Lab 10 – Linear functions

ENVX1002 Handbook

Semester 1, 2025

> ### 💡 Learning Outcomes
>
> - Fit simple linear models and obtain associated model summaries in R
> - Overlay fitted models onto scatterplots in R
> - Undertake hypothesis testing to determine if slope $\neq 0$
> - Check assumptions are met prior to assessing model output
> - Assess model summary in terms of fit and P-values

Before you begin

Create your Quarto document and save it as `Lab-10.Rmd` or similar. The following data files are required:

1) ENVX1002_wk10_practical_data_Regression.xlsx

Last week you fitted models in R, now it is time to understand what the output means.

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: Walkthrough - Fertiliser

Like last week, we will start off our R modelling journey by fitting a model to the fertiliser data.

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)
```

1.1 Scatterplot and correlation

To visually identify any trends or relationships, last week we created a scatterplot of the data. This is helps us visually understand our points so we know what we might expect from the model, and possibly identify if the relationship is looking non-linear.

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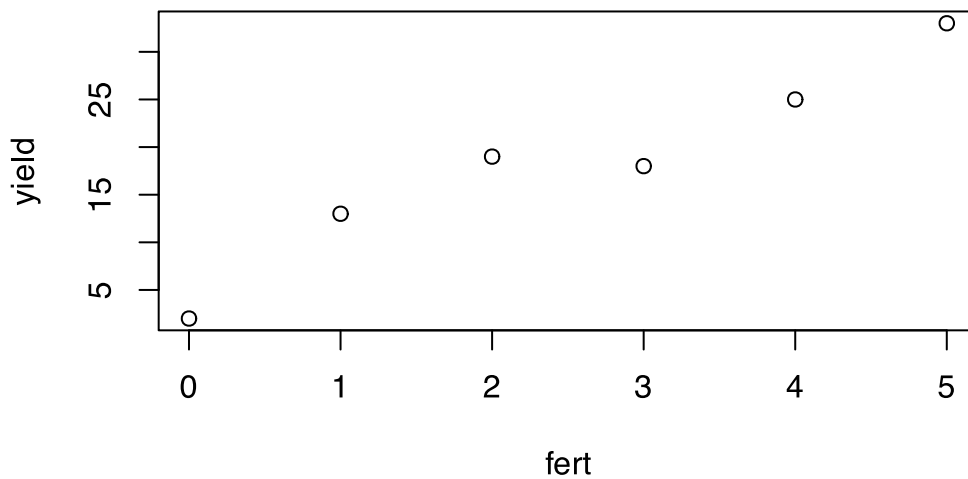

Remembering back to last week, we then calculated the correlation coefficient to numerically determine whether there was a relationship between fertiliser and yield.

Using the code below, we found there was quite a strong relationship between fertiliser and yield (0.964):

```
# Correlation coefficient
cor(fert, yield)
```

```
[1] 0.9636686
```

1.2 State hypotheses

Remembering back to the lecture and tutorial, the general equations for our hypotheses are:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

In the context of our data, the hypotheses would be:

H_0: Slope = 0; fertiliser is not a significant predictor of yield.

H_1: Slope $\neq$ 0; fertiliser is a significant predictor of yield.

If P > 0.05, we fail to reject the null hypothesis that the true slope (β_1) is equal to 0. If this is the case, it means our model does not predict better than the mean of our observations, and so there is no advantage to using our model over the mean of y ($\bar{y}$).

If we find there is a high probability of the slope not being equal to 0 (P < 0.05), we can reject the null hypothesis and conclude our model is better at predicting than the mean of our observations.

Now we understand what we are testing for, we can fit the model.

1.3 Fit the model

After checking the correlations and scatterplot, we need to fit the model 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)
```

1.4 Check assumptions

This time, before obtaining our model summary, we need to check our assumptions.

Smaller sample size (n = 6) makes it harder to check whether the assumptions have been met, but we will still run through the check.

Looking at each plot, we can see that the residual plots don't look the best;

- Residuals vs fitted: Will tell us if the relationship is linear. We are looking for an even scatter around the mean, and red line should be reasonably straight. In this case the red line is not too straight, but the scatter seems even.

- Normal Q-Q: If the residuals are normally distributed, most of the points should lie along the dotted line. Our points follow the line, but do not lie on it.

- Scale-Location: This is for testing whether the variance is equal in the residuals at each value of x. If the variance is equal, then we would expect to see an even scatter and no fanning. In this case, there is no fanning.

- Residuals vs Leverage: This will help us identify whether there are any single points influencing the slope or intercept of the model. We can see in the output plot there is a point sitting in the bottom-right corner, outside the dotted line, indicating that it may be having an influence on the model.

These plots are only useful as an example of how to obtain and interpret output. If we wanted to obtain a more reliable check of our assumptions (and a more reliable model), we would need a larger sample size (n > 10).

```
# Check your assumptions!!
par(mfrow = c(2, 2)) # sets plots to show as 2x2 grid
plot(model.lm)
```


In this case, we will assume the assumptions have been met and continue to assess the model output.

