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

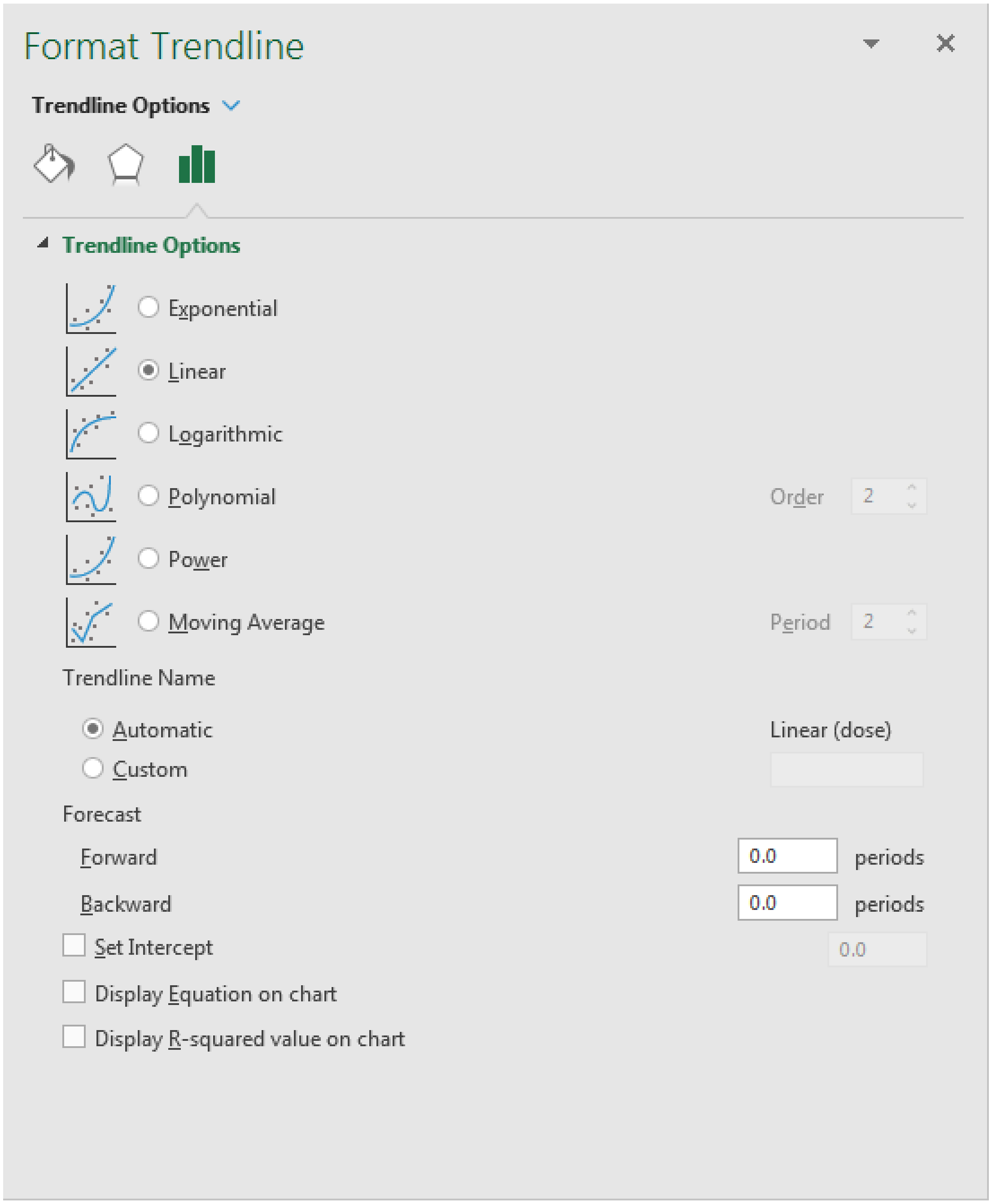
a) To see whether this is likely to be true, fit a model to the above data ‘by hand’ in Excel. To aid the calculation it is recommended to fill out the Excel table provided ENVX1002_practical_-data_Regression.xlsx, you can find it in the spreadsheet labelled *Horses*.

Solution

b0 = 4693.4 b1 = −2.40 SSxy = −24 SSxx = 10 mean(x) = 1946 mean(y) = 23

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**.



Format Trendline

Trendline Options

Trendline Options

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Linear (dose)

○ Custom

Forecast

Forward 0.0 periods

Backward 0.0 periods

□ Set Intercept 0.0

□ Display Equation on chart

□ Display R-squared value on chart

Figure 1: Screenshot: Format Trendline

Solution

Should be a nice plot in excel, equations should match your calculated b0 and b1

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?

