Lab 9 - Describing relationships

ENVX1002 Handbook

Semester 1, 2025

Learning Outcomes

- Calculate and interpret Correlation Coefficients in Excel and R
- Produce scatterplots in Excel and R
- Compare numerical and analytical model fitting methods in Excel
- Fit simple linear models and obtain associated model summaries in R
- Overlay fitted models onto scatterplots in R

Before you begin

Create your Quarto document and save it as `Lab-09.Rmd` or similar. The following data files are required:

1) ENVX1002_practical_data_Regression.xlsx

Exercise 1: Linear Modelling in Excel

This exercise focusses on fitting the model parameters and demonstrating two ways a model can be fitted - numerical or analytical;

- Analytical: equation(s) are used directly to find solution, e.g. estimate parameters that minimise residual sum of squares
- Numerical: computer uses "random guesses" to find set of parameters to that minimises objective function, in this case residual sum of squares

We mostly use R for modelling, but R does everything automatically. It is important to know what is going on 'behind the scenes', which is why we are starting in Excel. Similar to the tutorial, you will be calculating each component of the model parameter step by step in the exercises that follow.

1.1 Horses

This is our example of Analytical fitting method.

The number of horses on Canadian farms appeared to decrease after the war:

year	1944	1945	1946	1947	1948

horse	28	26	22	20	19

a) To see whether this is likely to be true, fit a model to the above data 'by hand' in Excel. To aid the calculation it is recommended to fill out the Excel table provided ENVX1002_practical_-data_Regression.xlsx, you can find it in the spreadsheet labelled *Horses*.

The table we have provided in Excel has broken the regression parameter equations (b0, b1) into smaller components so you can understand the underlying mechanisms and where these values come from.

b) Plot the two variables in Excel and fit a line. You can fit a number of models in Excel simply by right clicking on the scatter of points clicking **Add Trendline ...**. Within the add Tendline window (see screenshot below), a number of options are given, here we want **Linear** and we want to tick **display the Equation** and **Display r-squared on the chart**.

