ETC2420 - ETC5242 Final Exam, Semester 2, 2022

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

For each question asked in an SHORT ANSWER format, type your answer into the text box provided.

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.

A 1 page formula sheet is provided at the beginning of the exam.

Use of a calculator is not permitted.

Wherever a numerical answer is required, state your answer using numerical expressions that are as explicit as possible (simplify where easy to do so, there is no benefit in undertaking extensive calculations).

A three-page formula sheet may be downloaded for use during the examination time.

PART A : Description of the problem setting

Simpson, Olsen, and Eden (1975) “A Bayesian analysis of a multiplicative treatment effect in weather modification.” *Technometrics*, 17, pp 161-166, studied a controlled weather experiment that took place in Florida, USA, between 1968 and 1972. The experiments were designed to test whether by “seeding” (or infusing) an existing cloud with a particular chemical (silver nitrate) would have the effect of increasing precipitation (rainfall) measured in *acre-ft* (this is a unit of volume commonly used in the United States for situations such as the one here). After “seeding” a cloud (i.e. treating with the chemical) and selecting a corresponding “unseeded” cloud (i.e. a cloud with a similar chemical composition that has not been treated), the resulting rainfall from each was measured via radar as a ‘volume’ of rain. This process was repeated for 26 pairs of clouds, with results being loaded into an **R** tibble, named **df**.

Interest lies in both estimating the population (or “true”) mean difference:

$$\Delta = \mu_s - \mu_u$$

where μ_s and μ_u denote the average rainfall for seeded and unseeded clouds respectively, and in testing the hypotheses

$$H_0\colon \Delta = 0 \quad \text{versus} \quad H_1\colon \Delta > 0.$$

It is assumed that the value of Δ is constant, but unknown.

Several Tables and Figures are provided below and contain information relevant to **Question 1** through **Question 7**. A description of each is provided here, as follows:

- **Table A.1** displays a range of descriptive statistics for the variables in **df**.
- **Figure A.1** contains two plots showing the histograms and kernel densities of rainfall (in acre-ft), with the left panel relating to Seeded clouds, and the right panel relating to Unseeded clouds.
- **Table A.2** through **Table A.6** display the **R** output from five 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.

Note that you are not required to use all information provided.

Table A.1: Summary statistics for Rainfall data. Variable names ares shown in the first column (**Variable**). Each row contains the number of observations (**n**), the sample mean (**mean**), sample standard deviation (**sd**) and empirical quantiles corresponding to probabilities 2.5% (**Q2.5**), 25% (**Q25**), 50% (**median**), 75% (**Q75**) and 97.5% (**Q97.5**), respectively.

Type	n	mean	sd	Q2.5	Q25	median	Q75	Q97.5
Seeded	26	442.0	650.8	6.350	98.12	221.6	406.0	2090.7
Unseeded	26	164.6	278.4	3.438	24.82	44.2	159.2	969.8