1.5 Model output

Use the summary() function to obtain output for your model:

```
# Obtain model summary
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 **
```

4

```
---
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
```

In the model output obtained from `summary(model.lm)` the model parameters will be listed under 'Estimate' for the intercept and 'fert'. Last week we concluded the equation to be:

$$Yield = 4.7619 + 5.4286*fert$$

Furthermore, from our model estimate, we can say that as fertiliser increases by 1, yield will increase by 5.4286.

1.6 Is the model useful?

When looking at the model summary output, we obtain the p-value from the coefficients table. We are interested in the P-value for `fert` and not the intercept.

The significance of the intercept P-value depends on our scientific question. We only really look at our intercept P-value when we want to extrapolate our line to the intercept, and know if the intercept is equal to zero (H_0) or not (H_1). This depends on your dataset and whether it makes sense to do so.

Also notice how the p-values for the F-test at the bottom of the summary output, and the t-test p-values we are using are the same. The F-test gives us an idea whether our overall model is significant and in this case, as we are only using a single predictor, the P-values will be the same.

Therefore we can say the following:

Observing the model output, we can see that the P-value for fert is significant (P = 0.00196) and we can say that as P < 0.05, we reject the null hypothesis. We can conclude our slope is not 0 and our model is a better way to predict yield than the mean of our observations.

1.7 How good is the model?

To assess how well the model fits the data, we need to look at the Residual standard error (3.147) and the r-squared value (0.9287).

- We can say that our residual standard error is relatively low in terms of our response variable.

- Our r-squared indicates that fertiliser accounts for 92.9% of variation in yield. That's pretty good!

Note how Multiple R-squared and Adjusted R-squared are similar. For simple linear models we can opt for the multiple r-squared, but when using multiple predictors we need to use adjusted r-squared.

Finally, to visually present our results, we can provide a scatterplot with the model overlaid.

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


1.8 Our conclusions

Now we can put our interpretations together to form the conclusion:

Observing the model output, we can see that the P-value for fert is significant (P = 0.00196) and we can say that as P < 0.05, we reject the null hypothesis. We can conclude our slope is not 0 and our model is a better way to predict yield than the mean of our observations.

We can therefore conclude that fertiliser is a significant predictor of crop yield as the slope is not equal to zero (P < 0.05), and it accounts for 92.9% of the variation in yield.

Exercise 2: Toxicity in peanuts

Data: *Peanuts* spreadsheet

The data comprise of, for 34 batches, the average level of the fungal contaminant aflatoxin in a sample of 120 pounds of peanuts and the percentage of non-contaminated peanuts in the whole batch.

The data were collected with the aim of being able to predict the percentage of non-contaminated peanuts (Peanuts$percent) from the aflatoxin level (Peanuts$toxin) in a sample. We will now investigate whether this is the case.

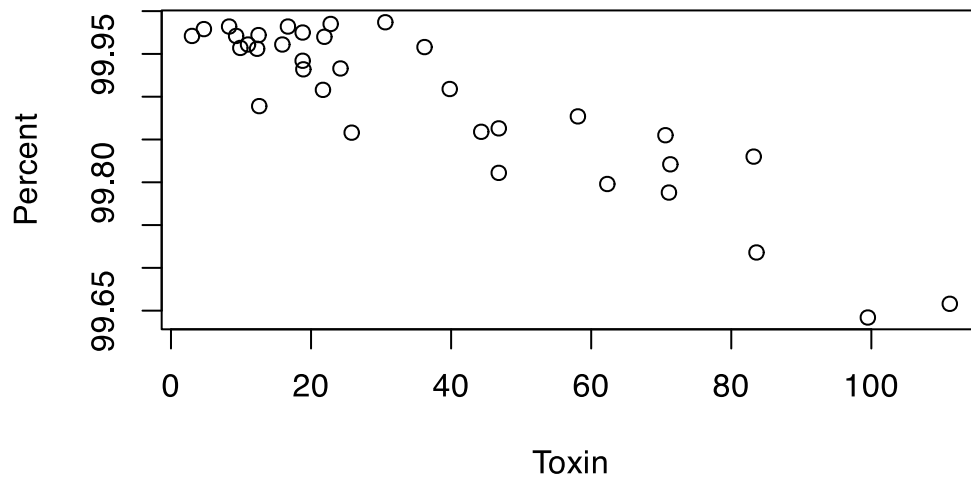First thing's first! Let's read in the data using read_xlsx command:

```
library(readxl)
Peanuts <- read_xlsx("data/ENVX1002_wk10_practical_data_Regression.xlsx", sheet
= "Peanuts")
head(Peanuts)
```

```
# A tibble: 6 × 2
  Percent Toxin
    <dbl> <dbl>
1   100.   3
2   100.   4.7
3   100.   8.3
4   100.   9.3
5   100.   9.9
6   100.  11
```

2.1 Scatter plot

Make a scatter plot of the data.

```
plot(Percent ~ Toxin, data = Peanuts)
```



```
#Alternate syntax:
#plot(Peanuts$Percent, Peanuts$Toxin)
```

a) Describe the relationship between the two variables.

Solution

a) There seems to be a strong, negative relationship between Percent and Toxin

b) Would you say that the percentage of non-contaminated peanuts in a batch could be predicted accurately from the level of aflatoxin in a sample via a linear relationship?