Solution
The intercept is 4693, suggesting at the year zero, Canadian farms had 4693 horses. Other than to obtain an intercept, it does not make sense to extrapolate beyond the years we have data for.

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 **)**.

Solution
r = −0.9798, meaning there is a strong relationship in the negative direction; i.e. as the years increase, the number of horses have decreased.

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.

Solution
The Pearson correlation coefficient is only useful for describing linear relationships. Based on the limited sample size we have, the scatterplot looks like a linear relationship, and so it would be ok to use the correlation coefficient in this case.

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.

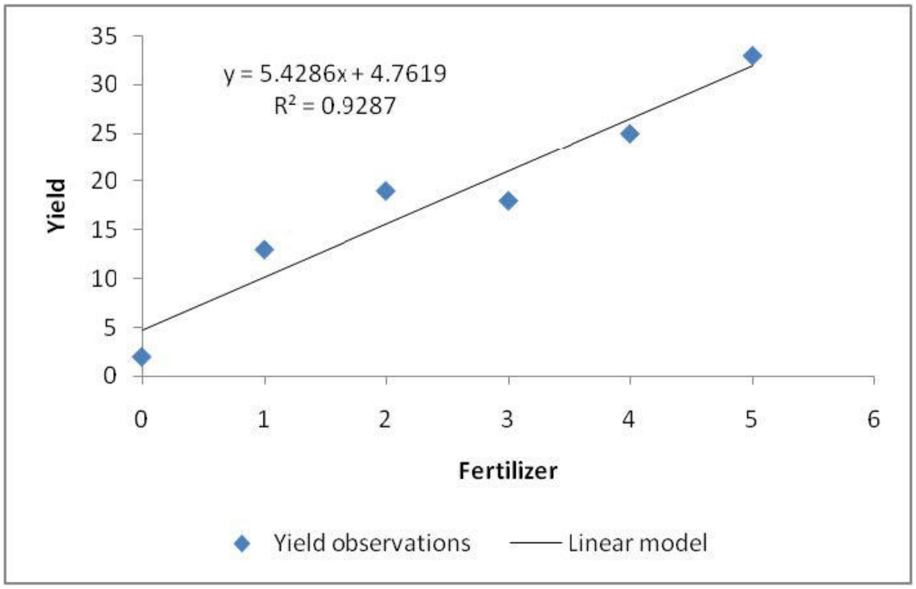


Figure 2: Figure 1: Plot of Yield-response to fertiliser

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 *i*th observation. In this case the predicted value is based on the linear model.

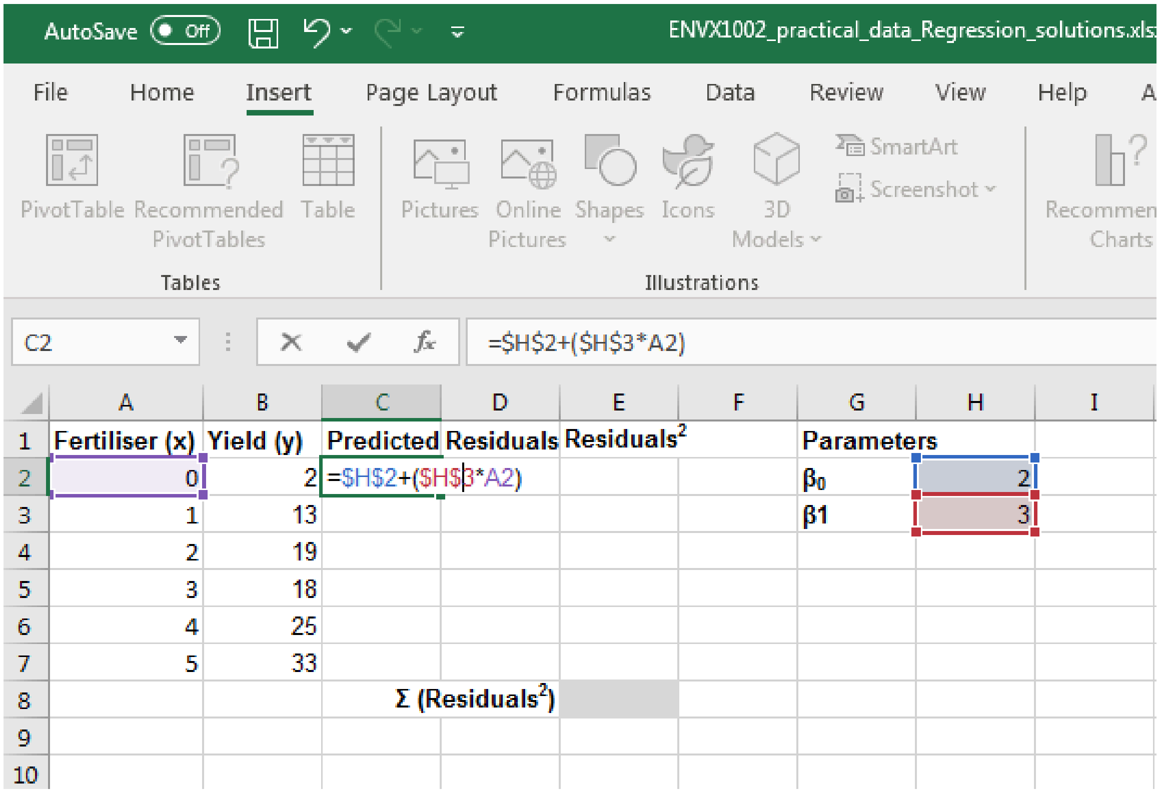
If we add up the square of the residuals for the n observations we get something called the Residual Sum of Squares (SS_{res}):