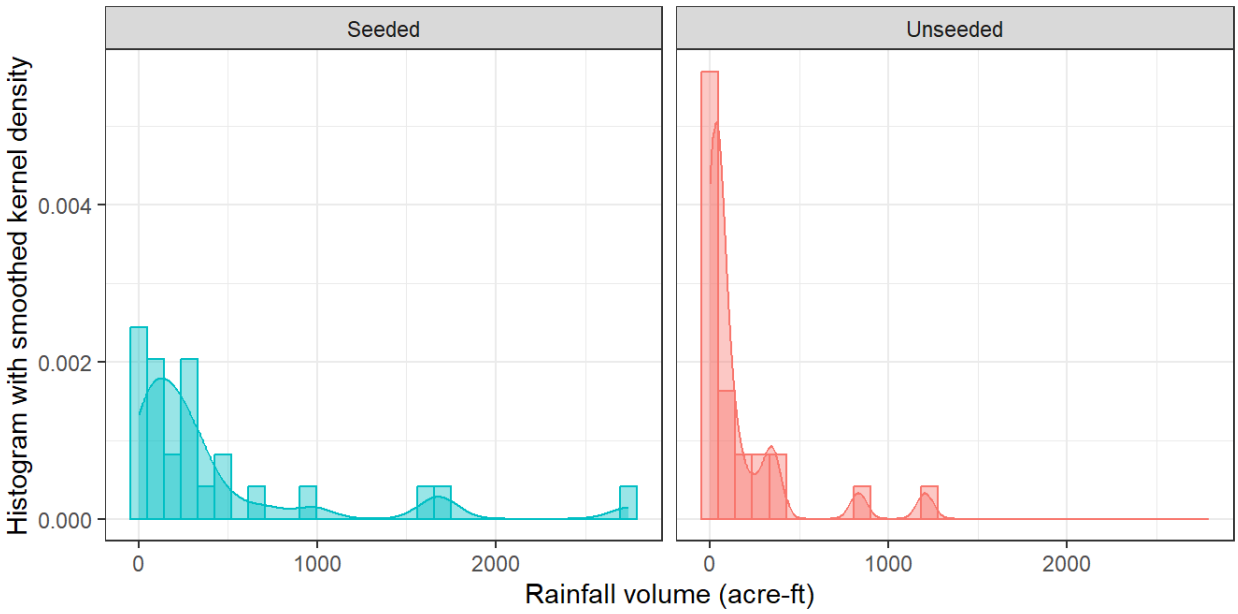

Figure A.1: Histograms and kernel density estimates of rainfall (in acre-ft) for Seeded (left), and Unseeded (right) clouds.

Table A.2: **R** output from **t.test**(x = df$Seeded, y = df$Unseeded, paired = FALSE, alternative = “greater”, conf.level = 0.95).

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
277.4	442	164.6	1.998	0.0269	33.86	42.63	Inf	Welch Two Sample t-test	greater

Table A.3: **R** output from **t.test**(x = df$Seeded, y = df$Unseeded, paired = TRUE, alternative = “greater”, conf.level = 0.95).

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
277.4	3.641	6e-04	25	147.2	Inf	Paired t-test	greater

Table A.4: **R** output from **t.test**(x = df$Seeded, y = df$Unseeded, paired = FALSE, alternative = “two.sided”, conf.level = 0.95).

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
277.4	442	164.6	1.998	0.0538	33.86	-4.764	559.6	Welch Two Sample t-test	two.sided

Table A.5: **R** output from **t.test**(x = df$Seeded, y = df$Unseeded, paired = TRUE, alternative = “two.sided”, conf.level = 0.95).

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
277.4	3.641	0.0012	25	120.5	434.3	Paired t-test	two.sided

Table A.6: **R** output from **t.test**(x = df$Seeded - df$Unseeded, alternative = “less”, conf.level = 0.95).

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
277.4	3.641	0.9994	25	-Inf	407.5	One Sample t-test	less

Part A Questions

Please use the description of the problem setting provided on page 3 to answer Part A questions 1 - 7.

Question 1 [True or False, 3 marks]

We should not use the difference of rainfall between Seeded and Unseeded clouds since it is not the same cloud that is being measured with and without seeding. The clouds are therefore independent.

Question 2 [True or False, 3 marks]

From the histograms in **Figure A.1**, it seems that the rainfall data are not close to Normal. We should be careful using a CLT based approach with this data given the small sample size.

Question 3 [Multiple Choice, 5 marks]

Based on the available **R** output shown in **Table A.2** through **Table A.6**, which of the following is the p-value of the most appropriate CLT-based test for the hypotheses of interest?

a. 0.0269
b. 0.0006
c. 0.0538
d. 0.0012
e. 0.9994
f. None of the above

Question 4 [Multiple Choice, 5 marks]

Consider a new variable constructed from the **Diff** variable, named **Good**, where the elements of **Good** are set equal to one if the corresponding element of **Diff** is larger than zero. That is, **Good** will be positive when the seeded cloud has higher rainfall than the unseeded cloud.

In addition, let g denote the sum of the values contained in the new variable **Good**.

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

a. The values in **Good** represent the outcomes from a sequence of Bernoulli Trials.

b. The value of g represents the outcome of a Binomial random variable.

c. Due to the Central Limit Theorem, the quantity g/n represents the outcome of an exactly normally distributed random variable.

d. If a continuous Uniform prior distribution is specified for p, where p denotes the chance that more rain will fall from when a cloud is seeded, then having observed the $n = 25$ pairs, the updated belief about p should be expressed as a $Beta(1 + g, 1 + n - g)$ distribution.

Question 5 [Short Answer, 8 marks]

Use the information available to report the outcome of the hypothesis test of interest at the 1% level of significance. Be sure to explain why you made this conclusion and what it means regarding the effectiveness of cloud seeding. Would your answer change if you assume that the clouds are independent? Give evidence to support your answer and explain why there is (or is not) a difference.

