Australian National University

Venue ______________________________

STUDENT NUMBER

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| U | | | | | | | |

## Research School of Finance, Actuarial Studies and Statistics

## Solutions to the 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.

*There is an obvious violation of the assumption that the errors are independent. There are definitely aspects of the underlying relationship that are not included in the model, suggesting that the model is not an appropriate model (which is also a violation of one of the more general assumptions underlying linear models). We could try a transformation to address the non-linearity or, as I do later in this question, we could try including a quadratic term in the model.*

**(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) | 1 | 5639.1 | 5639.1 | 4.004 | 0.068562 |
| Residual (Error) | 12 | 16908.1 | 1409.0 | | |
| Total | 13 | 22547.2 | | | |

*This is a simple linear regression model with just one predictor, fitted to a sample of size $n$ = 14, so the total degrees of freedom (df) equal $n$ – 1 = 13 and estimating just one slope coefficient requires* 1 *df, leaving* 12 *df to estimate the mean square error (MSE).*

*The total sum of squares (SS) is the total df times the variance of* Glucose (13 × 1734.4011) *and this can be divided between the model SS and the error SS using the R-squared value in the summary output (the rounded value of* 0.2501 *leads to some minor rounding errors).*

*You can then calculate the rest of the table and check that the MSE is the square of the residual standard error and the F statistic is the square of the t statistic for* Weeks*, with the same p-value (as this is a simple linear regression model).*

**(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.

*Using the following formulae and values:*

$\hat{Y}_i = 110.56 + 3.345 \cdot x_i \quad Y = \text{Glucose},\ x = \text{Weeks},\ i = 1, 2, \cdots n = 14$

$s = residual\ standard\ error = 37.54$

$x^* = \left\{ \bar{x} = 11 \quad \min(x) = 2 \quad \max(x) = 20 \right\}, \quad s_x^2 = 38.76923$

$t_{n-2}(1-\alpha/2) = t_{12}(0.975) = 2.1788$

$$\hat{Y} \pm t_{n-2}(1-\alpha/2)\, s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

| $x^*$ | $\hat{Y}$ | 95% *PI* | *observed Y* |
|---|---|---|---|
| $\bar{x}$ | 147.36 | (62.70, 232.02) | 99, 127 |
| $\min(x)$ | 117.25 | (26.46, 208.04) | 148, 167 |
| $\max(x)$ | 177.46 | (86.67, 268.25) | 212, 245 |

*The observed Y values fall within these very wide confidence intervals, but are well towards the lower end of the range for the mean number of weeks (and the predicted value is a lot larger than either observed value). The observed Y values also fall well towards the upper end of the range for both the minimum and maximum number of weeks (and the predicted values are a lot smaller than the observed values). So, even if we do not go the extra step of plotting the residuals, these systematic departures of the model from the observed data suggest there are major problems with the 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?

*As a general rule, when fitting higher order terms (quadratic or interaction terms), we should always include all lower order terms to allow maximum flexibility in how the model fits the data.*

*So, unless there is a good reason to assume a model of a particular form, a quadratic model involving* Weeks.sqd *such as* potatoes.lm2 *should include the linear term in* Weeks *and the constant* (Intercept) *term. Judging by the t test on the coefficient of* Weeks.sqd *in the summary output* ($t_{11}$ = 8.04, $p$ = 0.00000623) *this addition to the model is significant at* $\alpha$ = 0.05. *As this test is for the last term added to the model, we can square this test statistic to find the equivalent sequential F test in the ANOVA table* ($F_{1,11}$ = 64.64, $p$ = 0.00000623).

