ETC2420 - ETC5242 Final Exam, Semester 2, 2022, with Solutions

SECTION A

PART C : Description of the problem setting

Finance researchers typically characterise the (percentage) return of a financial investment (i.e. the money gained or lost from the investment) as being linearly related to other variables referred to as **financial factors**.

Ten years of monthly returns from an investment portfolio and the corresponding data regarding a set of four potential financial factors are available in an **R** tibble named **pf**. This data is used 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 $i = 1, 2, \ldots, n = 120$. Each of the four independent variables, x_1, x_2, x_3 and x_4 represent a different financial factor.

As you have become interested in understanding more about modelling financial returns, you are keen to use the available data to build a good linear regression model for the portfolio returns, using the financial factors as the regressors.

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 and financial investment scenario described above:

- **Table C.1** displays the results from the **R** command **head(pf)**.
- **Figure C.1** displays a histogram and kernel density estimate of $n = 120$ monthly financial (percentage) returns (y).
- **Table C.2** through **Table C.6** contain information produced in **R** using the **fitall()** function from the **meifly** package. Each table contains certain details of the so-called “best” fitting linear regression model, according one of the available criteria, as indicated in the relevant table caption.

Table C.1: Output from **R** function **head(pf)**.

y	x1	x2	x3	x4
3.75	2.96	0.505	-2.30	-2.87
4.27	2.64	0.937	-1.40	4.19
-1.31	0.36	0.632	-1.32	0.01
-3.98	-3.24	0.988	0.04	0.51
2.24	2.53	1.181	-0.20	-0.35
4.73	2.62	1.234	-0.04	-0.02

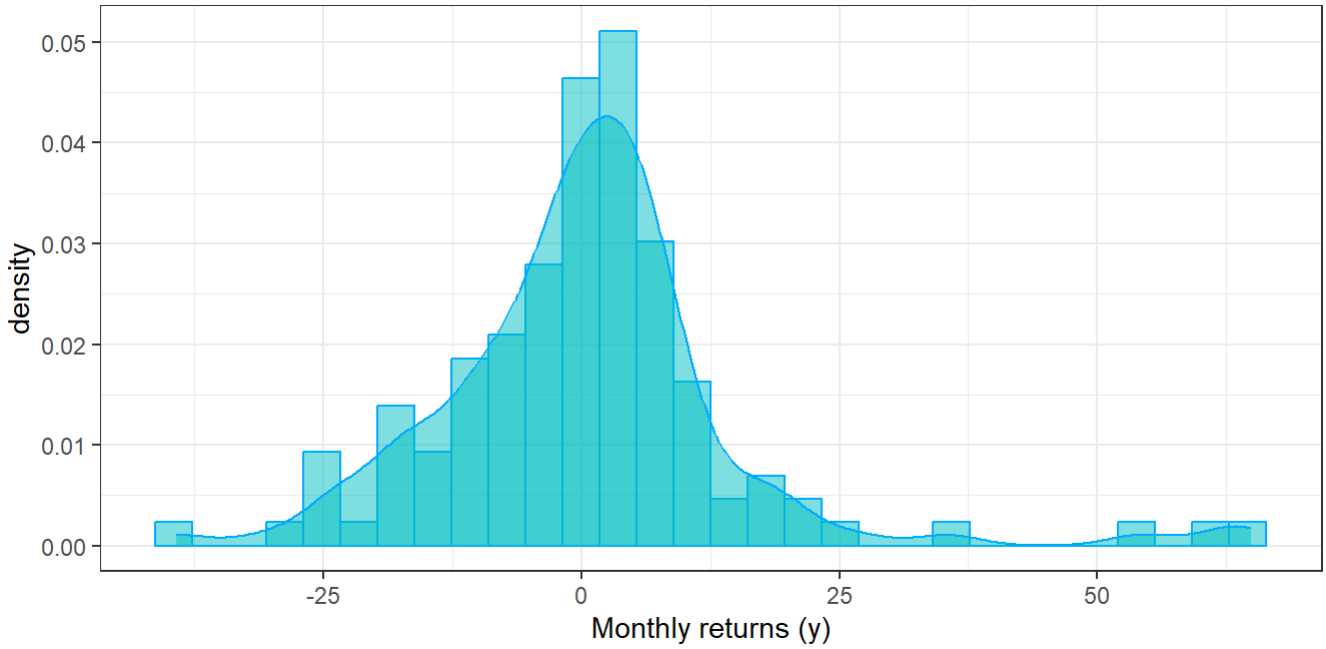

Figure C.2: Histogram and kernel density estimate of $n = 120$ monthly financial returns (y).

