ETC2420 - ETC5242 Practice Exam, Semester 2, 2023, with Solutions

Instructions

This exam is comprised of three sections, labelled as **Part A**, **Part B** and **Part C**. Each Part contains a description of a problem setting, and contains seven (7) individual questions that relate to that setting. The seven questions in each section include:

- 2 TRUE OR FALSE questions, each worth 3 marks,
- 2 MULTIPLE CHOICE questions, each worth 5 marks,
- 3 SHORT ANSWER/ CALCULATION questions, each worth 8 marks, and

The total number of marks in each section is 40 marks. All questions must be attempted to have the potential to receive full marks. The total number of marks in this exam is 120.

PART A : Description of the problem setting

Most commonly, if a person who maintains a regular fitness routine that includes running as a major component sustains an injury that interrupts their regular running fitness program, by the time that they eventually return to running after the injury has been resolved they struggle to return to their pre-injury fitness level.

A new sport physiotherapy treatment program, named *Rapid Return to Running* (RRR), provides a sequence of exercises aimed to help someone who has sustained an injury to maintain a high aerobic capacity while recovering from the injury. The theory is that if the person can maintain aerobic fitness over the recovery period they are more likely to return to their pre-injury running program more quickly than if their aerobic capacity has diminished due to not running during the recovery period.

A recent study was undertaken to assess such a new RRR treatment method. In total, 110 people were recruited to voluntarily participate in the study, all participants had been regular runners who had very recently sustained an injury such that the standard physiotherapy treatment would recommend no running for a minimum of twelve weeks. During this non-running recovery period, the participants would participate in the RRR program, with each participant's aerobic capacity measured by the maximum amount of oxygen they can process, known as Vo2max (in litres per minute, abbreviated as L/min). Vo2max measurements were undertaken on each participant at the start of the program, and at weekly intervals over the program period.

Vo2max data from each of the study participants the initial (week 0) assessment and the final (week 12) assessment are available, and are stored, as variables *Vo2max0* and *Vo2max12*, respectively in an **R** tibble named **dt**, with each row of **dt** corresponding to one of the 110 subjects who participated in the RRR study. A third variable, constructed as the difference *Diff = Vo2max12 - Vo2max0*, is also included in **dt**.

Interest lies in testing whether the *true* population average Vo2max in week 12, denoted as μ_{12}, is the same as the population average Vo2max at the start of the RRR program, denoted as μ_0. Finally, let $\Delta = \mu_{12} - \mu_0$.

Several Tables and Figures are provided below and contain information relevant to **Question 1** through **Question 8**. A description of each is provided here, as follows:

- **Table A.1** displays the results from the **R** command *head(dt)*.
- **Table A.2** displays a range of descriptive statistics for the variables in **dt**.
- **Figure A.1** contains a simple estimated density plot for the *Diff* variable. A vertical red line is included, located at the sample average value of *Diff*, and a black vertical line is positioned at zero.
- **Figure A.2** displays side-by-side violin plots of the two Vo2max measures, *Vo2max0* and *Vo2max12*.
- **Table A.3** through **Table A.5** display the **R** output from three different applications of the **t.test()** function. The function and its arguments that were used to produce each table are displayed in the table caption, following the table number.

Be aware that not all information provided is needed to answer the questions.

Table A.1: Output from the **R** command **head(dt)**.

case	Vo2max0	Vo2max12	Diff
1	37.15	37.30	0.1468
2	40.59	35.92	-4.6694
3	31.88	23.73	-8.1510
4	33.20	34.40	1.1956
5	47.60	29.49	-18.1159
6	34.37	35.31	0.9433

Table A.2: Summary statistics for individual variables named in **dt**, having variable name as shown in the first column (**Variable**). Each row contains the number of observations (**n**), the sample mean (**mean**), sample median (**median**), sample standard deviation (**SD**) and interquartile range (**IQR**), respectively.

Variable	n	mean	median	SD	IQR
Diff	110	-10.01	-8.913	14.679	20.03
Vo2max0	110	41.48	40.539	10.755	16.36
Vo2max12	110	31.46	31.569	7.733	11.31

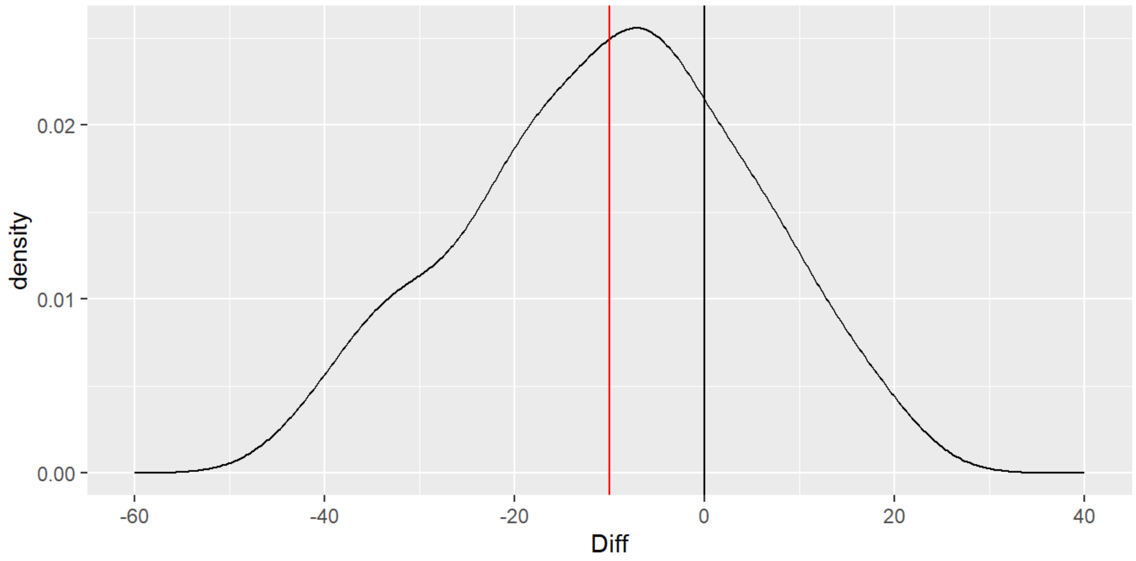


Figure A.1: Kernel density estimate of *Diff* variable (in Litres per minute), for the RRR study.