**(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?

*The* rstudent() *function in* R *is used to calculate the externally studentised residuals.*

*The residual vs fitted values plot provided can be used to assess most of the assumptions that the errors are independent and identically distributed with constant variance, the only aspect they don't really assess is whether or not the errors are normally distributed; which requires a normal quantile (qq) plot.*

**(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.

*For model* trees.lm*, the fitted value for observation* 31 *is:*

$$\hat{Y}_i = 0.102585 + 0.145290(20.6) + 0.016385(87) = 4.52$$

*So observation* 31 *is the one we are after. It has both a large externally studentised residual value, which is probably just outside* $t_{27}(0.025) = -2.0518$*, and as there is a relatively large vertical gap between this residual and the other residuals, it is also probably influential in the fit of the model. This is why the observation stands out from the other observations on the plot of Cook’s distances. This observation would probably require some sort of treatment (for example, exclusion), if it continues to be a problem once we are satisfied we have an otherwise appropriate model.*

**(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?

*The two models are not nested as one model is not a subset of the other, so we cannot use a nested model F test to decide between the two models. The two models do share the same response variable, so we can directly compare summary measures based on the residuals such as the residual standard error (smaller is better in terms of RSE). Even if they were not on the same scale, we could arguably still compare properly scaled summary measures such as the R-squared, the adjusted R-squared and the overall F statistic for the two models (larger is better for each of these other summary measures).*

*The second model* trees.lm2 a*ppears to be slightly better on all of the summary measures, but the main reason I would choose* trees.lm2 *over* trees.lm *is that it appears to be a better fit to all of the data (including observation* 31*, which no longer features prominently on the Cook's distance plot). Note the largest Cook's distance is much smaller for* trees.lm2 *than for* trees.lm.

**(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.

*The underlying population model is:*

$$\text{log(Volume)} = \beta_0 + \beta_1 \log\left(\text{Radius}\right) + \beta_2 \log\left(\text{Height}\right) + \varepsilon \quad \varepsilon \textit{ iid } N\left(0, \sigma^2\right)$$

*So the test for* log_Radius *is:*

$H_o : \beta_1 = 2 \quad H_a : \beta_1 \neq 2$

*Reject* $H_o$ *in favour of* $H_a$ *if observed test statistic* $\left(t\right)$:

$t < t_{28}\left(0.025\right) = -2.0484$ *or* $t > t_{28}\left(0.975\right) = 2.0484$

$$t = \frac{\hat{\beta}_1 - E\left[\beta_1 \middle| H_o\right]}{se\left(\hat{\beta}_1\right)} = \frac{1.98265 - 2}{0.07501} = -0.2313$$

*And the test for* log_Height *is:*

$H_o : \beta_2 = 1 \quad H_a : \beta_2 \neq 1$

*Same decision criteria as the previous test.*

$$t = \frac{\hat{\beta}_2 - E\left[\beta_2 \middle| H_o\right]}{se\left(\hat{\beta}_2\right)} = \frac{1.11712 - 1}{0.20444} = 0.5729$$

*In both tests, we do not reject the null hypothesis and can therefore conclude 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.

**(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?

*In mathematical terms, the fitted model* trees.lm2 *is:*

$$\ln\left(\hat{V}\right) = \hat{\beta}_0 + \hat{\beta}_1 \ln\left(r\right) + \hat{\beta}_2 \ln\left(h\right)$$

$$\Rightarrow$$

$$\begin{aligned}\hat{V} &= \exp\left[\hat{\beta}_0 + \hat{\beta}_1 \ln\left(r\right) + \hat{\beta}_2 \ln\left(h\right)\right]\\ &= e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1 \ln(r)} \cdot e^{\hat{\beta}_2 \ln(h)}\\ &= e^{\hat{\beta}_0} \cdot \left(e^{\ln(r)}\right)^{\hat{\beta}_1} \cdot \left(e^{\ln(h)}\right)^{\hat{\beta}_2}\\ &= e^{\hat{\beta}_0} \cdot r^{\hat{\beta}_1} \cdot h^{\hat{\beta}_2}\end{aligned}$$

*We have shown in part (d) that* $\hat{\beta}_1 \approx 2$ *and* $\hat{\beta}_2 \approx 1$*, so the above relationship is close to:*

$$\hat{V} = cr^2 h \textit{ where } c = e^{\hat{\beta}_0} = \exp(-0.33065) = 0.7185 \textit{ and } \pi \approx 3.1416, \frac{\pi}{3} \approx 1.0472$$

*So the fitted model is suggesting a fitted* Volume *of useable timber that is somewhat less than assuming that black cherry trees are either perfect cylinders or cones, although of those two alternatives, they are closer to fitting a conical tree model.*

