THE UNIVERSITY OF SYDNEY

ENVX2001 LAB
07 - REGRESSION MODEL DEVELOPMENT

Table of contents

> **Tip**
>
> Please work on this exercise by creating your own R Markdown file.

Exercise 1: Bird abundance

This is the *same* dataset used in the lecture.

Fragmentation of forest habitat has an impact of wildlife abundance. This study looked at the relationship between bird abundance (bird ha^{-1}) and the characteristics of forest patches at 56 locations in SE Victoria.

The predictor variables are:

- `ALT` Altitude (m)
- `YR.ISOL` Year when the patch was isolated (years)
- `GRAZE` Grazing (coded 1-5 which is light to heavy)
- `AREA` Patch area (ha)
- `DIST` Distance to nearest patch (km)
- `LDIST` Distance to largest patch (km)

Import the data from the "Loyn" tab in the MS Excel file.

```
library(readxl)
loyn <- read_xlsx("mlr.xlsx", "Loyn")
```

Often, the first step in model development is to examine the data. This is a good way to get a feel for the data and to identify any issues that may need to be addressed. In this case, we will examine the data using histograms and a correlation matrix.

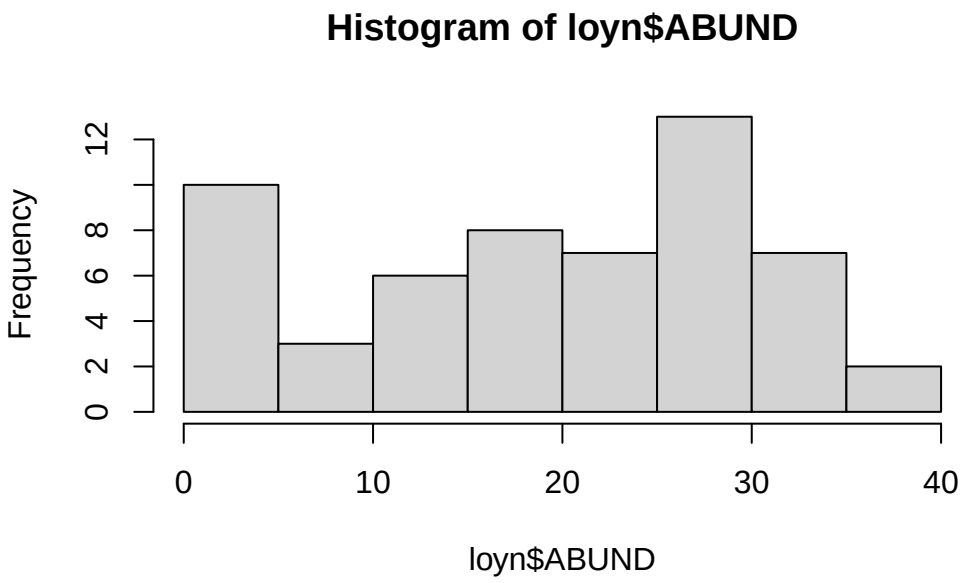
Histograms

There are a breadth of ways to create histograms in R. In each tab below you will find some different ways to create the same plot outputs.

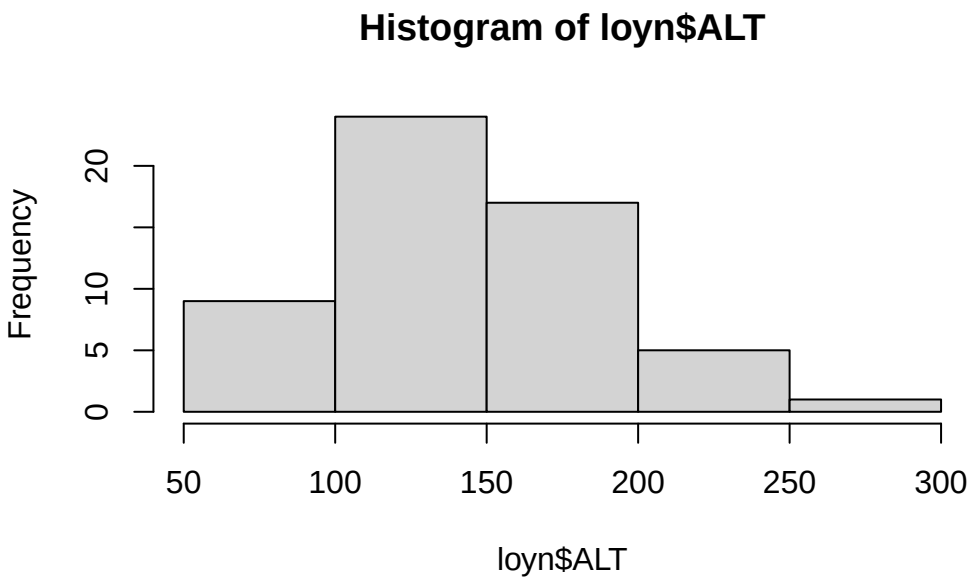
hist()

This is a straightforward way to create multiple histograms with `hist()`. The `par()` function is used to arrange the plots on the page. The `mfrow` argument specifies the number of rows and columns of plots.