Solution

b) The relationship on the plot does seem to follow a straight line and so we can say that percentage of non-contaminated peanuts in a batch could be predicted from the level of aflatoxin in a sample via a linear relationship.

2.2 State Hypotheses

a) What are the hypotheses we are testing? State them as the formulae and in the context of the study.

Solution

a) As above, the aim is to test whether it is possible to predict the percentage of non-contaminated peanuts from the aflatoxin level in a sample.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

H_0: Aflatoxin level is not a significant predictor of the percentage of non-contaminated peanuts and we are better off using the mean value.

H_1: Aflatoxin level is a significant predictor of the percentage of non-contaminated peanuts.
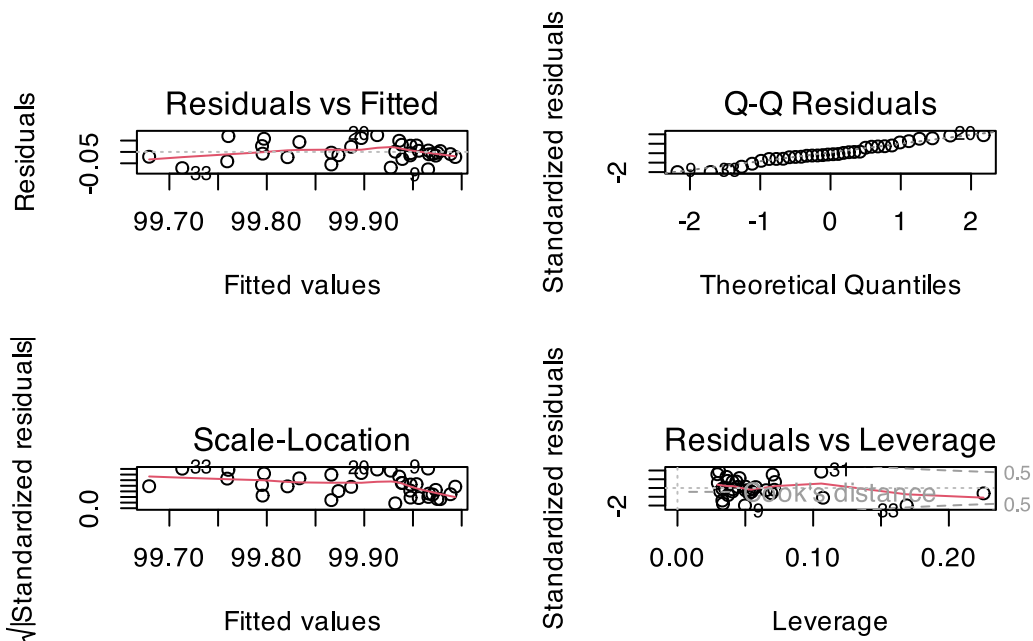
2.3 Fit a linear model

Use fit a linear model (`lm()`) to the Peanut data.

```
# fit a linear model using lm()
mod <- lm(Percent ~ Toxin, data = Peanuts)
```

2.4 Check assumptions

a) Inspect and comment on the residual plots- have the assumptions been met?

```
par(mfrow = c(2, 2))
plot(mod)
```

Solution

a) There seems to be no problem with the residual plots, the data are normally distributed and there is no indication of fanning (increasing variance) in the residuals.

2.5 Observe model output

Once you are certain the assumptions are met, you can proceed to look at the regression output.

a) Comment on the overall fit of the regression, i.e. Is the model fit good? Is the model significant, and how much variation in percentage of non-contaminated peanuts does aflatoxin level account for?

```
# Look at output with summary
summary(mod)
```

```
Call:
lm(formula = Percent ~ Toxin, data = Peanuts)

Residuals:
     Min        1Q    Median        3Q       Max
-0.076516 -0.020012 -0.004806  0.027094  0.073747

Coefficients:
```