*There could also be a problem with one of the many other assumptions made above, for instance, the height may have been measured to the tops of the tree rather than the tops of the trunk.*

**(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.]

*The problem is multicollinearity; there are strong relationships between the predictors in these models and the variance inflation factors (vifs) for the combination of explanatory variables involved in these models are high.*

*Judging by the scatterplot matrix and the variance inflation factors, the main culprit is the strong association between* Electoral_Roll *and* Attendance *which are two closely related measures of engagement with the Church of England. Again judging by the scatterplots, either of these measures is potentially a better predictor of* Annual_giving *than* Employment*, but* Electoral_Roll *and* Attendance *are so closely related that they will cause problems if both of them are included in the same model. As the ANOVA tables are fitted sequentially, the significance of* Employment *depends on which of these other variables have already been included in the model.*

**(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.

*The underlying population model for* church.lm2 *is:*

$$\text{Annual\_giving} = \beta_0 + \beta_1\text{Electoral\_Roll} + \beta_2\text{Attendance} + \beta_3\text{Employment} + \varepsilon$$

$$\varepsilon \; iid \; N\left(0, \sigma^2\right)$$

*So we can do a nested model F test for the addition of the two terms involving* Attendance *and* Employment *to a model that already includes* Electoral_Roll:

$$H_o : \frac{\sigma^2_{Addition}}{\sigma^2_{Error}} = 1 \quad H_a : \frac{\sigma^2_{Addition}}{\sigma^2_{Error}} > 1 \quad \text{or equivalently}$$

$$H_o : \beta_2 = \beta_3 = 0 \quad H_a : \text{at least one of } \beta_2, \beta_3 \neq 0$$

$$\text{Reject } H_o \text{ in favour of } H_a \text{ if observed test statistic } \left(F\right):$$

$$F > F_{2,16}\left(0.95\right) = 3.634$$

$$F = \frac{\left(64.60 + 189.88\right)/\left(1+1\right)}{MS_{Residual}} = \frac{127.24}{49.49} = 2.57$$

*So, we do not reject the null hypothesis and can therefore conclude that the additional terms are not a significant addition to the model.*

**(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.]

*In a purely observational situation, forward selection starts with a null model (involving just an intercept term). We then add the potentially most significant predictor (the one with largest sum of squares when added in first). Judging by the first three models, this would mean adding in* Electoral_Roll *with a SS of* 589.65 *as in model* church.lm2.

*Assuming this first step was a significant addition to the model, we would then add in whichever of the remaining predictors is the next most significant addition to the model. Comparing models* church.lm2 *and* church.lm2a*, the next most significant addition would appear to be* Employment *with a SS of* 251.91, *rather than* Attendance *with a SS of* 64.90. *Judging by the last sequential F-test in model* church.lm2b, *the addition of* Employment *to a model that already includes* Electoral_Roll *is significant* ($F_{1,17}$ = 5.3906, $p$ = 0.032923).

*Finally, it is apparent from model* church.lm2a *that the addition of the last remaining candidate predictor* Attendance *is not a significant addition* ($F_{1,16}$ = 0.0519, $p$ = 0.822615), *so we stop and choose model* church.lm2b, *a model which does not have a problem with multicollinearity (judging by the low vifs).*

**(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?

*The coefficient of* Electoral_Roll *suggests that on average* Annual_giving *declines by around* £4.01 *for each percentage increase on the* Electoral_Roll *and this decrease is significant* ( $t_{17}$ = –4.08, $p$ = 0.000779). *This is a little counter-intuitive as it suggests the communities with higher engagement with the Church of England tend to have lower* Annual_giving *than less engaged communities. However, it could be that the more engaged communities are the ones that are more likely to seek assistance from the church and are less able to give.*

*Similarly, the coefficient of* Employment *suggests that on average* Annual_giving *increases by around* £1.34 *as the* Employment *rate increases by* 1% *and this increase is significant* ( $t_{17}$ = 2.322, $p$ = 0.0032923). *Presumably communities with higher employment are the ones better placed to give to the church.*

*A negative intercept is possible in that it could be suggesting for the poorest communities that the church gives out more in charity than it receives in donations. However, this intercept is not significant* ( $t_{17}$ = –1.086, $p$ = 0.292531) *and the intercept is outside the range of both predictors, as* Electoral_Roll *ranges from* 1.9% *to* 8.7% *and* Employment *ranges from* 82.6% *to* 92.8%.