Question 6 [Short Answer, 8 marks]

Interpret the 95% confidence interval for the difference variable from **Table A.5** (this is the two-sided confidence interval). Is this consistent with the result of your hypothesis test? Explain why the confidence interval is not a probability statement about the unknown Δ.

Question 7 [Short Answer, 8 marks]

Explain how you would construct a 95% bootstrap-based confidence interval for Δ. In your answer,outline the reasons why the bootstrap approach may be more suitable than using a CLT-based approach in this context.

Note: For this question, you have the choice to either type your answer into a text box provided in this question or upload a picture containing your answer to Question 8.

PART B : Description of the problem setting

Lucky Insurance (the client we helped during semester) liked our work and has asked for some extra analysis. They are thinking of creating a new product for drivers over 55 years of age. They want to analyse the distribution of monthly premiums for existing customers.

Your training in business analytics gives you the right combination of skills to help the client understand some key attributes of their “senior” customers. They are interested in investigating the following research question:

What is the typical premium older customers pay?

There are 239 clients older than 55. To address this question, you first decide visualise the data by creating a histogram showing the empirical (sample) distribution for the premiums paid. You then decide to transform the data and find an appropriate probability distribution to describe this transformed data.

Answer **Question 8** through **Question 14** below, which relate to the analyses undertaken to help you understand more about the **Monthly premium for drivers over 55 years**. Several Tables and Figures are provided below that contain information you will need to answer these questions. These include the following items:

- **Figure B.1** displays a histogram and kernel density showing the empirical distribution of the **Monthly premium for drivers over 55 years**.

Due to the skewness evident in **Fig. B.1**, you decide to take the natural log of the data (**log(x)**).

- **Table B.1** displays output from the **MASS::fitdistr()** function, with argument **densfun = “normal”** applied to the log of the original data.
- **Figure B.2** displays the QQplot for your fitted distribution.

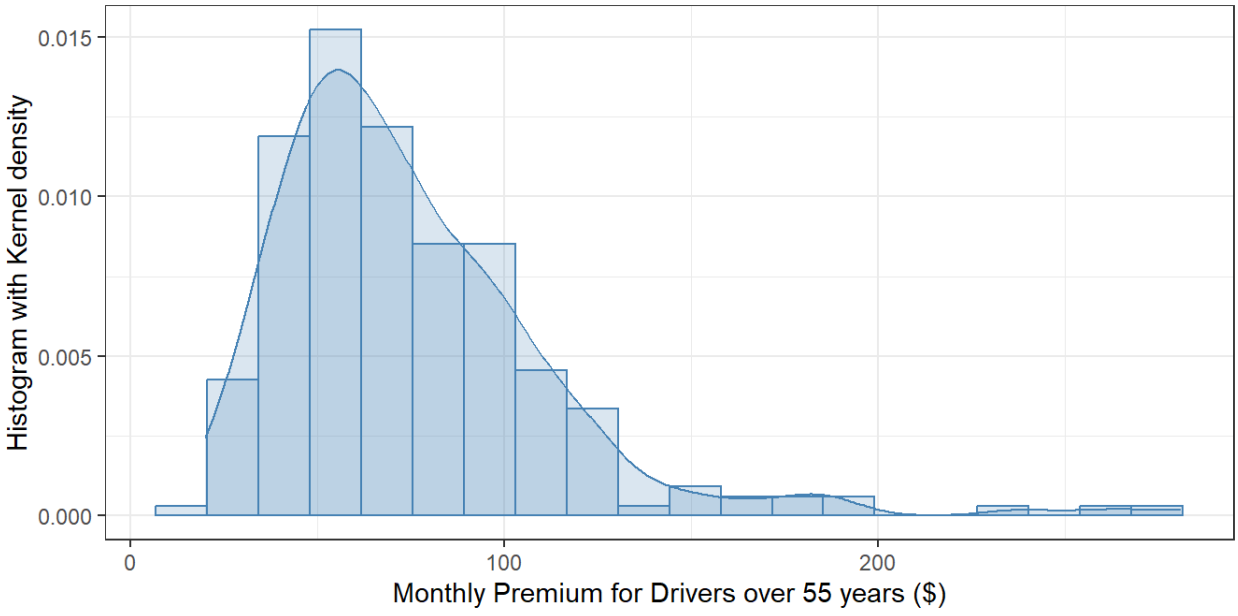

Figure B.1: Sample distribution of **Monthly premium for drivers over 55 years**

Table B.1: Tidy output from **fitdistr(log(x), “normal”)** function, where **x** refers to a vector containing the **Monthly premium for drivers over 55 years**.