9

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.000e+02  1.089e-02 9184.91  < 2e-16 ***
Toxin       -2.903e-03  2.335e-04  -12.44 8.54e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03933 on 32 degrees of freedom
Multiple R-squared:  0.8285,    Adjusted R-squared:  0.8232
F-statistic: 154.6 on 1 and 32 DF,  p-value: 8.538e-14
```

Solution

a) The overall fit is good, with 83% of the variation explained. The model is significant with a p-value for b_1 < 0.05 therefore indicating that the slope is significantly different from 0 and negative in this case (-0.0029).

b) Is toxin a significant predictor of percentage non-contaminated peanuts? If so, how can we tell?

Solution

b) yes, as the p-value = 8.54e-14, which is much smaller than 0.05

c) Interpret the slope parameter in terms of quantifying the relationship between toxin and percent.

Solution

c) An increase in toxin by 1 results in a decrease of 0.0029% of uncontaminated peanuts.

Exercise 3: Dippers

Data: *Dippers* spreadsheet

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. For the purpose of fitting a simple linear model, the dataset has been reduced to two variables:

- The number of breeding pairs of Dippers per 10 km of river
- The numbers of caddis fly larvae (Log(Number+1) transformed)

Now it is your turn to work through the steps as above. Does the number of caddis fly larvae influence the number of breeding pairs of Dippers?

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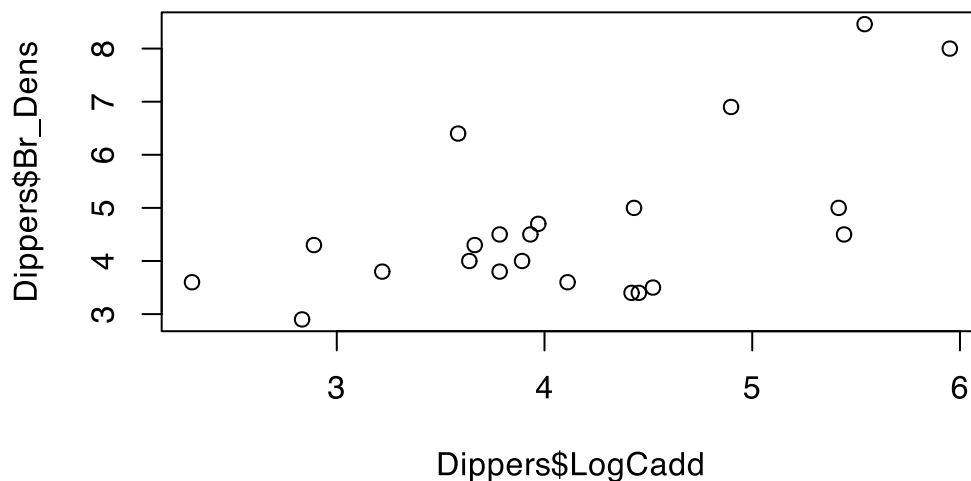
Solution

a) Use the following code:

```
library(readxl, quietly = TRUE)
Dippers <- read_xlsx(
  "data/ENVX1002_wk10_practical_data_Regression.xlsx",
  sheet = "Dippers"
)
```

b) Obtain a scatterplot, are there signs of a relationship between breeding pair density and caddis fly larvae?

Solution

b) Looking at the scatterplot there seems to be a weak positive relationship. We can follow this up using correlation, which indicates there is a moderate positive relationship of 0.613 between breeding pair density along the study area.

```
plot(Dippers$LogCadd, Dippers$Br_Dens)
```

```
cor(Dippers$LogCadd, Dippers$Br_Dens)
```

```
[1] 0.6126891
```

c) What are the hypotheses we are testing? State them as the formulae and in the context of the study.

Solution

c) As above, the aim is to test whether the number of caddis fly larvae influence the number of breeding pairs of Dippers

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

H_0: The number of caddis fly larvae is not significant predictor of dipper breeding pair density. OR could say: There is insufficient evidence to suggest the number of caddis fly larvae influence the number of breeding pairs of Dippers.

H_1: The number of caddis fly larvae is a significant predictor of number of dipper breeding pair density. OR could say: The number of caddis fly larvae has an influence upon the number of breeding pairs of dippers.

d) Let's investigate further. Run the model, but before looking at our model output, are the assumptions ok?

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

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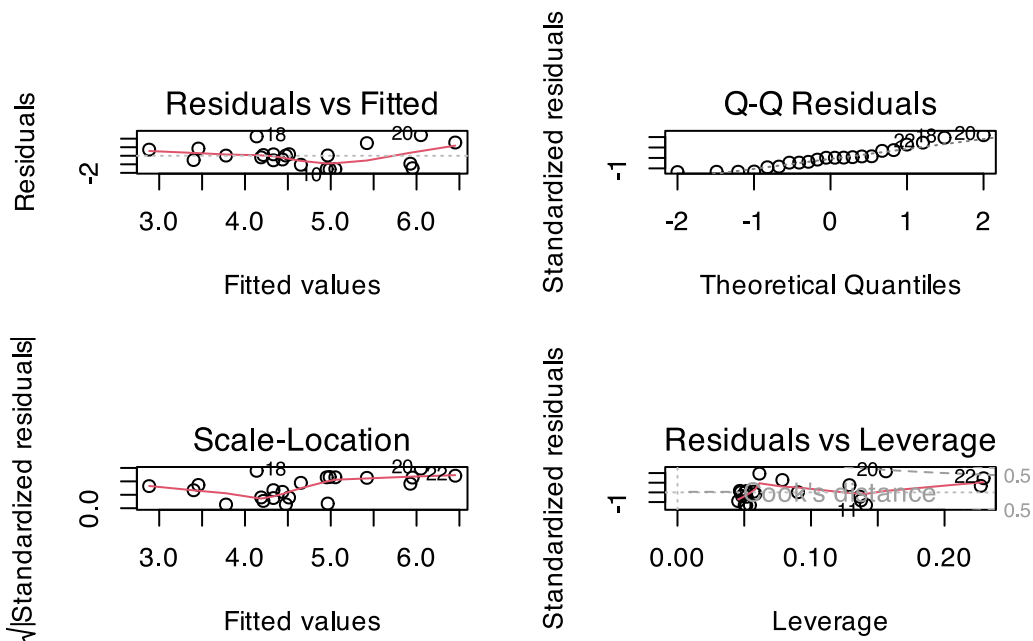
Solution

────────────────────────────────

d) Assumptions look ok:

- Residuals vs fitted: Points evenly scattered around mean
- Normal Q-Q: Points follow line
- Scale-Location: Points evenly scattered, no fanning.
- Residuals vs Leverage: No points occurring at top right or bottom right corners outside the dotted red lines. A couple of points are close, but not outside, so all is good.

We can continue to interpreting the output!

────────────────────────────────

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

Solution

────────────────────────────────