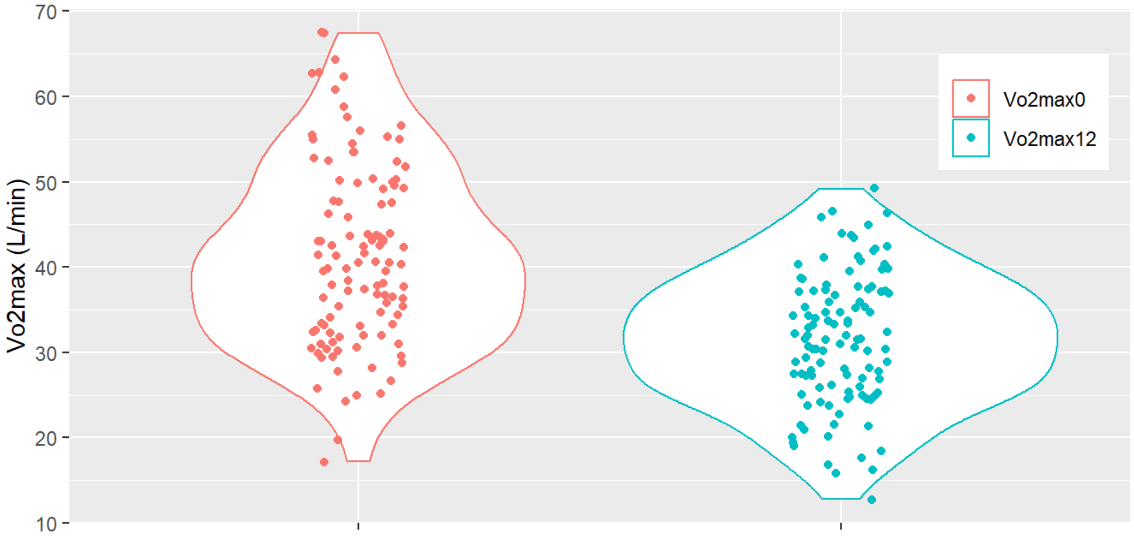


Figure A.2: Side-by-side violin plots of the *Vo2max0* and *Vo2max12* variables from the RRR study.

Table A.3: **R** output from **t.test**(x=dt$Diff).

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-10.01	-7.154	0	109	-12.79	-7.239	One Sample t-test	two.sided

Table A.4: **R** output from **t.test**(x=dt$Vo2max12, y=dt$Vo2max0, paired = TRUE, alternative = “less”).

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-10.01	-7.154	0	109	-Inf	-7.691	Paired t-test	less

Table A.3: **R** output from **t.test**(x=dt$Vo2max12, y=dt$Vo2max0, paired = FALSE, alternative = “greater”).

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-10.01	31.46	41.48	-7.928	1	197.9	-12.1	Inf	Welch Two Sample t-test	greater

Part A Questions

Question 1 [True or False, 3 marks]

Given the indicated question of interest, whether the *true* population average Vo2max in week 12 is the same as the population average Vo2max at the start of the RRR program, an hypothesis test of $H_0 : \Delta = 0$ vs. $H_1 : \Delta < 0$ is appropriate.

Answer: TRUE (We want to know if the program can maintain aerobic capacity in individuals, or if there is evidence to suggest that this capacity declines - we wouldn’t expect the capacity to increase, at least not on average, as participants are runners who are not able to do their normal running programs.)

Question 2 [True or False, 3 marks]

The sample average Vo2max measured in week 12 minus the sample average V02max measured in week 0, computed in **R** as

dt %>% summarise(diff_of_means = mean(Vo2max12) - mean(Vo2max0))

is equal to the average of differences, computed in **R** as

dt %>% summarise(mean_of_diffs = mean(Vo2max12 - Vo2max0)).

Answer: TRUE (The point estimate used in the two independent samples t-test and the paired t-test are the same (you can see it is true from the values in Table A.2., even if you don't already know this.)

Question 3 [Multiple Choice, 5 marks]

Which of the following statements is **FALSE**?

a. When considering $H_0 : \mu_0 - \mu_{12} = 0$ against a two-sided alternative, a paired t-test of based on the two variables *Vo2max0* and *Vo2max12* is equivalent to a one-sample t-test based on the *Diff* variable.

b. A p-value so close to zero that is reported as being equal is to zero provides strong evidence against the null hypothesis.

c. The pair of hypotheses $H_0 : \Delta = 0$ vs. $H_1 : \Delta < 0$ is equivalent to $H_0 : \mu_0 - \mu_{12} = 0$ vs. $H_1 : \mu_0 - \mu_{12} < 0$.

d. A p-value so close to one that is reported as being equal is to one provides does not provide strong evidence in favour of the alternative hypothesis.

Answer: c. (The roles of μ_0 and μ_{12} would need to be reversed for this to be true, as $\Delta = \mu_{12} - \mu_0$. Answer a. also has the means in the wrong order, but does not mention Δ, so it is OK.)

Question 4 [Multiple Choice, 5 marks]

What aspect of a distribution does the IQR measure?

a. location

b. scale

c. relative frequency

d. skewness

e. None of the above

Answer: b.

Question 5 [Short Answer, 8 marks]

Explain what is a 95% confidence interval for $\Delta = \mu_{12} - \mu_0$. Then, explain the motivation for the use of a bootstrap confidence interval, and detail how a bootstrap-based 95% confidence interval for Δ could be produced.

Answer: A 95% confidence interval for Δ refers to the range of Δ values that could reasonably be observed in 95% of all possible samples that could arise from being able to draw repeated independent samples of the same size, from same population. In reality we can’t do this repeated sampling, and we have only one sample of actual data available, and there are very few situations where the theoretical distribution can be mathematically derived, at least not exactly. The bootstrap method provides a method for obtaining an approximate 95% confidence interval for Δ. Here we use replicated samples of the observed dataset, using sampling with replacement, to obtain a large number ($R = 1000$, say) of different samples that are similar to the original one, but whose estimates of Δ (given by the average *Diff* variable) will vary. Once this bootstrap sample of $Diff$ values is available, the 2.5% and 97.5% empirical quantiles provide the approximate 95% confidence interval.

Question 6 [Short Answer, 8 marks]

Explain why it is important to either use the individual differences (contained in the *Diff* variable), or to use paired data (*Vo2max0* and *Vo2max12*), when devising an hypothesis test for Δ in the described setting.

Answer:

It is important to pair the data because people will naturally have different baseline levels of aerobic capacity, and so even if there is some change over the program period, two measurements of *Vo2max0* and *Vo2max12* on the same person could be more similar than if they were taken from two different people. Another way to say this is that the two measurements are correlated, as they are both related to the same person.

Without pairing or taking differences we would be losing the information about this correlation, and in fact would underestimate the standard error of the estimator, resulting in an inaccurate test.

