Assignment 2: statistical analysis

Code

Anthony Duong

This assignment focuses on problems relating to introductory statistics (Week 6), single-factor ANOVA (Week 7), and linear regression (Week 8). It involves identifying and applying appropriate statistical analyses to the data sets provided, interpreting the outcomes of analyses, and linking those outcomes back to the biological questions that motivated analyses. It forms 15% of the overall grade for BIO2010 and is **due 11:55 PM on Thursday 22 September (Week 9)**.

To complete the assignment:

- get your answers by entering and running the R code required in the code chunks provided (remember, you can run each chunk by clicking the green triangle at the top right of the chunk);
- type your answers in the spaces also provided (marks are based on your typed answers, but including your code helps markers understand how you got them);
- save your work as you go (press *Command-S* or *Control-S* keys, click *File* > *Save* in the menu bar, or click the disc icon in the toolbar)!

To create a PDF of the assignment for submission:

- create an HTML file and open it in a browser (click *Preview* in the toolbar and *Open in Browser* at the top left of the file), then save the file as a PDF (click *File* > *Print* > *Save as PDF* > *Save*); or
- create a Word file (click the arrow next to *Preview* in the toolbar and select *Knit to Word*), then save or export the file as a PDF.

Check that your PDF displays all information correctly, then upload it to the Moodle Dropbox by the due date.

Problem set 1: food and stalk-eyed flies

Stalk-eyed flies (*Cyrtodiopsis dalmanni*) have eyes at the ends of long stalks that extend horizontally from each side of the head. Longer eye spans (the combined lengths of stalks from tip to tip) enhance attractiveness to mates and success in competitive interactions. Is eye span affected by the quality of food that flies eat? To find out, researchers reared two groups of flies on different foods, then recorded each fly's eye span in millimetres. One group of flies was fed corn (a high-quality food), while the other group of flies was fed cotton wool (a low-quality food). Each fly per group was reared singly, and therefore represents an independent sampling unit. The data are contained in the file 'flies.csv', which you can load by running the chunk below.

Hide

```
# Run this chunk to load the data for Problem set 1
flies <- read_csv("./flies.csv",
                  col_names = TRUE, na = "NA",
                  col_types = list(food = col_factor(),
                                   eyeSpan = col_double()))
glimpse(flies, width = 75)
```

```
Rows: 45
Columns: 2
$ food    <fct> corn, corn, corn, corn, corn, corn, corn, corn, corn, cor…
$ eyeSpan <dbl> 2.15, 2.14, 2.13, 2.13, 2.12, 2.11, 2.10, 2.08, 2.08, 2.0…
```

EXERCISE 1: Based on the research description above, what are the null and alternate hypotheses of interest? (2 MARKS)

```
Null hypothesis: Quality of food eaten by flies does not affect their eye span.

Alternative hypothesis: Quality of food eaten by flies affects their eye span.
```

EXERCISE 2: Report (with standard errors) the estimates that are the focus of the hypotheses above. You can work out your answer using the chunk below. (2 MARKS)

```
Corn fed flies: μ = 2.05 σ = 0.016
Cotton fed flies: μ = 1.54 σ = 0.058
```

Hide

```
# Use this chunk to complete Exercise 2
flies %>%
  group_by(food) %>%
  summarise(mean = mean(eyeSpan),
              n = n(),
              sd = sd(eyeSpan, na.rm = TRUE),
              se = sd/sqrt(n))
```

food <fctr>	mean <dbl>	n <int>	sd <dbl>	se <dbl>
corn	2.047143	21	0.07470896	0.01630283
cotton	1.542917	24	0.28493293	0.05816169

2 rows

Hide

```
NA
```

EXERCISE 3: What is the appropriate statistical analysis for testing the null hypothesis? Give 3 reasons to justify your answer. Assume that the numerical variable has a normal distribution. You can work out your answer using the chunk below. (4 MARKS)

```
The most appropriate test to use is the two sample t-test as we are comparing the means of tw
o independent treatment groups where the both the explanatory and response variable are numer
ical.
```

Hide

```
# Use this chunk to complete Exercises 3, 4, and 5
t.test(eyeSpan ~ food,
       data = flies,
       alternative = "two.sided",
       var.equal =)
```

```
    Welch Two Sample t-test

data:  eyeSpan by food
t = 8.3477, df = 26.568, p-value = 6.666e-09
alternative hypothesis: true difference in means between group corn and group cotton is not e
qual to 0
95 percent confidence interval:
 0.3801943 0.6282581
sample estimates:
  mean in group corn mean in group cotton
            2.047143             1.542917
```