```
summary(dipper.lm)
```

```
Call:
lm(formula = Br_Dens ~ LogCadd, data = Dippers)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-1.58688 -0.82886  0.02255  0.58473  2.40971


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6295     1.1910   0.529  0.60292
LogCadd       0.9783     0.2822   3.467  0.00243 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 1.204 on 20 degrees of freedom
Multiple R-squared:  0.3754,    Adjusted R-squared:  0.3442
F-statistic: 12.02 on 1 and 20 DF,  p-value: 0.002434
```

f) What is the equation for our model, incorporating our coefficients?

Solution

f) Br_Dens = 0.6295 + 0.9783*LogCadd

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

Solution

g) The F-test statistic tells us whether our overall model is significant, which in this case it is (P = 0.002). In the case of simple linear models (with only 1 predictor variable), the hypothesis for the F-statistic will be the same as the hypothesis for the slope of our single predictor. Therefore we expect the P-values to be the same.

h) Is LogCadd a significant predictor of Dipper breeding pair density? How can we tell?

Solution

h) The P-value for LogCadd is less than 0.05 (P = 0.0024), and so we reject the null hypothesis that the slope is equal to zero. We can therefore say LogCadd a significant predictor of Dipper breeding pair density.

i) How good is the fit of our model?

Solution

i) Not too good. Our residual standard is (Residual SE = 1.204) and the r^2 is closer to 0 than 1 (r^2 = 0.375).

Our residual SE is a low value, but relative to the range of Br_Dens (min = 2.9, max = 8.4), it is not too good.

We can also back this up with a scatterplot and our model overlaid, notice how the points of the scatterplot follow the line somewhat, but not as tightly as we saw in the Peanuts data.

```
plot(Dippers$LogCadd, Dippers$Br_Dens)
abline(dipper.lm)
```
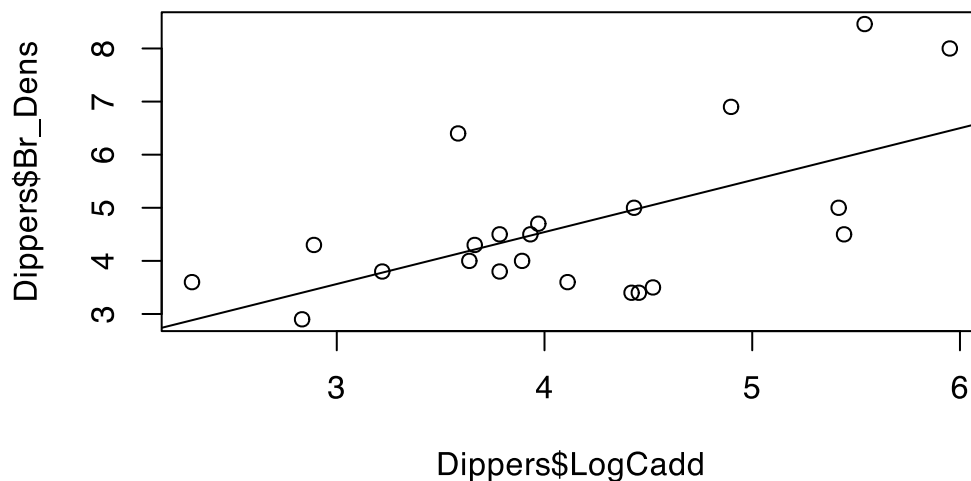

j) What conclusions can we make from this model output?

Solution

j) Although the number of caddis fly larvae (LogCadd) is a significant predictor of Dipper breeding pair density ($P < 0.05$), it only accounts for 37.5% of variation in breeding pair density.

Back to the original question: Does the number of caddis fly larvae influence the number of breeding pairs of Dippers? Yes, but based on our findings, this relationship is not very strong.

k) A final thought; Does our result make sense within the context? i.e. why might the Dipper breeding pair density be related to LogCadd?

Solution

k) Our findings make sense that there is a relationship between caddis fly larvae and Dipper breeding density, as Dippers feed on benthic invertebrates such as caddis fly larvae.

Why there is such a weak relationship may be because the caddis fly larvae is likely not to be the only type of invertebrate the Dippers eat, or perhaps the Dippers have a preference.

There are also other factors in play, but we would have to build some knowledge in this area through research.

Great work fitting simple linear models! Next week we step it up with *multiple* linear regression.

Bonus take home exercises

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

1. Scatterplot and correlation

2. State Hypothesis

3. Fit the model

4. Check assumptions

5. P-value and model fit

    a) Is the model significant?

    b) Is the predictor significant?

    c) How good is the model fit?

6. Conclusions

Exercise 1: Cars stopping distance

Use the cars dataset from last week to test if speed (mph) is a predictor of dist (stopping distance, fft).
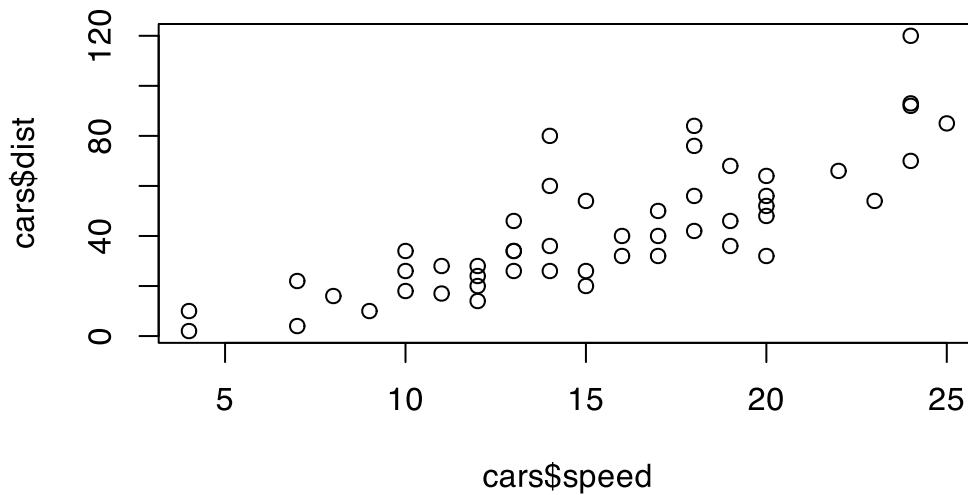
```
head(cars)
```

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

Solution

1. Scatterplot and correlation

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



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