Table C.2: Output from the **R lm()** function for the model for financial returns identified as having the largest maximised **adjusted R-squared**.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.479	0.378	-1.26	0.208
x1	1.115	0.052	21.63	0.000
x3	-0.114	0.082	-1.39	0.167
x4	0.684	0.076	9.01	0.000

Table C.3: Output from the **R lm()** function for the model for financial returns identified as having the largest **log-Likelihood**.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.865	0.589	-1.468	0.145
x1	1.107	0.053	21.049	0.000
x2	0.589	0.688	0.856	0.394
x3	-0.109	0.082	-1.327	0.187
x4	0.685	0.076	9.023	0.000

Table C.4: Output from the **R lm()** function for the model for financial returns identified as having the largest **negative AIC**.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.502	0.380	-1.32	0.189
x1	1.111	0.052	21.50	0.000
x4	0.657	0.074	8.92	0.000

Table C.5: Output from the **R lm()** function for the model for financial returns identified as having the largest **negative BIC**.

term	estimate	std.error	statistic	p.value

term	estimate	std.error	statistic	p.value
(Intercept)	-0.502	0.380	-1.32	0.189
x1	1.111	0.052	21.50	0.000
x4	0.657	0.074	8.92	0.000

Table C.6: Output from the **R lm** function for the model for financial returns identified with having the largest **R-squared**.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.865	0.589	-1.468	0.145
x1	1.107	0.053	21.049	0.000
x2	0.589	0.688	0.856	0.394
x3	-0.109	0.082	-1.327	0.187
x4	0.685	0.076	9.023	0.000

r.squared <dbl>	adj.r.squared <dbl>	sig… <dbl>	statistic <dbl>	p.value <dbl>	df <dbl>	logLik <dbl>	A.. <dbl>	B.. <dbl>	deviance <dbl>
0.925	0.922	4.13	355	1.04e-63	4	-338	688	705	1966

1 row | 1-10 of 12 columns

Part C Questions

Please use the description of the problem setting provided on page 11 to answer Part C questions 17 - 24.

Question 17 [True or False, 3 marks]

As is true for an R^2 measure, the value of the maximised log-Likelihood function associated with a linear regression model will not decrease when one or more additional variables are added to the model.

Answer: TRUE

Question 18 [True or False, 3 marks]

The value of the maximised log-Likelihood function associated with a multiple linear regression model is equal to the minimised sum of squared errors.

Answer: FALSE (The MLE for β and the estimate of β that minimises the MSE are the same, but the optimised objective functions need not have the same value, and generally won't.)

Question 19 [Multiple Choice, 5 marks]

Which of the following are **not ideal** properties for OLS residuals?

a. The residuals sum to zero.

b. The histogram of the residuals appears consistent with a normal distribution.

c. The residuals have detectable patterns when plotted against the fitted values.

d. The residuals are uncorrelated with the included factor variables x_1, x_2, x_3 and x_4.

Answer: c (There should not be any detectable patterns when plotting residuals against the fitted values.)

Question 20 [Multiple Choice, 5 marks]

In a multiple linear regression, **multicollinearity** refers to…

a. Non-zero correlation between two or more response variables.

b. Very large, non-zero correlation between two or more independent variables.

c. Positive correlation between two or more explanatory variables.

d. Large influence of two or more observations on the fitted regression line.

Answer: b.

Question 21 [Short Answer, 7 marks]

For the model identified as having the largest R^2 value, report a 95% CLT-based confidence interval for the regression coefficient β_4, and explain why this interval is **not** called a “probability” interval for β_4.

Answer

The model with the highest R^2 is the full model. The 95% CLT-based confidence interval is given generically by “[estimate + t(115, 2.5%)*std.error, estimate + t(115, 97.5%)*std.error]” We can use z(2.5%)=-1.96 and z(97.5%)=1.96 in place of t(n-5, 2.5%) and t(n-5, 97.5%), respectively, because the sample size n=120 is pretty large, and t(115, 2.5%)=-1.98 and z(2.5%)=-1.96 are very close.

