting salary for a student with a GPA of 3.0.

Part II: *SPSS*

1.

Based on the dataset *study_time.sav*, investigate whether there is any relationship between study time and exam scores or not.
Given a relationship, run a simple linear regression analysis at 5% level of significance representing the dependency of exam scores on study time.

2.

Ten cars between 1 and 6 years old were randomly selected from last year's sales records in a certain US American city. The following data were obtained:

x_i	6	6	6	4	2	5	4	5	1	2
y_i	125	115	130	160	219	150	190	163	260	260

where

x_i : age in years of car $i = 1, 2, ..., 10$

y_i : sales in hundreds of dollars in years of car $i = 1, 2, ..., 10$.

1. Create an *SPSS* data file. Call it *carprice.sav*.
2. Graph the data in a scatter plot to determine if there is a possible linear relationship.
3. Compute and interpret the linear correlation coefficient.
4. Determine the regression equation for the data.
5. Graph the regression equation and the data points.
6. Identify outliers and potential influential observations.
7. Compute and interpret the coefficient of determination.
8. Obtain the residuals and create a residual plot. Decide whether it is reasonable to consider that the assumptions for regression analysis are met by the variables in question.
9. At the 5% significance level, do the data provide sufficient evidence to conclude that the slope of the population regression line is not zero and, hence, that the age is useful as a predictor of sales price for cars?

10. Obtain and interpret a 95% confidence interval for the slope, β, of the population regression line that relates age to sales price for cars.
11. Obtain a point estimate for the mean sales price for all 4-year old cars.
12. Determine a 95% confidence interval for the mean sales price for all 4-year old cars.
13. Find the predicted sales price of someone's 4-year old car.
14. Determine a 95% prediction interval for the sales price of this someone's 4-year old car.

(Last revised: 09.12.2019)