THE UNIVERSITY of ADELAIDE

Primary Examination **!! PRACTICE EXAM !!**

MATHS 7107 Data Taming

Course Coordinator: Anthony Mays

Writing time: 120 mins

Number of questions: 7 Written, 16 Multiple Choice

Total marks: ??

Instructions for Candidates

- This examination is **closed book**.
- Examination materials must not be removed from the examination room.
- Answer all written questions in the answer book provided.
- Answer all multiple choice questions on the provided multiple choice sheet.

Permitted Materials

- Multiple Choice Question (MCQ) Sheet
- Paper translation dictionary
- Scientific calculator.
- One double-sided A4 page of handwritten or typed notes.

DO NOT COMMENCE WRITING UNTIL INSTRUCTED TO DO SO.
STOP WRITING IMMEDIATELY WHEN INSTRUCTED

Written Questions

1. A committee has been established to organise 200-year anniversary celebrations in each Australian capital city. A large statue has been built and needs to be moved to each city for the celebration. The following tibble contains some of the information that the committee has gathered

```
# A tibble: 8 × 5
  Name                          population urban_pop `Capital city`  Year
  <chr>                              <dbl>     <dbl> <fct>          <dbl>
1 New South Wales                  7759274      64.8 Sydney          1788
2 Victoria                         6179249      76.5 Melbourne       1851
3 Queensland                       4848877      48.7 Brisbane        1860
4 Western Australia                2558951      79.0 Perth           1829
5 South Australia                  1713054      77.3 Adelaide        1836
6 Tasmania                          517588      43.4 Hobart          1826
7 Australian Capital Territory      403468     100   Canberra        1913
8 Northern Territory                245740      59.4 Darwin          1911
```

The variables in the table are:

- Name: the name of the state or territory
- population: the number of people residing in the state
- urban_pop: the percentage of the state's population who lives in the state's capital city
- Capital city: the name of the state's capital city
- Year: the year the city became the state's capital city, which is used for planning where the statue needs to be.

(a) For each of the variables in the dataset identify the type of variable, ie. is it **quantitative continuous**, **quantitative discrete**, **categorical nominal** or **categorical ordinal**. Make sure you write a short description justifying your choice.

- Name: **Categorical nominal** since this is just the name of the state with no implied ordering.
- population: **Quantitative discrete** since this is the number of people in the state, and people are measured in whole units
- urban_pop: **Quantitative continuous** since this is a ratio of two integers it could be any rational number.
- Capital city: **Categorical nominal** since this is just the name of the city with no implied ordering.
- Year: **Categorical ordinal** since this is the name of the year, but the ordering of the years is important for planning.

Please turn over for page 3

(b) For each variable state whether the corresponding column in the tibble is the correct data type. If it is incorrect, say what the correct data type should be.

- Name: Incorrect. This should be a factor (`<fct>`).
- population: Incorrect. This should be an integer (`<int>`).
- urban_pop: Correct.
- Capital city: Correct.
- Year: Incorrect. This should be an ordered factor (`<ord>`).

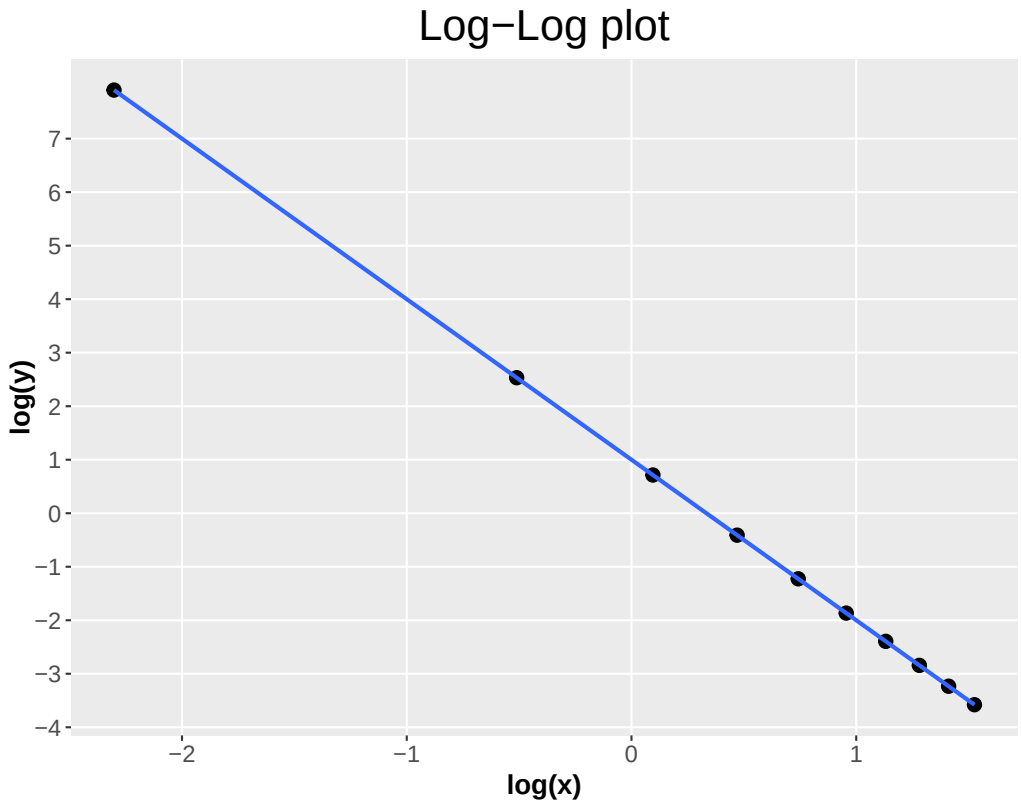

Figure 1: Log-Log plot of data for Question 2.

