Australian National University

Venue _____

STUDENT NUMBER

U							

Research School of Finance, Actuarial Studies and Statistics

PRACTICE FINAL EXAMINATION

Questions updated from previous exam papers in 2016

STAT3015/STAT4030/STAT7030 Generalised Linear Models

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

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

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-five (25) pages and there is a separate handout of R output which has a total of twenty-one (21) 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 collected at the end of the examination and destroyed, they will not be marked.
3. There are a total of six questions, worth a total of 65 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 either Q1 or Q1A, and each and every part of the other four questions.** This examination counts towards 65% of your final assessment.
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 20 and 21 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	Q1A	Q2	Q3	Q4	Q5	Total
Pages	2 to 7	8 to 10	11 to 15	16 to 18	19 to 22	23 to 25	
Marks	**13**	**13**	**13**	**13**	**13**	**13**	**65**
Score							

Question 1 **(13 marks)**

Ryan, B.F., Joiner, B.L. and Ryan, T.A. in their text, *MINITAB Handbook* (PWS-Kent: Boston, 1985, p.206), report an experiment conducted to assess the effect of the use of a supplement and the amount of whey on the quality of pancakes. Eight different batches of pancake mixture were made to different recipes both including and not including the **supplement** and using four levels of **whey** (0%, 10%, 20% and 30%). Three pancakes were baked from each batch and each pancake was given a **rating** by an expert, with higher ratings indicating better quality pancakes.

(a) Given the above description of the research question, what are the response and the explanatory variables for this experiment? What are the experimental treatments? How many replications were there for each treatment?

*The response variable consists of the **rating**s given by the expert.*

The explanatory variables are:
***supplement**, a factor variable with two levels (yes = included or no = not included); and*
*the level of **whey** (0%, 10%, 20% or 30%).*

The experimental treatments are the eight possible combinations of supplement and whey (the 8 different "recipes"):
no supplement with 0% whey; no/10%; no/20%; no/30%;
yes/0%; yes/10%; yes/20% and yes/30%.

There were 3 replications of each treatment, i.e. 3 pancakes per recipe.

Note: this is actually an example of pseudo-replication, as the experimental units are actually the different batches of each recipe and the pancakes are only sampling units, 3 from each batch, however, these concepts (pseudo-replication, experimental units and sampling units) are outside the scope of this unit (see either a good text on experimental design or try the appropriate statistics unit).

(2 marks)

Question 1 continued

R was used to fit an initial linear model `pancake.lm1` to the pancake data and the residual plot for this initial model is shown on page 1 of the `R` output.

(b) What problem with the initial model does the above residual plot suggest? What possible solution(s) could you investigate to fix this problem?

Heteroscedasticity: the variance is obviously not constant over the eight treatments, which violates one of the assumptions underlying the model.

One approach to correct non constant variance is to apply a transformation to the response variable, which is probably the best solution in this instance. Alternatively we could change the approach to modelling the error structure by using either a weighted least squares approach or possibly a generalised linear model with a non-normal error distribution.

(1 mark)

Question 1 continued

After examining the residual plot for the initial model, a monotonically increasing transformation was applied to the pancake rating values and a final linear model was chosen for the transformed data. Pages 2 to 4 of the R output shows some residual plots and summary output for this model (pancake.lm):

(c) Do the residuals plots suggest any problems with the underlying assumptions?

Are there any problem(s) shown on the "Residuals vs Fitted" plot on page 2?
If so describe the problem(s):

The main residual plot is definitely better for the new model. The groups with the larger fitted values have slightly larger variance than the other groups, but definitely not to the same extent as shown on the main residual plot for the earlier model.

Are there any problem(s) shown on the "Normal Q-Q" plot on page 2?
If so describe the problem(s):

The normal quantile plot also shows no obvious violations of the underlying assumptions. There is only one observations with a standardised residual value outside the range (-2, +2) *which is only* $1/24 \approx 4.2\%$ *of the data.*

Are there any problem(s) shown on the "Cook's distance" plot on page 2?
If so describe the problem(s):

The Cook's distance plot shows no real problems with outliers. The values of Cook's D are all reasonably small and no observations really stand out as being extreme relative to the other observations.

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 1 continued

(d) Discuss the following features of the model pancake.lm:

Would the model pancake.lm be best described as an ANOVA model or an ANCOVA model? Which of the explanatory variables have been treated as a factor and which, if any, are covariates?

Whey is fitted as a factor variable, so the chosen model is an analysis of variance (ANOVA) model involving two factors (supplement and whey) and an interaction between the factors. The model is therefore a two–way ANOVA with interaction.

Note: If whey had been fitted as a continuous covariate, then we would have the equivalent of two simple linear regression models, one for the observations where no supplement was used and another for the observations where the supplement was included. This would have been an example of an analysis of covariance (ANCOVA) model.