```
[1] 0.8068949
```
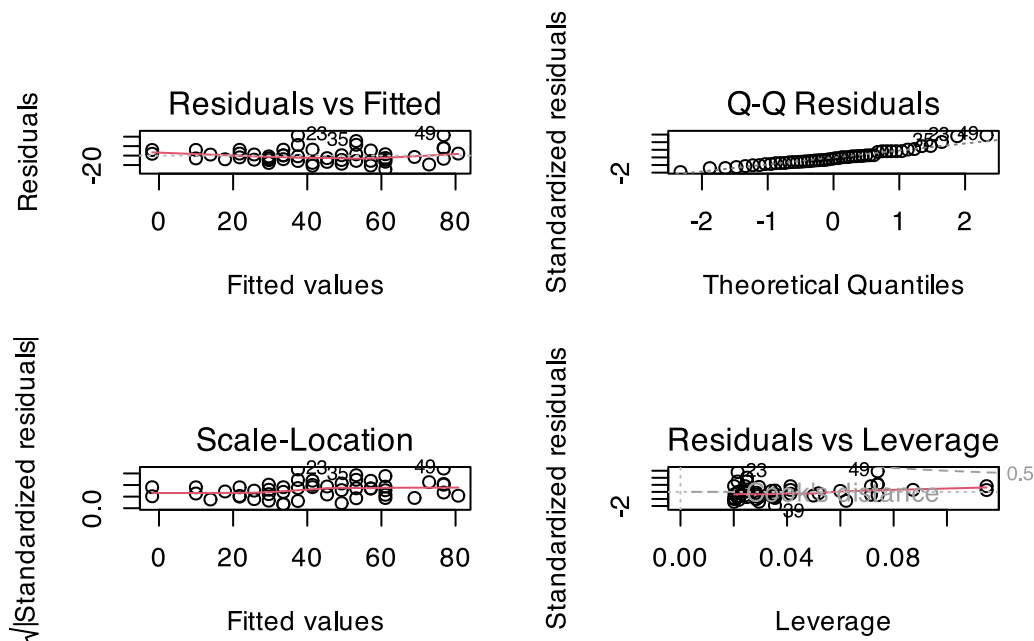
2. State Hypothesis

H_0: car speed is not a significant predictor of stopping distance H_1: car speed is a significant predictor of stopping distance

3. Fit the model

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

4. Check assumptions

```
par(mfrow = c(2, 2))
plot(cars_lm)
```

Residuals vs Fitted	Q-Q Residuals
Scale-Location	Residuals vs Leverage

Assumptions look ok:

- Residuals vs fitted: Points evenly scattered around mean
- Normal Q-Q: Points mostly follow line
- Scale-Location: Points evenly scattered, no fanning.
- Residuals vs Leverage: No points occurring at top right or bottom right corners outside the dotted red lines. A couple of points are close, but not outside, so all is good.

5. P-value and model fit

```
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
```

```
a)  Is the model significant? The F-test statistic tells us whether our overall
model is significant, which in this case it is (p < 0.05)

b)   Is the predictor significant? speed is a significant predictor (p < 0.05)

c)   How good is the model fit? The fit is decent though it could be better. The
R^2^ is 0.65, which tells us that the model accounts for 65% of the variation
in the data.
```

6. Conclusions

Speed is a statistically significant predictor of stoppingdistance. The model accounts for 65% of the data, so there may also be other factors that affect stoppping distance.

Exercise 2: Penguins

Use the palmer penguins dataset to test if `flipper_length` is a significant predictor of `bill_length`.

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