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.

Instead of typing ‘2’ and ‘3’ directly in your formula, you must use $H$2 and $H$3. The dollar signs lock these cells as reference points, so they won’t shift when you copy the formula. To apply the formula to other rows, either drag the small box at the bottom right of the cell down the column or double-click it.



=$H$2+($H$3*A2)

	A	B	C	D	E	F	G	H	I
1	Fertiliser (x)	Yield (y)	Predicted	Residuals	Residuals^2		Parameters		
2	0	2	=$H$2+($H$3*A2)				β_0	2	
3	1	13					β1	3	
4	2	19							
5	3	18							
6	4	25							
7	5	33							
8				Σ (Residuals^2)					
9									
10									

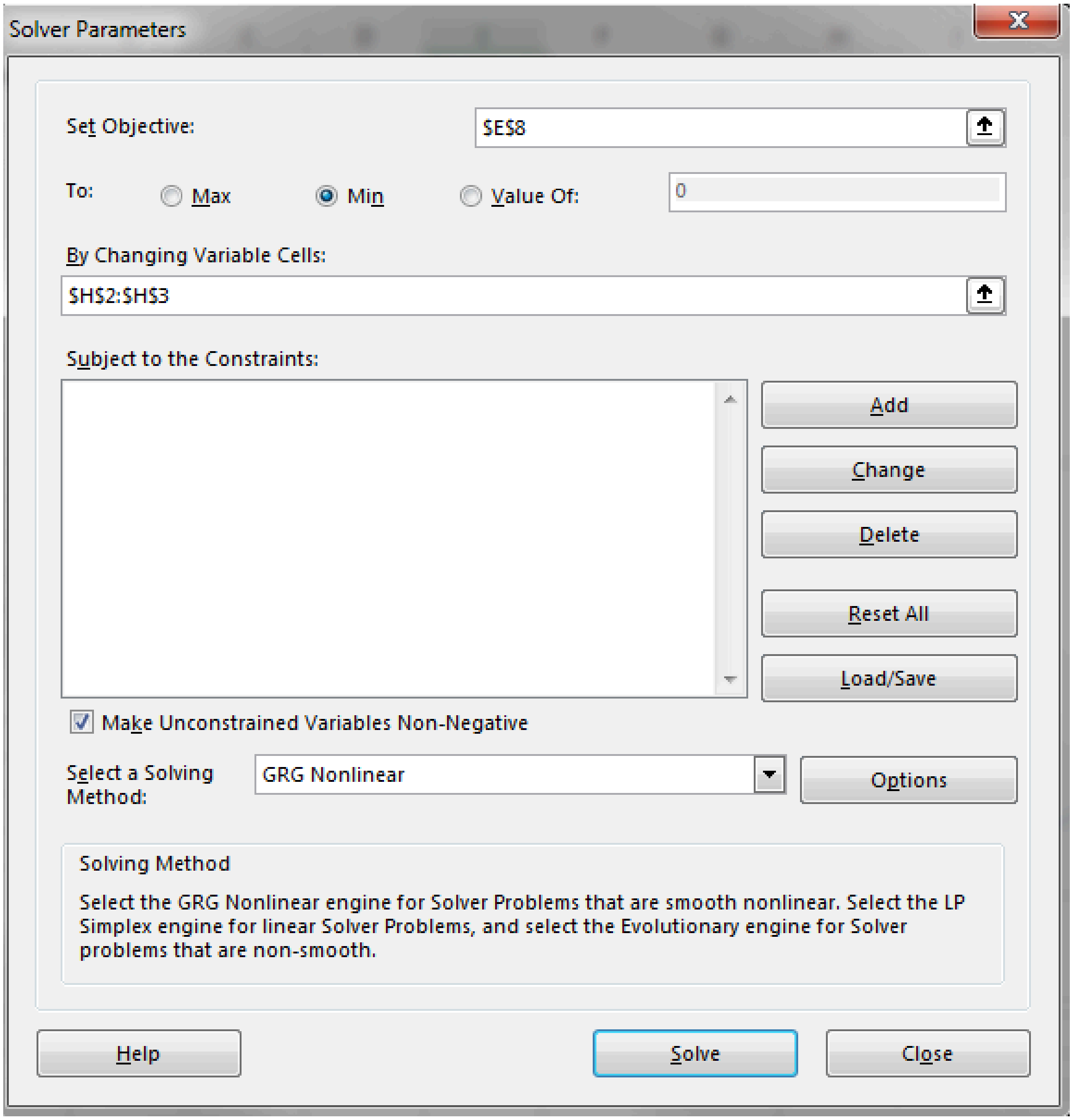
Figure 3: Screenshot of Predicted column input