Question 7 [Short Answer, 8 marks]

Now suppose an alternative 12-week long physiotherapy treatment program, named *Aerobic Conditioning Exercises* (ACE), was also available, and your study included an additional $n_{ACE} = 100$ patients randomly allocated to this treatment program in addition to the $n_{RRR} = 110$ patients previously discussed that were enrolled in the RRR program. And further, like for the RRR patients, the study collected weekly observations of Vo2max and you have available the paired observations of *Vo2max0* and *Vo2max12* from all of the ACE patients in the study.

Explain how you could use the expanded dataset, including the observations from both the RRR and ACE patients, to formally test whether the two programs have the same change in the average effect on aerobic capacity as each other, or if one program is actually better?

Be sure to state your null and alternative hypotheses, construction of the relevant test statistic and detail the type of test that you would undertake to address this question.

Answer: We can implement a independent samples (two-sample) t-test using the the paired difference data from the two separate groups to address this question. A two-sided alternative is most appropriate since we are interested to know if the two treatments have the same effect or different effects, without any preference given for one or the other. If $\Delta_{RRR} = \Delta$ denotes the true population change for the RRR treatment, and we let Δ_{ACE} denote the ACE treatment for the ACE treatment, then this approach would test $H_0 : \Delta_{RRR} = \Delta_{ACE}$ against $H_1 : \Delta_{RRR} \neq \Delta_{ACE}$.

PART B : Description of the problem setting

Vanessa owns and operates a large independent Veterinary clinic in Melbourne. Having noticed many of her clients buying pet insurance for their new pets, Vanessa was curious to know whether purchasing pet insurance occurred at the same rate for cat or dog owners - as most clients at her clinic were owners of either a cat or a dog.

Looking to establish some empirical evidence, Vanessa designed an experiment to track the choices of $n_{Cat} = 20$ clients who recently obtained a cat, and $n_{Dog} = 20$ clients who recently obtained a dog. Each of the 40 participants in this study were offered the same selection of three popular pet insurance policies and were asked to consider purchasing one of the policies so that if their new pet became unwell they would be able to have enough financial resources to pay for treatment.

Having appropriately randomised her sample, Vanessa collected the responses from each client in the sample indicating their decision regarding pet **Insurance** (either “Buy” or “Not Buy”). She also kept track of the **Pet Type** (either “Cat” or “Dog”), was eligible for the **Insurance** cover. Vanessa wants to now use this data to undertake testing the hypotheses given by:

$$H_0 : p_{Cat} - p_{Dog} = 0 \quad \text{vs.} \quad H_1 : p_{Cat} - p_{Dog} \neq 0,$$

where p_{Cat} denotes the population proportion of new “Cat” owners who would purchase pet **Insurance**, and similarly p_{Dog} denotes the population proportion of new “Dog” owners who would purchase pet **Insurance**.

Answer **Question 8** through **Question 14** below, which relate to Vanessa’s study. Several Tables and Figures are provided below that contain information you will need to answer these questions. These include the following items:

- **Table B.1** displays the cross-tabulation table between the **Pet type** (either “Cat” or “Dog”) and the **Insurance** decision chosen (either to Buy” or “Not Buy”).
- **Figure B.1** displays two bar charts, positioned side-by-side, with each showing the number of “Buy” and “Not Buy” responses for purchasing pet insurance, respectively, obtained during the study, corresponding to each **Pet Type**. The chart on the left relates to the **Insurance** decision of the cat owners, while the chart on the right relates to the corresponding decision of the dog owners.
- **Table B.2** displays the output of the **prop.test()** function, using information from the cross-tabulation table shown in **Table B.1**.
- **Figure B.2** displays a barplot that summarises the approximate sampling distribution of $\hat{p}_{Cat} - \hat{p}_{Dog}$, under H_0, obtained from a Randomisation (Permutation) Test. A vertical red line is positioned at $xobs = \hat{p}_{Cat} - \hat{p}_{Dog}$, the sample proportion of new cat owners who bought pet insurance minus the sample proportion of new dog owners who bought pet insurance. Note that the corresponding p-value for the two-sided test is 0.198.

Table B.1: Cross-tabulation of **Pet Type** (either “Cat” or “Dog”) and the pet **Insurance** decision (either to “Buy” or “Not Buy”).

	Insurance			
Pet Type	**Buy**	**Not Buy**	**Total**	**Prop**
Cat	9	11	20	0.45
Dog	14	6	20	0.70

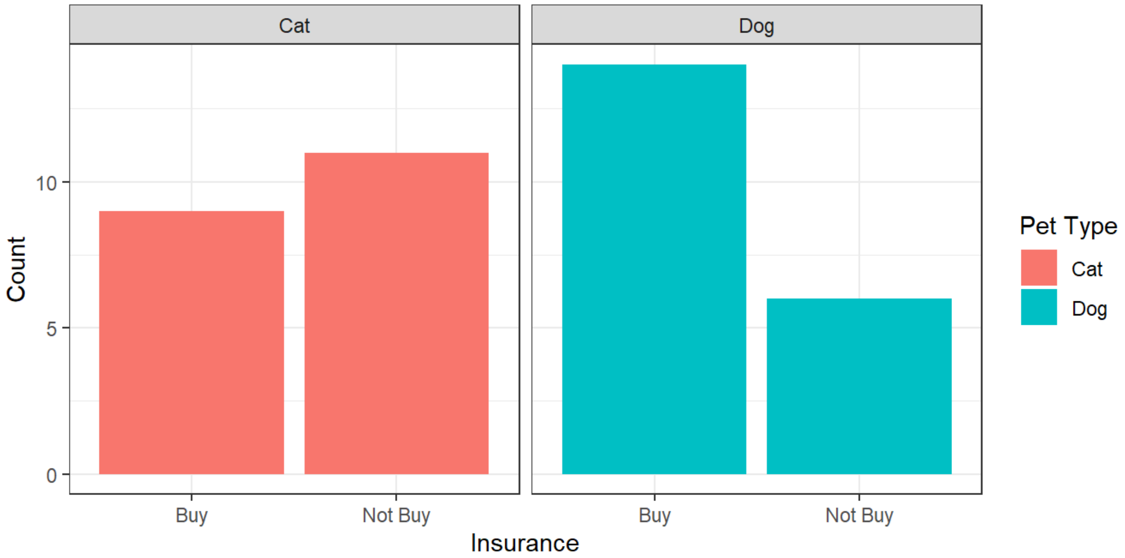


Figure B.1: Two barplots, each showing the counts of **Insurance** decision for study cases, according to the **Pet Type**.