term	estimate	std.error
mean	4.2112	0.0295

term	estimate	std.error
sd	0.4565	0.0209

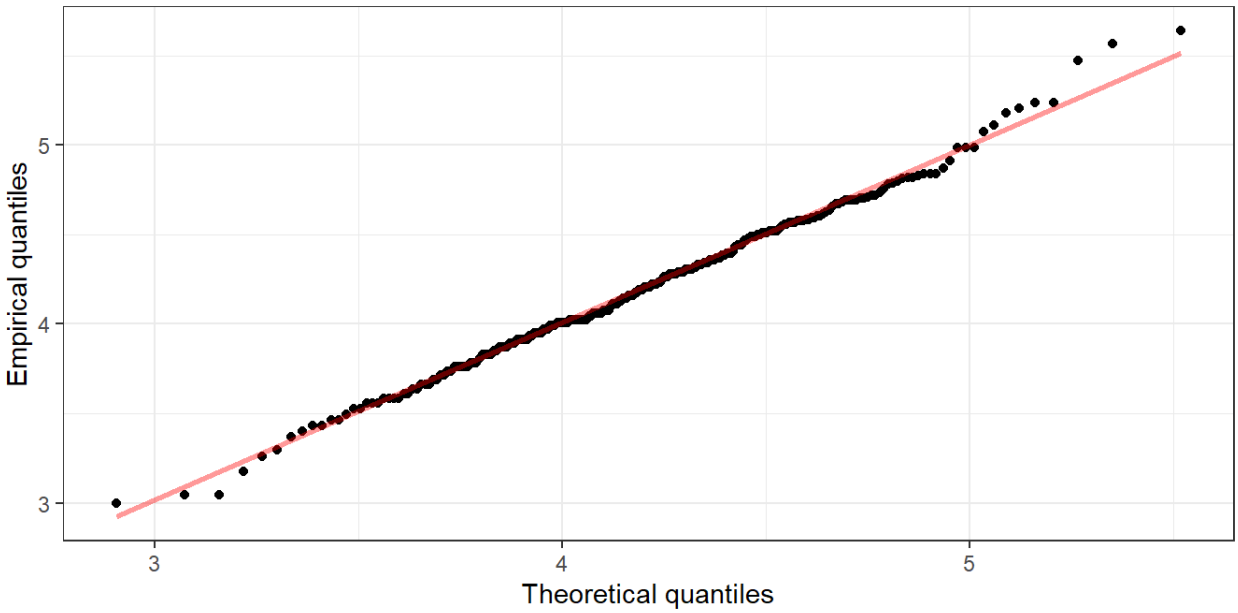

Figure B.2: QQplot for MLE - log(x) against Normal

Part B Questions

Please use the description of the problem setting provided on page 7 to answer Part B questions 8 - 14.

Question 8 [True or False, 3 marks]

We can never use the CLT to obtain a 95% confidence interval for the population standard deviation from the sample.

Question 9 [True or False, 3 marks]

The QQplot (shown in **Fig B.2**) shows that transforming the data and fitting a normal distribution was a reasonable decision.

Question 10 [Multiple Choice, 5 marks]

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

a. The MLE-based forecast distribution for the **Monthly premium for drivers over 55 years** of μ and σ is our prior distribution.

b. When evaluated at the Maximum Likelihood Estimator, $\hat{\theta}_{MLE}$, the log-Likelihood function has the same value as the Likelihood function.

c. We can only use the MLE estimate of μ because we must assume that the population variance σ^2 is known in order to estimate by maximum likelihood.

d. A Bootstrap-based confidence interval could be constructed for the MLE's of both μ_{MLE} and σ_{MLE}.

Question 11 [Multiple Choice, 5 marks]

Which of the following statements regarding Maximum Likelihood Estimates (MLEs) is **FALSE**?

a. The invariance property of MLEs allows us to get MLEs of functions of the parameters as well.

b. We can get CLT based confidence intervals and perform CLT based hypothesis tests using MLEs.

c. When using MLEs, parameters are treated as random.

d. MLEs typically have excellent large sample properties.

e. MLEs maximise the value of both the likelihood and the log-Likelihood functions.

Question 12 [Short Answer, 8 marks]

Again consider using a Normal(μ, σ^2) distribution to model the available $n = 239$ observations representing the log of the **Monthly premium for drivers over 55 years**.