2. (a) Assume that a data set consists of continuous variables x and y, which are related by the formula

$$y = \alpha x^k.$$

Using the logarithm laws, show that the Log-Log plot of this data will be a straight line.

$$y = \alpha x^k \Rightarrow \log(y) = \log(\alpha x^k) = \log(\alpha) + k\log(x)$$

which is of the form $y^* = mx^* + c$ for a straight line.

(b) Using the straight line in part (a), write down the:

i). gradient of the line,

ii). the value at which the line cuts the vertical axis.

Please turn over for page 5

i). gradient $= k$

ii). vertical axis intercept $= \log(\alpha)$

(c) Using the Log-Log plot in Figure 1 determine the explicit relationship between x and y. **(Make sure you show your working.)**

We identify the points $(0, 1)$ and $(1, -2)$ on the line. Using these points we calculate

- the gradient $= (-2 - 1)/(1 - 0) = -3 = k$
- the vertical axis intercept $= 1 = \log(\alpha)$

Plugging these into the equation in part (a) we get

$$y = \alpha x^k = ex^{-3}.$$

x_1	x_2	x_3	x_4	x_5
1	0	2	-1	3

Table 1: Dataset for Questions 3 and 4.

3. For this question, use the dataset in Table 1.

   (a) Transform the dataset in Table 1 to x^* by applying Min-Max scaling. Calculate your answers to 2 decimal places. **(Make sure you show your working.)**

$\min(x_1, x_2, x_3, x_4, x_5) = -1$ and $\max(x_1, x_2, x_3, x_4, x_5) = 3$

$$x_1^* = \frac{1-(-1)}{3-(-1)} = \frac{1}{2}$$
$$x_2^* = \frac{0-(-1)}{3-(-1)} = \frac{1}{4}$$
$$x_3^* = \frac{2-(-1)}{3-(-1)} = \frac{3}{4}$$
$$x_4^* = \frac{-1-(-1)}{3-(-1)} = 0$$
$$x_5^* = \frac{3-(-1)}{3-(-1)} = 1$$

   (b) What are the minimum and maximum values of the transformed dataset x^*? Give exact values for your answers. **(Make sure you show your working.)**

$\min(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*) = 0$ and $\max(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*) = 1$

4. For this question, use the dataset in Table 1.

   (a) Find the mean $\overline{x}$ and (sample) standard deviation σ_x of the dataset. Calculate your answers to 2 decimal places. **(Make sure you show your working.)**

$$\overline{x} = \frac{1+0+2+(-1+3)}{5} = \frac{5}{5} = 1.00$$

$$\begin{aligned}\sigma_x &= \sqrt{\frac{(1-1)^2+(0-1)^2+(2-1)^2+(-1-1)^2+(3-1)^2}{5-1}} \\ &= \sqrt{\frac{0+1+1+4+4}{4}} \\ &= \sqrt{\frac{10}{4}} = \sqrt{\frac{5}{2}} \approx 1.58\end{aligned}$$

(b) Standardise the dataset by calculating the z-scores x_j^* for each x_j. Calculate your answers to 2 decimal places. **(Make sure you show your working.)**

$$\begin{aligned}x_1^* &= \frac{1-1}{\sqrt{5/2}} = 0.00 \\ x_2^* &= \frac{0-1}{\sqrt{5/2}} = -\sqrt{2/5} \approx -0.63 \\ x_3^* &= \frac{2-1}{\sqrt{5/2}} = \sqrt{2/5} \approx 0.63 \\ x_4^* &= \frac{-1-1}{\sqrt{5/2}} = -2\sqrt{2/5} \approx -1.26 \\ x_5^* &= \frac{3-1}{\sqrt{5/2}} = 2\sqrt{2/5} \approx 1.26\end{aligned}$$

(c) What is the mean $\overline{x^*}$ and standard deviation σ_{x^*} of the transformed data set. Give exact values for your answers. **(Make sure you show your working.)**

$$\overline{x^*} = \frac{0-\sqrt{2/5}+\sqrt{2/5}-2\sqrt{2/5}+2\sqrt{2/5}}{5} = 0$$

$$\begin{aligned}\sigma_{x^*} &= \sqrt{\frac{1}{5-1}\left((0)^2+\left(-\sqrt{\frac{2}{5}}\right)^2+\left(\sqrt{\frac{2}{5}}\right)^2+\left(-2\sqrt{\frac{2}{5}}\right)^2+\left(2\sqrt{\frac{2}{5}}\right)^2\right)} \\ &= \sqrt{\frac{1}{4}\left(\frac{2}{5}+\frac{2}{5}+\frac{8}{5}+\frac{8}{5}\right)} \\ &= \sqrt{\frac{20}{20}} = \sqrt{1} = 1\end{aligned}$$

5. Show that when $x \neq 0$ the Box-Cox transformation is continuous at $\lambda = 0$, by showing that

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \log(x) \qquad (\text{for } x \neq 0).$$

Since $x^\lambda \to 1$ when $x \neq 0$ we have that

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda}$$

is of the form $\frac{0}{0}$, and so we can use L'Hôpital's rule.

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{\frac{d}{d\lambda}\left(x^\lambda - 1\right)}{\frac{d}{d\lambda}\lambda}$$

We calculate the derivatives $\frac{d}{d\lambda}\lambda = 1$ and

$$\begin{aligned} \frac{d}{d\lambda}\left(x^\lambda - 1\right) &= \frac{d}{d\lambda}\left(e^{\log(x^\lambda)} - 1\right) = \frac{d}{d\lambda}\left(e^{\lambda \log(x)} - 1\right) \\ &= \log(x) e^{\lambda \log(x)} = \log(x) e^{\log(x^\lambda)} = \log(x) x^\lambda \end{aligned}$$

and substitute back into the limit to find

$$\begin{aligned} \lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} &= \lim_{\lambda \to 0} \frac{\frac{d}{d\lambda}\left(x^\lambda - 1\right)}{\frac{d}{d\lambda}\lambda} \\ &= \lim_{\lambda \to 0} \frac{\log(x) x^\lambda}{1} \\ &= \log(x) \lim_{\lambda \to 0} x^\lambda \\ &= \log(x) \qquad (\text{for } x \neq 0). \end{aligned}$$

6. We are looking to measure the amount of pollen in the air. We have built 31 cubic boxes, each one with a pollen counter in it. The side lengths of each box is recorded in the list $(s_1, \ldots, s_{31})$, where the side lengths are measured in metres. We collect pollen counts from the air samples inside each box, and record the pollen counts in box j as p_j.