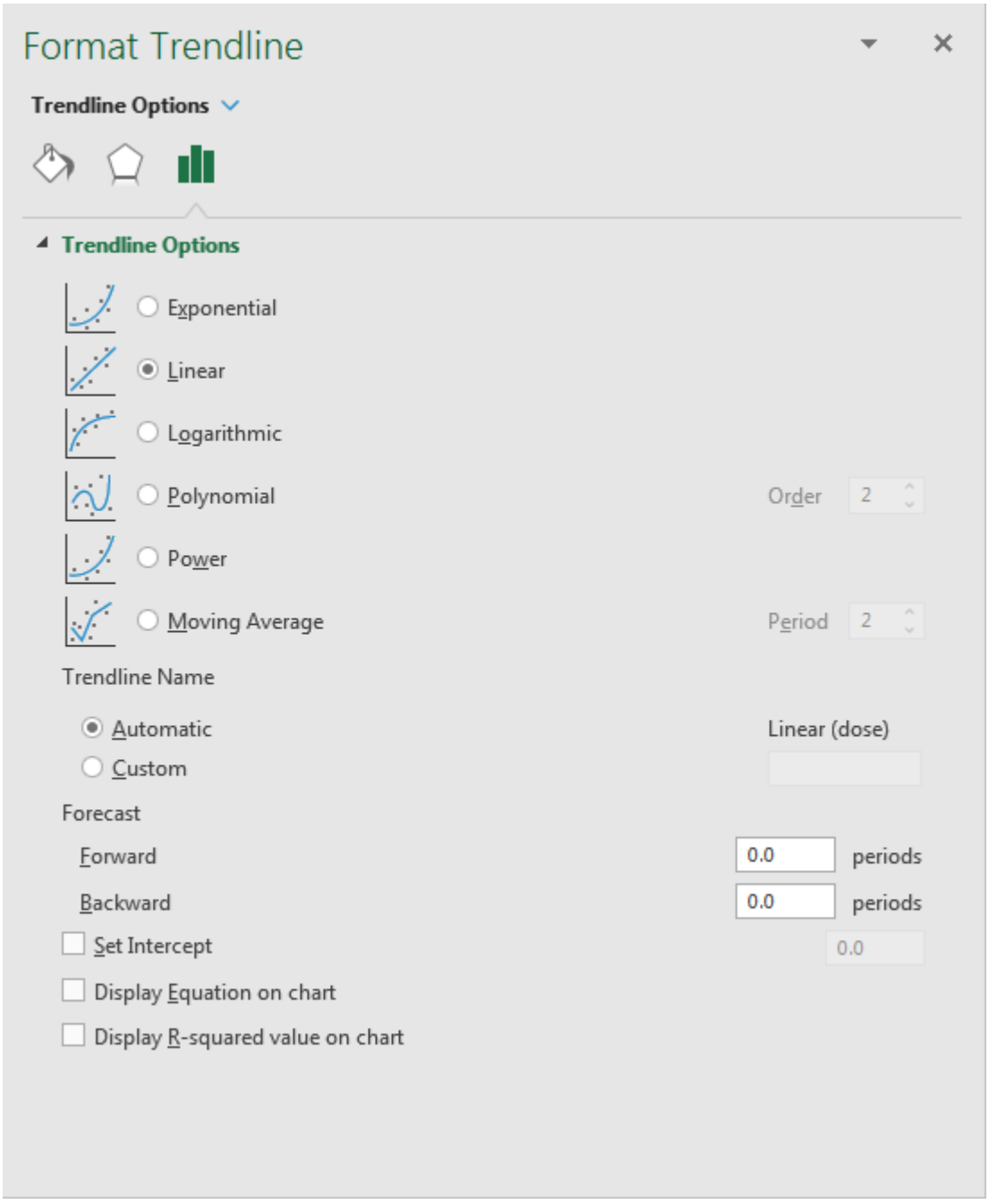


Figure 1: Screenshot: Format Trendline

The R-squared value is a measure of how well the model fits the data where 1.0 is a perfect fit; we will discuss this more in Week 10. The values which appear in the model equation should be the same as those obtained in your earlier calculations.

c) Although it is important for the model equation, do you think the intercept provides a realistic value in this particular case? What does it mean?

d) Calculate the correlation coefficient using the **=CORREL** function in Excel. Type **=CORREL(** and highlight the **Year** column, and then after a comma highlight the **Horses** column and close the brackets.

e) If the relationship was non-linear would this would be a good statistic to use to describe the relationship between horse and years? Explain your answer.

1.2 Fertiliser data

This is our example of numerical fitting of a model.

Figure 1 shows a plot of yield against fertiliser where a linear model is fitted through the scatterplot of raw observations. Intuitively you would draw this as a line that comes as close to possible to all observations which you may have come across as a 'line of best fit'. In this exercise we will explore how models can be fitted automatically based on least-squares estimation.

