Venue ______________________________

STUDENT NUMBER	U							

Research School of Finance, Actuarial Studies and Statistics

PRACTICE FINAL EXAMINATION

Questions updated from previous exam papers in 2017

STAT2008/STAT4038/STAT6038 Regression Modelling

Examination/Writing Time Duration: 180 minutes
Reading Time: 15 minutes

Exam Conditions:
Central Examination.
Students must return the examination paper at the end of the examination.
This examination paper is not available to the ANU Library archives.

Materials permitted in the exam venue: (No electronic aids are permitted e.g. laptops, phones)
Unannotated paper-based dictionary (no approval required),
One A4 page with notes on both side, Calculator

Materials to be supplied to Students:
Scribble Paper

Instructions to Students:

1. This examination paper comprises a total of twenty-three (23) pages and there is a separate handout of R output which has a total of nineteen (19) pages. During the reading time preceding the exam, please check that both documents have the correct number of pages.
2. All answers are to be written on this exam paper, which is to be handed in at the end of the exam. You may make notes on scribble paper (or on the R handout) during the reading time, but **do NOT write on this exam paper until after the start of the writing time.** If you need additional space, use the rear of the previous page and clearly indicate the part of the question that your answer refers to. The R handout and any scribble paper will be collected at the end of the examination and destroyed, they will not be marked.
3. There are four questions, worth a total of 60 marks. The parts of each question are of unequal value, with the marks indicated for each part. **You should attempt to answer all parts of Q1, Q2, Q3 and EITHER Q4 (STAT2008) or Q4A (STAT4038/6038).** This examination counts towards 60% of your final assessment (at least the real final exam will, rather than this practice exam).
4. **Please write your student number in the space provided at the top of this page.**
5. **Include a clear statement of the formulae you use to answer each question.**
6. Statistical tables (generated using R) are provided on pages 18 and 19 at the end of the handout of R output. Unless otherwise indicated, use a significance level of 5% and log x refers to the natural logarithm of x.

	Q1	Q2	Q3	Q4	Q4A	Total
Pages	2 to 5	6 to 10	11 to 15	16 to 19	20 to 23	
Marks	**15**	**15**	**15**	**15**	**15**	**60**
Score						

Question 1 **(15 marks)**

Sugar in Potatoes is example 3 from the appendix at the end of the lecture notes.
The data are shown in the R output for this question and were collected in an experiment designed to investigate the glucose content of potatoes during storage.

(a) An initial model (potatoes.lm) has been fitted to these data on page 1 of the R output. Does the residual plot shown on page 1 of the R output suggest a problem with one of the assumptions underlying the model? Do not discuss all of the assumptions, just decide whether or not there is a problem and if so, just choose the most important assumption associated with that problem and discuss that assumption.

(2 marks)

Question 1 continued

(b) Summary output from the initial model (potatoes.lm) is given at the top of page 2 of the R output, but details of the F statistic have been edited (replaced by question marks) and the analysis of variance (ANOVA) table is not shown. Fill in the details of the ANOVA table in the spaces shown below. Hint: you could do this by working with basic formulae from the data, but it is a lot easier to work from other items given in the R output – if you are worried about making mistakes, then as well as writing your answers below, give some details of how you obtained these answers in the space below, otherwise you will get no marks for any incorrect answers.]

Source	Degrees of Freedom	Sum of Squares	Mean Square	F statistic	p-value
Model (Regression)					
Residual (Error)				▨	▨
Total			▨	▨	▨

(5 marks)

Question 1 continued

(c) Use the initial model (potatoes.lm) to predict the glucose levels for the mean, minimum and maximum number of weeks in the data. Also calculate 95% prediction intervals for all three predictions and compare these intervals with the observed values of glucose in the data (the data includes observations taken at all three of these values). You should include some comments about what these comparisons tell you about the overall fit of this model.

(4 marks)

Question 1 continued

(d) There is also summary output for a second model (potatoes.lm2) on page 2 of the R output, which includes an additional term added to the initial model. If you are going to fit a model with this additional term, why should you still include the other terms from the initial model as well? Is this additional term a significant addition to the initial model? The ANOVA table is again not shown for this second model, but what would be the F statistic and degrees of freedom associated with this additional term?

(4 marks)

Question 2 **(15 marks)**

Black Cherry Trees is example 4 from the appendix at the end of the lecture notes. The data shown in the `R` output for this question were obtained from a sample of black cherry trees in order to examine the relationship between the `Volume` (measured in cubic feet) of wood in the tree and the `Height` (measured in feet) and the `Diameter` (measured in inches) of the trees.

(a) The first model shown on pages 3 and 4 of the `R` output is `trees.lm`, with residual plots shown on page 5. This is the model suggested in example 4 in the appendix at the end of the lecture notes. The Residuals vs Fitted Values plot for this model has been standardised – are the standardised residuals shown on the vertical axis; internally or externally studentised residuals? Which of the underlying assumptions of the model cannot be assessed using just the plots shown on page 5 of the output?