EXERCISE 4: After running the appropriate statistical analysis (which you can do in the chunk above), do you reject or fail to reject the null hypothesis? Why? Explain what this indicates about the relationship between food and eye span in stalk-eyed flies. (2 MARKS)

```
Since p = 7.345e-10, we reject the null hypothesis that the quality of food eaten by flies do
es not affect their eye span. This suggests that the difference between the eye span of corn
fed and cotton fed flies is not due to random chance and supports the alternative hypothesis.
```

EXERCISE 5: Based on the research description and your analysis of the data, write a brief report of the results as though it was your own study. Follow the guidelines in the week 7 workshop handout for reporting the results of a statistical analysis, providing all information necessary for readers to understand and properly evaluate the research outcomes. Be sure to use your own words (do not copy the week 7 workshop handout). There is no minimum or maximum word limit, but a few sentences should suffice. (4 MARKS)

```
The relationship ship between the quality of food and eye-span of Cyrtodiopsis dalmanni flies
was tested using a sample of 45 flies, 21 of which were fed cotton woll and the other 24 were
fed corn and their eye span was measured individually. It was found that there was a signific
ant difference between the eyespan of corn fed flies (μ = 2.05 σ = 0.016) and cotton fed flie
s (μ = 1.54 σ = 0.058) with a p value of 7.345e-10, where the data supports the hypothesis th
at the quality of food affects the eye span of the flies.
```

Problem set 2: parasites and traffic accidents

The parasite *Toxoplasma gondii* infects warm-blooded vertebrates, including 30–60% of humans, in most countries. A short phase of acute infection is followed by a latent phase (life-long presence of *Toxoplasma* cysts in nerve and muscle tissues), in which infected subjects are considered to be asymptomatic. Researchers noticed, however, that some subjects with latent infections had longer reaction times in laboratory tests, and speculated that latent infection might influence the performance of subjects in real-life situations.

To find out, the researchers surveyed a sample of 15- to 29-year-old drivers from Prague who had been involved in traffic accidents. Each driver was given a blood test at the time of the accident (to rule out intoxication), which revealed any latent infection by *Toxoplasma*. The researchers also surveyed a control sample of drivers of the same age living in the same area who had not been in any accidents. The data are contained in the file 'drivers.csv', which you can load by running the chunk below.

Hide

```
# Run this chunk to load the data for Problem set 2
drivers <- read_csv("./drivers.csv",
                    col_names = TRUE, na = "NA",
                    col_types = list(infectionStatus = col_factor(),
                                     accidentStatus = col_factor()))
glimpse(drivers, width = 75)
```

```
Rows: 308
Columns: 2
$ infectionStatus <fct> infected, infected, infected, infected, infected,…
$ accidentStatus  <fct> inAccidents, inAccidents, inAccidents, inAccident…
```

EXERCISE 6: Based on the research description above, what are the null and alternate hypotheses of interest? (2 MARKS)

```
Null hypothesis: Those with a latent infection of T.Gondii are not more likely to be involved
in a car accident than an uninfected person.

Alternate hypothesis: Those with a latent infection of T.Gondii are more likely to be involve
d in a car accident than an uninfected person.
```

EXERCISE 7: Based on the research description above, what is the response variable, what is the explanatory variable, and is each variable categorical or numerical? (2 MARKS)

```
The response variable is whether they have been in a car accident and the explanatory variabl
e is whethere they have a latent infection with T.Gondii. Both the response and explanatory v
ariable are categorical data.
```

EXERCISE 8: What is the appropriate statistical analysis for testing the null hypothesis, and why? Give 3 reasons to justify your answer. You can work out your answer using the chunk below. (4 MARKS)

```
Since both the explanatory and response variable are categorical data, a contingency chi squa
red test to determine if there is a relationship between latent infection and probability of
being in a car crash. Using a chi squared test, we can compare the results to the expected va
lues if the null hypothesis was true, then reject or fail to reject based on the result.
```

Hide

```
# Use this chunk to complete Exercises 8, 9, and 10
table <- table(drivers$infectionStatus, drivers$accidentStatus)

table
```

```
            inAccidents notinAccidents
  infected            21             38
  uninfected          38            211
```