   We expect to find a good linear fit with the model

   $$p_j = \beta_0 + \beta_1 z_j + \epsilon_j, \qquad N(0, \sigma) \tag{1}$$

   where $z_j = s_j^3$.

   We use R to fit our linear model to the data and we obtain the following output:

```
> pollen_lm <- lm(p ~ z, pollen)
> summary(pollen_lm)

Call:
lm(formula = p ~ z, data = pollen)

Residuals:
    Min      1Q  Median      3Q     Max
-22.195  -6.719  -0.049   6.826  20.877

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.23517    3.59069   1.736   0.0931 .
z            1.31177    0.05785  22.675   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.27 on ##### degrees of freedom
Multiple R-squared:  0.9466,Adjusted R-squared:  0.9448
F-statistic: 514.1 on 1 and ##### DF,  p-value: < 2.2e-16
```

   - *Note that we have replaced the number degrees of freedom with* `#####`*, and so you need to calculate this number.*

Please turn over for page 10

(a) Use the information in the output to write the equation for the line of best fit in Equation (1) with the appropriate values substituted. (Round off your answer to 4 decimal places.)

$$\hat{p}_j = 6.2352 + 1.3118 z_j$$

(b) From the output, what is the estimate of the standard deviation of the errors? (Provide your answer to 2 decimal places.)

$$\sigma \approx 11.27.$$

(c) With this model, what is the prediction for the pollen count in a box of side length $4.3m$? (Round off your answer to the nearest whole number of pollen grains.)

$$s = 4.3 \quad \Rightarrow \quad z = (4.3)^3 \approx 79.507$$

Substituting this into our model gives

$$\hat{p} \approx 6.2352 + 1.3118 \times 79.507 \approx 111.$$

(d) Write down the details for the two t-tests performed when fitting this linear model. State the null and alternative hypotheses, and the distribution being tested against (including the number of degrees of freedom).

R performed two t-tests, one for the intercept and one for the gradient. The hypotheses for the intercept test were

$$\begin{aligned} H_0 &: \quad \beta_0 = 0 \qquad \text{(null hypothesis)} \\ H_a &: \quad \beta_0 \neq 0 \qquad \text{(althernative hypothesis)} \end{aligned}$$

which were tested on a t-distribution with 29 degrees of freedom.
The hypotheses for the slope test were

$$\begin{aligned} H_0 &: \quad \beta_1 = 0 \qquad \text{(null hypothesis)} \\ H_a &: \quad \beta_1 \neq 0 \qquad \text{(althernative hypothesis)} \end{aligned}$$

which were tested on a t-distribution with 29 degrees of freedom.

(e) What are the relevant P-values for the tests? Are the estimates for the model parameters significant at the 95% level?

Please turn over for page 11

- intercept test: P-value $= 0.0931$. Not significant at 95% level.
- gradient test: P-value $< 2 \times 10^{-16}$. Significant at 95% level.

(f) Use the graphs in Figures 2 – 4 to determine if the assumptions for fitting a linear model are satisfied. (Make sure you refer to specific Figures in your answers.) In addition, for the **independence** assumption, come up with at least one reason why the assumption may not be satisfied.

i). Linearity: Satisfied. Using Figure 2 we see the red line is roughly flat and the points are roughly evenly distributed above and below it. [You might also argue that it is not satisfied, since the line is not really that flat. In this case, you should probably mention something like the data points in rows 1 and 4 might want to be investigated.]

ii). Constant variance (homoscedasticity): Satisfied. Using Figure 3 we see the red line is roughly flat and as we move from left to right the points have roughly the same distribution around the line.

iii). Normality: Satisfied. Using Figure 4 we see the bulk of the points are between -1 and 1, and in this section the points lie close to the diagonal.

iv). Independence: This can't be determined from a graph. We need more information about the details of the dataset to know if independence is satisfied. One problem for this assumption could be the placement of the boxes — if they are too close together, then the pollen inside one box could flow to the other box and be counted twice.

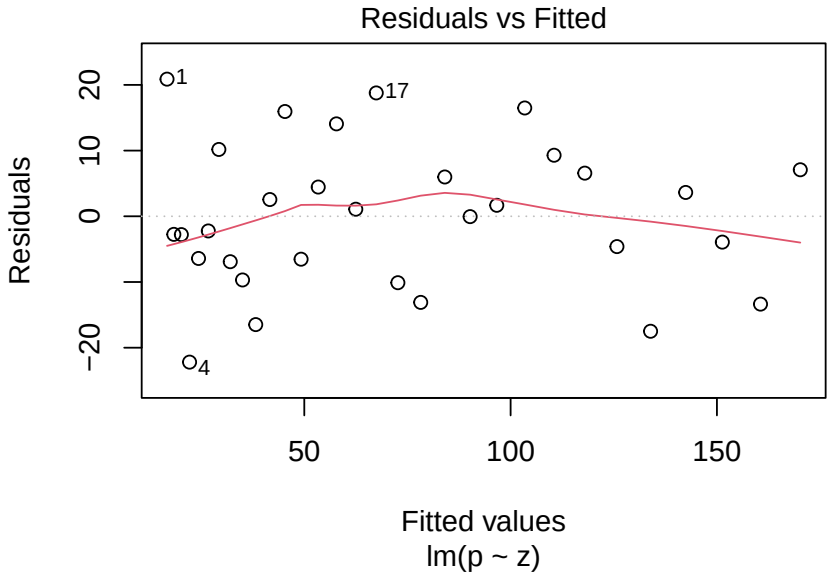

Figure 2: A residuals vs fitted graph for the linear model in Question 6.