c) Use this information to calculate (i) residuals (ii) residuals^2 (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.



Solver Parameters

Set Objective: $E$8

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells: $H$2:$H$3

Subject to the Constraints:

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

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

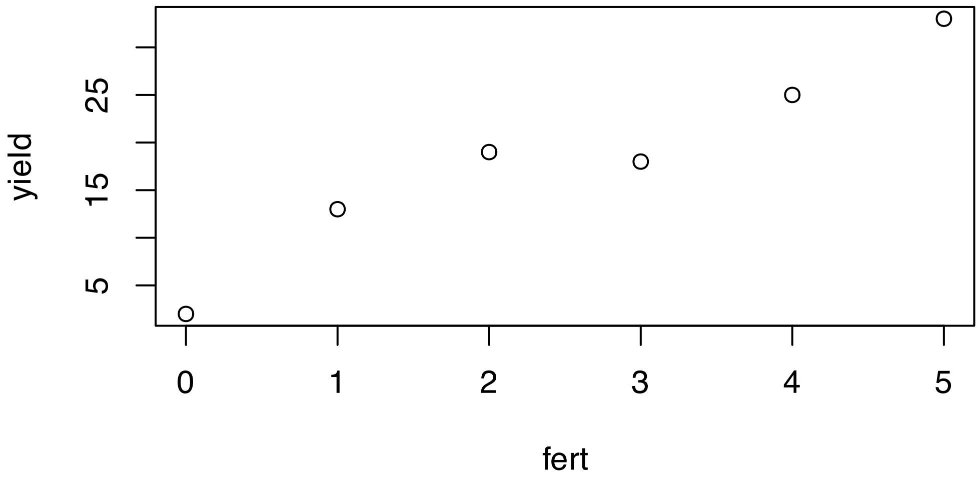
You will use the fertiliser data to fit a linear model in R. As we covered fitting linear models in the Tutorial, it is now your turn to have a go at fitting the models (with some hints along the way).

a) Read the following code into R:

```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?

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



Solution

The points in the scatterplot are showing a linear trend, increasing towards the top-right corner of the plot area.

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
```

Solution

The correlation coefficient is 0.964, indicating there is a strong relationship in the positive direction; i.e. as more fertiliser is applied, the yield increases.

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.

Solution

Intercept = 4.7619 = b_0 fert = 5.4286 = b_1 These coefficients should be the same as those calculated in Excel.

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

Solution

Substituting in parameters from R output:

