Lab 11 - Multiple Linear Regression

ENVX1002 Handbook

Semester 1, 2025

Learning outcomes

- Learn to perform MLR and interpret the results using R;
- Undertake hypothesis testing to determine if model is significant
- Undertake hypothesis testing to determine if the true partial regression slope $\neq 0$
- Check assumptions are filled prior to assessing model output
- Assess model summary in terms of fit and P-values
- Consider more parsimonious models

Before you begin

Create your Quarto document and save it as `Lab-11.Rmd` or similar. The following data files are required:

1) ENVX1002_wk11_practical_data_Regression.xlsx

Last week you explored simple linear regression and assessed the output of your models.

This week we will build upon this and venture into multiple linear regression.

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!

Exercise 1: Corn yields

Data: *Corn* spreadsheet

In this data

- y = P content of corn
- x_1 = inorganic P content of soil
- x_2 = organic P content of soil
- n = 17 sites (The original data had 18 sites, one is removed here.)

Aim of our investigation: Understand the relationship between Organic and Inorganic phosphorous contents in the soil, and the phosphorous content of corn. This will allow us to see which type of phosphorous is being taken up by the corn.

```
library(readxl)
Corn <- read_xlsx("data/ENVX1002_practical_wk11_data_Regression.xlsx", "Corn")
head(Corn)
```

```
# A tibble: 6 × 3
  CornP InorgP  OrgP
  <dbl>  <dbl> <dbl>
1    64    0.4    53
2    60    0.4    23
3    71    3.1    19
4    61    0.6    34
5    54    4.7    24
6    77    1.7    65
```

1.1 Examine the correlations

Some people find it difficult to visually interpret graphical summaries of data in more than 2 dimensions; however, 3-dimensional surface plots are reasonably common in statistics although not usually in descriptive statistics.

Initially we will examine the pairwise correlations to "get a feel" for the data. We will then make a 3-dimensional surface plot using the `lattice` package .

Using R, we can calculate the correlation matrix quite easily.

Note the use of `round()` to limit the number of significant digits.

```
round(cor(Corn),3)
```

```
       CornP InorgP  OrgP
CornP  1.000  0.720 0.212
InorgP 0.720  1.000 0.399
OrgP   0.212  0.399 1.000
```

a) What do the results of the correlation matrix tell you?

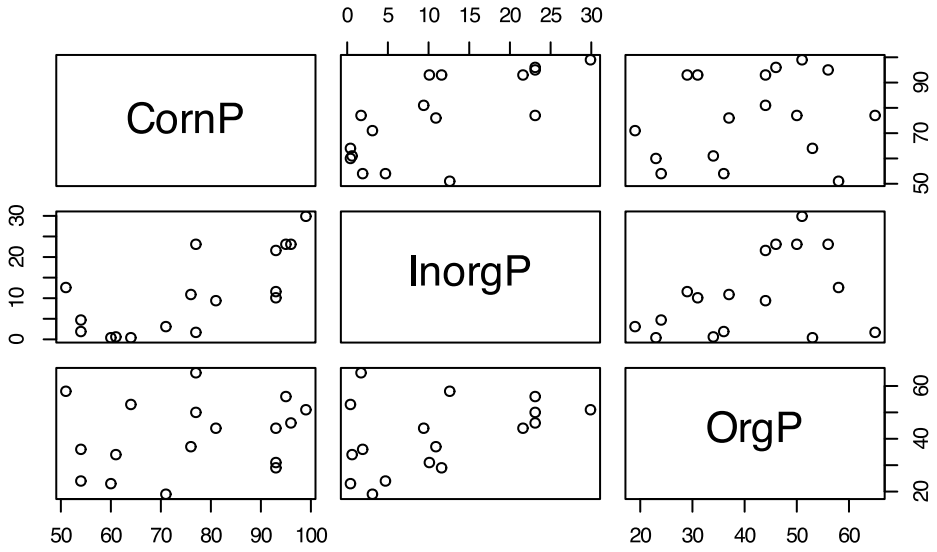
b) Based on the correlation matrix, if we were to fit a single predictor model involving EITHER InorgP OR OrgP, then which model would be more successful?

(Hint, the r^2 is exactly that for a single predictor regression, the square of the correlation, r).