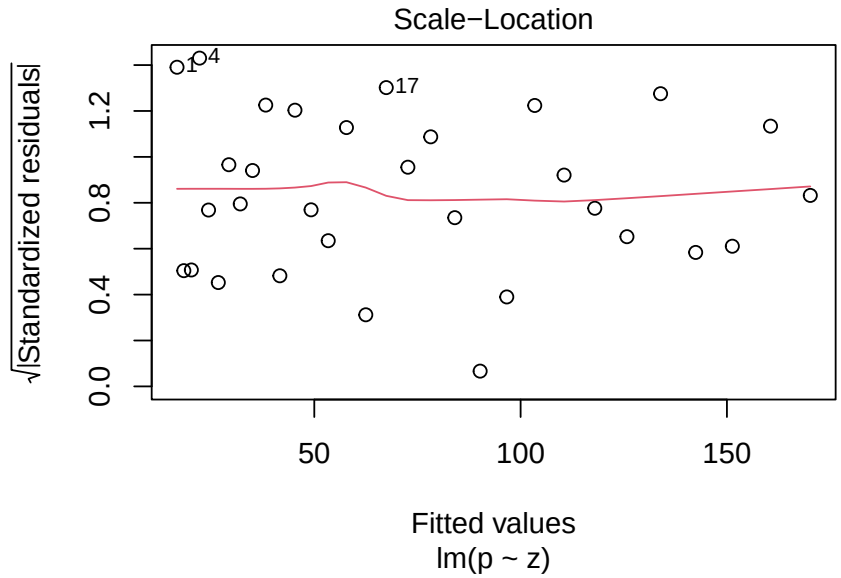

Figure 3: A scale–location graph for the linear model in Question 6.

Please turn over for page 13

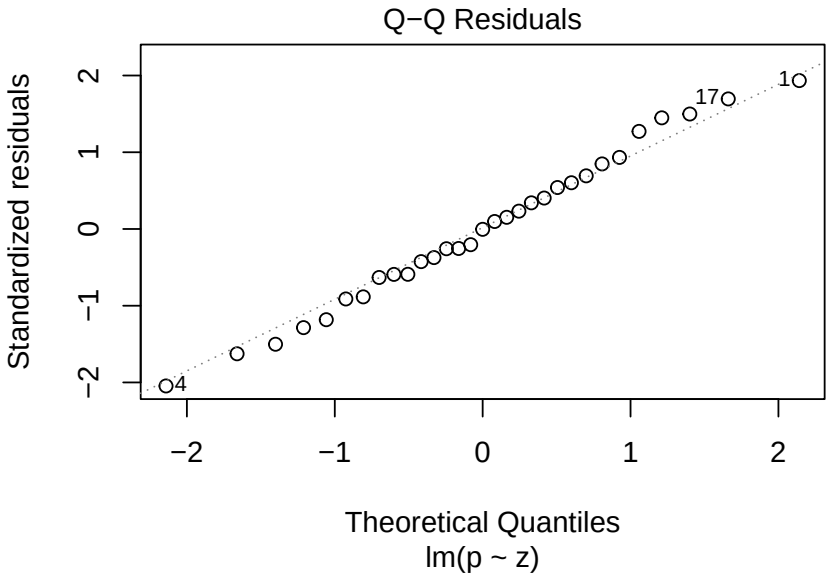

Figure 4: A Q-Q plot of residuals for the linear model in Question 6.

7. A medical test for a certain disease has the following confusion matrix:

	No disease	Has disease
Test negative	100	20
Test positive	50	180

A **true positive** is where a patient has the disease and the test is positive.

(a) Calculate the sensitivity of the test. Provide an exact answer or 3 significant figures of precision. **(Make sure you show your working.)**

$$\begin{aligned}\text{sensitivity} &= \frac{\#\text{ true positive tests}}{\text{total }\#\text{ of diseased patients}} \\ &= \frac{180}{180+20} = \frac{180}{200} = \frac{9}{10} = 0.900.\end{aligned}$$

(b) Calculate the specificity of the test. Provide an exact answer or 3 significant figures of precision. **(Make sure you show your working.)**

$$\begin{aligned}\text{specificity} &= \frac{\#\text{ true negative tests}}{\text{total }\#\text{ of healthy patients}} \\ &= \frac{100}{100+50} = \frac{100}{150} = \frac{2}{3} \approx 0.667.\end{aligned}$$

(c) Calculate the accuracy of the test. Provide an exact answer or 3 significant figures of precision. **(Make sure you show your working.)**

$$\begin{aligned}\text{accuracy} &= \frac{\#\text{ true positive tests} + \#\text{ true negative tests}}{\text{total }\#\text{ of tests}} \\ &= \frac{100+180}{50+100+180+20} = \frac{280}{350} = \frac{4}{5} = 0.800.\end{aligned}$$

(d) Based on your results from parts (a)–(c), would you estimate that the AUC (area under the curve) for the ROC curve is closest to:

- between 0.65 and 0.90
- between 0.45 and 0.55
- between 0.1 and 0.35
- between -0.35 and -0.1
- between -0.9 and -0.65

Justify your reasoning.

Please turn over for page 15

Between 0.65 and 0.90.
An AUC less than zero is not possible, and since the sensitivity and specificity are both clearly above 0.5, (assuming a smooth curve), the ROC curve must lie well above the diagonal. This gives us an AUC close to one.