Indeed, it would be OK just to round up to 2. So we have the 95% CLT-based confidence interval for β_4, the coefficient of x_4, given by:

students can replace crit values if they remember them, or use 2

using t(115,.): (0.675 -1.98*0.076 , 0.675 + 1.98*0.076)

using z(): (0.675 -1.96*0.076 , 0.675 + 1.96*0.076)

(Although probability is involved in the construction of the confidence interval, this probability relates to the repeated sampling distribution of the estimator under hypothetical repeated samples of size n from same population from which the data were drawn. Theoretically then, in the long run 95% of such confidence intervals will cover the true β_4.)

- **Discussion like this can be rewarded to compensate for missing marks elsewhere, but no penalty if not provided as it was not specifically asked for…**

We don’t refer to confidence intervals as probability intervals because while they are statements of “confidence” in the value of the parameter, the parameter itself is viewed as a fixed constant, and is not a random quantity.

Question 22 [Short Answer, 7 marks]

Explain (with words) what is a case-deleted residual, and when it is used.

Answer:

A case-deleted residual is the difference between a single observed response (dependent variable) value (y_i say) and the corresponding value *predicted* by the relevant regression model, but with that model estimated using all of the $n-1$ other data points available and leaving out the one for observation i. There are n individual case-deleted residuals for a given model and sample.

Case-deleted residuals are used to construct a leave-one-out cross validation (LOOCV) measure, which is the average of the individual squared case-deleted residuals from n observations, all obtained using the same model.

The LOOCV measure is used to assess the model's fit to the data.

Question 23 [Essay, 10 marks]

Explain in detail a general strategy for selecting the “best” (most preferred) linear regression model in the context of the model for monthly returns and financial factors, as described. Where possible, incorporate the information provided in the Tables and Figures to illustrate aspects of your approach, and also note any additional visualisations or analyses that you would typically undertake before determining your preferred model.

Answer:

In general there are many different strategies that one could employ. Here we are looking for a blend of automated computational techniques (such as starting with the largest model and then fitting all possible models to determine those that are “best” according to a collection of different penalised criteria, like AIC, BIC and adjusted R^2) and more heuristic methods, such as looking at residual plots (histograms, scatterplots, etc) to try to identify if there are detectable patterns remaining that could be exploited to improve the model. Also, we need to consider the purpose of the model - e.g. is it for prediction or policy analysis?

So starting with the available information in example provided, we can see that the model selection criteria AIC and BIC have selected the same model, the only that uses an intercept term plus x1 and x4. Adjusted-R^2 keeps the intercept term plus X1, X3 and X4. We would consider at least these two models as being contenders for the final choice.

So we might take these models, and look at the residual plots (against the fitted response, against other x-variables, against anything else we think might be relevant) to see if any information is obviously missing from one or the other.

We could also check for multi-collinearity using the variance inflation factors (though this could have been done earlier), and large leverage and Cook’s D to check for influential observations, as we want our model to be relatively robust to the influence of single data points.

Before making a final decision we would consider what the purpose of the model is for - if for forecasting investment returns, then probably it is good to leave out the additional regressors (conventional wisdom is smaller models tend to predict better). However, if I were wanting to better understand my investment risk, perhaps a larger model is more appropriate. So it just depends on the purpose of finding this “best” model.

SECTION B

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.