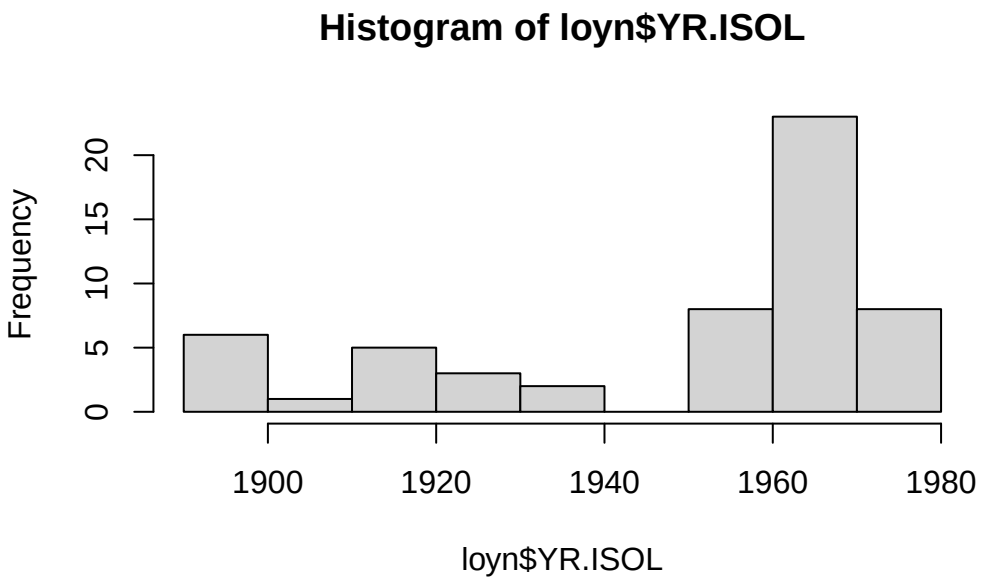
```
#par(mfrow=c(3,3))
hist(loyn$ABUND)
```