(e) What does an AUC value close to 0 tell you about your prediction model? If this was the case, how could we improve the model?

That the model is predicting the variables the wrong way around. To improve the model, we should switch the prediction to the other class (ie. instead of predicting a diseased patient, we should predict a healthy patient).

Multiple Choice Questions

Use the multiple choice answer sheet to record your answers.

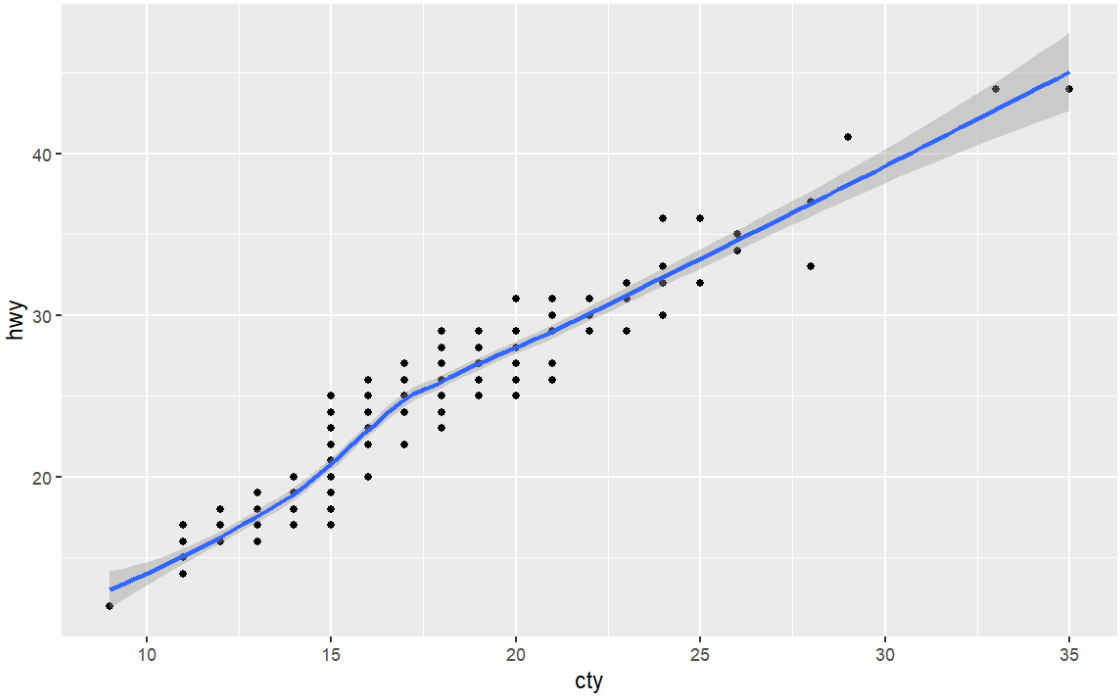

Figure 5: Plot of city fuel efficiency (`cty`) against highway field efficiency (`hwy`) for Multiple Choice Questions 1 and 2.

1. From the data in Figure 5 what is the relationship between city and highway fuel efficiency?
   (a) A strong positive relationship.
   (b) A weak negative relationship.
   (c) No relationship.
   (d) Impossible to tell from this plot.

   (a) correct

2. What R code could have generated the plot in Figure 5? (Select all the correct answers, and only the correct answers.)
   (a) `mpg %>% ggplot(aes(x=cty, y=hwy))+geom_point()+geom_smooth(method="lm")`
   (b) `mpg %>% ggplot(aes(x=cty, y=hwy))+geom_point()+geom_smooth()`
   (c) `ggplot(mpg, aes(x=cty, y=hwy))+geom_smooth()`
   (d) `ggplot(mpg, aes(x=cty, y=hwy))+geom_point()+geom_smooth()`