Hide

```
chisq.test(x = table, correct = FALSE)
```

```

    Pearson's Chi-squared test

data:  table
X-squared = 12.733, df = 1, p-value = 0.0003593
```

EXERCISE 9: After running the appropriate statistical analysis (which you can do in the chunk above), do you reject or fail to reject the null hypothesis, and why? Explain what this indicates about the relationship between *Toxoplasma* infection and traffic accidents. (2 MARKS)

```
A Chi-squared test was run on the results to determine if there is a relationship between a l
atent infection with T.Gondii and the probability of being in a car accident. The expected va
lues were determined assuming that the null hypothesis is true, that there is no differnce be
tween infection and uninfection. The chi-squared value of 12.733, df=1, gave p = 0.00036, mea
ning we fail to reject the null and conclude that there is no relationship between chance of
being in a car accident and a latent infection of T.Gondii
```

EXERCISE 10: Based on the research description and your analysis of the data, write a brief report of the results as though it was your own study. Follow the guidelines in the week 7 workshop handout for reporting the results of a statistical analysis, providing all information necessary for readers to understand and properly evaluate the research outcomes. Be sure to use your own words (do not copy the week 7 workshop handout). There is no minimum or maximum word limit, but a few sentences should suffice. (4 MARKS)

```
From a sample of 308 15 - 29 year olds who have been involved in a car accident, excluding th
ose under the influence of alchol, were sampled for a latent infection with Toxiplasma Gondi
i, to deterimine if infection has a higher chance of causing a car acciddent. There was no si
gnificant difference between the number of accidents by infected drivers and uninfected drive
rs, with a chi-squared score of 12.733 and p value of 0.00036, meaning we fail to reject the
null, thus there is no relationship between latent infection with T.Gondii and probability of
being in a car accident.
```

Problem set 3: maths and readers' research areas

Does adding maths to a scientific paper make readers think that it has more value, or does it depend on the reader's research area? Researchers sent two abstracts of scientific papers to each of 200 people with postgraduate degrees. Before sending the two abstracts to each person, the researchers chose one of the abstracts at random and added a meaningless sentence describing an unrelated mathematical model (the maths itself wasn't meaningless, but it had no conceptual link to the subject matter of the abstract). The other abstract had no maths added.

Participants were asked to score the quality of the research in each abstract on a scale from 1 to 100, and to also state the research area of their postgraduate degree: science including maths and technology (Sci); medicine (Med); humanities including social science (Hum); or other (Other). Then, for each person, the researchers calculated the perceived advantage of adding maths to a paper (mathsAdvantage) as the score for the abstract with maths added minus the score for the abstract without maths added. The data are contained in the file 'maths.csv', which you can load by running the chunk below.

Hide

```
# Run this chunk to load the data for Problem set 3
maths <- read_csv("./maths.csv",
                  col_names = TRUE, na = "NA",
                  col_types = list(degree = col_factor(),
                                   mathsAdvantage = col_double()))
glimpse(maths, width = 75)
```

```
Rows: 200
Columns: 2
$ degree         <fct> Hum, Hum, Hum, Hum, Hum, Hum, Hum, Hum, Hum, Hum, …
$ mathsAdvantage <dbl> -23, -27, -29, -25, -20, -20, -20, -15, -15, -13, …
```

EXERCISE 11: Based on the research description above, what are the null and alternate hypotheses of interest? (2 MARKS)

```
Null hypothesis: Adding maths to a scientific paper has no effect on the reader's valuation o
f the paper, regardless of their research area.

Alternative hypothesis: Adding maths to a scietific paper makes the reader think it has more
value, or the reader values a paper more depending on their area of research.
```

EXERCISE 12: Report (with standard errors) the estimates that are the focus of the hypotheses above. You can work out your answer using the chunk below. (2 MARKS)

```
The average of the maths advantage scores of each degree was used in the analysis. Humanities
has a mean of 6.8 and SE = 2.03. Medicine has a mean of 3.06 and SE = 4.00. Science had a mea
n on -0.77 and SE = 1.99. Other degrees were put into one category which had a mean on 11.65
and SE = 3.36
```