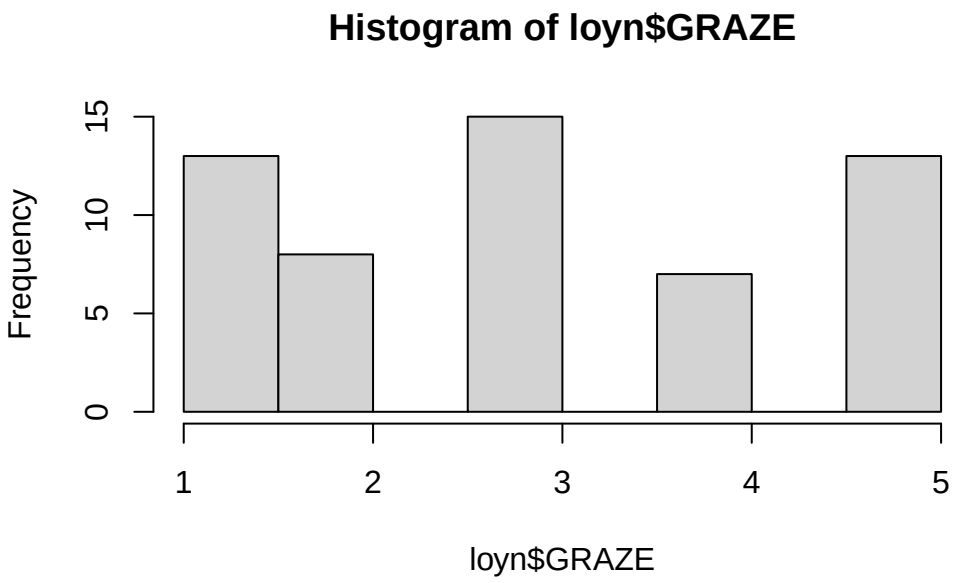

```
hist(loyn$ALT)
```



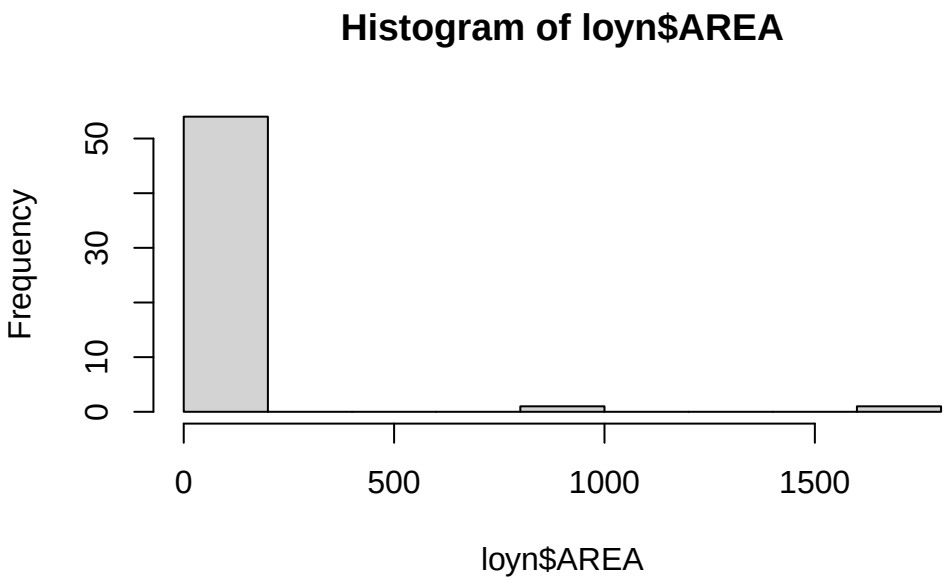

```
hist(loyn$YR.ISOL)
```



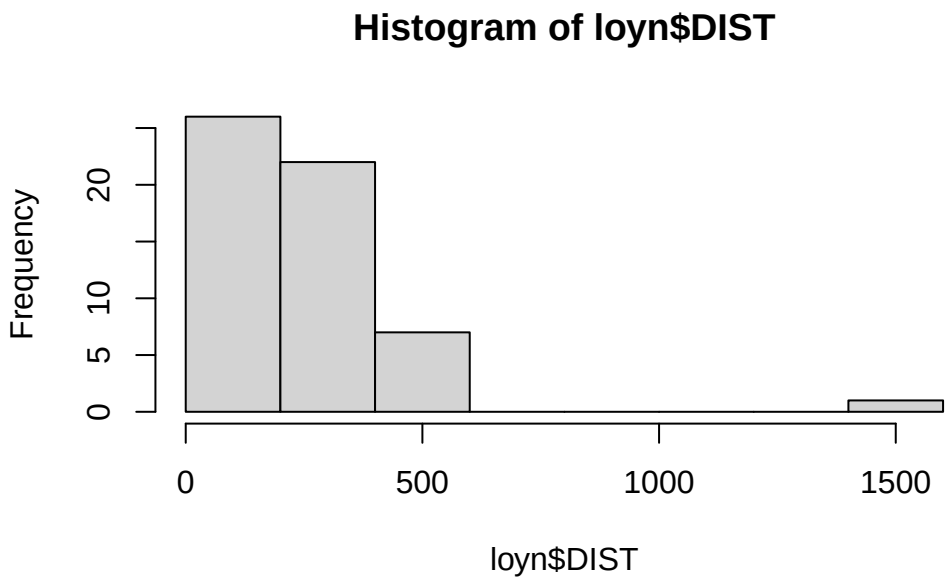

```
hist(loyn$GRAZE)
```



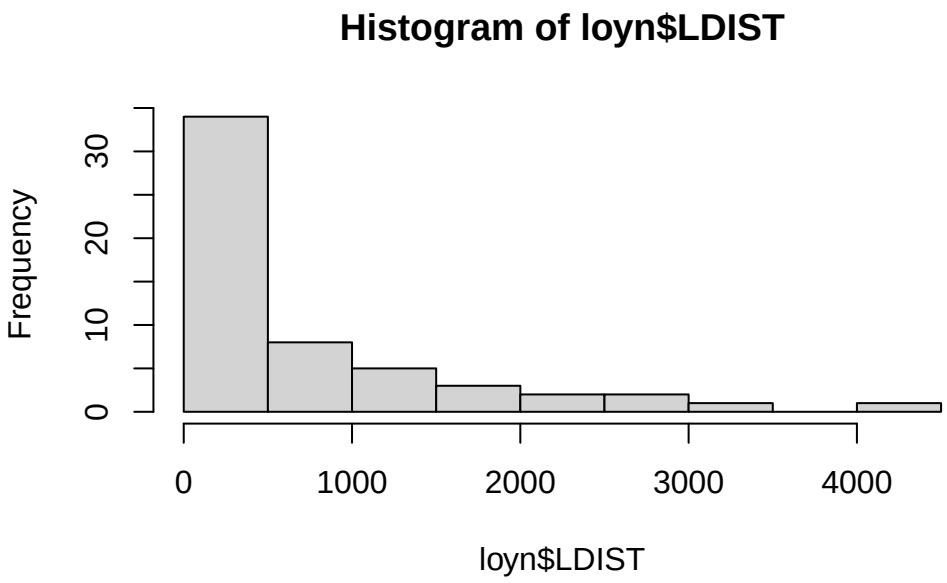

```
hist(loyn$AREA)
```




```
hist(loyn$DIST)
```




```
hist(loyn$LDIST)
```




```
#par(mfrow=c(1,1))
```

`hist.data.frame()` from `Hmisc`

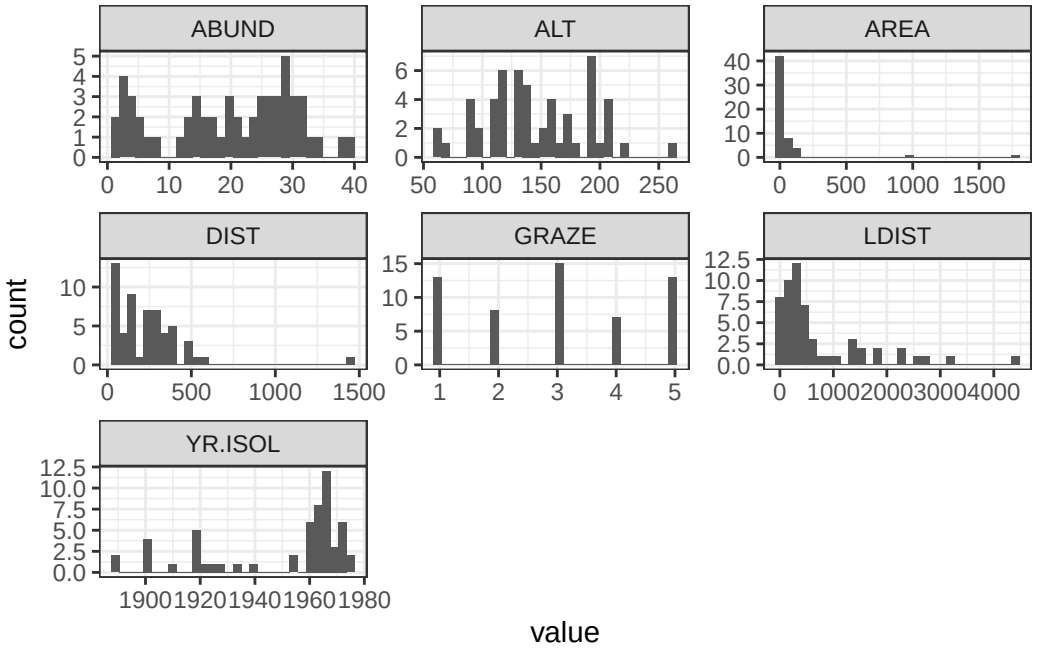
The `Hmisc` package provides a function `hist.data.frame()` that can be used to create multiple histograms, which can be called by simply using `hist()`. You may need to tweak the `nclass` argument to get the desired number of bins, as the default may not look appropriate.

```
# install.packages("Hmisc")
library(Hmisc)
hist(loyn, nclass = 50)
```