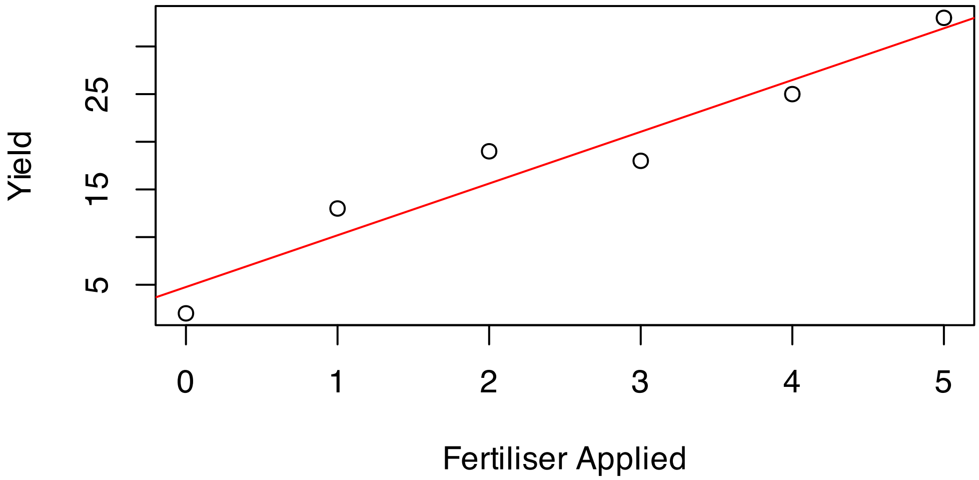
$$yield = 4.7619 + 5.4286 fert$$

This should be the same as the equation obtained 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.

```r
# 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.

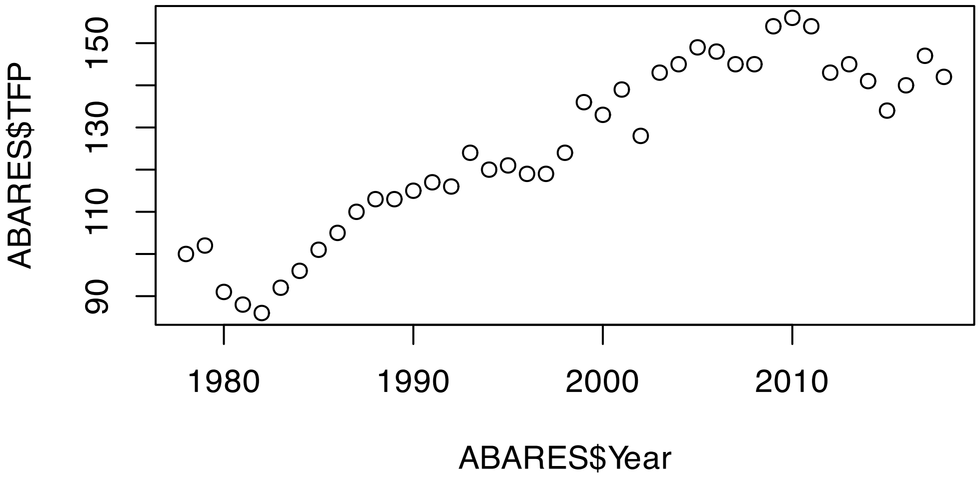
```
library(readxl)

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).

Solution

```
plot(ABARES$Year, ABARES$TFP)
```



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

Solution

There seems to be an overall positive trend in the plot

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

Solution

```
cor(ABARES$Year, ABARES$TFP)
```

```
[1] 0.9166122
```

The correlation coefficient is 0.917, indicating a strong relationship in the positive direction. This means there has been positive growth in the dairy industry over time.

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)?

Solution

```
# Fit linear model
abares.lm <- lm(TFP ~ Year, data = ABARES)

# Model summary
summary(abares.lm)
```

```

Call:
lm(formula = TFP ~ Year, data = ABARES)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.9166  -4.7782   0.9115   6.3601  12.7159 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2998.0129   218.1369  -13.74   <2e-16 ***
Year            1.5632     0.1092   14.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.271 on 39 degrees of freedom
Multiple R-squared:  0.8402,	Adjusted R-squared:  0.8361 
F-statistic:   205 on 1 and 39 DF,  p-value: < 2.2e-16
```

Based on the model summary output, our parameters are:

$b_0 = -2998.0129$ $b_1 = 1.5632$

f) What would the equation for this model be?

Solution

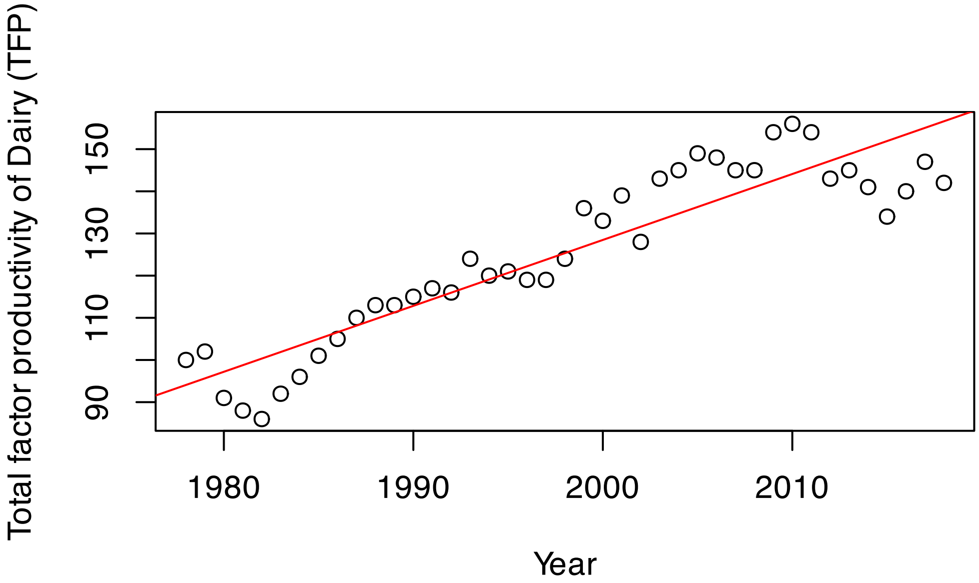
Use the coefficients from above and substitute into lm equation:

$$TFI = -2998.0 + 1.5632Year$$

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).