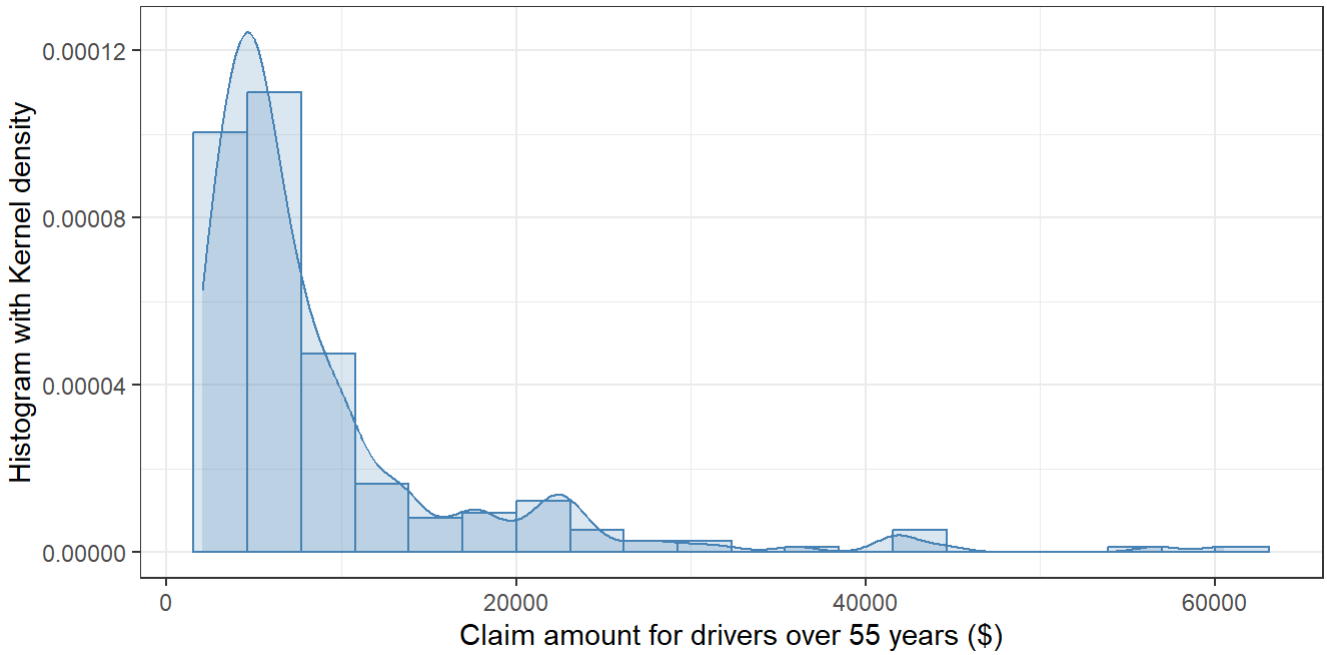

Figure C.1: Histogram and kernel density estimate of $n = 239$ claim amounts by drivers over 55 years (y).