`ggplot()`

A more modern approach is to use `ggplot()` with `facet_wrap()` to arrange multiple plots on a single page. To do this, the `pivot_longer()` function from the `tidyr` package is used to reshape the data into a tidy format.

```r
# tidy the data
loyn_tidy <- pivot_longer(loyn, cols = everything())

# plot
ggplot(loyn_tidy, aes(x = value)) +
  geom_histogram() +
  facet_wrap(~name, scales = "free") +
  theme_bw()
```

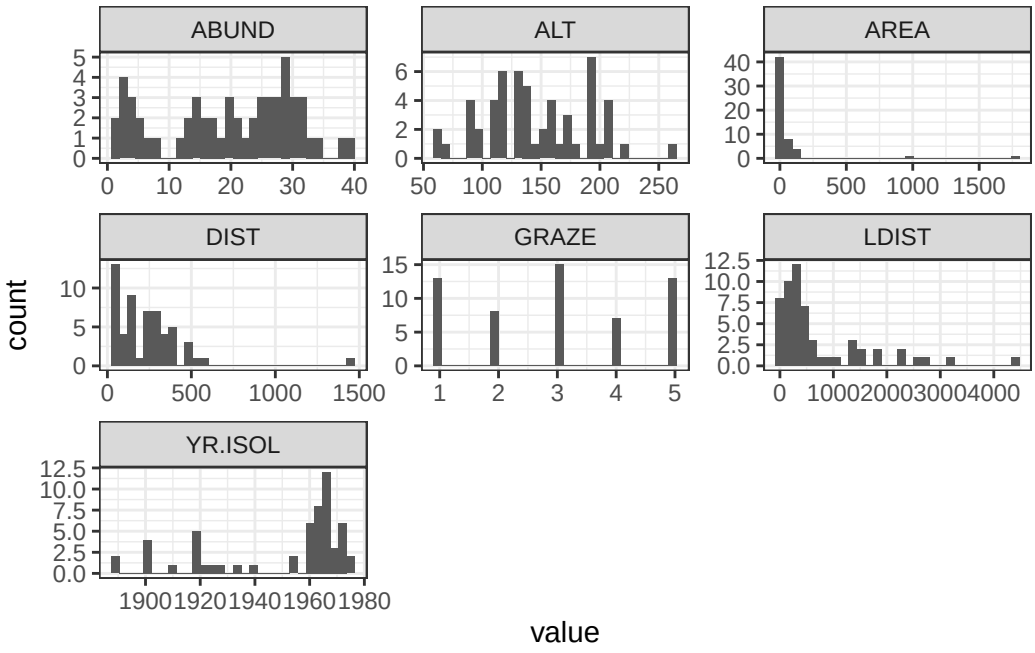


`ggplot()` with `dplyr`

Here we use the pipe operator `%>%` from `dplyr` to chain together a series of commands. The pipe operator takes the output of the command on the left and passes it to the command on the right (or below) the pipe. This means that we can create a series of commands that are executed in order.

```r
loyn %>%
  pivot_longer(cols = everything()) %>%
```

```
ggplot(aes(x = value)) +
geom_histogram() +
facet_wrap(~name, scales = "free") +
theme_bw()
```



Question 1

Comment on the histograms in terms of leverage. *Hint: what is the relationship between leverage and skewness?*

Answer

The histograms of `AREA`, `DIST` and `LDIST` are very skewed. The high values would have high leverage, this means that these would cause the residuals to be skewed. These would be candidates for transformation.

Correlation matrix

Calculate the correlation matrix using `cor(Loyn)`.

```
cor(loyn)
```

```
              ABUND         AREA      YR.ISOL       DIST       LDIST
ABUND    1.00000000  0.255970206  0.503357741  0.2361125  0.08715258
AREA     0.25597021  1.000000000 -0.001494192  0.1083429  0.03458035
YR.ISOL  0.50335774 -0.001494192  1.000000000  0.1132175 -0.08331686
DIST     0.23611248  0.108342870  0.113217524  1.0000000  0.31717234
LDIST    0.08715258  0.034580346 -0.083316857  0.3171723  1.00000000
GRAZE   -0.68251138 -0.310402417 -0.635567104 -0.2558418 -0.02800944
ALT      0.38583617  0.387753885  0.232715406 -0.1101125 -0.30602220
               GRAZE         ALT
ABUND   -0.68251138  0.3858362
AREA    -0.31040242  0.3877539
YR.ISOL -0.63556710  0.2327154
DIST    -0.25584182 -0.1101125
LDIST   -0.02800944 -0.3060222
GRAZE    1.00000000 -0.4071671
ALT     -0.40716705  1.0000000
```

> ⚠ Question 2
>
> Which independent variables are useful for predicting the dependent variable abundance? Is there evidence for multi-collinearity?

Answer

Some of the predictors are useful, but AREA has a low r.

The correlation between `GRAZE` and `YR.ISOL` is quite high (r = -0.63556710), suggesting multi-collinearity which may influence the model. If the relationship between these two variables was stronger, we would remove one of the variables to prevent this collinearity from affecting the model.