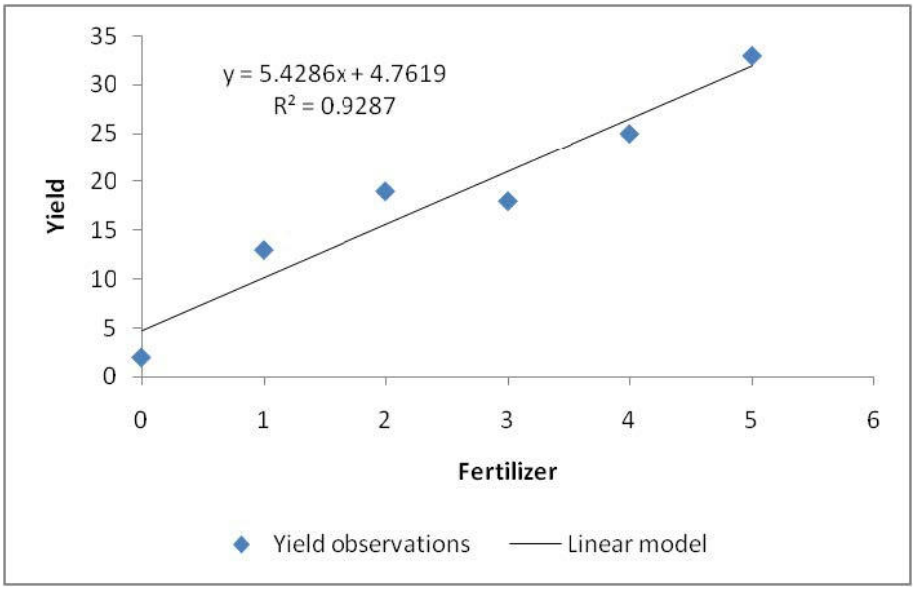


Figure 2: Figure 1: Plot of Yield-response to fertiliser

In Figure 1 you will notice that the line does not fit the data perfectly which is typical of biological and environmental data. A measure of how far the model is from the data is the residual.

Where y_i is the observed value for the ith observation and $\hat{y}_1$ is the predicted value for the ith observation. In this case the predicted value is based on the linear model.

If we add up the square of the residuals for the n observations we get something called the Residual Sum of Squares (SS_{res}):

The best fitting model will have the smallest RSS. The general method is called least-squared estimation. We will now use Excel to find the optimal model.

Enter values of 2 for the y-intercept (b_0) and 3 for the slope (b_1) in cells H2:H3. These are the initial guess values.

b) Now use these parameter values to create predictions for each value of fertiliser in the Predicted column.

Make sure that rather than writing in the value '2' and '3' for your predicted column, you refer to cells H2 and H3 (write as H2 and H3, see screenshot below). Once you have completed the equation, you can apply the equation to the other rows by clicking on the small box at the bottom right corner of the cell and drag it down the rows. Writing dollar signs into your references to H2 and H3 prevents your equation from moving down the row column.