The ANCOVA model assumes a linear relationship between the mean levels of whey for both levels of supplement (yes and no) – judging by the second of the two factor plots on page 4 *of the* R *output, this assumption is questionable.*

Specify the algebraic formula for this model. Clearly indicate how each of the variables have been included in the model, including any constraints applied to the parameters and the assumptions regarding the error distribution.

$\ln\left(\boldsymbol{rating}\right)_{ijk} = \beta_0 + \tau_j + \delta_k + \gamma_{jk} + \varepsilon_{ijk}$ *assumptions* ε_{ij} *i.i.d.* $N\left(0, \sigma^2\right)$

where j *indicates the levels of* ***supplement*** {0 =*"no"*, 1=*"yes"*},
k = 1,2,3,4 *indicates the level of* ***whey*** {0%, 10%, 20%, 30% },
i = 1, 2, 3 *for all j, k, and*
constraints $\tau_0 = 0$, $\delta_1 = \delta_{0\%} = 0$, $\gamma_{jk} = 0$ *for all* $j = 0$ *or* $k = 1$

(3 marks – 1 for the first part and 2 for the second part)

Question 1 continued

(e) Which of the explanatory variables in the model `pancake.lm` have a significant effect on the response variable? Which statistics in the `R` output are important in this question?

*The interaction term involving both supplement and whey is significant (in terms of effect on response variable), judging by the large F statistic (*41.38426*) and the very small associated p-value (*0.0000000913*) in the analysis of variance table, that is considerably less than the usual significance level of* 0.05*).*

This interaction is the "main story", as it indicates that particular combinations of supplement and whey produce significantly different values of the transformed ratings. However, the analysis of variance table also shows that the "main effect" terms for both supplement (F = 17.01389*, p-value <<* 0.05*) and whey (F =* 74.34722*, p-value <<* 0.05*) are also significant. As the interaction term is significant and involves both variables, we would not refine the model by removing these main effect terms, even if they were not significant.*

The large t statistics with small p-values in the table of coefficients also indicate that there are some significant differences between treatments (combinations of the levels of these factors), but if you want to tell a story about variables rather than differences between the levels within variables, then this is better done using the ANOVA table.

(1 mark)

(f) Which hypotheses can be tested using the F-statistic shown at the end of the summary output and what do you conclude as a result?

Overall F statistic ($F_{7,16}$ = 52.03*; p =* 0.000000000794*) is a test of:*

H_0 : *all* τ_j, δ_k & $\gamma_{jk} = 0$ *against* H_A : *not all* τ_j, δ_k & $\gamma_{jk} = 0$

Reject H_0 *in favour of* H_A *and conclude that at least one of the terms in the model involving the explanatory variables has a significant effect on the response.*

(1 mark)

Question 1 continued

(g) In the model pancake.lm, which combination of **supplement** and **whey** appears to produce the pancakes with the highest **rating**? Is this combination significantly better than all other combinations of supplement and whey?

Judging by the table of means, the recipe that produces pancakes with the best mean transformed rating includes the supplement and 30% whey. As the transformation applied to the ratings was monotonically increasing (a log transformation), this combination is also the one that produces the pancakes with the best mean untransformed rating.

To decide whether the apparent differences between any two of the mean transformed ratings shown in the tables of means are significant, we could calculate a 95% *confidence interval for the differences between these treatment means using the formula:*

$$(\bar{y}_1 - \bar{y}_2) \pm t_{residual\ df}(0.975)\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where $\bar{y}_1$ & $\bar{y}_2$ *represent the two treatment means,*

$t_{residual\ df}(0.975) = t_{16}(0.975) \approx 2.1199$ *from the tables,*

s^2 *= estimated error variance (MSE from the analysis of variance table),*

n_1 & n_2 *are the sample sizes for the two treatments.*

As there are 3 replications for each treatment $n_1 = n_2 = 3$ *regardless of which pair of treatments we are considering, so the width of the confidence interval is always:*

$$t_{residual\ df}(0.975)\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 2.1199\sqrt{0.03\left(\frac{1}{3}+\frac{1}{3}\right)} = 0.3$$

So, for any two treatment means that differ by more than 0.3*, a* 95% *confidence interval for the difference between the means will not include 0, so we can conclude (equivalent to doing a two-tailed hypothesis test), that the two treatments differ significantly.*

So, we can then answer the question that was asked by doing a single comparison (so there is no need to apply the Bonferroni correction for multiple comparisons), between the largest apparent mean of 5.433333 *(for yes/*30%*) and the next largest mean of* 5.033333 *(for yes/*20%*). The difference is* 0.4*, which is greater than* 0.3*, so we conclude that the yes/*30% *treatment is significantly better than the yes/*20% *treatment and therefore is significantly better than all other treatments.*

Note: In practice we probably would have planned a series of comparisons (and should therefore make the appropriate Bonferroni correction) to rank all of the possible recipes.