Please turn over for page 17

(b) & (d) correct

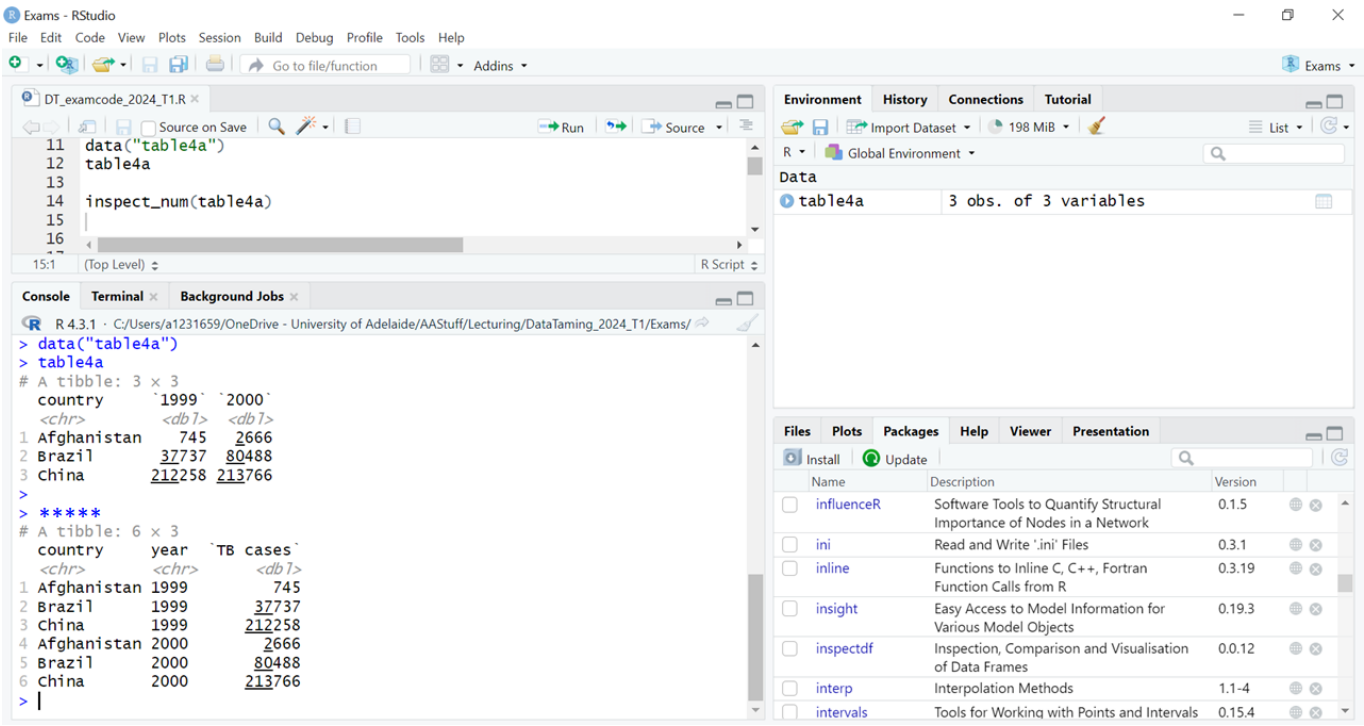

Figure 6: RStudio screenshot for Multiple Choice Questions 3–5.

3. Using the RStudio screenshot in Figure 6, what is the name of the Project?

   (a) `table4a`

   (b) `Console`

   (c) `DT_examcode_2024_T1`

   (d) `Exams`

   (d) correct

4. Look at the RStudio screenshot in Figure 6. If we name the tidy tibble `table4a_tidy`, which command will convert it back into wide form?

   (a) `gather(table4a_tidy, key = "year", value = "TB cases", ‘1999‘:‘2000‘)`

   (b) `inspect_df(table4a_tidy)`

   (c) `wide_form(table4a_tidy, year, ‘TB cases‘)`

   (d) `spread(table4a_tidy, key = year, value = ‘TB cases‘)`

   (d) correct

Please turn over for page 19

5. We would like to obtain summary statistics for the numerical variables in `table4a`. Based on the information in the RStudio screenshot in Figure 6, will the command `inspect_num(table4a)` work for this purpose?

   (a) Yes, `inspect_num` is the correct command.

   (b) No, `inspect_num` does not calculate summary statistics.

   (c) No, we must first run the command `library(inspectdf)`.

   (d) No, we must first run the two commands `install.packages("inspectdf")` and `library(inspectdf)`.

   (c) correct

6. Using the RStudio screenshot in Figure 6, what command will return a 1×1 tibble containing just the value `745`? (Select all the correct answers, and only the correct answers.)

   (a) `table4a %>% filter(country=="Afghanistan") %>% select("1999")`

   (b) `filter(table4a, country=="Afghanistan")`

   (c) `table4a %>% select("country"=="Afghanistan") %>% filter("1999")`

   (d) `table4a[1,2]`

   (a) & (d) correct

7. If we run the command `test_string <- "abc (15) xyz"`, what is the command to extract just the number `15`? (Select all the correct answers, and only the correct answers.)