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

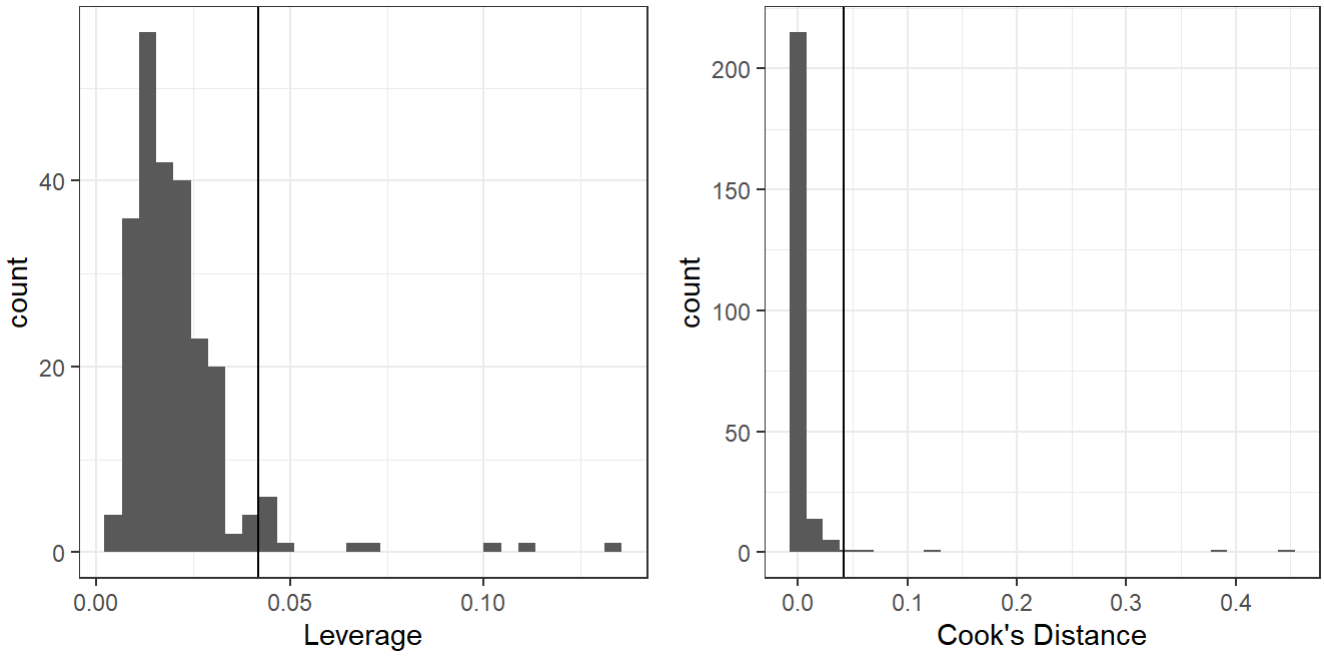

Figure C.2: Histogram of Leverage and Cook's Distance from the full model, with threshold.

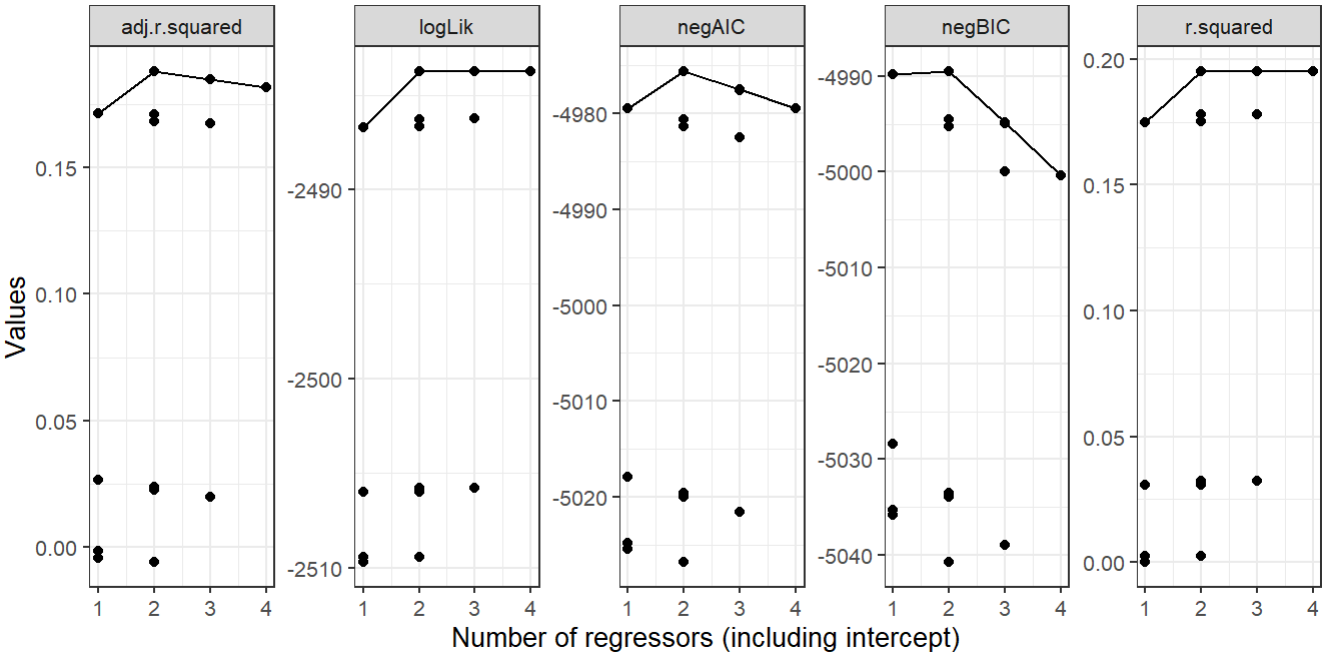

Figure C.3: Panel of visualisations of the model performance measures for all fitted models.