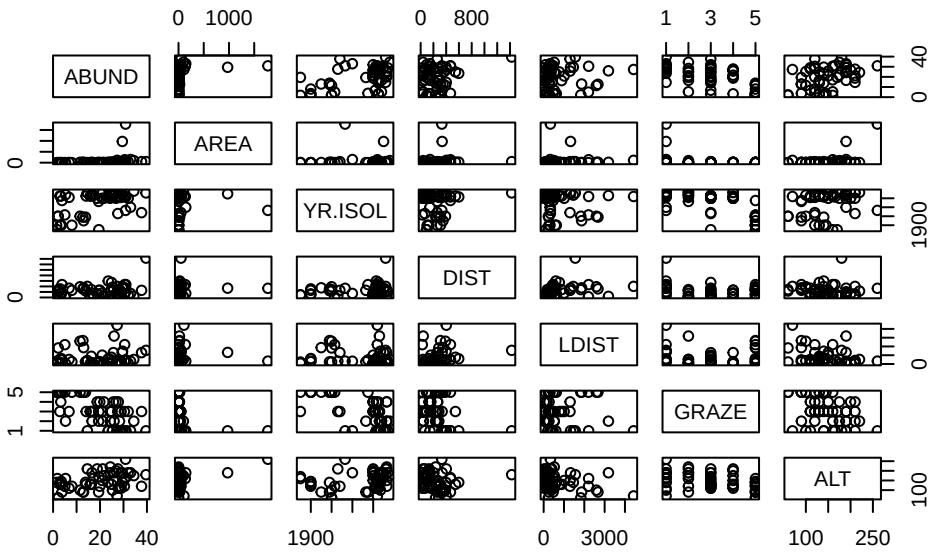
Note: For more information on collinearity and how it may impact the model, see Quinn & Keough p 127.

Plotting correlation

Examine correlations visually using `pairs()` or `corrplot()` from the `corrplot` package.

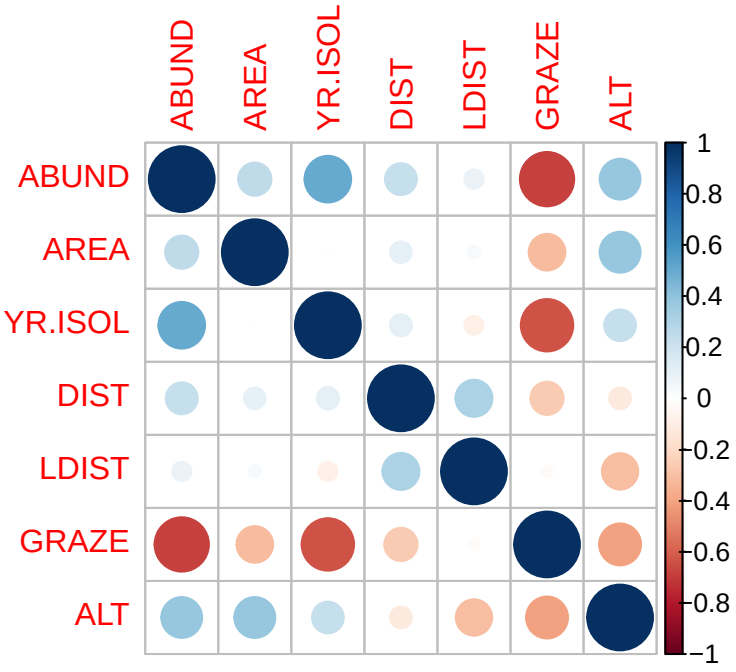
Scatterplot matrix

```
pairs(loyn)
```



Correlation matrix

```
library(corrplot)
corrplot(cor(loyn))
```



> **Question 3**
>
> Are there any trends visible from the plots?

Answer

Not really; the pairs plot reflects the strength of the linear relationship between each of the variables. There may be some stronger relationships occurring, but it is evident a few of the variables are skewed so it is harder to distinguish within the plots.

> **Tip**
>
> We can also bring in variance inflation factors (VIF) to help us identify multi-collinearity, but that is done only after we have selected a model.

Transformations

The AREA predictor has a small number of observations with very large values. Apply a $\log_{10}$ transformation and label the new variable `Loyn$L10AREA`.

```
loyn$L10AREA <- log10(loyn$AREA)
```

> ⚠ Question 4
>
> Why are we transforming AREA?

Answer

You do this to stabilise the variance of the regression to manage the leverage of the outliers in the variable. This reduces the skew. L10AREA is more likely to be a significant predictor.

> ⚠ Question 5
>
> Re-run `pairs(Loyn)` and create a histogram using the transformed value of AREA, how do the plots look?

```
hist(loyn$L10AREA)
pairs(loyn)
```

Answer

- Histogram looks better, less skewed
- Pairs plot shows a trend between ABUND and L10AREA

> ⚠ Question 6
>
> In preparation for modelling, transform the remaining skewed variables, DIST and LDIST the same way you did for AREA and examine the histogram and pairs plots using these new variables.

Make sure you end up with two new variables labelled `loyn$L10DIST` and `loyn$L10LDIST`.