The pairs plot creates scatterplots between each possible pair of variables. Like a single scatterplot the pairs plot allows us to visually observe any trends.

c) Observing the pairs plot below, do you see any strong trends? How well does this link to your correlation matrix?

```
pairs(Corn)
```



Simple 3-D plot

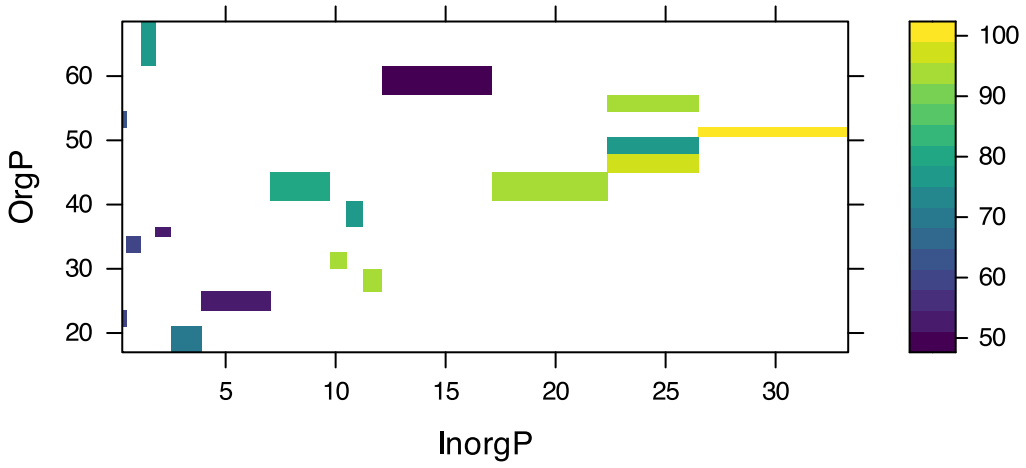
Unlike a simple plot we can creat for a simple linear regression, it is a bit more complex to visualise a model with more predictors. One way we can visualise the relationship is with a 3-D plot, which can be made using the function `levelplot()` in lattice.

Here we plot the OrgP and InorgP in the axes and the levels in the plot are CornP.

Note the package Viridis has been called, this is through personal choice.

The Viridis package has a range of assembled colour ramps which are easier for the reader to differentiate the colours, especially when printed in grayscale, or if the reader is colourblind.

```
library(lattice, quiet = T)
library(viridis, quiet = T) # will need to install.packages("viridis")

levelplot(CornP ~ InorgP + OrgP, data = Corn
          , col.regions = viridis(100))
```



The level plot shows us the x (InorgP) and y variable (OrgP), with the colour scale representing the z variable, which in this case is Phosphorous being taken up by the corn (cornP). From the plot we can see that with higher levels of orgP and inorgP in the soil, the Phosphorous content in the corn is generally higher.

It is clear that the 3-D surface plot does not have colours everywhere, but this relates of course to the underlying data. In this case we don't have continuous data in both directions, so the response (the colour) is only plotted where we have input variables.

If we did have continuous data in both directions the plot would look more like a heatmap, here are some examples.

1.2 Fit the model

We will now use regression to estimate the joint effects of both inorganic phosphorus and organic phosphorus on the phosphorus content of corn.

$CornP = \beta_0 + \beta_1 InorgP + \beta_2 OrgP + error$

This is fairly simple and follows the same structure as simple linear regression and uses `lm()`.