Since you had to transform the data, you know this corresponds to a log-normal distribution for the original monthly premiums. Previous research suggests that older drivers pay an average of $50 per month for insurance, and so you choose a mean of 3.412 for your prior distribution (since $\mu_0 = ln(50) - 1/2) = 3.412$). You decide to choose a variance equal to 1 for this prior and assume the variance of the logged premium is known to be 0.16. Use the information in **Table B.1** to determine the posterior distribution for the population mean log monthly premium.

[Note: You are **not** required to formally derive the general form of the posterior distribution, rather you will need to identify the specific form of this distribution.]

Question 13 [Short Answer, 8 marks]

Given the posterior distribution of μ, as detailed in Question 13, what is the **Bayes estimator** of μ under the squared error loss function? Identify the credibility factor and explain what it tells us.

Question 14 [Short Answer, 8 marks]

Now that you have a posterior distribution, explain how you would obtain a 90% credible interval for the posterior mean. Explain what this tells you and how it can be used to answer Lucky Insurance's research question. Lucky Insurance are also concerned about the subjective nature of Bayesian analysis compared to Maximum Likelihood. Write a brief note (3-4 sentences only) to address this concern.

[Note that you are not expected to produce the interval, only to explain what it is and how it works.]

PART C : Description of the problem setting

Lucky Insurance also wants you to identify predictors of the claim size (y_i) of drivers over 55 year of age. They hope that this will give them valuable insights to help develop their new product. You think that the monthly premium (x_1), the number of months since the last claim (x_2) and age (x_3) would be good predictors. You also use a dummy variable (x_4) for medium-sized cars (=1 if a medium sized vehicle, 0 otherwise). You use these variables to construct the multiple linear regression model (referred to as the **full model**) given by:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + \varepsilon_i,$$

where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, for $i = 1, 2, \ldots, n = 239$. You will use the available data to build a good linear regression model for Lucky Insurance.

Several Tables are provided below and contain information relevant to **Question 15** through **Question 21**, including output of estimated linear regression models using the data supplied by Lucky Insurance:

- **Figure C.1** displays a histogram and kernel density estimate of $n = 239$ claim amounts for drivers over 55 years of age (y).
- **Table C.1** displays the regression results for the full model.
- **Figure C.2** displays a histogram of the leverage and Cook's distance measures, with threshold, for the regression from **Table C.1**.
- **Figure C.3** displays a panel of visualisations of the model performance measures for all fitted models.
- **Table C.2** displays the regression results for the model with the best fit values.
- **Figure C.4** displays some residual diagnostics from the regression from **Table C.2**, the preferred model.
- **Figure C.5** displays a histogram of the leverage and Cook's distance measures for the model with the best fit statistics.