Hide

```
# Use this chunk to complete Exercise 12
maths %>%
  group_by(degree) %>%
  summarise(mean = mean(mathsAdvantage),
            n = n(),
            sd = sd(mathsAdvantage),
            SE = sd/sqrt(n))
```

degree <fctr>	mean <dbl>	n <int>	sd <dbl>	SE <dbl>
Hum	6.8333333	84	18.61780	2.031369
Med	3.0625000	16	15.98945	3.997362

degree <fctr>	mean <dbl>	n <int>	sd <dbl>	SE <dbl>
Other	11.6451613	31	18.69857	3.358363
Sci	-0.7681159	69	16.56995	1.994787

4 rows

EXERCISE 13: What is the appropriate statistical analysis for testing the null hypothesis? What assumptions does the analysis make, and do you think those assumptions are adequately met in this case? Explain why or why not. You can work out your answer using the chunk below. (4 MARKS)

An F-test (ANOVA) is the most suited to this dataset as there are more than one variable that needs to be tested. The assumptions made are that the data for all degrees are normally distributed and have the same variance. It also assumes that all data is independent.
A Q-Q plot and the residuals on the plot are roughly symetrically spaced above and below the horizontal axis indicate that the data is roughly normally distributed. The residual plot also shows that the data has equal variance as the residuals have roughly equal spread on both sides of the horizontal. The data does not look to be associated with any other data points and thus we can assume that the data is independent. This means that we can use the F-test without any modification to the data.

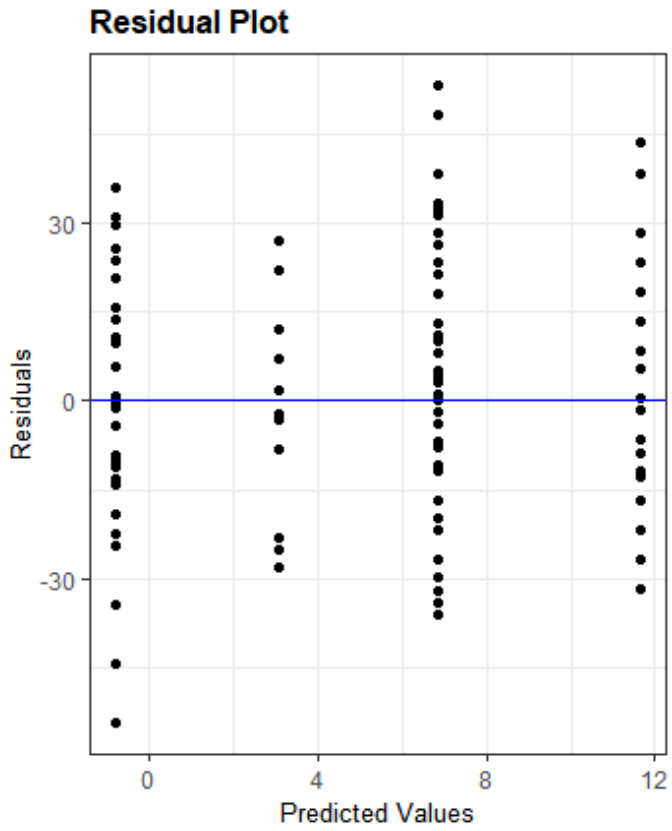
Hide

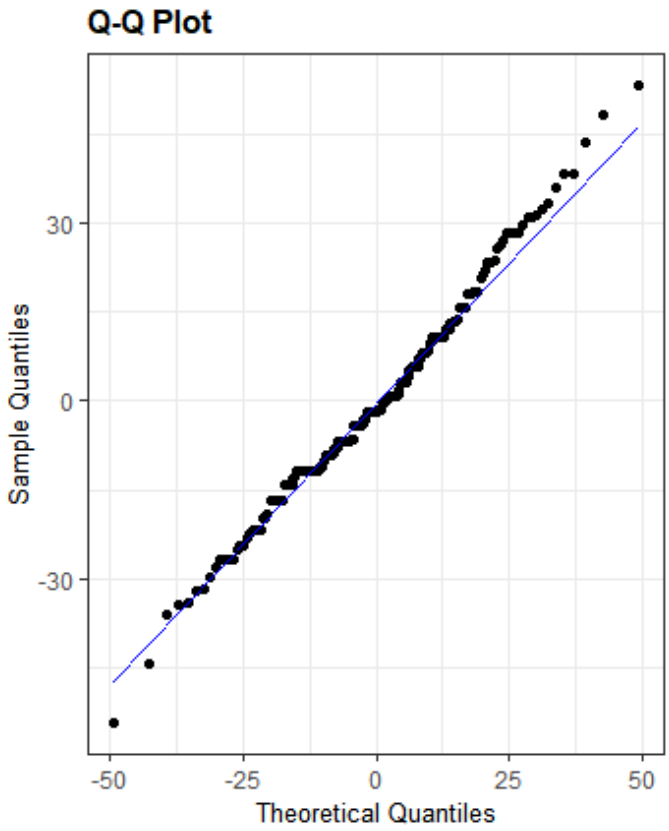
```
# Use this chunk to complete Exercises 13 and 14 (results here will also help you to answer Exercise 16)
maths.mod <- lm(mathsAdvantage ~ degree,
                data = maths,
                contrasts = list(degree = "contr.sum"))

maths.mod
```

```

Call:
lm(formula = mathsAdvantage ~ degree, data = maths, contrasts = list(degree = "contr.sum"))

Coefficients:
(Intercept)      degree1      degree2      degree3
      5.193        1.640       -2.131        6.452
```

