**–Correlation and Regression**

Note: **Linear Regression**

The linear regression is a straightforward approach for predicting a quantitative response in one variable (Variable Y) on the basis of another variable changing (X Variable). It assumes that there is approximately a linear relationship between X and Y, where one unit increase in one variable (X) will affect the other variable (Y). Here are some examples.

**Question 1**

You have been placed on the data analysis team for a large ecommerce site in Netherlands. The management believe there is a relationship between the amount spent on advertising and sales. You have been tasked to investigate this. To do this, you are asked to use the variable, advertising costs, to predict sales. You have collect the entire used a stratified sampling technique across the company’s product lines to get a well-covered sample of the cost of advertising on each product line and the sales. You have collected the following data:

The average weekly advertising costs were €7.50, with a standard deviation of €2.50. The average weekly sales were €20, with a standard deviation of €2.00. The largest week of spending on advertising was the first week of December, which totalled €75. The smallest week of advertising costs totalled €10. R = 0.72

**a.** What is the correlation coefficient?

R = 0.72

b. What can you say about the direction of the correlation coefficient?

The correlation coefficient is positive

c. Calculate the slope

slope = r\*(sy/sx) 🡪 0.72\* (2/2.5) = 0.576

d Calculate the intercept

intercept = mean y-slope\*mean x 🡪 20-(0.576\*7.50) = 15.68

e Give the regression equation

Sales= 15.68 + 0.576 \* Euro spent on advertising

f. Predict the weekly sales when advertising is €10.00

15.68 + 0.576 \* 10 = 21.44

g. Predict the weekly sales when advertising is €20.00

15.68 + 0.576 \* 10 = 27.20

g. What is the expected weekly sales when advertising is €0

15.68

i. How much of the variability in weekly sales is determined by advertising spend?

.72 \* .72 🡪 51.84% of the variability in weekly sales are explained by advertising

**Question 2**

The marketing manager of a large supermarket chain would like to use shelf space to predict the weekly sales of pet food. A random sample of 12 equal-sized stores is selected. The researcher gathered the following data;

The average shelf space found was 11 feet, with a standard deviation of 3.5 feet, and the average weekly sales of pet food were €14, with a standard deviation of €2. The largest shelf space found in the data was 20 feet and the smallest shelf space was 3 feet. R = 0.83

**a.** What is the correlation coefficient?

R = 0.83

b. What can you say about the direction of the correlation coefficient?

The correlation coefficient is positive

c. Calculate the slope

slope = r\*(sy/sx) 🡪 0.83\* (2/3.5) = 0.474

d Calculate the intercept

intercept = mean y-slope\*mean x 🡪 14-(0.474\*11) = 8.79

e Give the regression equation

Sales of per food = 8.79 + 0.474 \* shelf feet

f. Predict the weekly sales when the shelf space is 17 feet

8.79 + 0.474 \* 17 = 16.8

g. Predict the weekly sales when the shelf space is 9 feet

8.79 +0.474 \* 9 = 13.1

h. Predict the weekly sales when the shelf space is 34 feet

You cannot, it is outside of the range of the regression equation (max shelf space was 20 feet)

i. How much of the variability in weekly sales is determined by the shelf space?

0.83 \*0.83 = 0.6889 🡪 68.89% of the variability in sales are explained by shelf space

**Question 3**

The owner of a moving company typically has his most experienced manager predict the total number of labour hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more precise method of predicting labour hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved as a predictor of labour costs. The owner collected data of 36 moves in which the origin and destination were within the borough of Manhattan in New York City and in which the travel time was an insignificant portion of the working hours.

In a preliminary research it is found that r = 0.943. The average number of cubic feet moved in the dataset is 307 with a standard deviation of 94.3. The average number of labour hours is 13, with a standard deviation of 5. The lowest cubic feet moved in the sample was 22 and the highest cubic feet moved was 700.

1. What is the correlation coefficient?

R = 0.943

1. What can you say about the direction of the correlation coefficient?

It is positive (and strong!)

1. Calculate the slope and the intercept and give the regression equation

slope = r\*(sy/sx) 🡪 0.943\* (5/94.3) = 0.05

intercept = mean y-slope\*mean x 🡪 13-(0.05\*307) = -2.35

Labour hours = -2.35 + **(**0.05\* cubic feet moved)

1. Predict the hours of labour if the cubic feet moved is 680

Labour hours = -2.35 + **(**0.05\* 680) = 31,65

1. Predict the hours of labour if the cubic feet moved is 190

Labour hours = -2.35 + **(**0.05\* 190) = 7.15

1. Predict the hours of labour if the cubic feet moved is 20

You cannot predict this, it is outside the range of the regression equation.

1. How much of the variability in hours of labour is determined by cubic meters moved?

0.943\*0.943 = 0.89 🡪 89% 🡪 89% of the variability in labour hours is determined by cubic metres moved

**Question 4**

Your Methodology and Data-analysis teacher has noticed a pattern between students who complete or don’t complete the weekly exercises and the student’s final grades. You have been asked to predict the grades of students who reported to have completed the weekly exercises. You used a simple random sampling method and collected self-reported data from students, asking them how many weekly exercises they completed. You then compared this with their final grades. There were 10 exercises in total during the quarter. The maximum grade a student received in the quarter was 10, the lowest grade was 1. In total you collected a sample of 20 students.

From the data collected, you found the average amount of exercises a student completed was 7, with a standard deviation of 1.5 exercises. The average grade of the sample was also 7, with a standard deviation of 2. You found that r = 0.971

1. What is the correlation coefficient?

R = 0.971

1. What can you say about the direction of the correlation coefficient?

It is positive (and strong!)

1. Calculate the slope and the intercept and give the regression equation

slope = r\*(sy/sx) 🡪 0.971\* (1.5/2) = 0.728

intercept = mean y-slope\*mean x 🡪 7-(0.728\*7) = 1.90

Grade = 1.90 + **(**0.728\* weekly exercises completed)

1. Predict the grade when the student has completed all 10 exercises

Grade = 1.90 + **(**0.728\* 10) = 9.18

1. If the student completes half of the exercises (5 exercises) will the student pass or fail? Show your regression calculations.

Pass

Grade = 1.90 + **(**0.728\* 5) = 5.54

1. Predict the grade when the student has completed 7 exercises

Grade = 1.90 + **(**0.728\* 7) = 6.94

1. How much of the variability in grades is explained by the students completion of the weekly exercises?

0.971\*0.971 = 0.942 🡪 94% 🡪 94% of the variability in a students grade can be explained through their completion of the weekly exercises