   (a) `str_match(test_string, "abc \\((\\d*)")[,2]`

   (b) `str_match(test_string, "abc \\((\\d)")[,2]`

   (c) `str_match(test_string, "abc (\\(\\d*)")[,2]`

   (d) `str_match(test_string, "abc \\((.*)\\)")[,2]`

   (a) & (d) correct

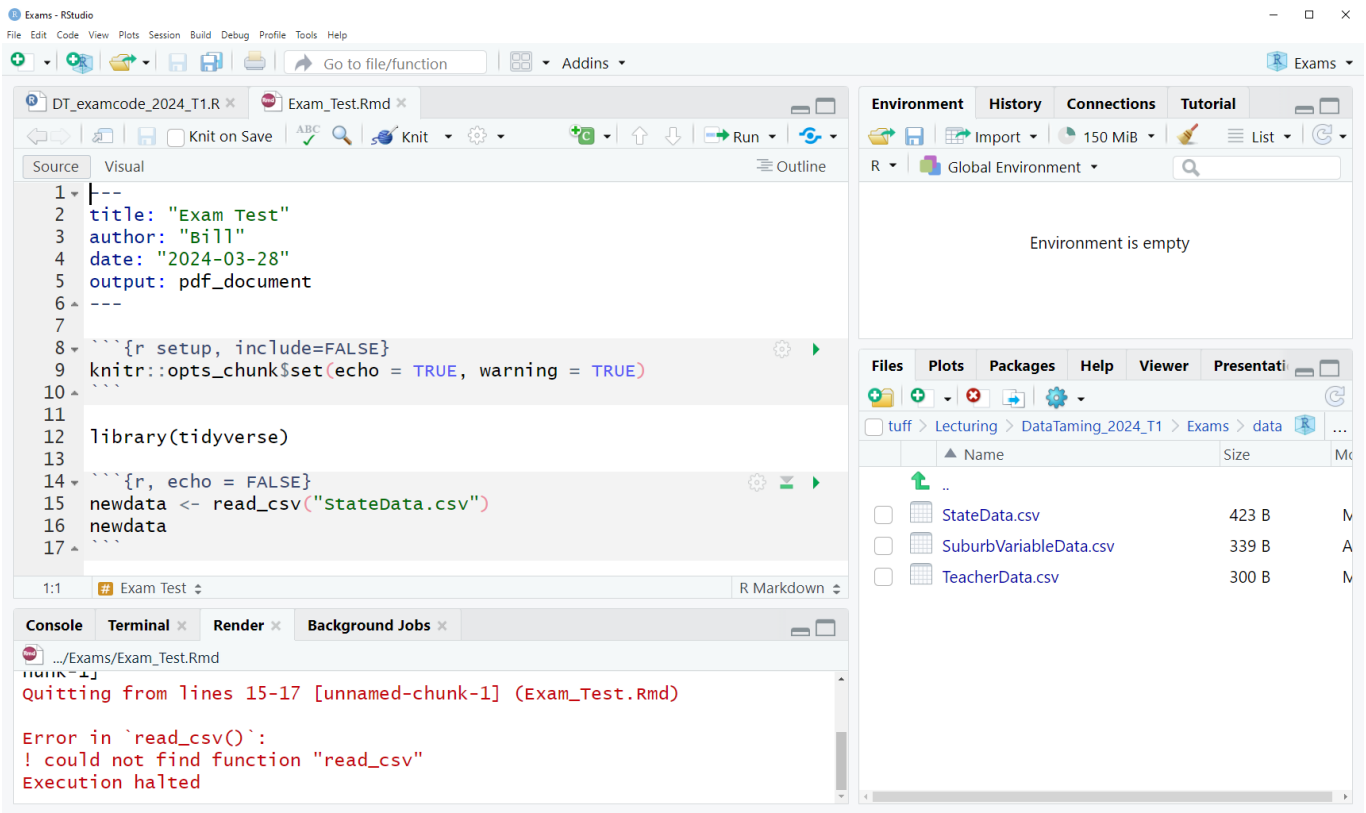

Figure 7: RStudio screenshot for Multiple Choice Questions 8–9.

8. We have hit an error when knitting the R Markdown file in Figure 7. We definitely have the `tidyverse` package installed, so what problems will prevent our file from knitting successfully?

   (a) We have not specified the correct directory for the `csv` file and we need to click "Knit on Save".

   (b) `library(tidyverse)` is not inside a code chunk, and we have not specified the correct directory for the `csv` file.

   (c) We should have `warning = FALSE` and we need to click "Knit on Save".

   (d) `library(tidyverse)` is not inside a code chunk and we should have `warning = FALSE`.

   (b) correct

9. Once we get the R Markdown file in Figure 7 to knit, we find that the output is being displayed in the PDF. However, the code `newdata <- read_csv("StateData.csv")` runs successfully, but produces many warnings that we don't want to see. What changes could we make to prevent R Markdown from showing any warnings in the whole document?

   (a) Include the command `print(!warnings)` after line 16

   (b) Include the command `print(read_csv == FALSE)` after line 9

   (c) Change `warning = TRUE` to `warning = FALSE` on line 9

   (d) Change `echo = FALSE` to `echo = FALSE, warning = FALSE` on line 14

   (c) correct

See the attached page "mpg2 data set" for Questions 10 to 16.

For the remaining questions we are going to be working with the mpg2 data set, which is a simplified version of mpg.

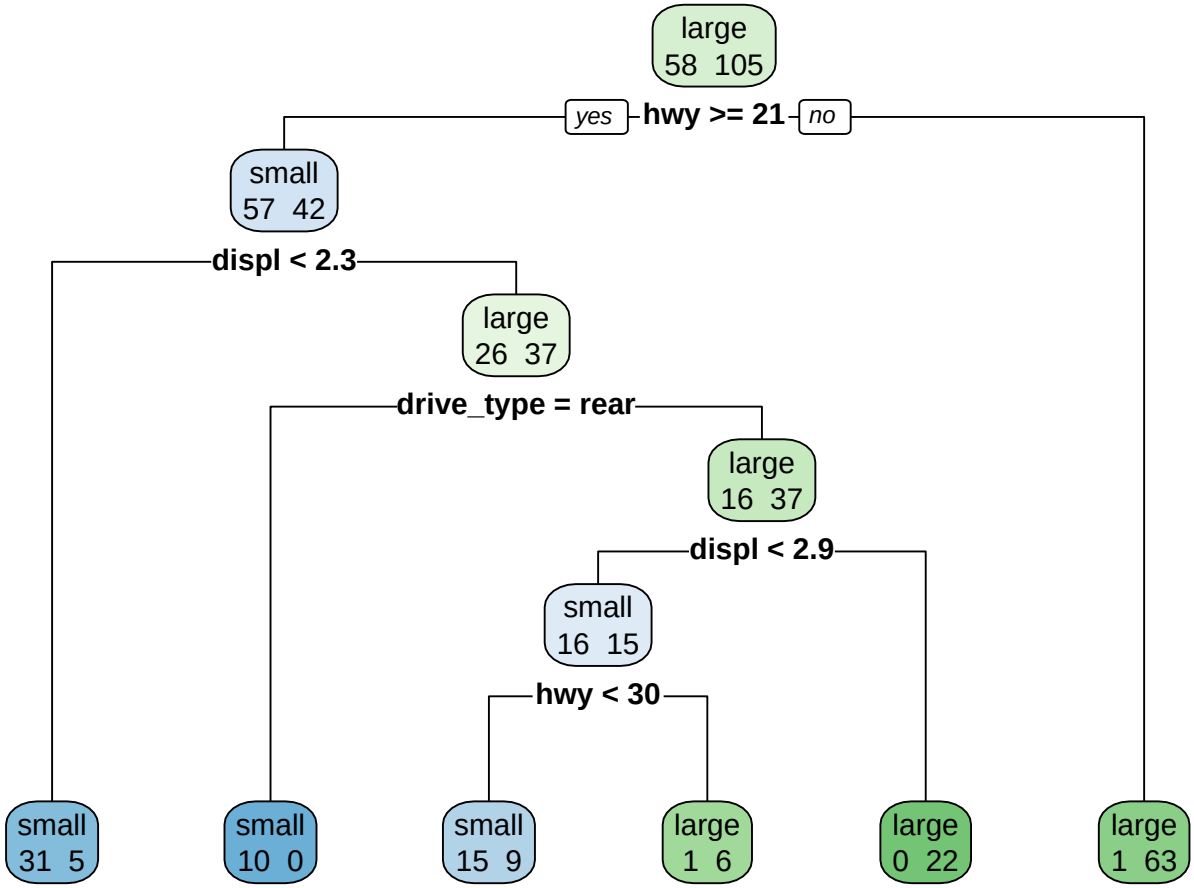

Figure 8: Decision tree diagram for Multiple Choice Questions 10–13.