Table B.2: Tidy output of the **prop.test(test_x, alternative = “two.sided”)** function, where **test_x** is a (2 x 2) matrix of cross-tabulated counts comprised of the two columns extracted from **Table B.1** indicating the **Insurance** decision tallies.

estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
0.45	0.7	1.637	0.2008	1	-0.5964	0.0964	2-sample test for equality of proportions with continuity correction	two.sided

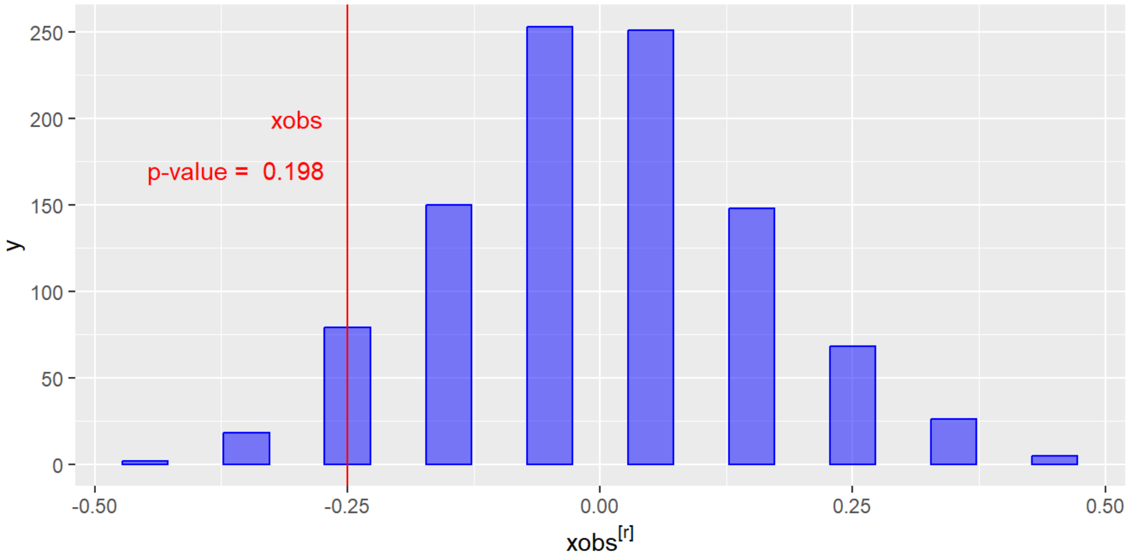


Figure B.2: Approximate sampling distribution of $\hat{p}_{Cat} - \hat{p}_{Dog}$, under H_0, obtained using Randomisation (Permutation) Test. A vertical red line is positioned at “xobs”, which is equal to the sample proportion of new cat owners who bought pet insurance minus the sample proportion of new dog owners who bought pet insurance. Note that the corresponding p-value for the two-sided test is

0.198.

Part B Questions

Question 8 [True or False, 3 marks]

A 95% CLT-based confidence interval for $\hat{p}_{Cat} - \hat{p}_{Dog}$ is given by [-0.5964, 0.0964].

Answer: FALSE (Since the parameters should not have “hats” ($\hat{}$’s) on them!)

Question 9 [True or False, 3 marks]

A bar chart is appropriate to display the approximate sampling distribution of $\hat{p}_{Cat} - \hat{p}_{Dog}$ obtained from a Randomisation test (as shown in **Figure B.2**) because the replicated test values $\hat{p}_{Cat}^{(r)} - \hat{p}_{Dog}^{(r)}$ can only take on values from the finite set $\{-1,\ -0.95,\ -0.9,\ \ldots,\ 0.9,\ 0.95,\ 1\}$.

Answer: TRUE (Other plots might also be appropriate, but certainly a barchart is OK here.)

Question 10 [Multiple Choice, 5 marks]

Which of the following statements is **TRUE**?

a. The Central Limit Theorem provides the exact sampling distribution of $\hat{p}_{Cat}$.

b. The combined distribution of the number of clients who decide to Buy pet **Insurance** is $Binomial\left(40,\ \frac{1}{2}p_{Cat} + \frac{1}{2}p_{Dog}\right)$

c. A Randomisation (Permutation) Test may be used to evaluate if there is evidence against $H_0 : p_{Cat} = p_{Dog}$.

d. None of the above.

Answer: c. (Answer a. is FALSE because the CLT provides an approximate, not an exact, sampling distribution for $\hat{p}_{Cat}$. Answer b. is FALSE in general, though it could be true if H_0 holds or if the population is evenly split between dog and cat owners. Answer c. is TRUE, and is one of the main reasons we covered this type of test, so Answer d. is FALSE.)

Question 11 [Multiple Choice, 5 marks]

Which of the following statements is **FALSE**?

a. If $X_1 \sim Bernoulli(p)$, $X_2 \sim Bernoulli(p)$, and X_1 and X_2 are independent, then $X_1 + X_2 \sim Binomial(2, p)$.

b. An *RMarkdown* file enables one to combine executable code with formatted text in a single document.

c. Bayes theorem provides a rational framework for updating belief in light observed evidence or data.

d. If $X_1 \sim N(\mu_1,\ \sigma_1^2)$ and $X_2 \sim N(\mu_2,\ \sigma_2^2)$, then $X_1 + X_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$.

e. The *tidyverse* is an **R** package that contains a set of other **R** packages.

Answer: d. (d. would be true if X_1 and X_2 are independent, but if not then the variance of the sum should include the covariance term.)

Question 12 [Short Answer, 8 marks]

Explain how to use the $n = n_{Cat} + n_{Dog}$ observations to find the Maximum Likelihood Estimate (MLE) of the common proportion $p = p_{Cat} = p_{Dog}$ that would hold when H_0 is true. Report the value of this MLE and explain what its value suggests about new pet owners buying pet **Insurance** in this setting.

Indicative Answer: Under H_0, the binary outcomes of all client decisions depend on the common proportion $p = p_{Cat} = p_{Dog}$. Then the $n = 40$ pet owners that decide to "Buy" pet **Insurance** has a $Binomial(n = 40, p)$ distribution, and the MLE is given by the overall sample proportion of "Buy", which here is $\hat{p} = \frac{9+14}{20+20} = \frac{23}{40} = 0.575$. This value is greater than 0.5, so we estimate that more clients would "Buy" pet **Insurance** than "Not Buy" pet **Insurance**. (This is only an estimate. We could construct a confidence interval or do a formal test here too, but these are not requested.)

Question 13 [Short Answer, 8 marks]

Figure B.2 displays a barchart of the approximate sampling distribution of $\hat{p}_{Cat} - \hat{p}_{Dog}$, under H_0, obtained using Randomisation (Permutation) Test. Explain the steps involved to generate the "data" needed to produce such a plot. In addition, explain why the randomisation ensures that approximate sampling distribution obtained is consistent with H_0.

[Note: In your answer, focus on the construction of the "data" used to generate the plot, not on the **R** code used to produce the actual plot itself.]

