Statistics & Probability
Statistical Programming with R
Class 4 – Simple Linear Regression

- In this lab class we will go through the steps to fit a simple linear regression model to a data set.

- A fire station in Dublin records, for a number of fires it has attended, the distance of the fire from the fire station and also the cost of the damage caused in thousands of Euro. Go to the class webpage and download and save the dataset firedamage.txt in your folder for this course.

- Import the data into R. See previous lab sheets or ask the demonstrators if you need assistance with this.

- We want to know whether fire damage depends on distance from the fire station. First, print the dataset by typing

```
firedamage
```

- Looking at the raw data doesn't tell us very much. Instead, let's draw a scatterplot. This will give us a visual representation of the relationship between the distance to the nearest fire station and the amount of damage done.

- We can do this in R using a function called `plot`. This function takes two vectors containing numeric data and draws a scatterplot. You can also specify labels for the axes and a title for the whole graph. As always, you can find out more about the function by typing `?plot`. We use it like this:

```
plot(firedamage$distance, firedamage$damage, main=‘‘Fire Damage vs.
Distance’’, xlab=‘‘Distance (km)’’, ylab=‘‘Damage (1000 euro)’’)
```

- Notice that we put the predictor (or dependent) variable first and the response (or independent) variable second.

- Remember to copy the graph into another document before continuing as it will be overwritten when you produce another graph! Right clicking on the graph window and selecting "copy as bitmap" should allow you to then paste it into a Word document for your records.

Complete part A of your answer sheet

- Correlation is a way to assess the strength of the linear relationship between two variables. There is a function to calculate the correlation coefficient (we called this r in lectures) it is called `cor`. It takes two columns of numeric data (it does not matter which order you put them in) and returns the correlation between them.

- To caluculate the correlation between `distance` and `damage` for the `firedamage` data set type

```
cor(firedamage$distance, firedamage$damage)
```

- Remember that the correlation coefficient takes values between -1 and +1. Values close to -1 or +1 indicate that there is a strong linear correlation between the variables (negatively or positively correlated respectively). Values close to 0 indicate that there is a weak linear relationship between the variables.

Complete part B of your answer sheet

- Linear regression is another way to investigate the relationship between two variables. This means drawing a line to fit through all of the points and estimating its intercept (the point where it passes through the y-axis) and slope (the expected increase in y when x increases by 1).

- You might recall from lectures that working out the slope and intercept by hand takes a lot of (tedious!) calculation. In R, we can do it quickly and easily, using the `lm` function. Type:

```
reg <- lm(damage ~ distance, data=firedamage)
summary(reg)
```

- This code fits a simple linear regression model to the `firedamage` data and stores the results in an object called `reg`. A summary of this information is then printed to screen.

- The first table in the output summarises the residuals. The minimum, first quartile, median, 3rd quartile and maximum are presented. These numbers allow us to see if there is anything unusual about the residuals such as are there any outliers.

- The second table presents the model coefficients. The estimates for the intercept (β_0) and the slope (β_1) are given in the first column of this table. Note that the second row is not labelled 'slope' but named after the independent variable, `distance` in this case.

- The t value for the utility test discussed in class is given in the second row of the third column. Recall that the null and alternative hypotheses are:

$$H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0$$

  You can see that the t value is quite large here at 12.525. The corresponding p-value is given in the second row of the fourth column and you can see that it is almost 0. If we specify our desired confidence level (α) to be 0.05 then we would reject the null hypothesis that the slope is equal to 0 and conclude that there is a significant relationship between distance to the nearest fire station and the amount of damage done.

- Lastly, the coefficient of determination (R^2) is given in the text at the bottom of the output. The value given here is 0.9235 meaning that 92% of the variation in `damage` can be explained by its linear relationship with `distance`.

- Note that this value should be the square of the coefficient of correlation that you caluculated earlier. You can calculate the square root of the presented R^2 value using the `sqrt()` function.

Complete part C of your answer sheet

- We can use the `predict` function to predict the response corresponding to a new value of the dependent variable (within the range of the original data!). To predict the amount of damage done to a house which is 4.2 km from the nearest fire station type:

```
predict(reg, data.frame(distance=4.2))
```

- We will draw the scatterplot again, with the regression line added to it. Type:

```
plot(firedamage$distance, firedamage$damage, main = "Fire Damage vs
Distance", xlab = "Distance (km)", ylab = "Damage (1000 euro)")
lines(firedamage$distance, reg$fitted.values)
```

  Complete part D of your answer sheet

- Fit a linear regression model to the `ojuice.txt` data on blackboard. It is of interest to investigate if the sweetness of the juice depends on the amount of pectin.

  Complete part E of your answer sheet

Make sure you save your R script to a USB key or to Google drive before you log off.

Statistics & Probability
4 – Simple Linear Regression

A. Sketch the scatteplot of `distance` vs. `damage`.

What sort of relationship exists between `distance` vs. `damage`?

Do you think a linear regression model looks appropriate here?

B. What is the value of the correlation coefficient? Interpret this value.

C. What is the equation of the regression line?

D. How much damage would you expect to be done to a building which is 4.2 km from the nearest fire station? Perform the calculation by hand and make sure it agrees with the **R** output.

E. What is the equation of the regression line fitted to the `ojuice` data?

Using the `predict` function, predict the sweetness of orange juice with 243 units of pectin. Give the code you used.