10. In Figure 8 we have fit the model `class ~ year + cyl + drive_type + displ` to a training subset of the mpg2 dataset for predicting the type of car, either small or large. What type of model have we fit?

    (a) Logistic tree

    (b) General linear model tree

    (c) Regression tree

    (d) Classification tree

(d) correct

11. From the diagram in Figure 8, determine how decisions the model made.
    (a) 5
    (b) 6
    (c) 11
    (d) 21

    (a) correct

12. From the diagram in Figure 8, what proportion of the whole dataset was chosen for our training set?
    (a) 58%
    (b) 105%
    (c) 70%
    (d) 15%

    (c) correct

13. Using the model in Figure 8, what size do we predict for a car with an engine size of 3L, highway fuel efficiency of 23 miles per gallon, four wheel drive and built in 2008?
    (a) small
    (b) large
    (c) It can't be determined from this diagram
    (d) The model was inconclusive

    (b) correct

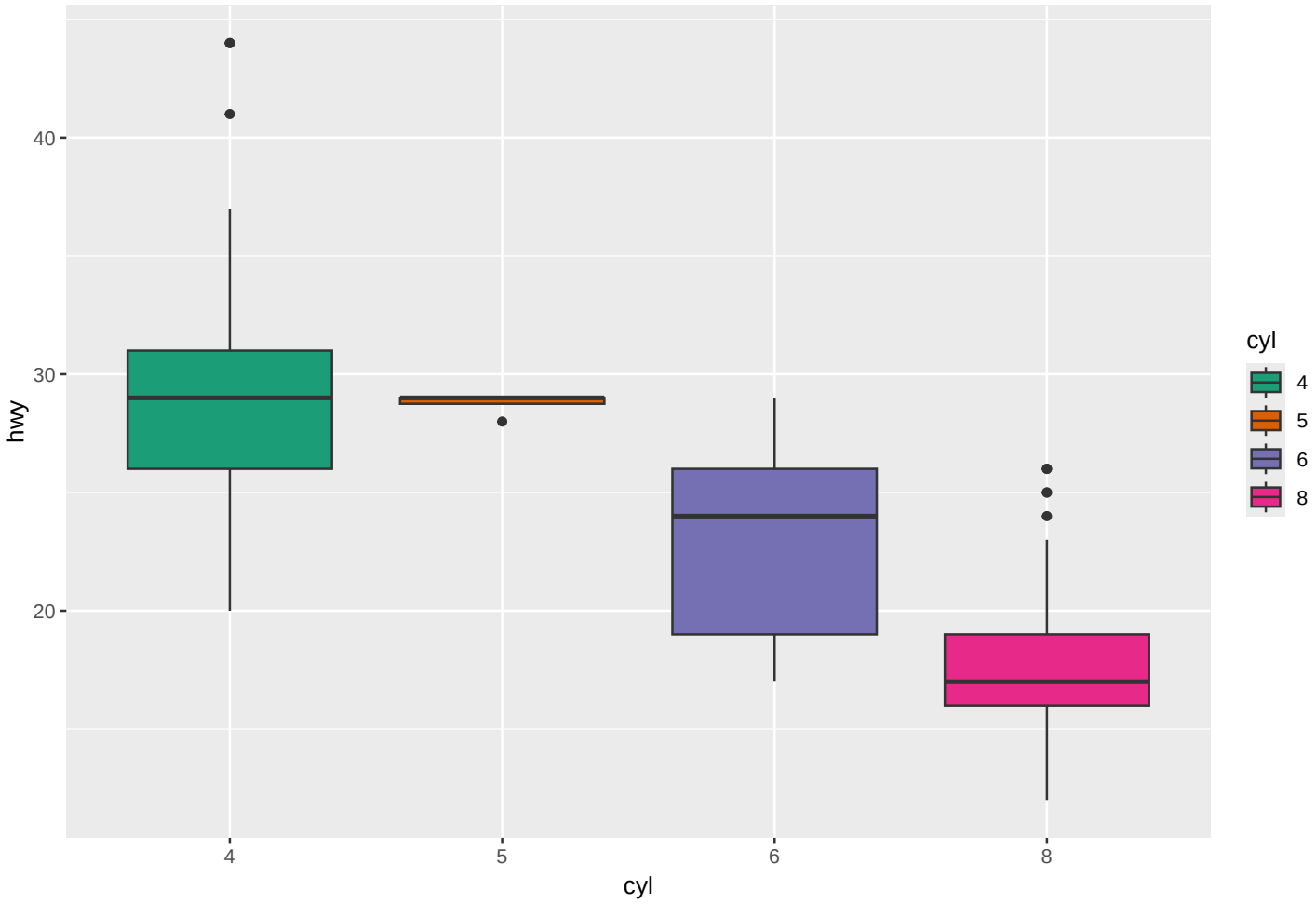

Figure 9: Boxplot of highway fuel efficiency `hwy` of a car against the number of cylinders in the car's engine `cyl` for Multiple Choice Questions 14–16.

14. According to the box plot in Figure 9, which number of cylinders has the smallest median highway fuel efficiency?

    (a) 4

    (b) 5

    (c) 6

    (d) 8

    (d) correct

15. According to the box plot in Figure 9, which number of cylinders has the smallest interquartile range?

    (a) 4

    (b) 5

    (c) 6

    (d) 8

    (b) correct

16. According to the box plot in Figure 9, which number of cylinders has the largest number of outliers?

    (a) 4

    (b) 5

    (c) 6

    (d) 8

    (d) correct