(2 marks)

Question 2 continued

(b) Which observation is in the bottom right-hand corner of the Residuals vs Fitted Values plot, with a fitted value just greater than 4.5 and a standardised residual value below –2? [Hint: take a guess based on the Cook's distance plot and then confirm your guess by calculating the fitted value for that observation and show the details of this calculation below]. Is this observation a potential problem? If so, describe the nature of the problem.

(3 marks)

Question 2 continued

(c) There is also a second model (trees.lm2) shown on page 4 of the R output, with residual plots shown on page 6. Are trees.lm and trees.lm2 nested models? Which measures from the two models are directly comparable? Use these measures and the residual plots to compare the two models. Which model do you think is the better fitting model: trees.lm or trees.lm2?

(3 marks)

Question 2 continued

(d) For the trees.lm2 model, test the hypotheses that the coefficient of log_Radius is not significantly different from 2 and that the coefficient of log_Height is not significantly different from 1. Give full details of both these hypotheses tests.

(4 marks)

Question 2 continued

(e) The trees.lm2 model is fitted with all three variables on the log scale. What does this model suggest is the relationship between the variables on the original scale: Volume (V in cubic feet); Radius_ft (r in feet); and Height (h in feet)? Assume that most of the useable volume of wood is located in the trunk of a tree and that the height is measured to the top of the trunk. So, if the volume of a cylinder is $V = \pi r^2 h$, and the volume of a cone is $V = {}^1/_3(\pi r^2 h)$, are the tree trunks of the black cherry trees closer to being cylinders or cones under this model?

(3 marks)

Question 3 **(15 marks)**

Giving in the Church of England is example 5 from the appendix at the end of the lecture notes. The data shown in the R output for this question record the amount of annual giving in ₤ (pounds sterling) per church member (Annual_giving) in a sample of 20 dioceses in the Church of England (a diocese is an administrative region usually containing a number of churches). Three other potentially relevant factors are also recorded for each diocese: employment rate as a percentage (Employment); the percentage of the population on the electoral roll of the church (Electoral_Roll); and the percentage of the population who usually attend church (Attendance).

(a) Explain what is going on with the three models church.lm1, church.lm2 and church.lm3 shown on page 8 of the R output. How can Employment be marginally insignificant (at $\alpha = 0.05$) if fitted last in the model, significant if fitted second, but insignificant if fitted first? [Hint: there are 2 marks attached to this question, so a one-word answer, simply naming the problem, may be enough to get you one mark, but it will not be sufficient detail to get you both marks.]

(2 marks)

Question 3 continued

(b) In the context of model church.lm2 on page 8 of the R output, are Attendance and Employment (grouped together) a significant addition to a model that already contains Electoral_Roll? Give full details of an appropriate hypothesis test.

(4 marks)

Question 3 continued

(c) Compare all 5 models shown on pages 8 and 9 of the R output (church.lm1, church.lm2, church.lm3, church.lm2a and church.lm2b) and explain how a forward selection version of a model selection process would end up choosing model church.lm2b. [Hint: there are 3 marks here, as the selection process consists of at least 3 steps.]

(3 marks)

Question 3 continued

(d) Interpret the values and significance of the estimated partial regression coefficients in the summary output for the model `church.lm2b` shown on page 9 of the `R` output. Does the intercept coefficient have a sensible interpretation in the context of this model?

(3 marks)

Question 3 continued

(e) The summary output for the model church.lm2b shown on page 9 of the *R* output has been edited with various summary statistics replaced by question marks. Calculate the value of these missing summary statistics and their associated degrees of freedom. Do not estimate the missing *p*-value, but do interpret the meaning and significance of the overall F statistic.

(3 marks)

Question 4 (STAT2008) **(15 marks)**

The Galapagos Islands are located in a remote part of the Pacific Ocean (1,000km off the coast of Ecuador) and are a fertile laboratory for studying the factors that influence the development and survival of different plant species. The data frame gala in the faraway library contains information for 30 different islands on the number of plant Species, the number of species that occur on only that island (Endemics), the Area (km^2) of the island, the highest Elevation on the island (metres), the distance from the Nearest island (km), the distance from Santa Cruz (Scruz, also measured in km), and the area of the Adjacent (nearest) island (km^2).

Santa Cruz is the central and most heavily populated island (in terms of human population). At the time of this study (early 1970s), only 5 of the islands had regular human inhabitants (Baltra, Isabela, San Cristobal, Santa Cruz and Santa Maria).

The goal of the analysis is to assess the factors that influence diversity, as measured by some function of the number of species and the number of endemic species. One suggestion for measuring diversity is Diversity = Species – Endemics.

(a) An initial model (gala.lm) has been fit to these data on page 10 of the R output. The response variable is log(Diversity + 1) rather than log(Diversity). Why was it necessary to add 1 before taking logs?

(1 mark)

(b) Using the ANOVA table for gala.lm on page 10 of the R output, conduct a nested *F* test to determine if any of the explanatory variables: Elevation, Nearest, Scruz and log(Adjacent) are significant additions to a model which already includes log(Area)? Give full details of this hypothesis test.

(4 marks)

Question 4 continued

(c) Residual plots for a reduced model (gala.lm2) are shown on page 11 of the R output. Do these plots suggest any problems with the underlying assumptions?