**(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.

$$\textit{Residual standard error} = \sqrt{MS_{Residual}} = \sqrt{46.73} = 6.836 \textit{ on } 17 \textit{ degrees of freedom}$$

$$\textit{Multiple R-squared} = \frac{SS_{Model}}{SS_{Total}} = \frac{(589.65+251.91)}{(589.65+251.91+794.44)} = \frac{841.56}{1636} = 0.514$$

$$\textit{Adjusted R-squared} = 1 - \frac{MS_{Residual}}{SS_{Total}/df_{Total}} = 1 - \frac{46.73}{1636/19} = 0.457$$

$$F = \frac{SS_{Model}/df_{Model}}{SS_{Residual}/df_{Residual}} = \frac{(589.65+251.91)/(1+1)}{MS_{Residual}} = \frac{420.78}{46.73} = 9.004$$

*The above F statistic is on* 2 *and* 17 *degrees of freedom, so can be compared with* $F_{2,17}(0.95) = 3.592$ *and we can conclude that at least one of the terms in the model is significant, either the term involving* Electoral_Roll *or the term involving* Employment *or, as shown in part (d), in this instance, both terms are significant.*

**(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?

*A few of islands (Caldwell, Enderby and Onslow) have no non endemic species (i.e.* Species = Endemics*), so the value of our chosen measure of* Diversity *is* 0.

log(0) *is not defined, which means these islands would be treated as missing and not included in the estimation of the regression model. To avoid losing these observations, we need to add a small positive constant before applying the log transformation.*

**(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.

*The underlying population model is:*

$$\text{log(Diversity + 1)} = \beta_0 + \beta_1 \text{log(Area)} + additional\ terms + \varepsilon \quad \varepsilon\ iid\ N\left(0,\sigma^2\right)$$

*Where the additional terms are:* $\beta_2$Elevation + $\beta_3$Nearest + $\beta_4$Scruz + $\beta_5$log(Adjacent)

*The nested F-test on this group of additional terms:*

$$H_o : \frac{\sigma^2_{addition}}{\sigma^2} = 1 \quad H_a : \frac{\sigma^2_{addition}}{\sigma^2} > 1$$

$$\left[\equiv H_o : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \quad H_a : at\ least\ one\ \beta_j \neq 0, j = 2,3,4,5\right]$$

$$Reject\ H_o\ in\ \ favour\ of\ H_a\ \ if\ \ observed\ test\ statistic\ F > F_{4,24}\left(0.95\right) = 2.776$$

$$F = \frac{MS_{Addition}}{MS_{Residuals}} = \frac{\left(0.139 + 1.599 + 1.621 + 0.052\right)/4}{21.401/24} \approx 0.96$$

*So do not reject the null hypothesis and conclude that the additional terms are not a significant addition to a model which already includes* log(Area).

**(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):

*No obvious departure from the assumption of independence on the main residuals vs fitted plot, though there is a suggestion of decreasing (non-constant) variance, as the fitted values increase. As usual,* 3 *observations have been labelled (by default). There is some space in the vertical direction between observation* 7 *(Daphne Minor) and the other observations, so it might be a possible outlier, but this is definitely not the case for the other two labelled observations (*3 *and* 13*).*

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

*Similarly, there is no obvious departure from the assumption of normality on the normal quantile plot. Only observation number* 7 *out of* 30 *observations has an internally studentised residual outside* (–2, 2) *and it is only just outside this range.*

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

*The Cook's distance for observation* 7 *appears large relative to the other observations, however, the vertical scale on this plot only goes to just over* 0.2*, which is relatively small for Cook's distances, even with a relatively small sample size.*

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)?

*I suspect the apparent decreasing variance on the main plot and the status of observation 7 as a possible outlier are linked. The log transformation, applied to the response and some of the explanatory variables, appears to have been too strong, as it has over-corrected the observations with larger fitted values, which leaves some of the observations with smaller fitted values looking like potential outliers.*

*We could experiment with weaker transformations, such as a square root transformation, which might be appropriate, given the nature of variables such as* Area.