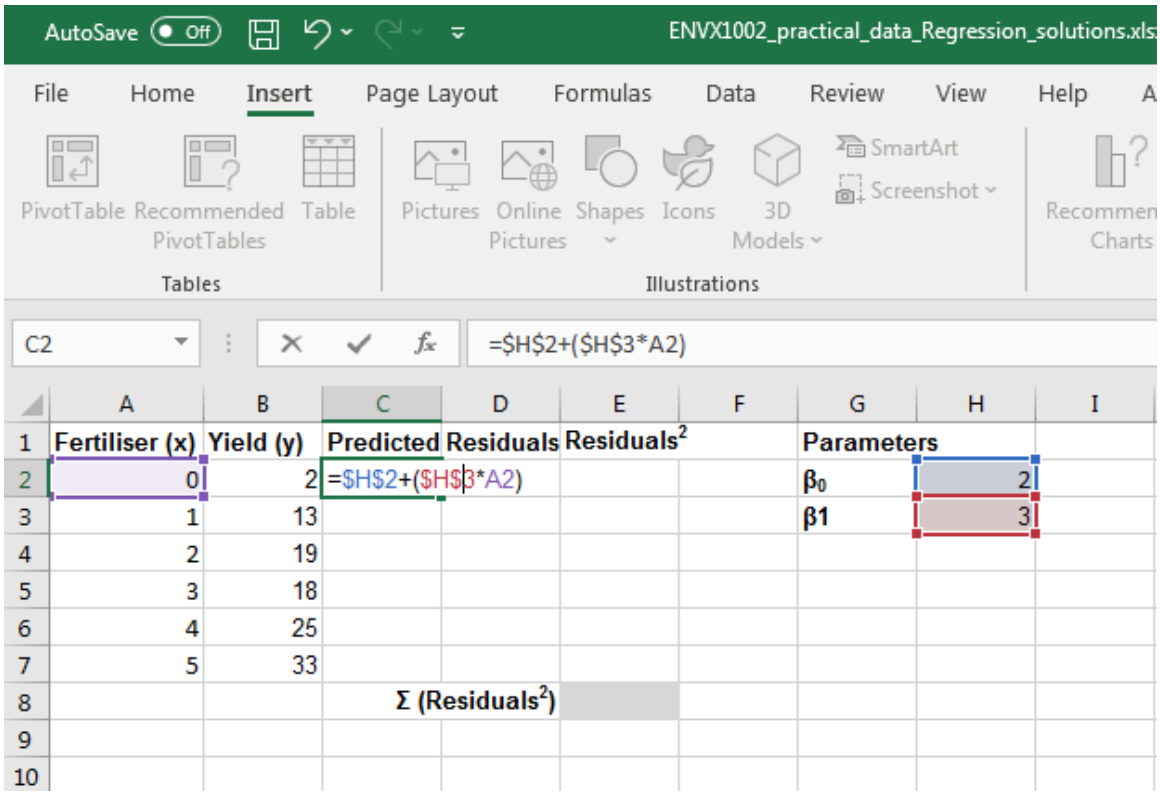


Figure 3: Screenshot of Predicted column input

c) Use this information to calculate (i) residuals (ii) residuals² (iii) RSS.

d) Create a plot similar to Figure 1 where the observations are plotted as symbols and the model predictions are a line. You should have your spreadsheet set up so that if you change the values of the parameters the plotted line changes as well. Try to fit the line manually. This can be difficult, especially for non-linear models.

e) Follow instructions provided in the Tutorial, or in the file How to install Solver to ensure you have Solver ready to use in Excel.

Once you have added Solver, click on the tab **Data >> Solver**, and you will see the following (see screenshot below). For **Set Objective**, you need to select the cell where your RSS value has been calculated. We wish to minimize this so we click on Min, and we do this by Changing Cells where

the parameters of the model are found, in this case the y-intercept and slope. Before clicking **Solve**, make sure you can see your calculated values so you can see how your how it all changes.