Hide

```
resid_panel(maths.mod, plots = c("resid", "qq"))
```





Hide

```
Anova(maths.mod)
```

```
Anova Table (Type II tests)

Response: mathsAdvantage
          Sum Sq  Df F value   Pr(>F)
degree      3983   3  4.2134 0.006482 **
Residuals  61764 196
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

EXERCISE 14: After running the appropriate statistical analysis (which you can do in the chunk above), do you reject or fail to reject the null hypothesis in EXERCISE 11, and why? Explain what this indicates about the effect of a reader’s research area on the perceived advantage of adding maths to a paper. (2 MARKS)

```
Since the p value = 0.0065, we reject the null and state the not not all group means are the
same and that adding maths to a paper has an effect on some degrees more than others.
```

EXERCISE 15: Does the perceived advantage of adding maths to a paper differ significantly, on average, between readers from different research areas? Your answer should explain which readers differ and how (i.e., who perceives maths more or less positively than others). You can work out your answer using the chunk below. (4 MARKS)

```
There is no significant difference between humanities and medicine (p = 0.86), humanities and other degrees (p = 0.57), medicine and other degrees (p = 0.40) and medicine and science (p = 0.86). There was a significant differnce between humanities and science (p = 0.04) where people with humanities degree perceive maths in papers as higher quality research compared to people with science degrees. There was also a signidicant difference between other degrees and science degrees (p = 0.008) wherer other degrees favour papers with paths more than those with science degrees.
```

Hide

```
# Use this chunk to complete Exercises 15 and 16
means <- emmeans(maths.mod, specs = "degree")

pairs(means, adjust = "tukey")
```

```
 contrast    estimate   SE  df t.ratio p.value
 Hum - Med       3.77 4.84 196   0.779  0.8640
 Hum - Other    -4.81 3.73 196  -1.290  0.5705
 Hum - Sci       7.60 2.88 196   2.636  0.0445
 Med - Other    -8.58 5.46 196  -1.571  0.3978
 Med - Sci       3.83 4.93 196   0.778  0.8645
 Other - Sci    12.41 3.84 196   3.234  0.0078

P value adjustment: tukey method for comparing a family of 4 estimates
```

EXERCISE 16: Based on the research description and your analysis of the data, write a brief report of the results as though it was your own study. Follow the guidelines in the week 7 workshop handout for reporting the results of a statistical analysis, providing all information necessary for readers to understand and properly evaluate the research outcomes. Be sure to use your own words (do not copy the week 7 workshop handout). There is no minimum or maximum word limit, but a few sentences should suffice. (4 MARKS)

```
The effects of adding maths to a reseach paper on people with varying degrees was tested with a sample of 200 people with varying degrees. An F-test was run to determine if there is any difference in how it affects each degree. No significant difference between humanities and medicine (p = 0.86), humanities and other degrees (p = 0.57), medicine and other degrees (p = 0.40) and medicine and science (p = 0.86). However, there was a significant differnce between humanities and science (p = 0.04) where people with humanities degree perceive maths in papers as higher quality research compared to people with science degrees. There was also a signidicant difference between other degrees and science degrees (p = 0.008) wherer other degrees favour papers with paths more than those with science degrees.
```

Problem set 4: Metabolism and body size

The relationship between size and energy metabolism has fascinated biologists for over a century. An organism's metabolic rate is the amount of energy that the organism expends on life processes in a given time period, and is expected to scale with body size because larger bodies have more metabolising tissue. To understand the relationship between metabolism and body size in primates, a researcher gathered data on metabolic rate (in watts) and body size in terms of mass (in grams) for 17 primate species. The data are contained in the file 'primates.csv', which you can load by running the chunk below.

Hide