Solution

```r
plot(ABARES$Year, ABARES$TFP, xlab = "Year", ylab = "Total factor productivity
of Dairy (TFP)")
abline(abares.lm, col = "red")
```



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

Bonus take home exercises

Exercise 1: Cars stopping distance

For this exercise we will use the inbuilt dataset cars to see if there is a relationship between a car's `speed` (mph) and `dist`)stopping distance, fft).

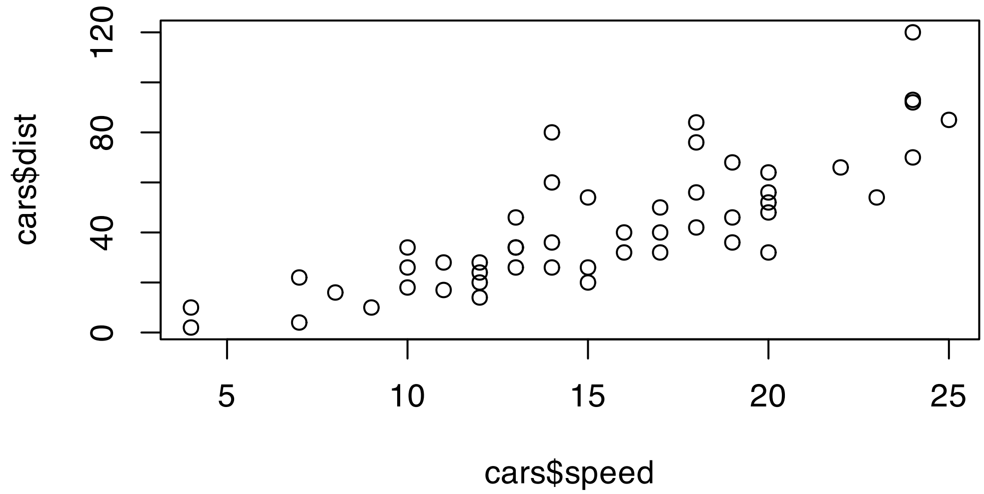
```
head(cars)
```

```
  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 `dist`.

Solution

```
plot(cars$speed, cars$dist)
```



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

Solution

There is a positive relationship between `speed` and `dist`.

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

Solution

```
cor(cars$speed, cars$dist)
```

```
[1] 0.8068949
```

The correlation coefficient is 0.8, which indicates a very strong positive relationship

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

Solution

```
cars_lm <- lm(dist ~ speed, data = cars)

summary(cars_lm)
```

```

Call:
lm(formula = dist ~ speed, data = cars)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$b_0 = -17.5791$ and $b_1 = 3.9324$.

e) What would the equation of the line be?

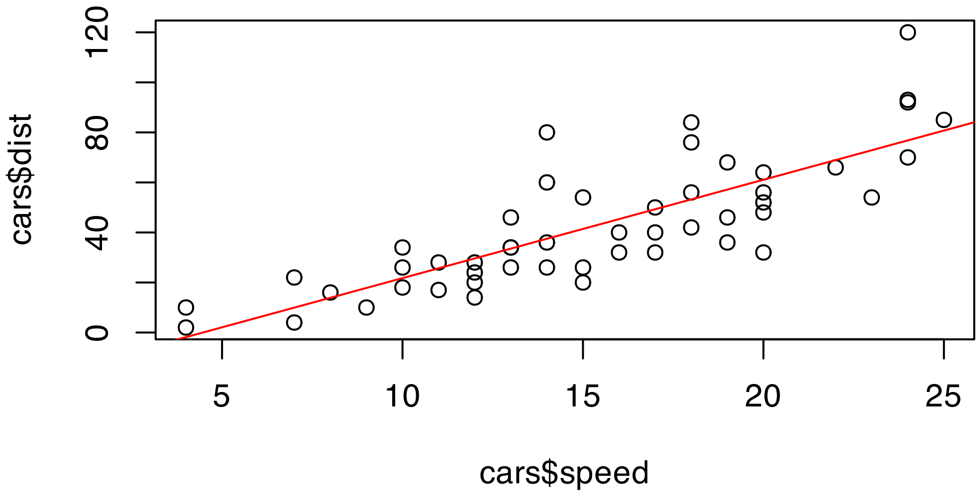
Solution

Distance = 3.9324*Speed −17.5791

f) Overlay your model onto the scatterplot you produced earlier.

Solution

```r
plot(cars$speed, cars$dist)
abline(cars_lm, col = "red")
```



```
### In ggplot
# ggplot(cars, aes(x = speed, y = dist)) +
#   geom_point() +
#   geom_line(y = predict(cars_lm), col = "red")
```

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 (`bill_length_mm`, `flipper_length_mm`).