Are there any problem(s) shown on the "Residuals vs Fitted" plot on page 11?
If so describe the problem(s):

Are there any problem(s) shown on the "Normal Q-Q" plot on page 11?
If so describe the problem(s):

Are there any problem(s) shown on the "Cook's distance" plot on page 11?
If so describe the problem(s):

What is your overall assessment? (select just ONE of the following options)

□ Residuals are not independent (obvious pattern)
□ Residuals do not have constant variance (heteroscedasticity)
□ Residuals are not normally distributed
□ There are possible outliers and/or influential observations
□ More than one of the above problems
□ No obvious problems

(2 marks – 0.5 for each section)

Question 4 continued

(d) Can you suggest some possible modification to the model that might remedy all of the problems you identified in part (c)?

(1 mark)

(e) Summary output for the reduced model (gala.lm2) are shown on page 12 of the R output. Suppose there was an additional island not included in the original study, which has an Area of 2.59 km^2. Use the reduced model (gala.lm2) to predict the expected Diversity on this island and also find an appropriate 95% interval (confidence or prediction) for this estimate.

(4 marks)

Question 4 continued

(f) Given the goal of the analysis, do you think the researchers will be happy with a model (gala.lm2) for species diversity that only involves the size (Area) of each island and doesn't include any of the other variables?

[Note this question is really asking for a brief, but sensible discussion of the underlying research question and whether this model really helps to address that question, so very short answers will not get any marks, nor will long answers that fail to address the issues.]

(2 marks)

(g) Added variable plots for each of the other possible explanatory variables (other than Area) are shown on page 13 of the R output. What is the purpose of an added variable plot and do these plots appear to be useful in this instance?

(1 mark)

Question 4A (STAT4038/STAT6038) (15 marks)

The data frame baycheck in the Using R library contains estimated populations for a variety of Bay Checkerspot butterflies near California. A common model for environmental population dynamics is the Ricker model, for which *t* is time in years:

$$N_{t+1} = aN_t e^{bN_t} W_t \qquad (1)$$

Where *a* and *b* are parameters and W_t is a log-normal multiplicative error. This can be turned into a linear regression model by dividing by N_t and then taking logs of both sides to give:

$$\log\left(\frac{N_{t+1}}{N_t}\right) = \log a + bN_t + \varepsilon_t \qquad (2)$$

Let y_t be the left-hand side of equation (2), n_t be the estimated populations (N_t) for all available years (but excluding the last year), *r* represent the unconstrained intrinsic growth rate, *K* represent the environmental carrying capacity and then equation (2) can be written as:

$$y_t = r\left(1 - \frac{n_t}{K}\right) + \varepsilon_t \qquad (3)$$

(a) Page 14 of the R output shows details of how to reorganise the data in order to fit model (3) as a linear model and page 15 shows the estimated coefficients from this model (baycheck.lm) and a plot of the reorganised data with the model superimposed. Use the estimated partial regression coefficients from the model to estimate *r* and *K* and interpret these estimates.

(2 marks)

(b) Identify the observation in the bottom right hand corner of the plot on page 15 of the R output. To which year in the original data does this observation correspond and what was different about that year? Do you think this observation is causing a problem in the context of the model baycheck.lm? What other diagnostics should you check?

(3 marks)

Question 4A continued

(c) Residual plots for a reduced model with observation 17 excluded (`baycheck.lm2`) are shown on page 16 of the `R` output. Do these plots suggest any problems with the underlying assumptions?

Are there any problem(s) shown on the “Residuals vs Fitted” plot on page 16?
If so describe the problem(s):

Are there any problem(s) shown on the “Normal Q-Q” plot on page 16?
If so describe the problem(s):

Are there any problem(s) shown on the “Cook’s distance” plot on page 16?
If so describe the problem(s):

What is your overall assessment? (select just ONE of the following options)

□ Residuals are not independent (obvious pattern)
□ Residuals do not have constant variance (heteroscedasticity)
□ Residuals are not normally distributed
□ There are possible outliers and/or influential observations
□ More than one of the above problems
□ No obvious problems

(2 marks – 0.5 for each section)

Question 4 continued

(d) Can you suggest some possible modifications to the model that might remedy the problems you identified in part (c)?

(1 mark)

(e) Summary output for the reduced model (baycheck.lm2) are shown on page 17 of the R output. Again estimate *r* and *K* and compare these estimates with the ones in part (a). Use the reduced model to predict the expected value of y_t for the final year (1986) and also find an appropriate 95% interval (confidence or prediction) for this estimate. What does this prediction suggest the estimated population (N_t) will be in 1987?

(4 marks)

Question 4 continued

(f) Given your analysis of the residual plots in part (c), do you think this is an appropriate model for these data? Do the estimated populations of Bay Checkerspot butterflies really follow a Ricker model?

[Note this question is really asking for a brief, but sensible discussion of the underlying research question and whether this model really helps to address that question, so very short answers will not get any marks, nor will long answers that fail to address the issues.]

(3 marks)

END OF EXAMINATION