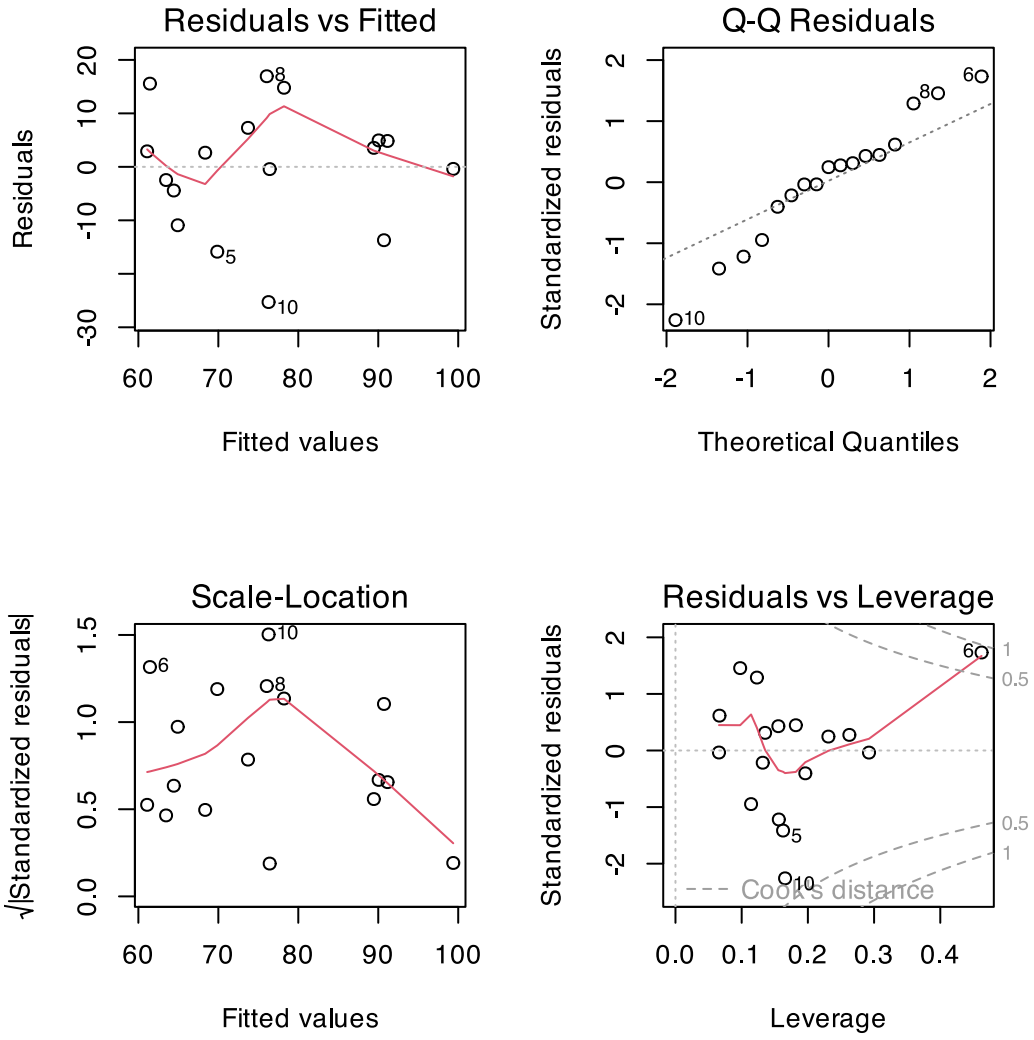
```
MLR.Corn <- lm(CornP ~ InorgP + OrgP, data=Corn)
```

1.3 Check assumptions

Let's check the assumptions of regression are met via residual diagnostics.

a) Are there any apparent problems with normality of CornP residuals or equality of variance for this small data set?

```
par(mfrow=c(2,2))
plot(MLR.Corn)
```



1.4 Model output

After checking our assumptions and we are happy with them, we can interpret our model output.

a) Incorporating the partial regression coefficient estimates, what is the model equation?

```
summary(MLR.Corn)
```

```
Call:
lm(formula = CornP ~ InorgP + OrgP, data = Corn)

Residuals:
    Min      1Q  Median      3Q     Max
-25.282  -4.428   2.645   4.949  16.946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.4654     9.8496   6.748 9.35e-06 ***
InorgP        1.2902     0.3428   3.764  0.00209 **
OrgP         -0.1110     0.2486  -0.447  0.66195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.25 on 14 degrees of freedom
Multiple R-squared:  0.5253,    Adjusted R-squared:  0.4575
F-statistic: 7.746 on 2 and 14 DF,  p-value: 0.005433
```

Simple linear regression allowed us to describe the relationship incorporating our regression coefficient estimate. We would interpret it as follows:

"As x increases by 1, y decreases by b_1 units" (depending on the direction of the relationship).

This week it is a bit different because we are dealing with **partial** regression coefficients instead.

Instead, we would say:

"as x_1 increases by 1, y decreases by b_1 units given all other partial regression coefficients are held constant".

Applied to our model, if we wanted to describe the relationship between InorgP and CornP, we would say:

"As InorgP increases by 1, CornP increases by 1.2902, given OrgP is held constant."

b) Given the above, how would you interpret the relationship between OrgP and CornP?

1.5 Is the model useful?

Remember now that the F-test and T-test are testing slightly different things.