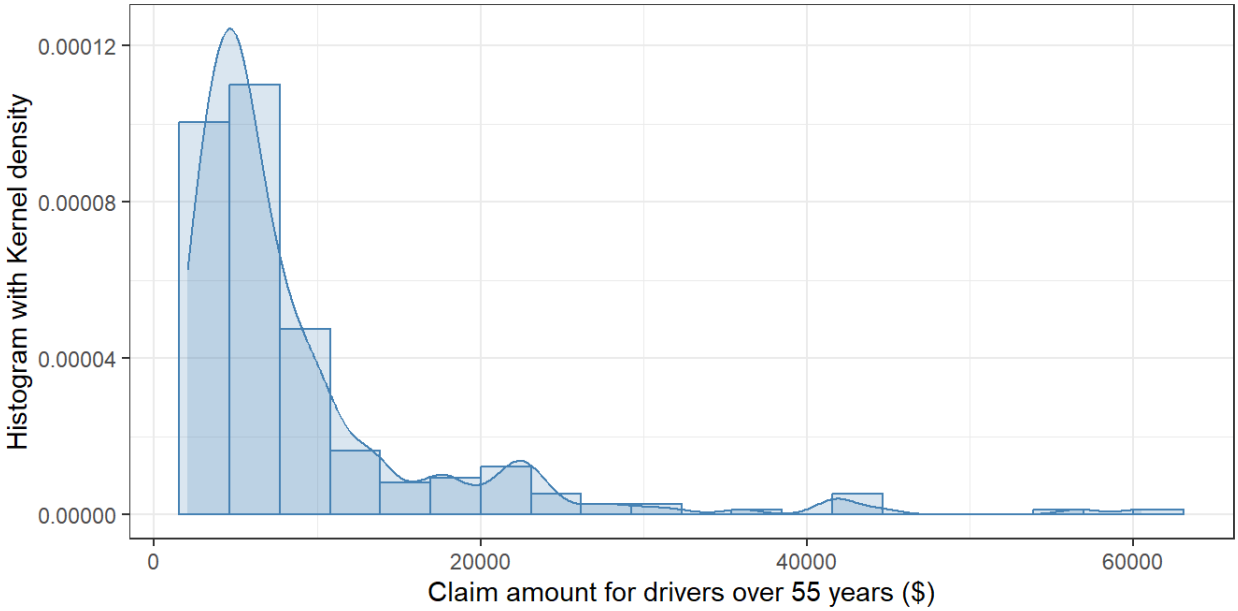

Figure C.1: Histogram and kernel density estimate of $n = 239$ claim amounts by drivers over 55 years (y).

Table C.1: Tidy regression output for the full model.

term	estimate	std.error	statistic	p.value
(Intercept)	-19294.9	17231.2	-1.120	0.264
x1	94.1	13.7	6.890	0.000
x2	794.6	354.7	2.240	0.026
x3	-88.4	334.5	-0.264	0.792
x4	131.9	1197.0	0.110	0.912

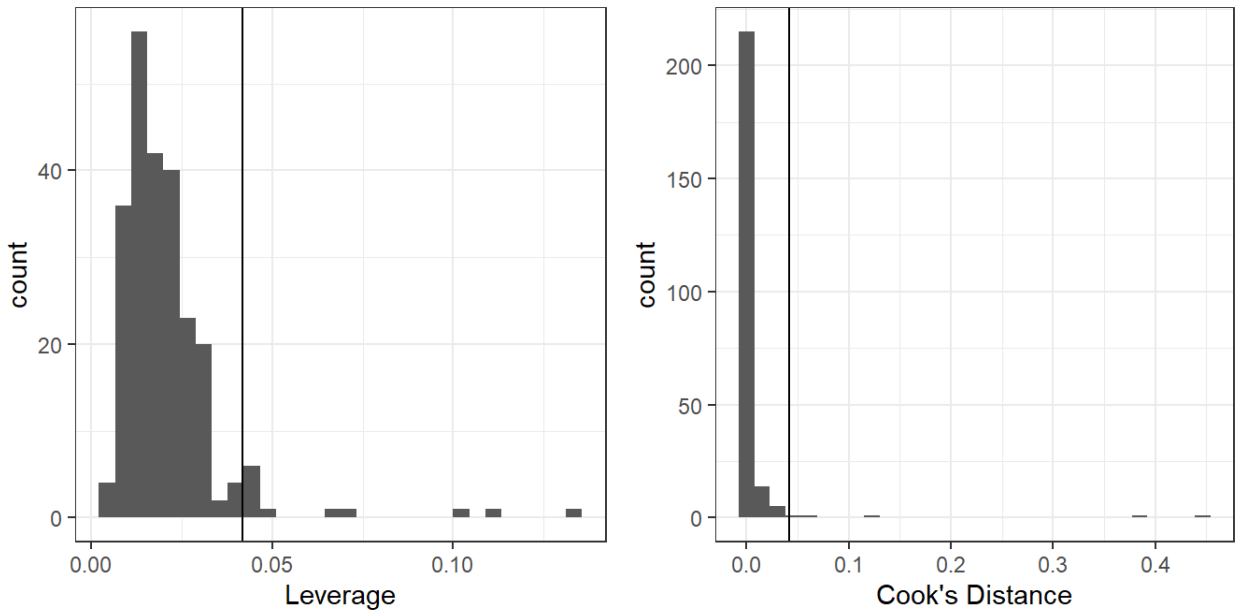

Figure C.2: Histogram of Leverage and Cook's Distance from the full model, with threshold.