```
# Run this chunk to load the data for Problem set 3
primates <- read_csv("./primates.csv",
                     col_names = TRUE, na = "NA",
                     col_types = list(species = col_factor(),
                                      mass = col_double(),
                                      metabolic_rate = col_double()))
glimpse(primates, width = 75)
```

```
Rows: 17
Columns: 3
$ species        <fct> Alouatta palliata, Aotus trivirgatus, Arctocebus c…
$ mass           <dbl> 4670.0, 1020.0, 206.0, 190.0, 105.0, 300.0, 261.5,…
$ metabolic_rate <dbl> 11.57, 2.56, 0.73, 0.86, 0.55, 1.10, 1.20, 2.93, 0…
```

Hide

```
primates
```

species <fctr>	mass <dbl>	metabolic_rate <dbl>
Alouatta palliata	4670.0	11.57
Aotus trivirgatus	1020.0	2.56
Arctocebus calabarensis	206.0	0.73
Callithrix jachus	190.0	0.86
Cebuella pygmaea	105.0	0.55
Cheirogaleus medius	300.0	1.10
Euoticus elegantilus	261.5	1.20
Galago crassicaudatus	1039.0	2.93
Galago demidovii	61.0	0.42
Galago elegantulus	261.5	1.20

1-10 of 17 rows Previous **1** 2 Next

EXERCISE 17: Based on the research description above, what is the response variable of interest and what is the explanatory variable of interest? (2 MARKS)

```
Explanatory variable is the mass of primates. The response variable is their metabolic rate.
```

EXERCISE 18: Based on the research description above, what are the null and alternate hypotheses of interest, and what parameter estimate do hypotheses focus on? (2 MARKS)

```
Null hypothesis: The mass of primates does not affect their metabolic rate

Altnernate hypothesis: Mass of primates changes their metabolic rate.
```

EXERCISE 19: Fit the appropriate linear model for testing the null hypothesis using untransformed data, and check the model's assumptions. Do you think they are adequately met in this case? Give 2 reasons to justify your answer. You can work out your answer using the chunk below. (3 MARKS)

```
The assumptions are not met as the data is not normally distributed as shown by the outlier p
oints on the qq plot. The data also does not have equal variance as the residuals are not eve
nly distributed above and below the horizontal line.
```

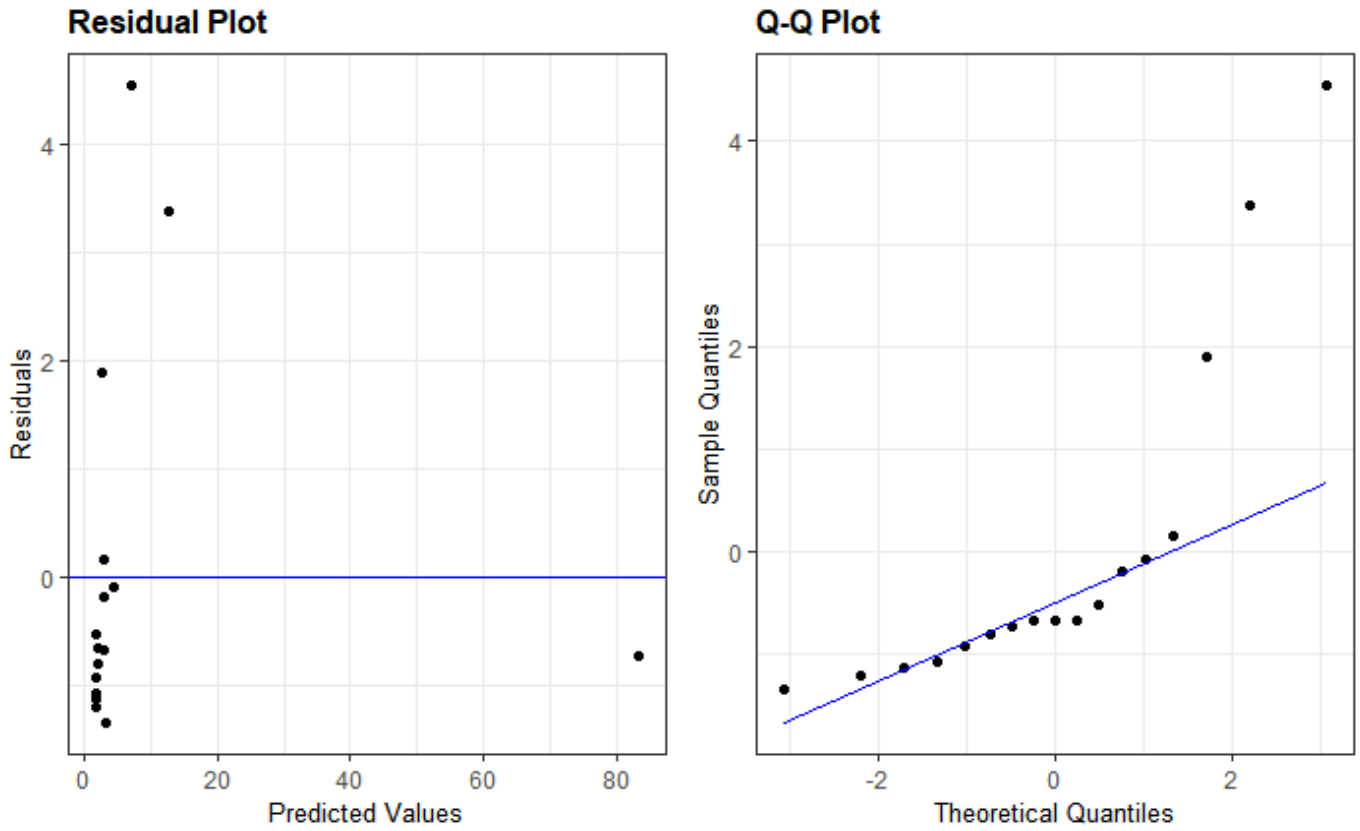
Hide