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

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

head(penguins)
```

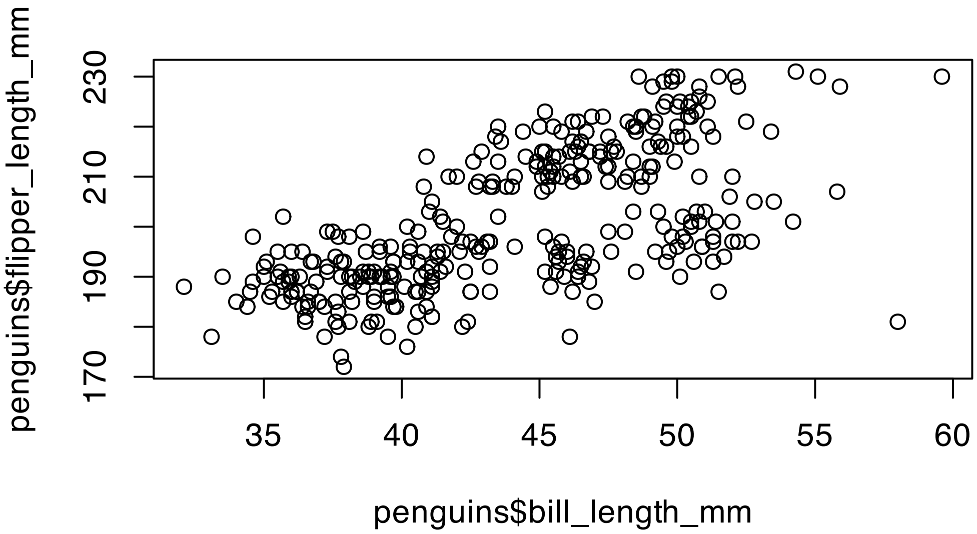
```
# A tibble: 6 × 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           39.1          18.7               181        3750
2 Adelie  Torgersen           39.5          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_mm` vs `flipper_length_mm`.

Solution

```r
plot(penguins$bill_length_mm, penguins$flipper_length_mm)
```



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

Solution

There is a positive trend

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

Solution

```r
cor(penguins$bill_length_mm, penguins$flipper_length_mm)
```

```
[1] 0.6530956
```

The correlation coefficient is 0.65, which indicates a medium positive relationship.

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

Solution

```
penguin_model <- lm(bill_length_mm ~ flipper_length_mm, data = penguins)

summary(penguin_model)
```

```

Call:
lm(formula = bill_length_mm ~ flipper_length_mm, data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.6367 -2.6981 -0.5788  2.0663 19.0953 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -7.21856    3.27175  -2.206    0.028 *  
flipper_length_mm  0.25482    0.01624  15.691   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.148 on 331 degrees of freedom
Multiple R-squared:  0.4265,	Adjusted R-squared:  0.4248 
F-statistic: 246.2 on 1 and 331 DF,  p-value: < 2.2e-16
```

$b_0 = -7.21856$
$b_1 = 0.25482$

e) What would the equation of the linear model be?

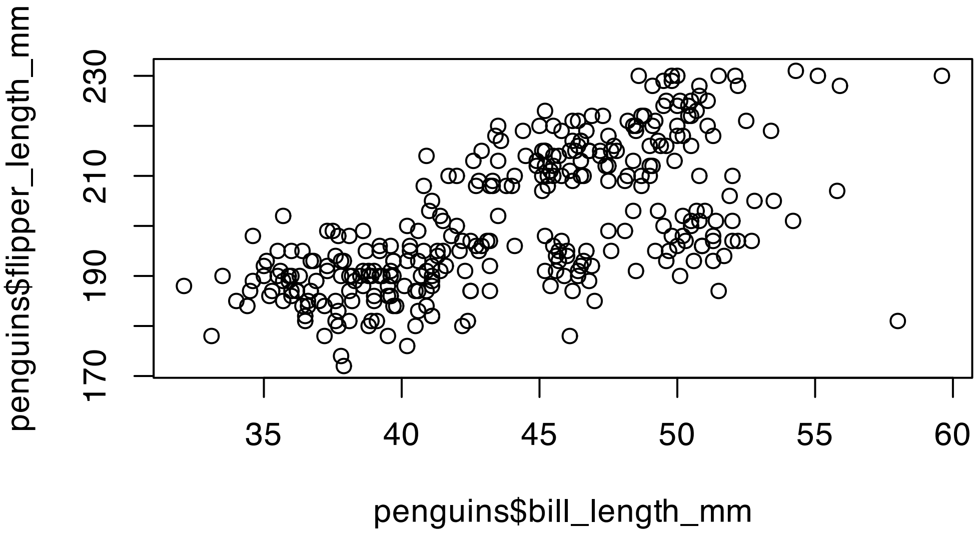
Solution

$ = −7.21856 + 0.25482* $

f) Overlay your model onto the scatterplot you produced earlier.