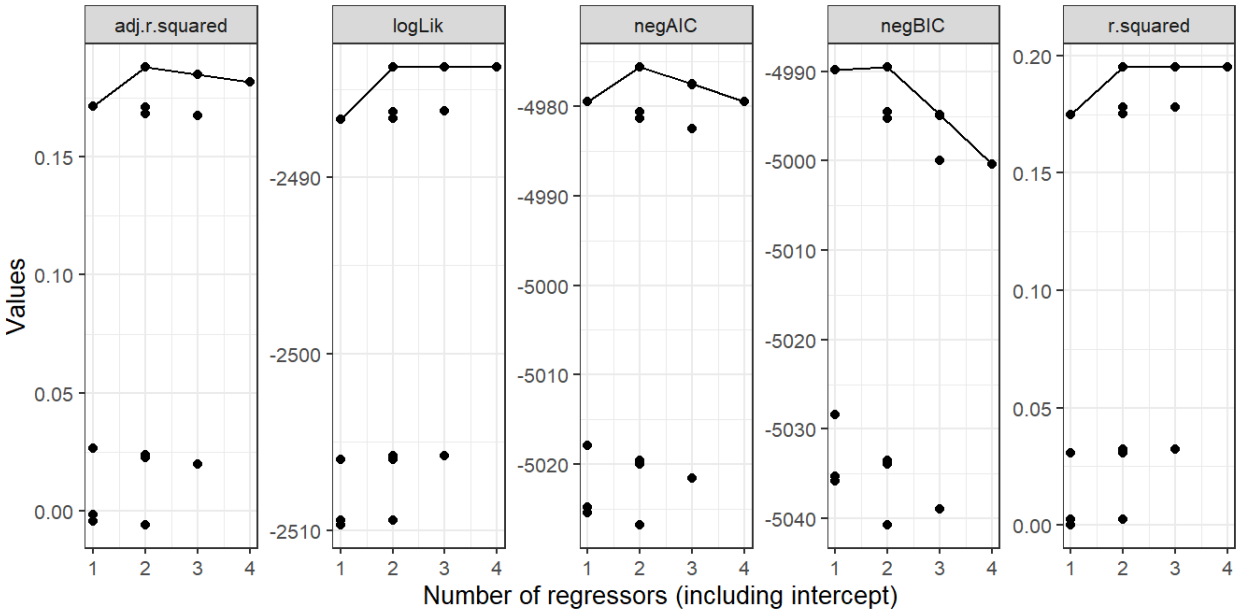

Figure C.3: Panel of visualisations of the model performance measures for all fitted models.

Table C.2: Tidy regression output of model with best fit statistics.

term	estimate	std.error	statistic	p.value
(Intercept)	-22885.2	10210.8	-2.24	0.026
x1	94.3	13.6	6.94	0.000
x2	750.7	309.9	2.42	0.016

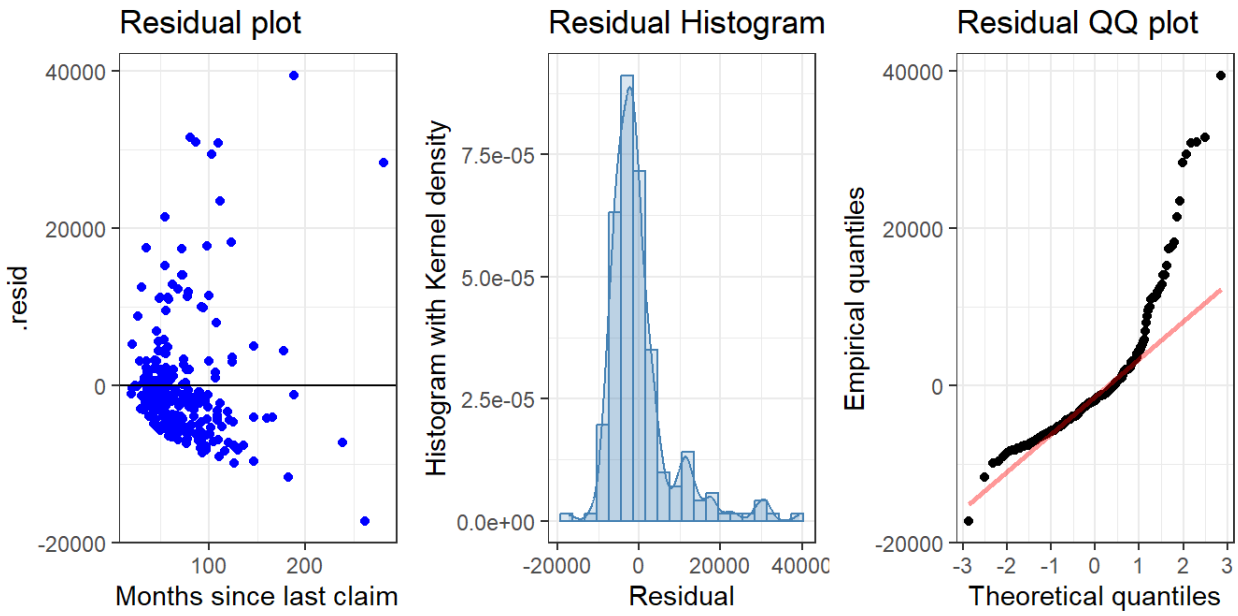

Figure C.4: Residual plots for best fitting model.