Answer

Histogram for both look better, less skewed Pairs plot shows potential trend between ABUND and L10DIST, and ABUND with L10LDIST compared to untransformed DIST and LDIST

Exercise 2: Modelling bird abundance

We will now use the transformed data in `loyn` for this exercise. If you have not already figured out how to perform the transformation, or if something is wrong, you may use the `loyn` tab in the `mlr.xlsx` MS Excel document. Alternatively, the code to convert the data is below.

```
# reset the data import just in case it has been modified
loyn <- read_xlsx("mlr.xlsx", "Loyn")
# make transformations

loyn <- loyn %>%
  mutate(L10AREA = log10(AREA),
    L10DIST = log10(DIST),
    L10LDIST = log10(LDIST))

# check
glimpse(loyn)
```

```
Rows: 56
Columns: 10
$ ABUND    <dbl> 5.3, 2.0, 1.5, 17.1, 13.8, 14.1, 3.8, 2.2, 3.3, 3.0, 27.6, 1.~
$ AREA     <dbl> 0.1, 0.5, 0.5, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 2.0, 2~
$ YR.ISOL  <dbl> 1968, 1920, 1900, 1966, 1918, 1965, 1955, 1920, 1965, 1900, 1~
$ DIST     <dbl> 39, 234, 104, 66, 246, 234, 467, 284, 156, 311, 66, 93, 39, 4~
$ LDIST    <dbl> 39, 234, 311, 66, 246, 285, 467, 1829, 156, 571, 332, 93, 39,~
$ GRAZE    <dbl> 2, 5, 5, 3, 5, 3, 5, 5, 4, 5, 3, 5, 2, 1, 5, 5, 3, 3, 3, 2, 2~
$ ALT      <dbl> 160, 60, 140, 160, 140, 130, 90, 60, 130, 130, 210, 160, 210,~
$ L10AREA  <dbl> -1.0000000, -0.3010300, -0.3010300, 0.0000000, 0.0000000, 0.0~
$ L10DIST  <dbl> 1.591065, 2.369216, 2.017033, 1.819544, 2.390935, 2.369216, 2~
$ L10LDIST <dbl> 1.591065, 2.369216, 2.492760, 1.819544, 2.390935, 2.454845, 2~
```

THE UNIVERSITY OF SYDNEY

Best single predictor?

> **Question 1**
>
> Obtain the correlation between ABUND and all of the predictor variables using `cor()`. Based on these, what would you expect to be the best single predictor of ABUND?

```
cor(loyn)
```

Answer

```
cor(loyn)
```

```
              ABUND         AREA      YR.ISOL       DIST       LDIST
ABUND     1.00000000  0.255970206  0.503357741  0.2361125  0.08715258
AREA      0.25597021  1.000000000 -0.001494192  0.1083429  0.03458035
YR.ISOL   0.50335774 -0.001494192  1.000000000  0.1132175 -0.08331686
DIST      0.23611248  0.108342870  0.113217524  1.0000000  0.31717234
LDIST     0.08715258  0.034580346 -0.083316857  0.3171723  1.00000000
GRAZE    -0.68251138 -0.310402417 -0.635567104 -0.2558418 -0.02800944
ALT       0.38583617  0.387753885  0.232715406 -0.1101125 -0.30602220
L10AREA   0.74003580  0.584651024  0.278414517  0.3047850  0.33680642
L10DIST   0.12672333  0.163054319 -0.019572228  0.8233190  0.29365797
L10LDIST  0.11812448  0.101607829 -0.161116108  0.4968169  0.82059568
               GRAZE        ALT    L10AREA     L10DIST    L10LDIST
ABUND    -0.68251138  0.3858362  0.7400358  0.12672333  0.11812448
AREA     -0.31040242  0.3877539  0.5846510  0.16305432  0.10160783
YR.ISOL  -0.63556710  0.2327154  0.2784145 -0.01957223 -0.16111611
DIST     -0.25584182 -0.1101125  0.3047850  0.82331904  0.49681692
LDIST    -0.02800944 -0.3060222  0.3368064  0.29365797  0.82059568
GRAZE     1.00000000 -0.4071671 -0.5590886 -0.14263922 -0.03399082
ALT      -0.40716705  1.0000000  0.2751428 -0.21900701 -0.27404380
L10AREA  -0.55908864  0.2751428  1.0000000  0.30216662  0.38247952
L10DIST  -0.14263922 -0.2190070  0.3021666  1.00000000  0.60386637
L10LDIST -0.03399082 -0.2740438  0.3824795  0.60386637  1.00000000
```

The best single predictor would be `L10AREA` as this has the highest r (r = 0.74)