```
# Use this chunk to complete Exercise 19
primates.mod <- lm(metabolic_rate ~ mass,
                   data = primates)

primates.mod
```

```

Call:
lm(formula = metabolic_rate ~ mass, data = primates)

Coefficients:
(Intercept)         mass
   1.554912     0.001171
```

Hide

```
resid_panel(primates.mod, plots = c("resid", "qq"))
```



EXERCISE 20: Re-fit the linear model using ln-transformed data (where 'ln' means natural log) for one or both variables as necessary to adequately meet the model's assumptions. Which variable or variables did you end up transforming, and why? Explain your answer with reference to the

model's assumptions. You can work out your answer using the chunk below. (3 MARKS)

Both mass and metabolic rate needed to be transformed in order to linearise the relationship between the variables since the relationship of the untransfored data seems to be exponentaial based on the untransformed qq graph, making a log-log graph by taking the natural log of both variables linearises the data. The qq plot now shows that the values are approximately normally distributed and the residual plot shows that there is roughly equal variance in the data.

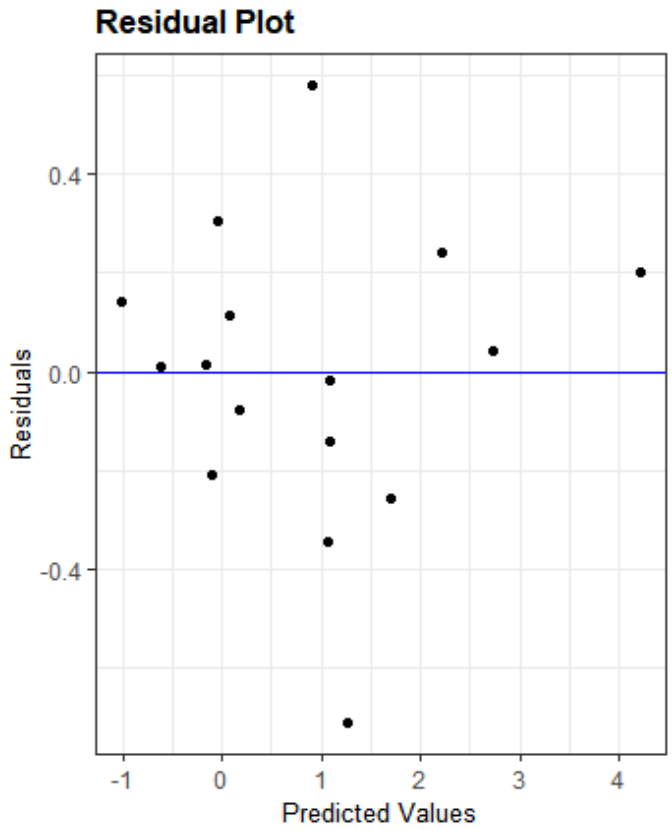
Hide

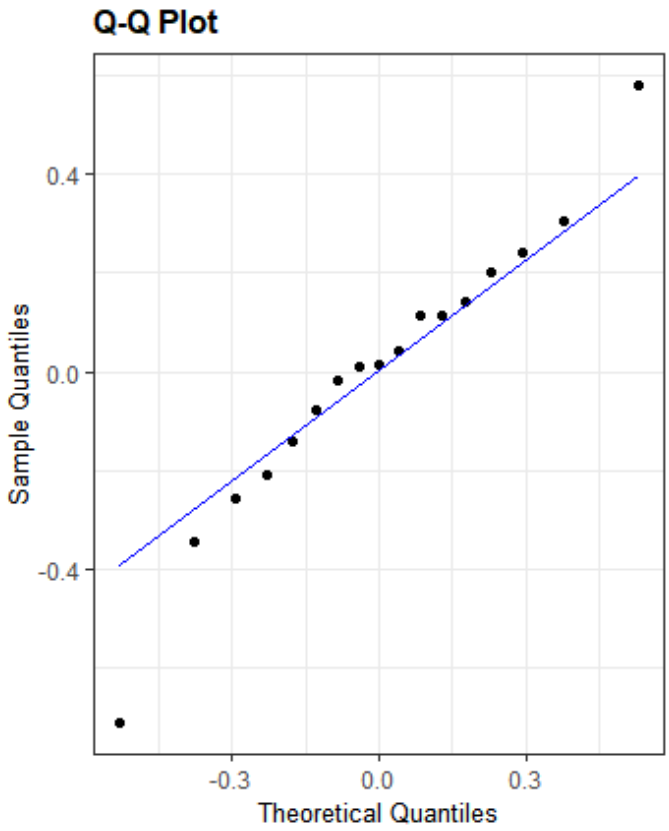
```
# Use this chunk to complete Exercise 20
primates.trans <- primates %>%
  mutate(metabolic_rate_ln = log(metabolic_rate),
         mass_ln = log(mass))

primates.mod2 <- lm(metabolic_rate_ln ~ mass_ln,
                data = primates.trans)

primates.mod2
```

```

Call:
lm(formula = metabolic_rate_ln ~ mass_ln, data = primates.trans)

Coefficients:
(Intercept)      mass_ln
    -4.0579       0.7416
```

Hide

```
resid_panel(primates.mod2, plots = c("resid", "qq"))
```





Hide

```
summary(primates.mod2)