**(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.

*The predicted value is:*

$$\log(\widehat{\text{Diversity}} + 1) = 2.33602 + (0.41277)\log(2.59) \approx 2.72884$$

*Back-transformed and rounded to the nearest whole number:*

$$\widehat{\text{Diversity}} = e^{2.72884} - 1 \approx 14$$

*A* 95% *prediction interval for this estimate is:*

$$\log(\widehat{\text{Diversity}} + 1) \pm t_{28}(0.975) \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{[\log(2.59) - mean(\log(\text{Area}))]^2}{(n-1) \cdot var(\log(\text{Area}))}}$$

$$= 2.7288 \pm (2.0484)(0.9414)\sqrt{1 + \frac{1}{30} + \frac{(\log(2.59) - 1.554093)^2}{29(12.25517)}}$$

$$\approx (0.77, 4.69)$$

*Again, back-transformed to the nearest whole numbers:*

$$(e^{0.77} - 1, e^{4.69} - 1) \approx (1, 108)$$

*Which is not a very precise prediction.*

**(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.]

*The researchers presumably collected the other explanatory variables as they had prior assumptions that factors such as: the nature of the terrain (*Elevation*); the proximity (*Nearest*) and size (*Adjacent*) of the nearest island; and how far the island is from the main centre of human population (*Scruz*), were all likely to play a role. The researchers will probably not be happy to find that none of these variables make a difference and that the only significant factor in explaining species diversity is the size of the island. I also suspect there are a number of other possible variables that might affect diversity that have not been measured and included in the study.*

*With the variables that have been measured, the lack of significance could be the result of measurement issues, in that the collected variables may only be poor proxy measures for the factors that the researchers are really interested in. For example, is the suggested response variable really the best way to measure species diversity? Another possibility is that we have not used the best approach to incorporate the information in these variables into the model. Added variable plots may suggest different scales for some of the exploratory variables – see part (g), below.*

**(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?

*Added variable plots are used to assess whether or not an additional variable (a candidate predictor) should be added to a multiple regression model. They might even suggest a certain functional form (e.g. a particular transformation or a quadratic term) to model the relationship between the unexplained part of the response variable and a candidate predictor (also adjusted for the effects of the other variables).*

*However, none of the added variable plots on page* 13 *suggest any linear or other relationships, so they do not appear to help in this instance. The closest to suggesting some relationship is the added variable plot for* Scruz*, but the apparent reasonably strong negative slope is heavily influenced by just a couple of observations (probably Wolf and Darwin, the two remote islands located more than* 250 *kms from Santa Cruz).*

**(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.

*From the table of partial regression coefficients, the intercept* 0.3458097 *is r and the slope* –0.0004088 *is equal to –r/K, so K* = –0.3458097/–0.0004088 ≈ 845.91.

*As the name suggests, r is the growth rate from year to year, i.e. in most years, the population increases by a multiplicative factor of* $e^r$ = exp(0.3458097) ≈ 1.41, *however, the intercept term is not significantly different from* 0 *(t* = 1.4, *p* = 0.17*).*

*Equation* (3) *suggests that when the population grows to exceed K, the carrying capacity, then r, the growth rate, becomes negative and the population will decline.*

**(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?

*Observation* 17 *is the only observation with an* $n_t$ *value greater than* 7,000, *and this observation is for the year* 1976. *The* 1976 *population of* 7,227 *does represent a substantial increase on the previous year,* 1975, *but the population for that year,* 1,819, *was already in excess of the estimated carrying capacity (K). Even though the population did collapse after* 1976, *to* 852 *in* 1977, *the reduced population was still greater than the estimated K.*

*This observation has definitely been highly influential in the fit of the model, though it only appears to have a relatively small raw (unstandardised) residual and will probably not be classed as a vertical outlier.*

*We should check the usual residual and Cook's D plots and various influence diagnostics. I would expect observation* 17 *to have high leverage; large DFFITS, DFBETAS and COVRATIO and possibly a large externally studentised residual.*