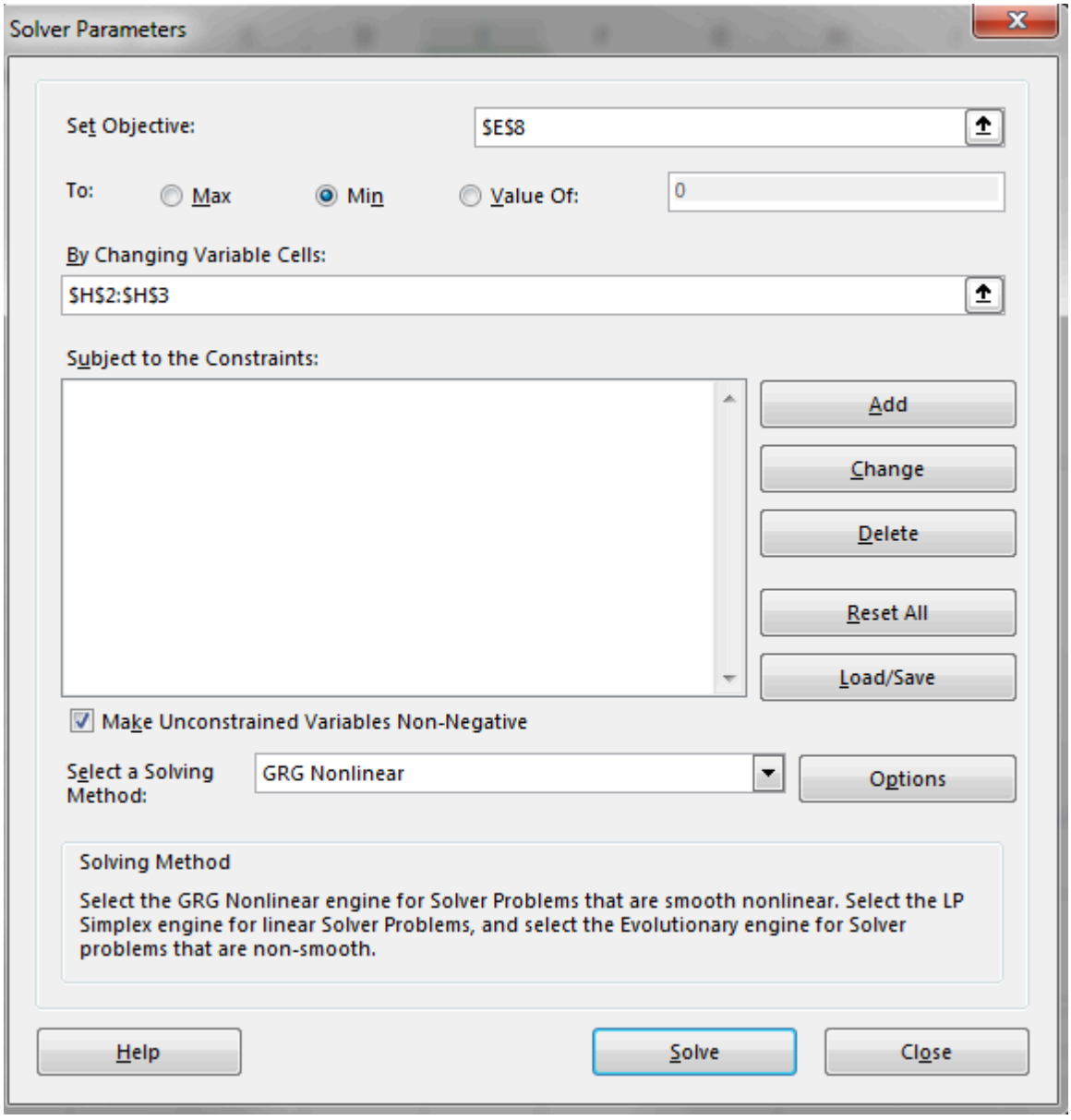


Figure 4: Screenshot of solver with input values

When ready, click on **Solve** and it should find a solution for the minimum RSS. Solver uses an iterative procedure to find the minimum RSS which means it successively guesses values until it finds the optimal value. This is a numerical solution to the problem of model fitting.

Your 'SOLVED' parameters should be the same as what appears in your trendline equation.

Exercise 2: Fitting a model in R

Now we have a deeper understanding of what is going on behind the scenes, we can fit linear models in R.

Before you begin, ensure you have a project set up in your desired folder. Then open up a fresh R markdown and save the file within this folder.

Don't forget to save as you go!

2.1 Have a go - Fertiliser data

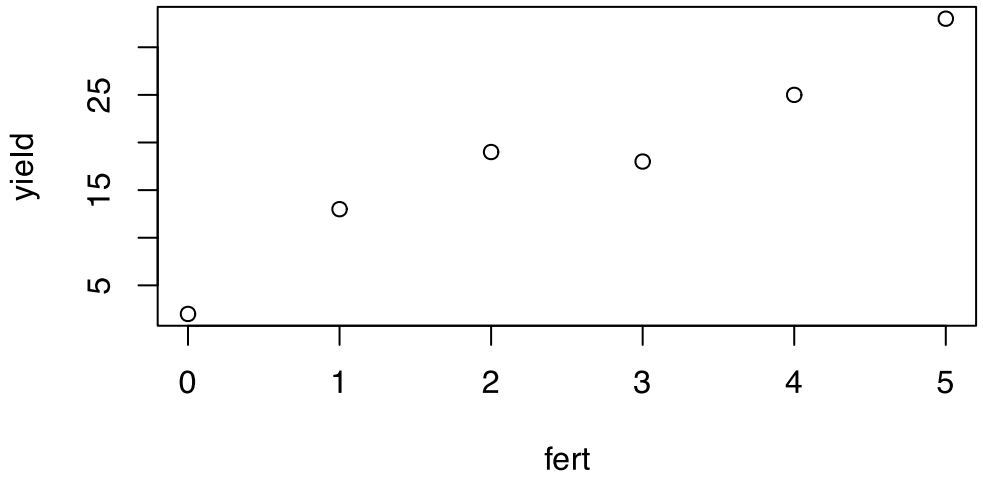
You will use the fertiliser data to fit a linear model in R. As we covered fitting linear models in the Tutorial, it is now your turn to have a go at fitting the models (with some hints along the way).