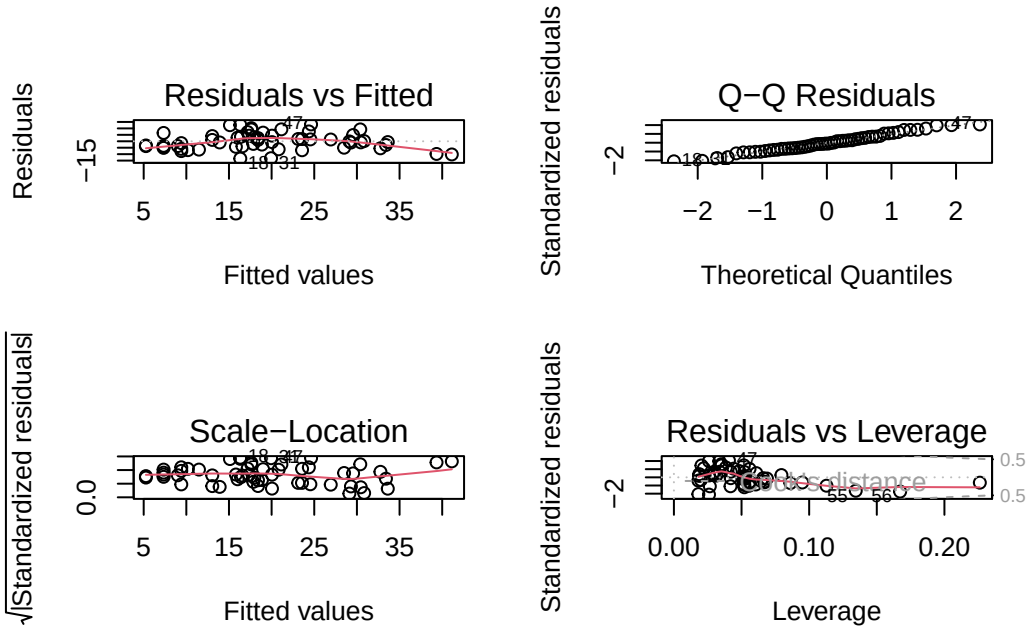
Assumptions and interpretation

> **Question 2**
>
> Use multiple linear regression to see whether ABUND can be predicted from L10AREA and GRAZE. Are the assumptions met? Is there a significant relationship? *Note: we are using these 2 predictors as they have the largest absolute correlations. Use `lm()` and specify the model as `ABUND ~ L10AREA + GRAZE`.*

```
lm.mod1 <- lm(ABUND~GRAZE + L10AREA, data=loyn)

par(mfrow=c(2,2))
plot(lm.mod1)
par(mfrow=c(1,1))

summary(lm.mod1)
```

Answer

```
lm.mod1 <- lm(ABUND~GRAZE + L10AREA, data=loyn)

par(mfrow=c(2,2))
plot(lm.mod1)
```




```r
par(mfrow=c(1,1))

summary(lm.mod1)
```

```
Call:
lm(formula = ABUND ~ GRAZE + L10AREA, data = loyn)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4296  -4.3186  -0.6323   4.1273  13.0739

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.6029     3.0917   6.987 4.73e-09 ***
GRAZE        -2.8535     0.7125  -4.005 0.000195 ***
L10AREA       6.8901     1.2900   5.341 1.98e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.444 on 53 degrees of freedom
Multiple R-squared:  0.6527,    Adjusted R-squared:  0.6396
F-statistic: 49.81 on 2 and 53 DF,  p-value: 6.723e-13
```

This is a significant model as both b1 and b2 are significant and the model is significant.

The residuals look reasonable. They are approximately normally distributed (both right hand plots), but possibly the variance is not totally constant and there are possibly a few values with high leverage (left hand plots).

> ⚠ Question 3
>
> How good is the model based on the (i) r^2 (ii) adjusted r^2? Use `summary()`.

```
summary(lm.mod1)$r.squared
summary(lm.mod1)$adj.r.squared
```

Answer

The Adjusted r^2 is lower than the r^2, but we would opt for the adjusted r^2 as it takes the number of predictors into account. Overall the model is ok, explaining 64.0% of variation in Abundance.

> ⚠ Question 4
>
> Which variable(s) has the most significant effect(s)? *(Refer specifically to the t probabilities in the table of predictors and their estimated parameters or coefficients in the output of* `summary()`*).* Interpret the p-values in terms of dropping predictor variables.

Answer

Both `L10AREA` and `GRAZE` are highly significant, `L10AREA` is the most significant. In terms of effect, a 1 unit change in `GRAZE` results in a -2.9 decrease in abundance (with `L10AREA` remaining constant), while a 1 unit change in `L10AREA`, (therefore a 10 unit change in `AREA`) results in a 6.9 increase in abundance (`GRAZE` holding constant).

Question 5

Repeat the multiple regression, but this time include YRS.ISOL as a predictor variable (it has the 3rd largest absolute correlation). This will allow you to assess the effect of YRS.ISOL with the other variables taken into account.

Answer

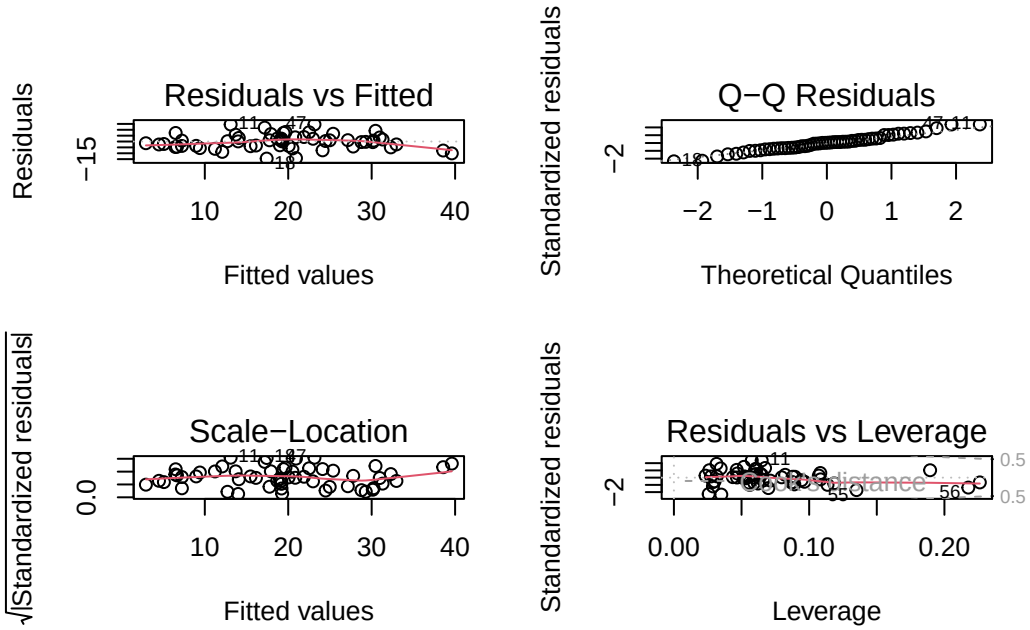
```
lm.mod2 <- lm(ABUND ~ GRAZE + L10AREA + YR.ISOL, data=loyn)
```

Question 6

Check assumptions, do the residuals look ok? If you are happy with the assumptions, you can proceed to interpret the model output.