Indicative Answer:
We have two different **Pet Type** groups in our sample, one made up of "Cat" owners and the other made up of "Dog" owners. Each member of these groups has to decide whether to "Buy" or "Not Buy" pet **Insurance**. The corresponding two variables, **Pet Type** and **Insurance**, are used to construct the test statistic, $\hat{\Lambda} = \hat{p}_{Cat} - \hat{p}_{Dog}$, by finding the sample proportion of "Cat" owners who chose to "Buy" minus the sample proportion of "Dog" owners who chose to "Buy".

To create one replicate sample under H_0, the relationship between the **Insurance** choice and the **Pet Type** group, the **Pet Type** variable is permuted (shuffled, or resampled without replacement) in order to "break" the connection (if one exists) in the data between **Pet Type** and **Insurance** by arbitrarily mixing up the responses, and thereby ensure the sampling distribution is consistent with H_0.

The corresponding estimate, $\hat{p}^{(r)}_{Cat} - \hat{p}^{(r)}_{Dog}$, is generated under the constraint that H_0 is true. Repeating this process a large number of times, e.g. R=1000, we end up with a collection $\{\hat{p}^{(r)}_{Cat} - \hat{p}^{(r)}_{Dog}, r = 1, \ldots, R\}$, that is treated as a sample from the approximate sampling distribution of $\hat{p}_{Cat} - \hat{p}_{Dog}$ under H_0. These $R = 1000$ values are then plotted using functions from the ggplot package, specifically using the *geom_barplot* function. To ensure we always get the same results (for a given analysis) we can use **R**'s *set.seed()* command before the permutations are generated (reproducibility).

Question 14 [Short Answer, 8 marks]

Using only the information relating to the Cat owners, and assuming a Uniform prior over p_{Cat}, detail the form of the posterior distribution for p_{Cat}. Then, explain how to use this posterior distribution to construct a 95% credible interval for p_{Cat}, and report the corresponding **Bayes estimator** of p_{Cat} under the squared error loss function.

Be sure to include text to describe any mathematical or numerical expressions reported in your answer.

Indicative Answer:
First, the number of Cat owners in the study that decide to “Buy” pet **Insurance** has a $Binomial(n = 20, p = p_{Cat})$ distribution. Second, the $Uniform(0, 1)$ distribution is also known as a Beta(1,1) distribution. Third, the $Beta(\alpha = 1, \beta = 1)$ distribution is conjugate for the Binomial likelihood, and so the posterior will be a $Beta(1 + 9, 1 + 14) = Beta(10, 15)$ distribution.

There are infinitely many 95% credible intervals, since any interval containing 95% posterior probability would be acceptable. But common ones are $[0,\ q_{0.95}]$, where $q_{0.95}$ is the 95% quantile of the $Beta(10, 15)$ distribution (the posterior). Others could be $[q_{0.025}, q_{0.975}]$ or $[q_{0.05},\ 1]$.

The Bayes estimator under squared error loss is always the posterior mean. Since the mean of a $Beta(\tilde{\alpha}, \tilde{\beta})$ distribution is $\frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}}$, the Bayes estimator here will be $\frac{10}{25} = 0.4$.

PART C : Description of the problem setting

A dataset, referred to as the *Swiss fertility data*, contains information about women's fertility across 47 Swiss provinces in 1888, a time when fertility rates were still relatively high but were beginning to decline due to major economic and social changes.

The variables contained in the Swiss fertility data include a well-established and standardised fertility measure, denoted by I_g and expressed in percentage terms. Higher values of this fertility index indicate a greater number of births per married woman, on average, compared with lower index values. Other measures available for each province reflect economic and social aspects from 1888, and are named *Agriculture*, *Examination*, *Education*, *Catholic* and *Infant.Mortality*. A description of each of these variables is provided below, along with convenient labels given by **y**, **x1**, **x2**, **x3**, **x4** and **x5**, respectively.

- **y** $=$ Fertility: I_g marital fertility index (percentage)
- **x1** $=$ Agriculture: Percentage of males involved in agriculture as an occupation
- **x2** $=$ Examination: Percentage of male army draftees with highest mark on army examination
- **x3** $=$ Education: Percentage of male army draftees with education beyond primary school
- **x4** $=$ Catholic: Percentage population identifying as 'Catholic' (as opposed to 'Protestant')
- **x5** $=$ Infant.Mortality: Percentage of live births who live less than one year

Equation (C.1) below contains the linear regression model that was fitted to the Swiss fertility data, using the **lm()** function in **R**:

$$\mathrm{y} \;=\; \beta_0 \;+\; \beta_1\, x_1 \;+\; \beta_2\, x_2 \;+\; \beta_3\, x_3 \;+\; \beta_4\, x_4 \;+\; \beta_5\, x_5 \;+\; \varepsilon. \tag{C.1}$$

Several Tables and Figures are provided below and contain information relevant to Question 15 through Question 21. A description of each is provided here, as follows:

- **Table C.1** displays some summary statistics relating to the variables in the Swiss fertility data set.
- **Figure C.1** displays two plots resulting from suggested regression model for the Swiss fertility index **y**.
- **Figure C.2** displays six residual plots from the **lm()** fit of the regression equation to the Swiss fertility data.
- **Figure C.3** displays a histogram and QQ-plot resulting from the fitted regression equation to the Swiss fertility data.
- **Tables C.2 - C6** report model estimates and related information for the model with the highest *adjusted* R^2 (**adj-r-squared**), *log-Likelihood* (**log-Lik**), *negative AIC* (**negAIC**), *negative BIC* (**negBIC**) and R^2 (**r.squared**) values, respectively.

Table C.1: Summary statistics for Swiss fertility dataset.

n	mean(y)	sd(y)	mean(x1)	sd(x1)	mean(x2)	sd(x2)	mean(x3)	sd(x3)	mean(x4)	sd(x4)	mean(x5)	sd(x5)
47	70.14	12.49	50.66	22.71	16.49	7.978	10.98	9.615	41.14	41.7	19.94	2.913