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

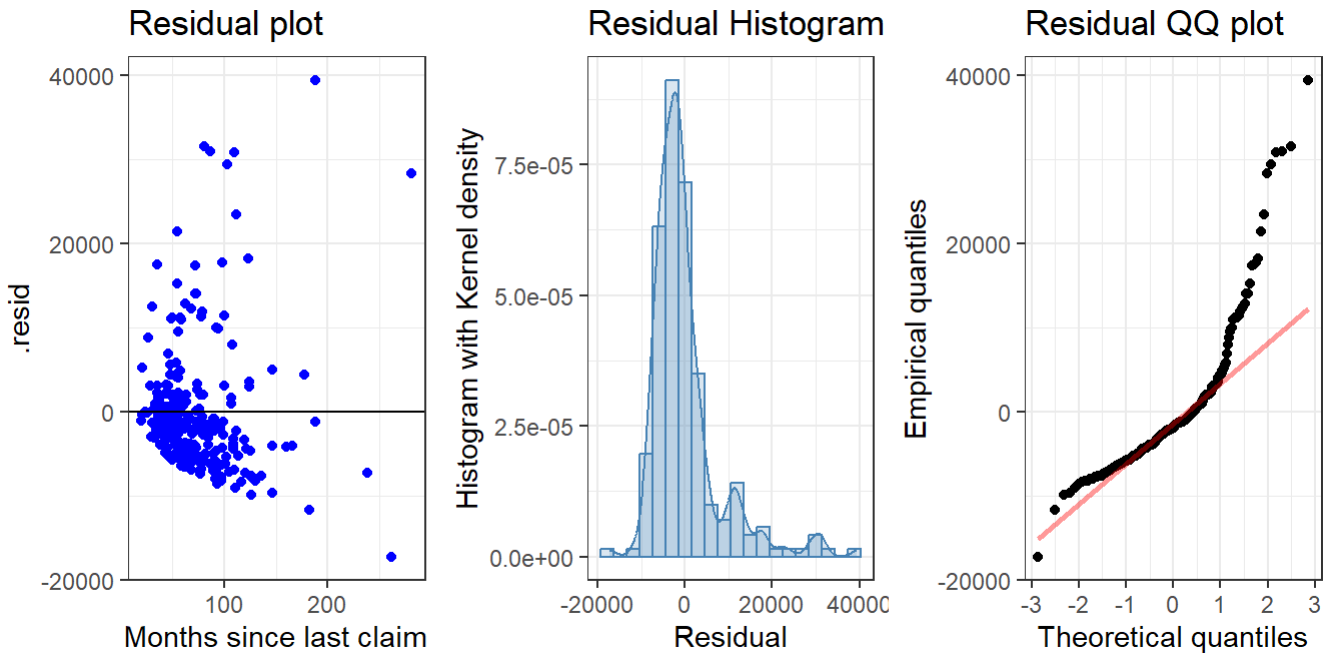

Figure C.4: Residual plots for best fitting model.