F-test

H_0 : all $\beta_k = 0$, i.e. $\beta_1 = \beta_2 = 0$

H_1 : at least 1 $\beta_k \neq 0$, i.e. our model is significant

We will find the P-value for this test at the end of the summary output.

t-test

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0$

Where β_k refers to one of the model partial regression coefficients. In the case of our model we only have 2: b_1 (InorgP) and b_2 (OrgP).

Now it is your turn:

a) Looking at the `summary()` output, is our overall model significant?

b) Which independent variable is a significant predictor of corn yield?

1.6 How good is the model?

a) How much of the variation in CornP content is explained by the two independent variables?

b) Run the model again but this time with *only* the significant independent variable. How do the model performance criteria (r^2-adj, P-values, Residual Standard Error) change?

1.7 Our conclusions

Writing up conclusions for multiple linear regression are similar to simple linear regression, just with a couple of extra P-values to state.

We would first mention that our overall model is significant as we rejected the null hypothesis (P = 0.05). We could then describe the hypothesis test results for our predictor variables. Finally, we would describe the model fit and our adjusted-r^2.

Remember the scientific conclusion then relates our findings back to the context, answering aims.

a) What would our statistical conclusion be?

b) What would our Scientific conclusion be?

Exercise 2: Water quality

Data: *Microbes* spreadsheet

This exercise will use data from the NOAA Fisheries data portal. The dataset contains the results of microbial study of Pudget Sound, an estuary in Seattle, U.S.A.

The dataset contains the following variables:

- total_bacteria = Total bacteria (cells per ml) –> this will be our response variable
- water_temp = Water temperature (℃)
- carbon_L_h = microbial production (µg carbon per L per hour)
- NO3 = Nitrate (µm)

First thing to do, is read in the data. This time we will be using the *Microbes* spreadsheet:

```
mic      <-      read_xlsx("data/ENVX1002_practical_wk11_data_Regression.xlsx",
"Microbes")
```

```
# Check the structure
str(mic)
```

```
tibble [55 × 4] (S3: tbl_df/tbl/data.frame)
 $ total_bacteria: num [1:55] 498353 529135 529911 603798 632847 ...
 $ water_temp    : num [1:55] 11.2 8.5 8.6 11.2 8.4 8.9 11.3 8.5 9 8.3 ...
 $ carbon_L_h    : num [1:55] 0.286 0.37 0.324 0.287 0.637 ...
 $ NO3           : num [1:55] 21.1 23.5 23.5 21.8 23 ...
```

2.1 Examine the correlations

For this dataset we may expect to see some correlations;

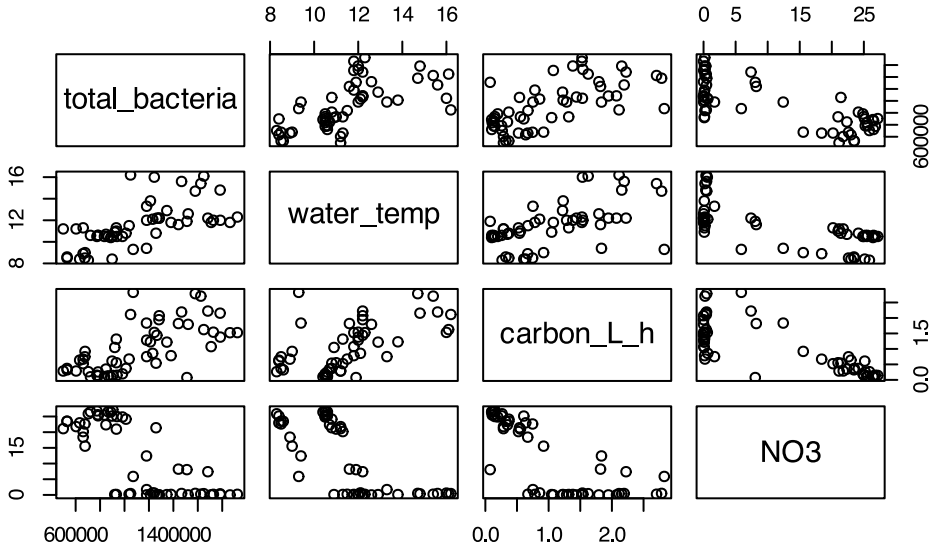
- Warmer water temperature we would expect to see a higher amount of bacterial growth
- Carbon is a proxy for microbial production, so if we see a higher rate of carbon production, we would expect to see higher levels of bacteria
- NO3 (Nitrate) is an essential nutrient for plants and some bacteria species metabolise this