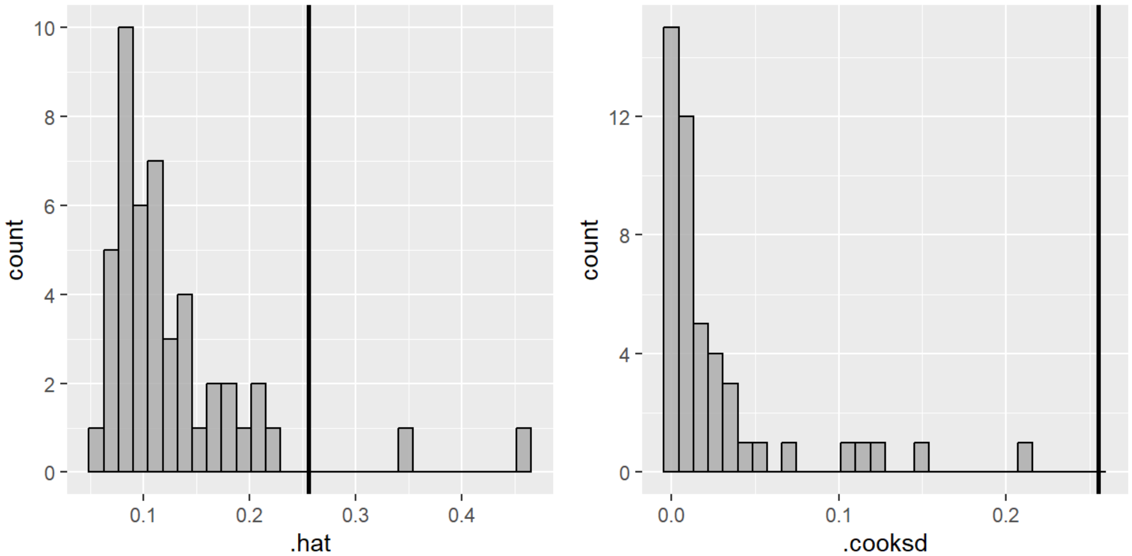
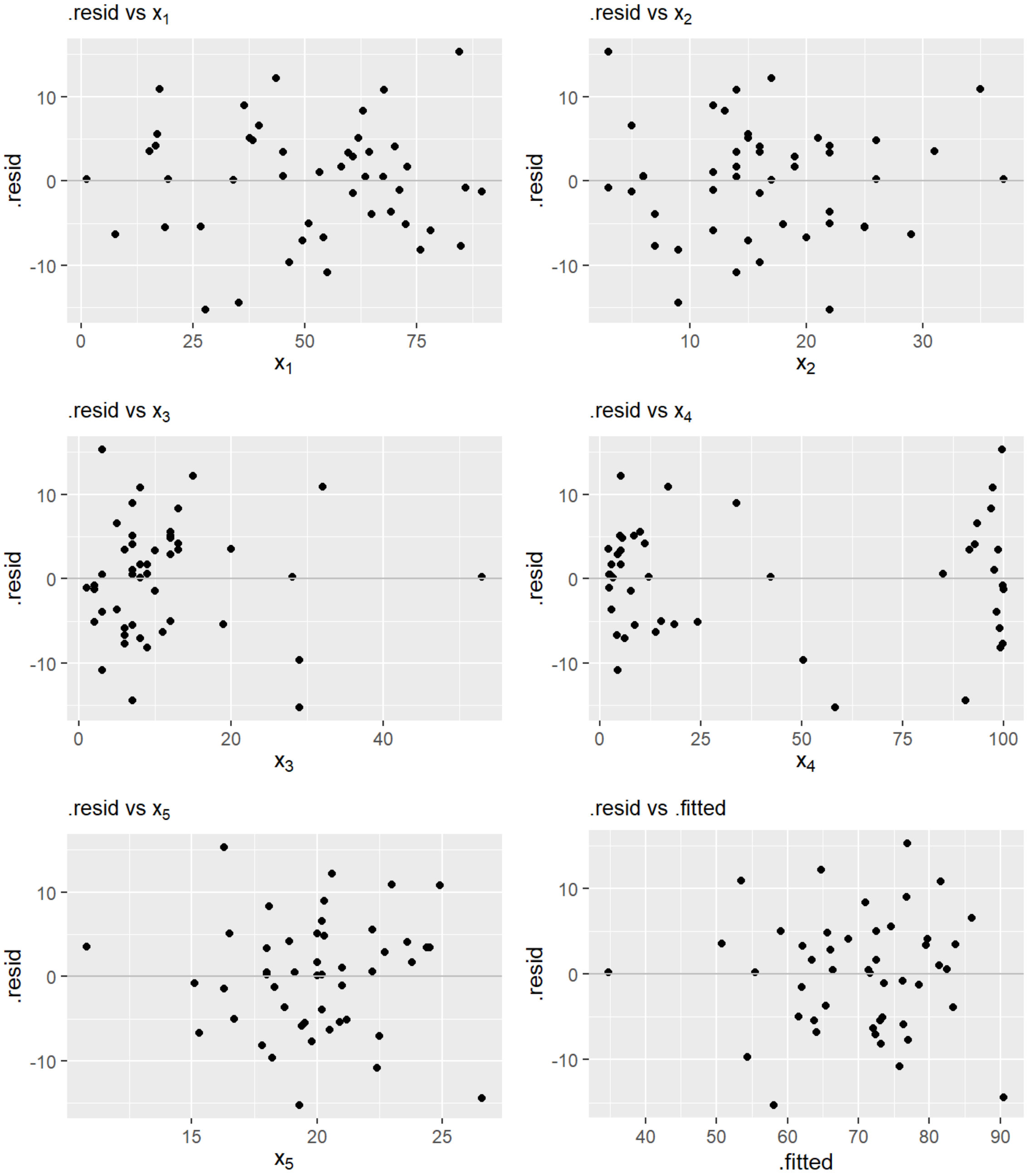


Figure C.1: Output from fitted regression model shown in Equation (C.1) for the Swiss fertility index. The vertical line in each panel corresponds to a ‘rule of thumb’ value.



.resid vs x_1

.resid vs x_2

.resid vs x_3

.resid vs x_4

.resid vs x_5

.resid vs .fitted

Figure C.2: Residual plots from fitted regression model shown in Equation (C.1) for the Swiss fertility index.