a) Read the following code into R:

```
# add the data to R Studio
fert <- c(0, 1, 2, 3, 4, 5)
yield <- c(2, 13, 19, 18, 25, 33)
```

b) To visually identify any trends or relationships, create a scatterplot of fertiliser vs yield. From the scatterplot you see, are there any relationships or trends evident?

```
# Create a scatterplot
plot(fert, yield)
```



c) To numerically determine whether there is a relationship, calculate the correlation coefficient. (assume data is normally distributed). Does the correlation coefficient indicate a relationship between fertiliser and yield?

```
# To calculate the correlation coefficient:
cor(fert, yield)
```

```
[1] 0.9636686
```

d) You can now fit the model in R using the `lm()` function. Remember to tell R the name of the object you want to store it as (in this case, `model.lm <-`), then state the name of the function. The arguments within the function (i.e. between the brackets) will be `yield ~ fert`, with `yield` being the response variable and `fert` being the predictor.

```
# Run your model
## yield = response variable (x)
## fert = predictor variable (y)
model.lm <- lm(yield ~ fert)

# Obtain model summary - In here you can obtain the model parameters
# Look for Intercept Estimate and fert Estimate
summary(model.lm)
```

```

Call:
lm(formula = yield ~ fert)

Residuals:
     1      2      3      4      5      6
-2.762  2.810  3.381 -3.048 -1.476  1.095

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.7619     2.2778   2.091  0.10476
fert          5.4286     0.7523   7.216  0.00196 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.147 on 4 degrees of freedom
Multiple R-squared:  0.9287,    Adjusted R-squared:  0.9108
F-statistic: 52.07 on 1 and 4 DF,  p-value: 0.001956
```

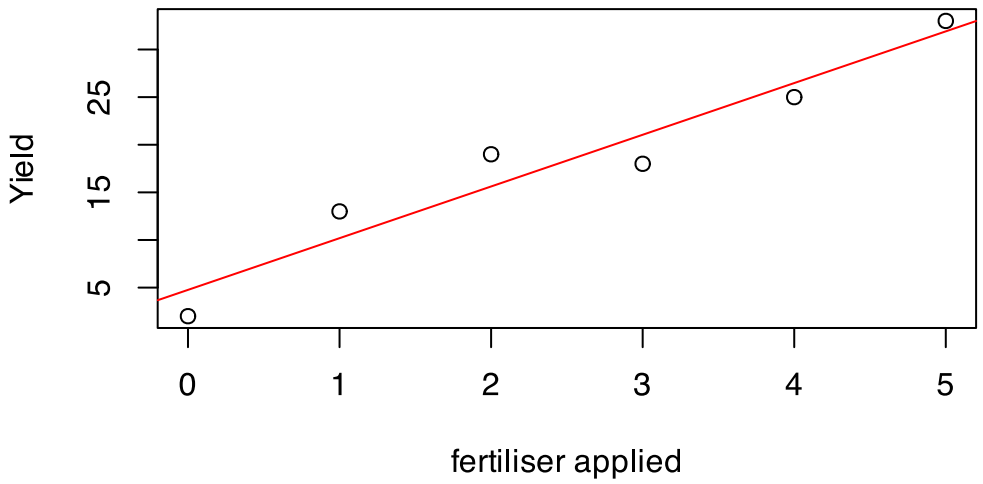
e) In the model output obtained from `summary(model.lm)` the model parameters will be listed under 'Estimate' for the intercept and 'fert'. Compare these values to what you have calculated in Excel.

f) Based on this output, what would the model equation be? Does it match your findings in Excel?

g) You can now fit your model to the scatterplot you created previously using the `abline()` function. Make sure you run the plot function and the abline function in one go. If the lines are run separately, an error may appear saying "plot.new hasn't been called yet"; this is because the abline function requires a current plot on which it can overlay the line.