```
── Attaching core tidyverse packages ──────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ──────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use  the  conflicted  package  (<http://conflicted.r-lib.org/>)  to  force  all
conflicts to become errors
```

```
#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>
```

Solution

1. Scatterplot and correlation

```
#scatter plots
plot(penguins$flipper_length_mm,penguins$bill_length_mm)
```
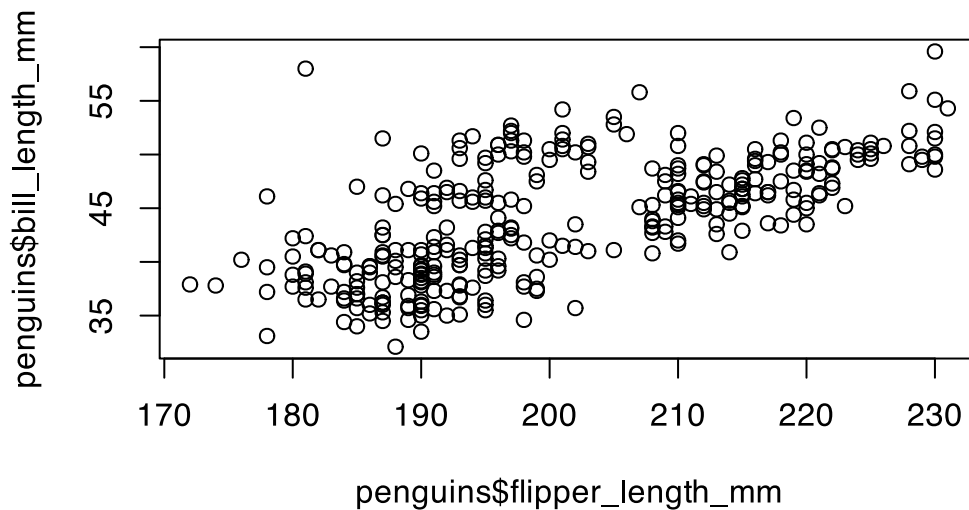


```
#correlation
cor(penguins$bill_length_mm, penguins$flipper_length_mm)
```
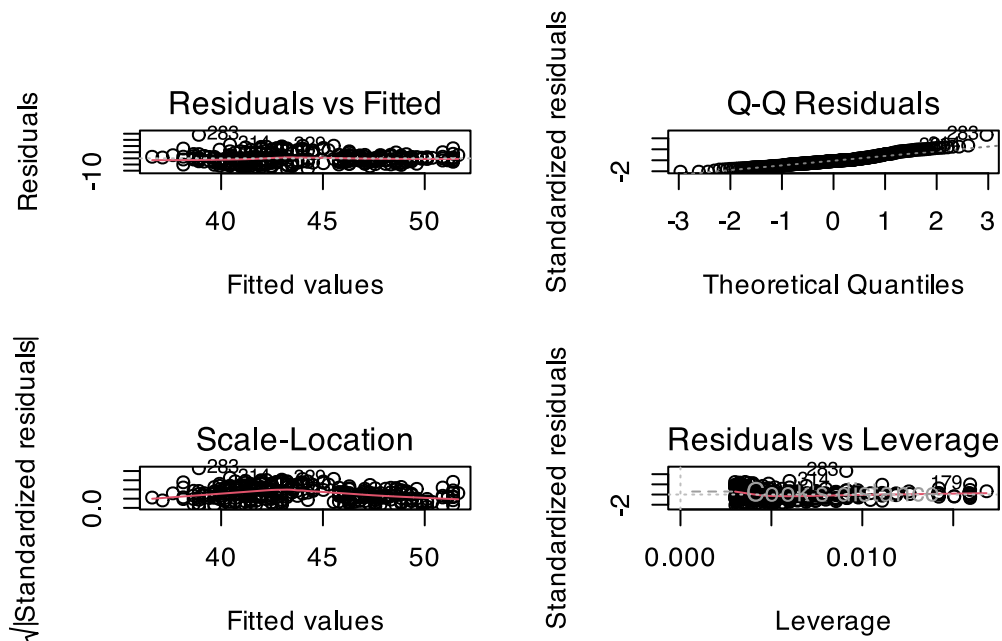
```
[1] 0.6530956
```

2. State Hypothesis

H_0_: flipper length is not a significant predictor of bill length H_1_: flipper length is a significant predictor of bill length

3. Fit the model

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

4. Check assumptions

```
par(mfrow = c(2, 2))
plot(penguin_model)
```


Assumptions look ok:

- Residuals vs fitted: Points evenly scattered around mean
- Normal Q-Q: Points follow the line
- Scale-Location: Points evenly scattered, no fanning.
- Residuals vs Leverage: No points occurring at top right or bottom right corners outside the dotted red lines.

5. P-value and model fit