Answer

```
par(mfrow=c(2,2))
plot(lm.mod2)
```




```
par(mfrow=c(1,1))
```

Question 7

Compare the r^2 and adjusted r^2 values with those you calculated for the 2 predictor model, Which is the better model? Why?

```
summary(lm.mod2)
```

Answer

Both of these are greater than for model in step 3, so this is a better model.

At your own time: California streamflow

> **Note**
>
> This additional exercise can be done at your own time. Most of the code are provided. You will need to run the code and interpret the results.

The following dataset contains 43 years of annual precipitation measurements (in mm) taken at (originally) 6 sites in the Owens Valley in California. I have reduced this to three variables labelled `L10APSAB` (Lake Sabrina), `L10OBPC` (Big Pine Creek), `L10OPRC` (Rock Creek), and the dependent variable stream runoff volume (measured in ML/year) at a site near Bishop, California (labelled `L10BSAAM`). There is also a variable `Year` but you can ignore this.

Note the variables have already been log-transformed to increase normality of the residuals in the regressions.

Start with a full model and manually remove the variables one at a time, checking every time whether removal of a variable actually improves the model.

```
# read in the data
s.data <- read_xlsx("mlr.xlsx", "California_streamflow")
names(s.data)
```

```
[1] "L10APSAB" "L10OBPC"  "L10OPRC"  "L10BSAAM"
```

```
s.mod_full <-lm(L10BSAAM~L10APSAB + L10OBPC + L10OPRC, data=s.data)
s.mod_full <-lm(L10BSAAM~., data=s.data) ## you can also use the . to indicate use all
 ↪  variables
summary(s.mod_full)
```

```
Call:
lm(formula = L10BSAAM ~ ., data = s.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09885 -0.03331  0.01025  0.03359  0.09495
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.25716    0.12360  26.352  < 2e-16 ***
L10APSAB     0.05631    0.03756   1.499  0.14185
L10OBPC      0.21085    0.06756   3.121  0.00339 **
L10OPRC      0.43838    0.08798   4.983 1.32e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04861 on 39 degrees of freedom
Multiple R-squared:  0.8817,    Adjusted R-squared:  0.8726
F-statistic: 96.88 on 3 and 39 DF,  p-value: < 2.2e-16
```

Partial F-Tests

The above analysis tells us that both `L10OBPC` & `L10OPRC` are significant, according to the t-test, in the model and `L10APSAB` is not? This involves performing Partial F-Tests as discussed in the lecture.

This can be done in **R** by using `anova()` on two model objects. To be able to compare the models and run the anova, you need to make objects of all the possible model combinations you want to compare.

```r
s.mod_reduced <- lm(L10BSAAM ~ L10OPRC + L10OBPC, data=s.data)
anova(s.mod_reduced, s.mod_full)
```

The last row gives the results of the partial F-test.

> ⚠ Question 1
>
> Should we remove `L10APSAB` from the model?

Answer

Yes, we should remove L10APSAB as the p-value is > 0.05 and opt for the simpler model.

> ⚠ Question 2
>
> Is the p-value for the f-test the same as for the t-test?

Answer

Yes, P-values for the t-statistic and for the Partial F-statistic are related (Partial F = t^2)

> ⚠ Question 3
>
> Write out the hypotheses you are testing.

Answer

H_0: $\beta_{L10APSAB} = 0$
H_1: $\beta_{L10APSAB} \neq 0$

Perform a Partial F-Test to work out if the removal of `L10APSAB` and `L10OBPC` improves upon the full model.

```
s.mod_reduced2  <- lm(L10BSAAM ~ L10APSAB + L10OBPC,data=s.data)
anova(s.mod_reduced2, s.mod_full)
```

```
Analysis of Variance Table

Model 1: L10BSAAM ~ L10APSAB + L10OBPC
Model 2: L10BSAAM ~ L10APSAB + L10OBPC + L10OPRC
  Res.Df      RSS Df Sum of Sq     F    Pr(>F)
1     40 0.150845
2     39 0.092166  1   0.05868 24.83 1.321e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> ⚠ Question 4
>
> Which variable should be added to the model containing L10OPRC?

Answer

L10APSAB does not improve the model with only L10OPRC ($\beta_{L10APSAB} = 0$), so we can say that we should add L10OBPC to the model containing L10OPRC.

Remember: H0: No difference between the models, so choose the simplest H1: Full model is better

> ⚠ Question 5
>
> Could things be even simpler? Perform a partial F-Test to see if a model containing L10OPRC alone could be suitable.

```
s.mod_reduced3  <- lm(L10BSAAM ~ L10OPRC,data=s.data)
anova(s.mod_reduced3, s.mod_full)
```

Answer

Fitting with only L10OPRC does not improve model fit (P<0.05) and so we can conclude that the better model is the one with L10OBPC and L10OPRC as predictors, with L10APSAB removed.

> ⚠ Question 6
>
> What is your optimal model?

Answer

The best model is: $L10BSAAM = \beta_0 + \beta_1 L10OPRC + \beta_2 L10OBPC + error$

That's it for today! Great work fitting simple and multiple linear regression! Next week we jump into stepwise selection and predictive modelling!