a) Let's test this. Observe the correlation matrix and pairs plots. Do you notice any strong correlations?

```
cor(mic)
```

```
               total_bacteria water_temp carbon_L_h        NO3
total_bacteria      1.0000000  0.6445878  0.6629503 -0.7587849
water_temp          0.6445878  1.0000000  0.5883947 -0.6958635
carbon_L_h          0.6629503  0.5883947  1.0000000 -0.7653497
NO3                -0.7587849 -0.6958635 -0.7653497  1.0000000
```

```
pairs(mic)
```



2.2 Fit the model

We can now fit the model to see how much these predictors account for the variation in total bacteria.

$$totalbacteria = \beta_0 + \beta_1 watertemp + \beta_2 carbon + \beta_3 NO3 + error$$

There are two forms the lm code can take; you can either specify which variables you want to include by naming each one, or if only your desired variables are within your dataset, you can use the ~. to specify all columns.

```
names(mic) # tells us column names within the dataset
```

```
[1] "total_bacteria" "water_temp"     "carbon_L_h"     "NO3"
```

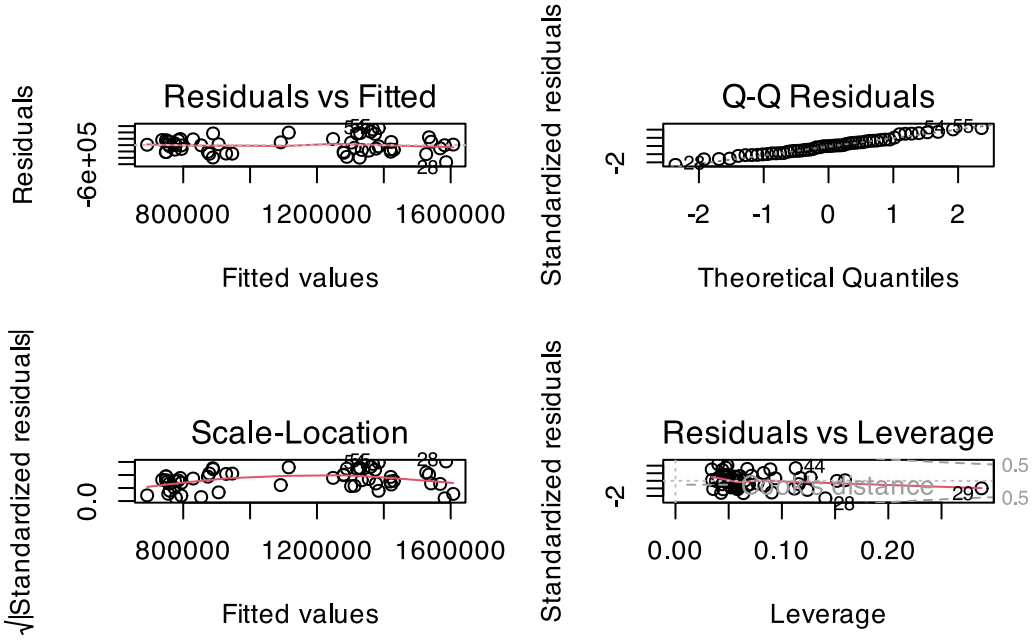
```
# Form 1:
mic.lm <- lm(total_bacteria ~ water_temp + carbon_L_h + NO3, data = mic)

# Form 2
mic.lm <- lm(total_bacteria ~ ., data = mic)
```

2.3 Check assumptions

Let's check the assumptions of regression are met via residual diagnostics.

a) Are there any apparent problems with normality of total_bacteria residuals or equality of variance for this data set?

```
par(mfrow=c(2,2))
plot(mic.lm)
```



2.4 Model output

After investigating the assumptions, they seem to be ok, so we can move onto the model summary.

```
summary(mic.lm)
```

```
Call:
lm(formula = total_bacteria ~ ., data = mic)

Residuals:
    Min      1Q  Median      3Q     Max
-536948 -191145   -2584  154144  539394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   766294     314957   2.433  0.01852 *
water_temp     40051      23547   1.701  0.09505 .
carbon_L_h     84425      67427   1.252  0.21624
NO3           -16586       5256  -3.156  0.00268 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 250500 on 51 degrees of freedom
Multiple R-squared:  0.614, Adjusted R-squared:  0.5913
F-statistic: 27.04 on 3 and 51 DF,  p-value: 1.318e-10
```

a) Incorporating the partial regression coefficient estimates, what is the equation for this model?

b) Like you did in Exercise 1.4, how would you interpret the relationship between total_bacteria and water_temp?