```

```
Call:
lm(formula = metabolic_rate_ln ~ mass_ln, data = primates.trans)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71136 -0.13959  0.01586  0.14175  0.57990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.05787    0.28770  -14.11 4.61e-10 ***
mass_ln      0.74160    0.04198   17.66 1.89e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2983 on 15 degrees of freedom
Multiple R-squared:  0.9541,    Adjusted R-squared:  0.9511
F-statistic:   312 on 1 and 15 DF,  p-value: 1.891e-11
```

EXERCISE 21: Based on your final model, do you reject or fail to reject the null hypothesis? Report the relevant test results. Explain what the test indicates about the relationship between metabolic rate and body mass in primates. (2 MARKS)

```
From the linearised data, we reject the null hypothesis that there is no relationship between mass and metabolic rates in primates. The relationship has a slope of 0.74 (p = 1.9e-11) with an adjusted R-squared value of 0.9511. This means that there is an exponential relationship b etween mass an metaolic rate of primates after accounting for the linearisation that was perf ormed on the dataset.
```

EXERCISE 22: What is the estimated change in metabolic rate per change in body mass? Explain what the estimate indicates about the relationship between metabolic rate and body mass in primates. Report results on the transformed scale. (1 MARK)

```
The estinated change is 0.74 watts per gram
```

EXERCISE 23: What fraction of the among-species variation in metabolic rate is explained by its relationship with body mass? Report results on the transformed scale. You can work out your answer using the chunk below. (1 MARK)

```
0.95 as R-squaredd value is 0.95.
```

Hide

```
# Use this chunk to complete Exercise 23
```