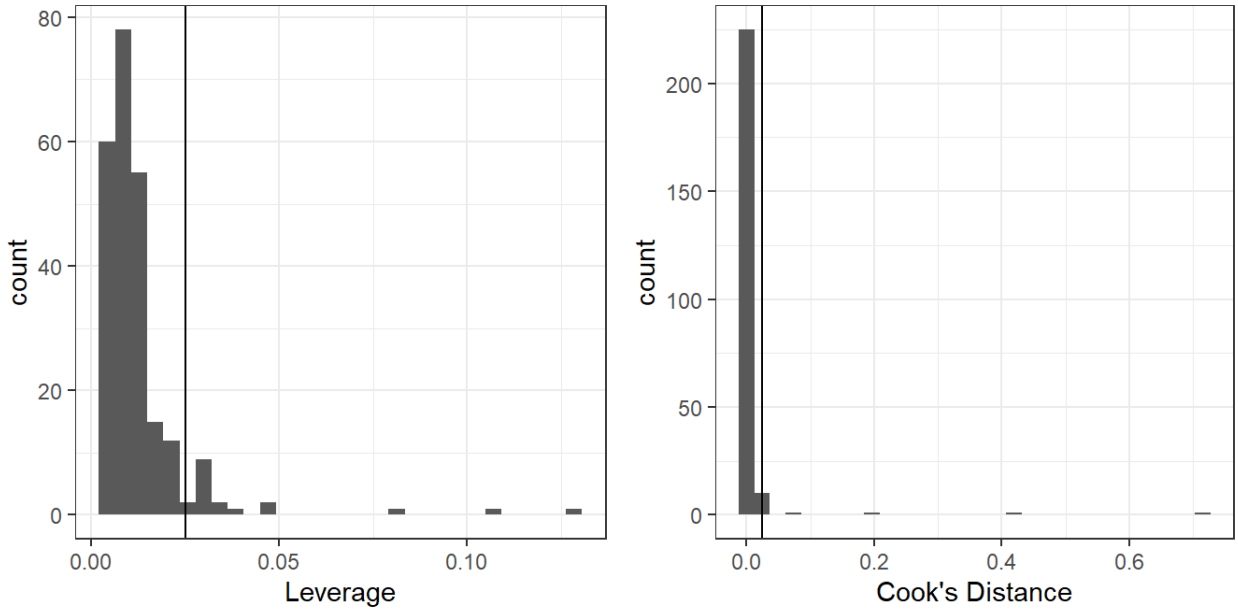

Figure C.5: Histogram of Leverage and Cook's Distance with threshold for model with best fit statistics.

Part C Questions

Please use the description of the problem setting provided on page 15 to answer Part C questions 17 - 24.

Question 15 [True or False, 3 marks]

To construct a leave-one-out cross validation (LOOCV) measure, which is the average of the n squared case-deleted residuals, we do not need to fit n different models.

Question 16 [True or False, 3 marks]

Variance Inflation Factors greater than 10 are an indication of multicollinearity in multiple regression.

Question 18 [Multiple Choice, 5 marks]

In a multiple linear regression, the **BIC**…

a. Always increases with extra regressors and can't be used to compare the fit of different models.

b. Penalises extra regressors and can be used to compare the fit of different models.

c. Identifies potentially influential observations.

d. Uses case-deleted residuals to assess the model fit.

Question 19 [Short Answer, 8 marks]

Using **Fig C.2**, provide an expression for the rule of thumb threshold. Approximately how many data points have high Cook's distance? How many have high Leverage values? Briefly outline what these statistics measure and what we should do given this information.

Question 20 [Short Answer, 8 marks]

The regression in **Table C.1** was not the preferred regression. Explain the process you would use to attain the preferred regression in **Table C.2** using the information provided. Explain why the preferred model may not necessarily be the best model. Note any additional visualisations or analyses that you would typically undertake before determining the model you would present to Lucky Insurance.

Question 21 [Short Answer, 8 marks]

We expect that x_1 (Monthly premium) will be positively related to the claim amount in general (more expensive cars cost more to repair, and so have higher premiums).
You decide to test the hypothesis

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 > 0$$

using a **Permutation test**.

Explain how you would implement a permutation test and obtain a p-value. Provide a description of the steps involved to execute the test in the described setting, using a Type I error rate of 5%.

[Note that you are not expected to produce the test, only to explain what it is and how it works.]