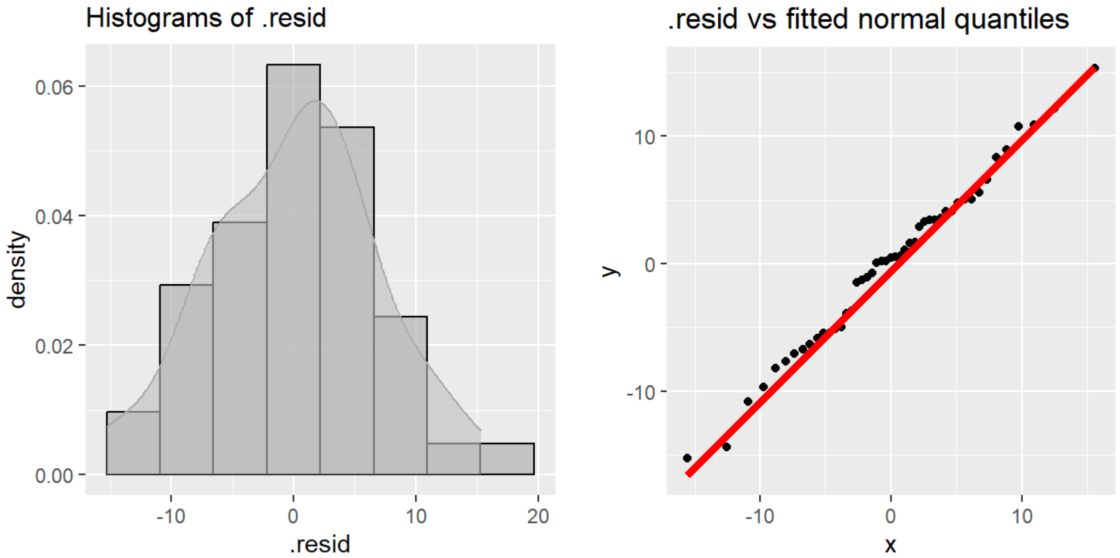


Figure C.3: Histogram (left panel) and QQ-plot (right panel) from fitted regression model shown in Equation (C.1) for the Swiss fertility index.

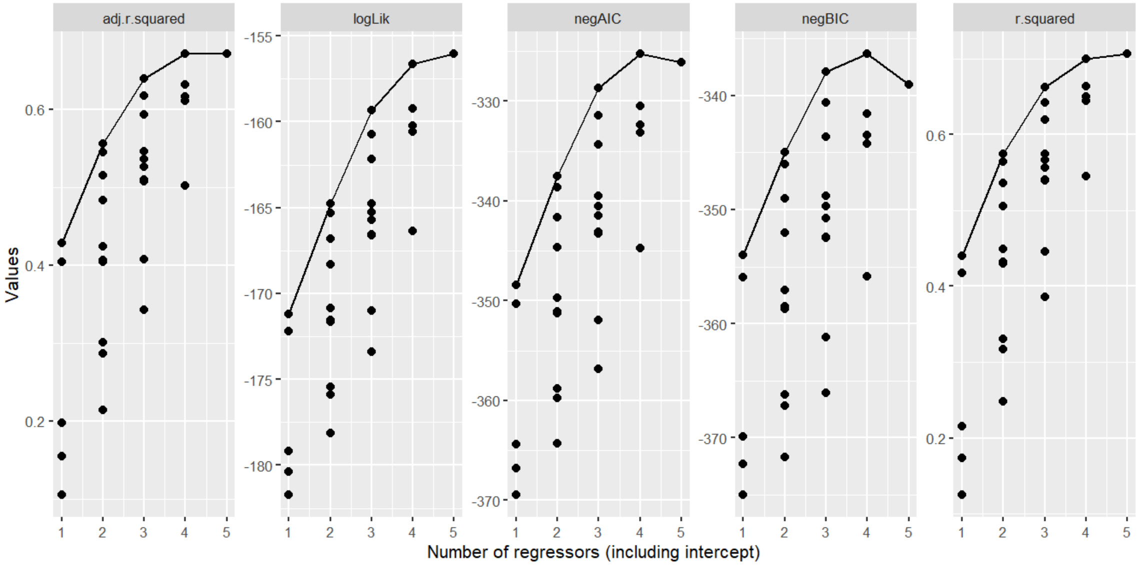


Figure C.3: Output containing information produced using the **meifly::fitall()** function corresponding to the multiple linear regression model for the Swiss fertility data shown in Equation (C.1).

Table C.2: **R lm()** function output for the model with the highest **adj.r.squared** value.

term	estimate	std.error	statistic	p.value
(Intercept)	66.9152	10.7060	6.250	0.0000
x1	-0.1721	0.0703	-2.448	0.0187
x2	-0.2580	0.2539	-1.016	0.3155
x3	-0.8709	0.1830	-4.758	0.0000
x4	0.1041	0.0353	2.953	0.0052
x5	1.0770	0.3817	2.822	0.0073

Table C.3: **R lm()** function output for the model with the highest **logLik** value.

term	estimate	std.error	statistic	p.value
(Intercept)	66.9152	10.7060	6.250	0.0000
x1	-0.1721	0.0703	-2.448	0.0187
x2	-0.2580	0.2539	-1.016	0.3155
x3	-0.8709	0.1830	-4.758	0.0000
x4	0.1041	0.0353	2.953	0.0052
x5	1.0770	0.3817	2.822	0.0073

Table C.4: **R lm()** function output for the model with the highest **negAIC** value.

term	estimate	std.error	statistic	p.value
(Intercept)	62.1013	9.6049	6.466	0.0000
x1	-0.1546	0.0682	-2.268	0.0286
x3	-0.9803	0.1481	-6.617	0.0000
x4	0.1247	0.0289	4.315	0.0001
x5	1.0784	0.3819	2.824	0.0072

Table C.5: **R lm()** function output for the model with the highest **negBIC** value.

term	estimate	std.error	statistic	p.value
(Intercept)	62.1013	9.6049	6.466	0.0000
x1	-0.1546	0.0682	-2.268	0.0286
x3	-0.9803	0.1481	-6.617	0.0000
x4	0.1247	0.0289	4.315	0.0001
x5	1.0784	0.3819	2.824	0.0072

Table C.6: **R lm()** function output for the model with the highest **r.squared** value.

term	estimate	std.error	statistic	p.value
(Intercept)	66.9152	10.7060	6.250	0.0000
x1	-0.1721	0.0703	-2.448	0.0187
x2	-0.2580	0.2539	-1.016	0.3155
x3	-0.8709	0.1830	-4.758	0.0000
x4	0.1041	0.0353	2.953	0.0052
x5	1.0770	0.3817	2.822	0.0073

Part C Questions

Question 16 [True or False, 3 marks]

There are two influential observations in the fit of the multiple linear regression model shown in Equation (C.1) for the Swiss fertility data.

Answer: False. (Although there are two observations flagged by the leverage (".hat") rule of thumb, neither of these points (nor any others) are flagged according to Cook's D (".cooksd"). Also, both of these diagnostic measures only identify *possible* problematic data points, and additional investigation is required before labelling the points as definitely being "influential". Here, in addition to Cook's D not showing any problem, we also don't notice any individual points as being particularly peculiar in the residual plots.)

Question 17 [True or False, 3 marks]

The **.hat** values shown in **Figure C.1** may be used, together with the residuals of the fitted model (whose histogram is shown in the final plot of **Figure C.2**) to calculate $n = 82$ individual Leave One Out Cross Validation (LOOCV) statistics corresponding to each of the observations.

Answer: FALSE. (The .hat values are also referred to as "leverage" values, and are denoted by h_{ii}, for $i = 1, 2, \ldots, n = 82$ in the lecture slides. Since there are $n = 82$ observations, there are $n = 82$ possible *case-deleted residuals* - and these can be used to calculate the **single LOOCV** statistic, which would be constructed by taking the *average* of the $n = 82$ squared case-deleted residuals.)