**(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):

*The main residuals vs fitted plot does not really show a departure from the assumption of independence, and the variance looks reasonably constant, however the large horizontal gap suggests that the model is predicting distinctly different fitted values for a group of* 4 *observations. This group includes observation* 16, *which refers to* 1975, *which had the equal second largest population after the now removed* 1976. *This observation appears to have been promoted to potential problem status, now that we have removed observation* 17. *The other members of this group are likely to be the other years with big populations (*1972, 1981 *and* 1983*).*

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

*The normal quantile plot shows only minor departures from the assumption of normality, given the relatively small sample size. However, the problem with observation* 16 *is again apparent as a potential vertical outlier in the upper tail of the residual distribution. The internally studentised residual for observation* 2 *is not less than* –2, *suggesting less of a problem in the lower tail.*

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

*The Cook's distance plot suggests problems with observation* 16 *and with the point with the second most negative internally studentised residual, observation* 23, *which is almost certainly the point with the large negative residual at the bottom of the group of* 4 *observations in the left side of the main residual plot. Following the removal of* 1976, *observation* 23 *now refers to* 1983, *which was another year with a relatively big population. Both of these points also probably have relatively high leverage, if we check some of the influence diagnostics.*

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)?

*I would be very reluctant to remove observation* 16 *(as well as observation* 17*) as we already have a small sample size and I suspect removal of any observation would just promote another observation to potential problem status (observation* 23*).*

*The data do appear to divide into two distinct groups (usual years and big population years) and it may not be reasonable to fit the same model to both groups. We could try modelling the two groups separately, by including an indicator variable in the model, but we have only a relatively small number of observations in the big population group, depending on how we define "big".*

**(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?

*The estimated growth rate is now r* = 0.5400918 *and the population is expected to increase each year by a multiplicative factor of* $e^r$ = exp(0.5400918) ≈ 1.72, *compared to* 1.41, *under the previous model.*

*On the other hand, the estimated carrying capacity has decreased from* 845.91 *down to K* = –0.5400918/–0.0007721 ≈ 699.51.

*Using the observed value of* $N_t$ *in* 1986 *as the value of* $n_t$*, the predicted value is:*

$$\hat{y}_t = 0.5400918 - (0.0007721)(94) \approx 0.4675$$

*A* 95% *prediction interval for this estimate is:*

$$\hat{y}_\mathrm{t} \pm t_{23}(0.975)\cdot s\cdot\sqrt{1+\frac{1}{n}+\frac{(n_t-\bar{n}_t)^2}{(n-1)s_{n_t}^2}}$$

$$= 0.4675 \pm (2.0687)(1.074)\sqrt{1+\frac{1}{25}+\frac{(94-586.52)^2}{(24)358{,}119}} = (-1.8280, 2.7631)$$

*The estimated population for* 1987*:*

$$\hat{y}_{1986} = \log\left(\frac{N_{1987}}{N_{1986}}\right) \approx 0.4675 \;\; \textit{so}\; N_{1987} = N_{1986}e^{0.4675} \approx 150$$

*With 95% prediction interval:*

$$(N_{1986}e^{-1.8280} \approx 15,\; N_{1986}e^{2.7631} \approx 1{,}490)$$

*Which is not a very precise prediction (a multiplicative factor of around* 10 *either side of the estimate).*

**(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.]

*There are no obvious problems with the assumptions of independence, constant variance and normality in the residual plots in part (c), so whilst the model is arguably an appropriate model for most of the data, there is definitely an issue with the group of observations with big populations identified in the discussion in part (d). This issue means the model is far from being an adequate description of the data, as evidenced by the highly imprecise prediction in part (e).*

*It is very hard to believe that just one model of this type can actually cover both of the groups in the data (and* 1976*, which is arguably yet another group of size* 1*). The usual years may well follow a Ricker model of regular growth, followed by a collapse once the population exceeds some optimal carrying capacity. However, there is definitely something different going on in the big population years, when the population appears to explode well beyond the supposed carrying capacity threshold. Include an indicator variable for type of year (big population or usual) may be enough to allow for different models for the two groups, but there may also be some environmental covariates we could measure (e.g. temperature, rainfall) that might suggest why the carrying capacity seems to suddenly increase in certain years.*

**(3 marks)**

**END OF EXAMINATION**