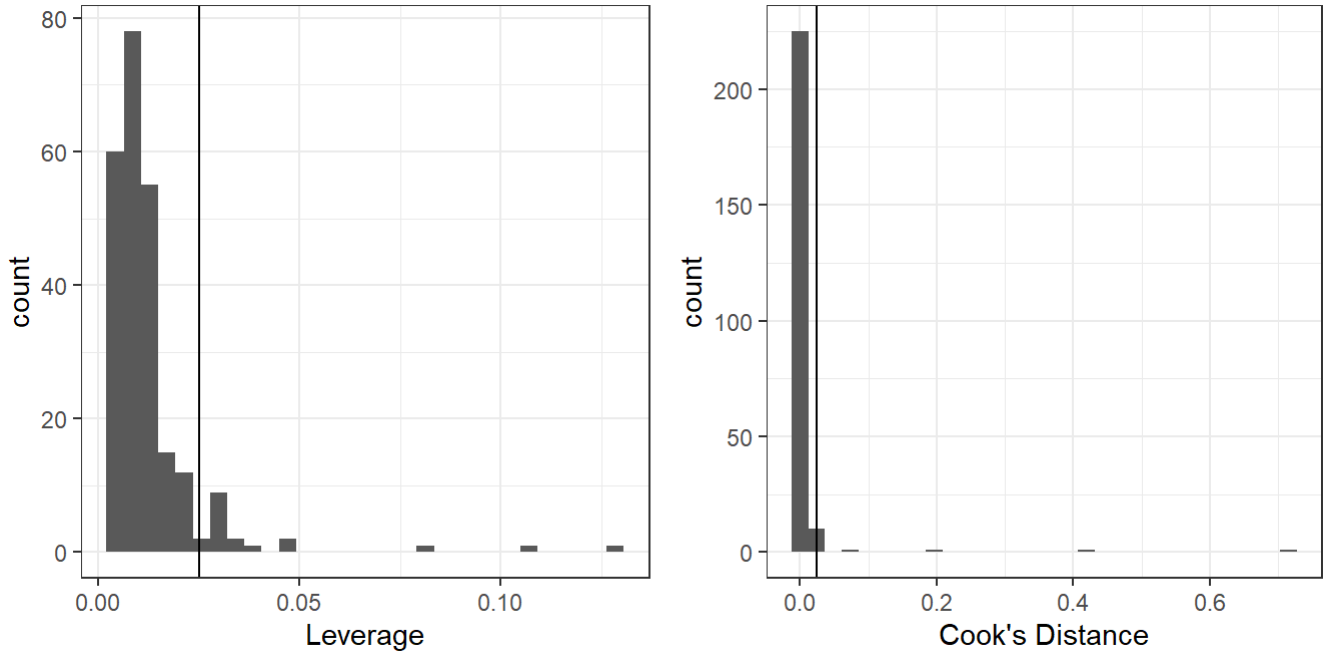

Figure C.5: Histogram of Leverage and Cook's Distance with threshold for model with best fit statistics.

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.

Answer: TRUE (The LOOCV measure uses the residuals and leverage, so only one model needs to be estimated)

Question 16 [True or False, 3 marks]

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

Answer: TRUE (VIF's > 10 indicate a strong correlation between the variables.)

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.

Answer: b.

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.

Answer

The model has 5 parameters including the constant. The threshold is therefore 10/239.

About 5 observations have high Cook's distance. There are about 12 with high Leverage (the histograms are hard to read, so numbers close to this are OK).

The two measures give an indication of potential influential observations. The leverage statistic uses the hat matrix to identify X values far from the mean. It only looks in the horizontal direction.

Cook's Distance takes into account leverage and fit by using the residual. It looks in both the vertical and horizontal directions, and may give a fairer idea as to the influence an observation may have.

High leverage does not necessarily mean a high Cook's distance, and vice versa.

Once identified, we need to investigate possible reasons explaining the influence these variables have. We may remove them, transform the data or add new variables (among other things). The point is we need to investigate them first - we don't automatically remove them from the regression.

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.

Answer:

Using the information, we estimated the “large” model. First of all, we can see that the coefficients of x3 and x4 are insignificant. Next we can use the fit statistics. R^2 and the log-likelihood cannot be used as they will increase with extra regressors. The other measures penalise extra regressors and each identify that the best fit is from a regression with two variables (including the constant). Given the insignificance of x3 and x4, the obvious choice is to use x1 and x2 as regressors.

The preferred model is not necessarily the “best” model as we have based the selection mainly upon fit. The preferred model has some potential issues. When plotted against x1, the residuals exhibit a “flared” pattern, suggesting that there may be heteroskedasticity present.

The histogram and QQplot show that the residuals do not seem to be normally distributed. We could therefore consider some transformation of the data or some other model specification (eg WLS).

Other variables could also be considered.

We would also want to check the VIF (not provided) for potential multicollinearity.

The preferred model may not meet our purpose, so we would investigate other specifications before presenting a “final” model.

- **Must give a suggested solution & must mention purpose**

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.]

Answer:

The permutation test is a way of approximating the sampling distribution of the β_1 assuming the H_0 is true. So we must "break" the relationship between x1 and y, *while holding x2, x3 and x4 constant*.

We can break the relationship between x1 and y by permuting the **x1 column only**. So we shuffle x1 by sampling *without* replacement from our sample. Once we do this, we can estimate the regression under the constraint that $\beta_1 = 0$.

Repeating this process a large number of times, e.g. R=1000, we end up with a collection replicated regressions and therefore a collection of R $\hat{\beta}_1$'s statistics, which are treated as a sample from the approximate sampling distribution of $\hat{\beta}_1$ under the null hypothesis. Note that we always set the random seed for the permutations so that the analysis can be reproduced exactly later, if needed.

To obtain the p-value, we consider the proportion of these replicated $\hat{\beta}_1$ values that are greater than or equal to the observed $\hat{\beta}_1$ from the original data. This gives an approximate p-value for the upper one-sided hypothesis test. If this p-value is less than 0.05, we reject the null hypothesis for a 5% significance test that the two proportions are equal, and conclude that they are not equal.