Question 18 [Multiple Choice, 5 marks]

Which of the following statements concerning the residuals of the fitted model is **TRUE**?

a. There **is** a strong systematic pattern between the residuals and regressor **x1**.

b. There **is** a strong systematic pattern between the residuals and regressor **x2**.

c. There **is** a strong systematic pattern between the residuals and regressor **x3**.

d. There **is** a strong systematic pattern between the residuals and regressor **x4**.

e. There **is** a strong systematic pattern between the residuals and regressor **x5**.

f. There **is** a strong systematic pattern between the residuals and the fitted regression values.

g. The distribution of the residuals **is not** relatively symmetric.

h. The distribution of the residuals **is not** relatively "bell-shaped".

i. None of the above.

Answer: i. (The residual plots for **x1**, **x2**, **x5** and the fitted values appear to suggest there is nothing systematic remaining in the residuals. The residuals appear to be approximately symmetric around zero and have a "bell-shaped" distribution. There appears that there could bee some sort of pattern in the locations of **x3** and **x4**, but it is not strong. So all of the statements are False)

Question 19 [Multiple Choice, 5 marks]

Which of the following statements is **FALSE**?

a. Under an ordinary least squares (OLS) setting, the regression errors should be approximately normally distributed.

b. Under an ordinary least squares (OLS) setting, the regression errors will *ideally* be approximately normally distributed.

c. When assuming normal regression errors under a Maximum Likelihood Estimation (MLE) setting, the regression errors should be approximately normally distributed.

d. If regression errors are not approximately normally distributed, by transforming one or more of the regressor variables, or the response variable, the regression errors for the corresponding regression using the transformed variables could be approximately normally distributed.

e. None of the above.

Answer: a. (OLS does not require normal errors. However, normal errors remain the *ideal* situation because they would indicate that there is only unpredictable zero-mean “noise” that remains once the fitted model is removed from the data. Sometimes by transforming some of the variables in the data set (as we did with the Olympics regression example) we can get a better description of the data - the regression line really captures the relationship between y and the x-variables, without needing any qualifying statement about there remaining some kind of skewness in the residual distribution.)

Question 20 [Short Answer, 8 marks]

Explain why **Table C.3** and **Table C.6** may be disregarded when attempting to determine which regressors should be retained in a linear regression model for **y =** Swiss fertility index.

Indicative Answer:
Table C.3 shows the model with the highest *logLik* value and **Table C.6** shows the model with the highest r.squared value. However, before even looking at the various models we know that the model that contains all of the regressors will have the highest **log-Lik** value and the highest **r.squared** value. That is, we know this model will always be the “full model”, and so these two criterion functions (logLik and r-squared) are not useful for determining which regressors should be retained in the model. We know the answer even without looking at the data, so there is no need to even look at these two Tables - they can be disregarded.

Question 21 [Short Answer, 8 marks]

Explain how to construct a 95% CLT-based confidence interval for β_2, for the model reported to have the largest *log-Likelihood* (**logLik**) value.

Note that although you do not need to report the numerical values for this interval, you should be as precise as possible in terms of describing its construction.

Indicative Answer:
This model is just the full model, given in Equation (C.1). The 95% CLT-based confidence interval is given generically by “[estimate + t(n-p, 2.5%)*std.error, estimate + t(n-p, 97.5%)*std.error]”. Here n=47, as shown on Table C.1, and p=6, since there are five non-constant regressors plus one intercept term. So we also need the (lower) 2.5% and 97.5% critical values of the Student-t distribution with n-p=47-6=41 degrees of freedom, denoted as t(41, 2.5%) and t(41, 97.5%), respectively. Students don’t need to find these values, but they will be approximately equal to -2 and +2 since this t-distribution will be close to a N(0,1), albeit having slightly fatter tails. (If **R** is used for this calculation, we find t(41, 2.5%)=-2.02 and t(41, 97.5%) = 2.02.)

Then we just need the estimated value of the coefficient of **x2**, and its (estimated) standard error. These values are available from Table C.3, and are equal to -0.2580 and 0.2539, respectively. Hence the interval would be something like: [-0.2580 +t(41, 2.5%)*0.2539, -0.2580 + t(41, 97.5%)*0.2539], or [-0.2580 - 2*0.2539, -0.2580* + 20.2539] (which simplifies to [-0.7658, 0.2498]).

Question 21 [Short Answer, 8 marks]

Discuss the considerations that an analyst would consider when making a choice between the model given in Equation (C.1) for the Swiss fertility index, and the slightly smaller model that excludes the regressor **x2**. In addition, comment on any additional calculations or visualisations that could be undertaken to help make this decision.

Indicative Answer:
The model that includes only the regressors **x1**, **x3**, **x4**, and **x5**, is the one selected using the **negAIC** and **negBIC** criteria. This alone provides some justification for selecting the smaller model. In addition, if the full model is considered, we notice that the estimated coefficient for **x2** is not statistically significant as its p-value (from Table C.1, for example) is 0.3155, which is much higher than 0.05. The estimated coefficient is only one (estimated) standard deviation away from zero.

On the other hand, the full model in Equation (C.1) is selected by the **adj.r-squared** criterion, which suggests **x2** does contain at least some degree of explanatory power even after accounting for the number of other parameters in the model, and the available sample size.

One would then need to consider other reasons to prefer a smaller or larger model, which could come from the purpose of the regression analysis and how the final model will be used. For example, if the model is to be used to forecast (or predict) a new fertility index value, one might prefer the smaller model as it is commonly thought that smaller models are better for this purpose. (Noting that this sample size is relatively small here, perhaps with more data we would in fact find that the additional variable contributes to the statistical explanation of **y**.) Alternatively, if this is a preliminary analysis for some future study, or if this analysis is to help guild a policy decision of some kind, then it may well be worthwhile to retain the additional regressor **x2** in the model.

To help make the decision, the analyst might consider doing a leave-one-out-cross-validation comparison to see if one model has a smaller **LOOCV** value (this might be particularly helpful if the model is to be used for prediction) or to look at all possible residual plots resulting from the slightly smaller model - particularly the one that plots these residuals against **x2** - to see if one can visually detect anything that might explain why this variable seems to have some predictive power. Similarly, leverage and Cook’s D could be considered, as these will differ slightly for the reduced model, and one could also check the variance inflation factors to be sure there is no evidence of substantial overlap (correlation) between the different regressors. (If this is present then the analyst might want to re-think the entire analysis!) Through careful consideration of all such information and preferences, it is up to the discretion of the analyst (or their boss!) as to which model to ultimately use.