Solution

```
plot(penguins$bill_length_mm, penguins$flipper_length_mm)
abline(penguin_model, col = "red")
```



Exercise 3: Old Faithful Geyser Data

For this exercise, we will be looking at the relationship between geyser eruption time (`eruptions`) and the time between eruptions (`waiting`), using the inbuilt data set `faithful`.

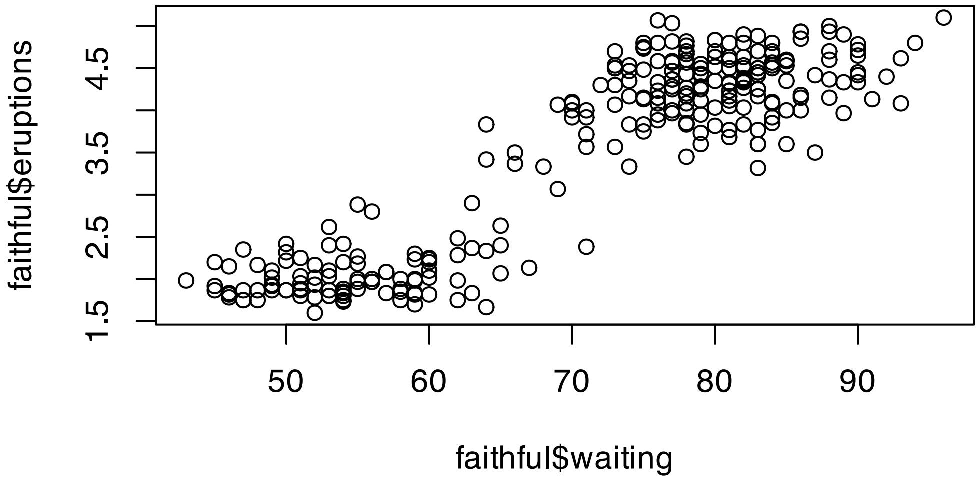
```
head(faithful)
```

```
  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).

Solution

```
plot(faithful$waiting, faithful$eruptions)
```



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

Solution

There is a positive relationship between eruptions length and waiting time.

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

Solution

```
cor(faithful$eruptions, faithful$waiting)
```

```
[1] 0.9008112
```

The correlation coefficient is 0.9, which indicates a very strong positive relationship.

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

Solution

```
faithful_lm <- lm(eruptions ~ waiting, data = faithful)

summary(faithful_lm)
```

```

Call:
lm(formula = eruptions ~ waiting, data = faithful)
```

```

Residuals:
     Min       1Q   Median       3Q      Max 
-1.29917 -0.37689  0.03508  0.34909  1.19329 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,	Adjusted R-squared:  0.8108 
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

$b_0 = -1.874016$ $b_1 = 0.075628$

e) What would the equation of the line be?

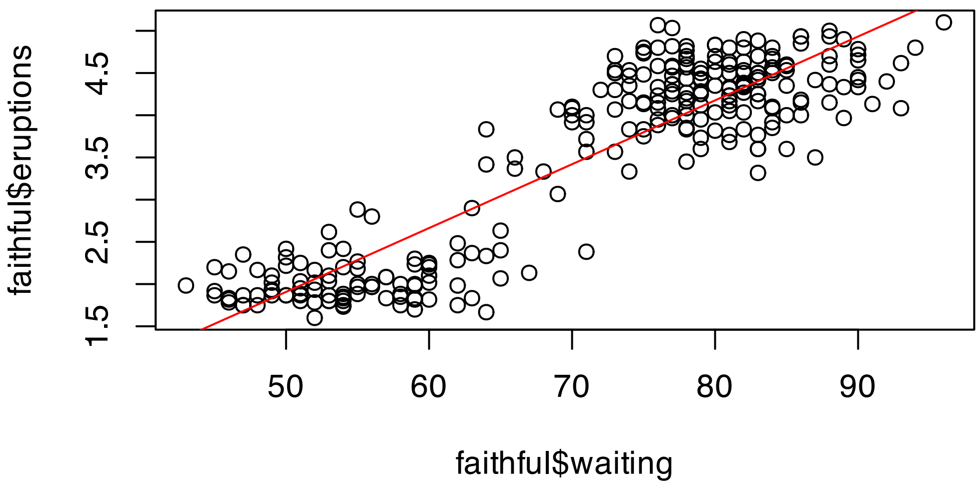
Solution

eruptions = 0.075628* waiting −1.874016

f) Overlay your model onto the scatterplot you produced earlier.

Solution

```r
plot(faithful$waiting, faithful$eruptions)
abline(faithful_lm, col = "red")
```


faithful$eruptions
1.5
2.5
3.5
4.5
50
60
70
80
90
faithful$waiting