```
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
```

```
a)   Is the model significant? The F-test statistic tells us whether our overall
model is significant, which in this case it is (p < 0.05)

b)   Is the predictor significant? flipper length is a significant predictor of
bill length (p< 0.05)

c)   How good is the model fit? The model has an R^2^ of 0.42, which tells us
that the model only explain 42% of the variation in the data. It could be worse,
but it's not particularly good.
```

6. Conclusions > Flipper length is a statistically significant predictor of bill length. However, only 42% of the variation in the data was explained by the model. There are likely to be other variables that also help predict bill length. For example, the dataset has three different penguin species, which could also be adding noise to our model.

Exercise 3: Old Faithful Geyser Data

Using the inbuilt `faithful` dataset, test whether waiting time (`waiting`) is a significant predictor of eruption time (`eruption`).
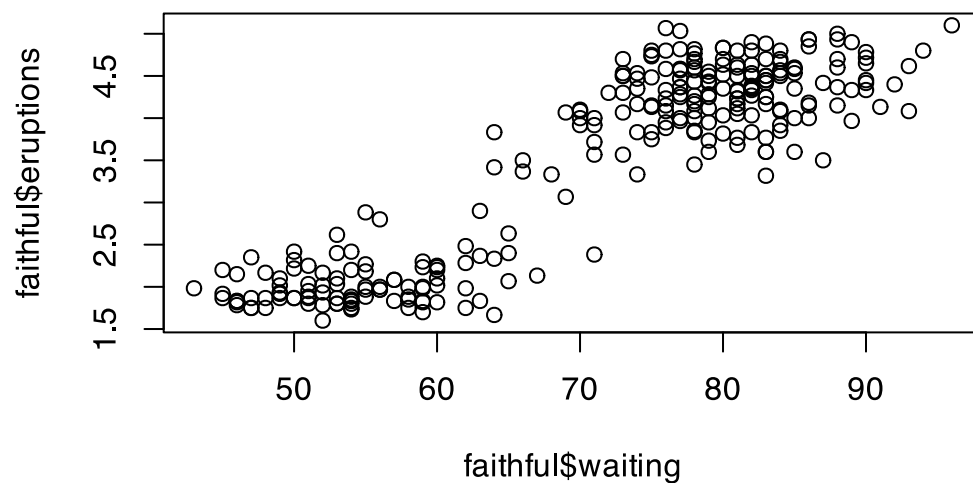
```
head(faithful)
```

```
  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
```

Solution

1. Scatterplot and correlation

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



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

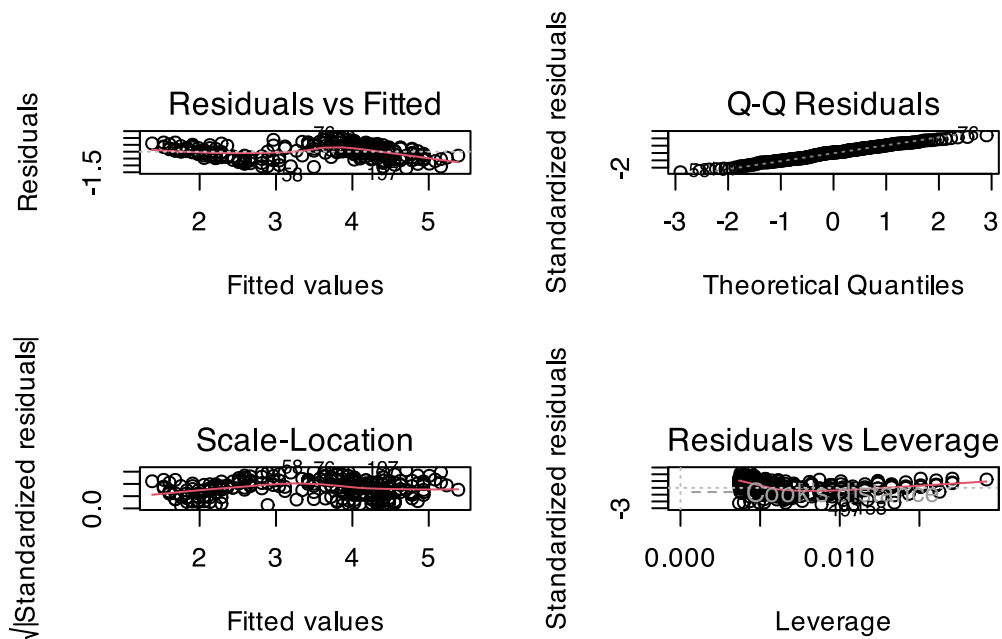```
[1] 0.9008112
```

2.  State Hypothesis

H_0_: Waiting time is not a significant predictor of eruption time H_1_: Waiting time is a significant predictor of eruption time

3. Fit the model

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

4. Check assumptions

```
par(mfrow = c(2,2))
plot(faithful_lm)
```


Assumptions look ok:

- Residuals vs fitted: Looks a little funky, but the red line is still reasonably straight and the dots are vaguely spread above and below the line
- Normal Q-Q: Points follow the line
- Scale-Location: Points evenly scattered, no fanning.
- Residuals vs Leverage: No points occurring at top right or bottom right corners outside the dotted red lines.

5. P-value and model fit

```
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
```

a) Is the model significant? The F-test statistic tells us whether our overall model is significant, which in this case it is ($p < 0.05$)

b) Is the predictor significant? waiting is a significant predictor of eruption ($p < 0.05$)

c) How good is the model fit? The model has an R^2 of 0.81, which tells us that the model explains 81% of the varation in the data. This is an excellent fit!

6. Conclusions Time between eruptions is a statistically significant predictor of eruption time.