2.5 Is the model useful?

a) Observing the P-value of the F-statistic in the summary, can we say our model is significant?

b) Are any predictors significant?

2.6 How good is the model?

a) How much of the variation in total bacteria is explained by the three independent variables?

b) Run the model again but this time excluding the variable with the largest P-value. How do the model performance criteria (r^2-adj, P-values, Residual Standard Error) change?

```
summary(lm(total_bacteria ~ water_temp + NO3, data = mic))
```

```

Call:
lm(formula = total_bacteria ~ water_temp + NO3, data = mic)

Residuals:
    Min      1Q  Median      3Q     Max
-506706 -203093  -11950  157707  533145

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   858466     307901   2.788  0.00739 **
water_temp     43611      23502   1.856  0.06917 .
NO3           -20620       4175  -4.938 8.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 251900 on 52 degrees of freedom
Multiple R-squared:  0.6021,    Adjusted R-squared:  0.5868
F-statistic: 39.34 on 2 and 52 DF,  p-value: 3.927e-11
```

2.7 Conclusions

a) What would the statistical conclusion be for this model?

b) What would our scientific conclusion be?

Exercise 3: Dippers

Data: *Dippers* spreadsheet

We will revisit the Dippers dataset from last week, but now incorporating other factors which may be influencing the distribution.

The file, Breeding density of dippers, gives data from a biological survey which examined the nature of the variables thought to influence the breeding of British dippers.

Dippers are thrush-sized birds living mainly in the upper reaches of rivers, which feed on benthic invertebrates by probing the river beds with their beaks.

Twenty-two sites were included in the survey. Variables are as follows

- water hardness
- river-bed slope
- the numbers of caddis fly larvae
- the numbers of stonefly larvae
- the number of breeding pairs of dippers per 10 km of river

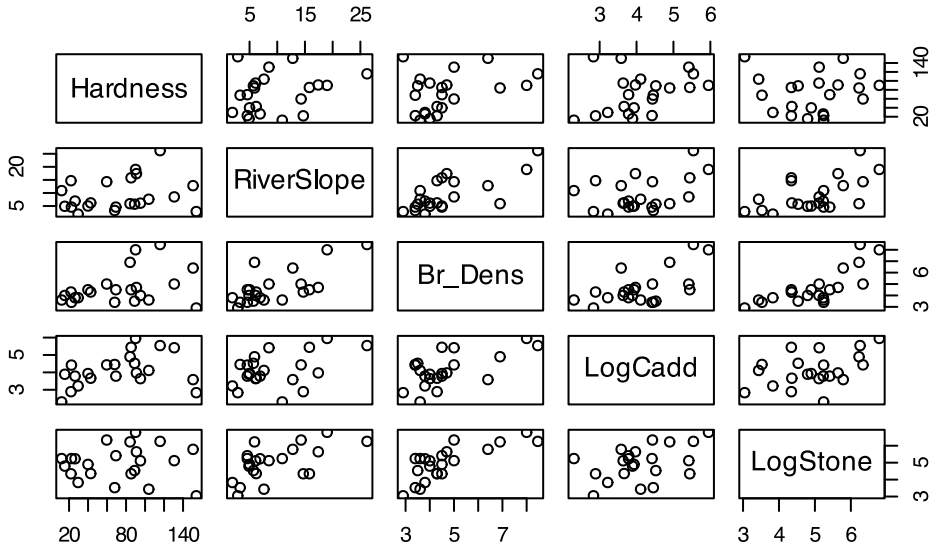
In the analyses, the four invertebrate variables were transformed using a Log(Number+1) transformation.