Also remember, when presenting plots (e.g. in a report), they should be able to stand alone and be self-explanatory. We therefore need to make sure there are clear axis labels. This can be done using 'xlab' and 'ylab' arguments.

```
# Add the linear model to your scatterplot
plot(fert, yield, xlab = "fertiliser applied", ylab = "Yield")
abline(model.lm, col = "red")
```



2.2 ABARES data

In this final example we will be using a dataset obtained from the Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES). The dataset provides a measure of productivity growth (TFP; Total Factor Productivity) in the Australian dairy industry from the years 1978 to 2018.

More information about the ABARES dataset and productivity can be found here.

a) Read in the data from the Excel file for today's practical.

Because we have such a large dataset this time, it is better to read the data straight from Excel than read in each individual value. Reading straight from the source file in Excel saves time and reduces chance of input error.

```
library(readxl)

ABARES <- read_excel("data/ENVX1002_practical_data_Regression.xlsx", sheet =
"ABARES")
```

b) Create a scatterplot of `Year` against `TFP`. Dont forget the format will be different now - instead of only mentioning the object name, e.g. plot(yield, fert), you will need to refer to the specific columns within the ABARES dataset. (i.e. ABARES$Year).

c) Can you see a trend between TFP and Year? Or are the points evenly scattered?

d) Calculate the correlation coefficient between these two variables. Is there a strong relationship?

e) Fit a model to your data and obtain the model summary. Year will be our predictor and TFP will be our response variable. What are the model parameters (i.e. b_0 and b_1)?

f) What would the equation for this model be?

g) Overlay your model onto the scatterplot you produced earlier. When plotting make sure you refer to the column names as you did for the model (e.g. ABARES$Year).

That's it! Great work today. Next week: interpreting linear models!

Bonus take home exercises

Exercise 1: Cars stopping distance

For this exercise we willuse the inbuilt dataset cars to see if there is a relationship between a cars speed (mph) and stopping distance (fft).

```
head(cars)
```

```
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

a) Create a scatterplot of speed vs distance

b) Is there are trend? Or are the point evenly scatered?

c) Calculate the correlation coefficient between these two variables. Is there a strong relationship?

d) Fit a model to your data and obtain the model summary.

e) What would the equation of the line be?

f) Overlay your model onto the scatterplot you produced earlier

Exercise 2: Penguins

For this exercise, we will be using the palmer penguin data set to see if there is a relationship between bill and flipper length.

```
#Load libraries
library(palmerpenguins)
library(tidyverse)

#Clean data
penguins <- penguins%>%
  na.omit()#remove missing data

head(penguins)
```

```
# A tibble: 6 × 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           39.1          18.7               181        3750
2 Adelie  Torgersen           39.5          17.4               186        3800
3 Adelie  Torgersen           40.3          18                 195        3250
4 Adelie  Torgersen           36.7          19.3               193        3450
5 Adelie  Torgersen           39.3          20.6               190        3650
6 Adelie  Torgersen           38.9          17.8               181        3625
# ℹ 2 more variables: sex <fct>, year <int>
```

a) Create a scatterplot of bill length vs flipper legth

b) Is there are trend? Or are the point evenly scatered?

c) Calculate the correlation coefficient between these two variables. Is there a strong relationship?

d) Fit a model to your data and obtain the model summary. What are the parameters?

e) What would the equation of the line be?

f) Overlay your model onto the scatterplot you produced earlier

Exercise 3: Old Faithful Geyser Data

For this exercise, we will be looking at the relationship between geyser eruption time and time between erupttions, using the inbuilt data set faithful

```
head(faithful)
```

```
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

a) Create a scatterplot of eruptions vs waiting time

b) Is there are trend? Or are the point evenly scatered?

c) Calculate the correlation coefficient between these two variables. Is there a strong relationship?

d) Fit a model to your data and obtain the model summary.

e) What would the equation of the line be?

f) Overlay your model onto the scatterplot you produced earlier