Now it is your turn to work through the steps as above. What other factors are influencing the number of breeding pairs of Dippers?

a) Read in the data from today's Excel sheet, the corresponding sheet name is "Dippers"

b) Investigate a correlation matrix and pairs plot of the dataset, are there signs of a relationship between breeding pair density and other independent variables?

```
pairs(Dippers)
```

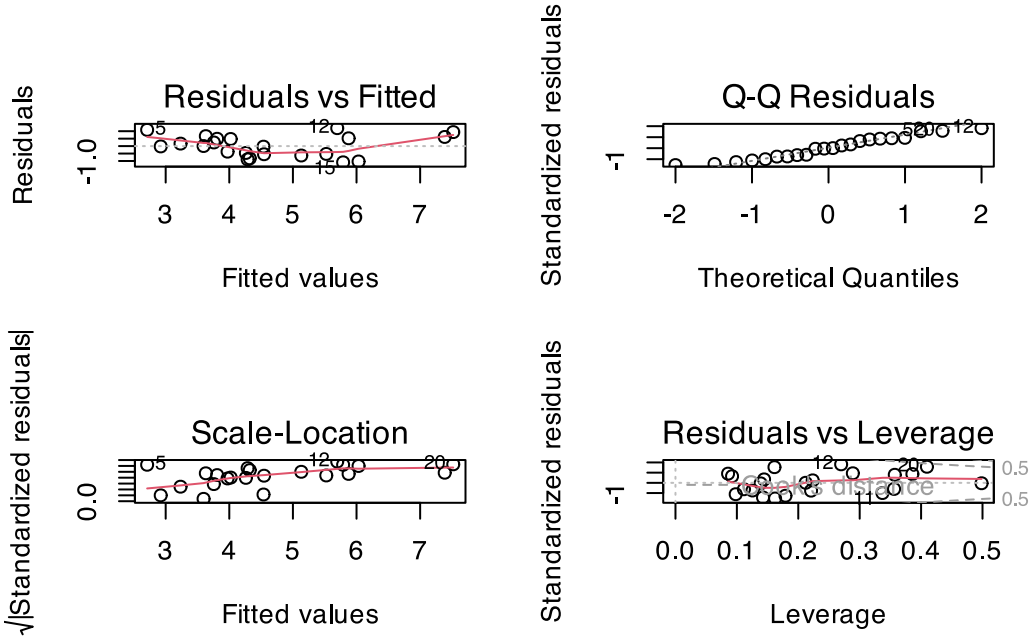



```
cor(Dippers)
```

```
             Hardness RiverSlope   Br_Dens   LogCadd   LogStone
Hardness   1.00000000  0.2420874 0.3503482 0.3337586 0.02951218
RiverSlope 0.24208735  1.0000000 0.7102216 0.4313081 0.57479242
Br_Dens    0.35034817  0.7102216 1.0000000 0.6126891 0.76294367
LogCadd    0.33375857  0.4313081 0.6126891 1.0000000 0.44347440
LogStone   0.02951218  0.5747924 0.7629437 0.4434744 1.00000000
```

c) Let's investigate further. Run the model incorporating all of our predictors, but before looking at our model output, are the assumptions ok?

```
# Run model
dipper.lm <- lm(Br_Dens ~ ., data=Dippers)

# Check assumptions
par(mfrow = c(2,2))
plot(dipper.lm)
```



Once you are happy assumptions are good, you can interpret the model output using `summary()`.

d) What is the equation for our model, incorporating our partial regression coefficients?

e) Based on the F-statistic output, is the model significant? How can we tell? Is it different to the significance of LogCadd this time?

f) Is LogCadd still a significant predictor of Dipper breeding pair density?

g) What are the significant predictors of this model?

h) How good is the fit of our model?

i) What might we do to improve the model fit?

j) What statistical and scientific conclusions can we make from this model output?

Another week of linear models done! Great work fitting Multiple Linear Regression! Next week we break away and explore non-linear functions.

Bonus Take Home Exercises

Use the template below to test the hypotheses for each exercise.

1. Scatterplot and correlation
2. Fit the model
3. Check assumptions
4. P value and model fit

a) Is the model significant?

b) Are the predictors significant?

c) How good is the model fit?

5. Conclusions

Exercise 1: House Prices

Data: - Housing

This exercise will use a dataset from kaggleto explore the variables affecting house prices.

Exercise 2: Energy use

Data:

- energy

Use the dataset from kaggle to explore which variables affect energy consumption.

Exercise 3: Sales vs advertising budget

Data:

- advertising budget

Use the dataset from kaggle to explore how different advertising budgets affect sales.