(3 marks)

Question 1A (mixed effects model) **(13 marks)**

Cook R.D. and Weisberg, S. in their text, *Applied Regression including Computing and Graphics*, (Wiley, New York, 1999, problem 16.8 on page 395), report an experiment conducted to compare the rubber yield of seven varieties of guayule, a desert shrub that contains rubber. The experimental area consisted of 35 plots arranged in a 5 x 7 grid; the rows of the grid forming 5 randomized complete **Block**s, each containing 1 plot of each **Variety**. The **Yield** is the total grams **P1** + **P2** of rubber for two selected plants from each plot.

Pages 5 to 7 of the R output show the results of fitting a linear mixed effects model (rubber.lme) to these data.

(a) What are the response and the explanatory variables for this experiment? For each explanatory variable, describe whether it should be treated as a fixed or random effect. How many experimental treatments are there? How many replications are there of each treatment?

*In this model, **Yield** (the sum of **P1** and **P2**) is the response variable, the explanatory variables include a fixed factor, **Variety** and a random effect, **Block**. The experimental treatments are the seven different varieties of guayule, i.e. the 7 levels of the factor variable **Variety**. The experiment is designed to compare the **Yield**s for these 7 different treatments and the experiment is replicated 5 times, once within each **Block**.*

*So the **Block**s represent the 5 replications of each of the 7 treatments and are a random source of variation in the experiment, i.e. the experimenters want to control for any differences between the **Block**s when making the main comparison of interest in this experiment, which is how the **Yield**s differ for the different varieties of guayule.*

(2 marks)

(b) Specify the algebraic formula for this model. Clearly indicate how each of the variables have been included in the model, including any constraints applied to the parameters and the assumptions regarding the error distribution.

The underlying population model is: $Y_{ijk} = \mu + \eta_j + \lambda_k + \varepsilon_{ijk}$

*where Y is the **Yield** (the response variable = **P1** + **P2**),*

*j indicates the levels of the explanatory factor **Variety** = {1, 2, 3, 4, 5, 6, 7},*
*k indicates the different **Block**s = {1, 2, 3, 4, 5} and*
there is only i = 1 observation for each combination of j and k;
for a total of 1 x 7 x 5 = 35 *observations.*

*The constraint applied to the factor **Variety** is* $\sum_j \eta_j = 0$.

*We are assuming that the **Block**ing term* λ_k *consists of a series of random effects which are independently and identically distributed N(0,* σ_λ^2*), but which are also independent of the residual variation in the model (the errors), which are independently and identically distributed N(0,* σ_ε^2*).*

(3 marks)

Question 1A continued

(c) A residual plot for the model rubber.lme is shown on page 6 of the R output. Does the residual plot suggest there are any problems with the fitted model? What other diagnostics could you examine to decide whether or not there really is a problem with the model?

The plot looks a little strange as there is a gap in the fitted values; however, the important aspect is the distribution of the residuals in the vertical direction. In the vertical direction there are no obvious patterns indicating that the errors might not be independent and the variance appears to be reasonably constant, so there are no obvious problems with the underlying assumptions.

The residuals have also been standardised and there is only one residual lying more than 2 standard deviations away from the mean, and even that one is reasonably close to the other observations, so there are no apparent outliers.

It would also be nice to produce a normal quantile plot to assess the assumption of normality of the errors and also a plot of Cook's distances (or a similar "outlier" plot, such as a leverage bar plot) that we could use to assess the influence of particular observations in determining the fit of the model.

(2 marks)

(d) Does the rubber **Yield** vary depending on the **Variety** of guayule? Present an appropriate hypothesis test to support your conclusion.

*In the ANOVA table, the small p-value ($F_{6,24} = 3.6612$, $p = 0.0101$) associated with the **factor(Variety)** term is smaller than $\alpha = 0.05$, which indicates that there are significant differences in the **Yield**s between the different varieties of guayale, i.e. we should reject the following null hypothesis in favour of the alternative:*

$H_0: \quad \eta_1 = \eta_2 = \eta_3 = \eta_4 = \eta_5 = \eta_6 = \eta_7 = 0 \quad vs$

$H_A: \quad not\ all\ \eta = 0 \ \ (at\ least\ one\ \eta \neq 0)$

As the model was fitted using sum contrasts, the η's represent deviations ($\mu_j - \mu$) from the overall mean, μ, so the above hypotheses are equivalent to testing:

$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu \quad vs$

$H_A: \quad at\ least\ one\ \mu_j\ differs\ from\ \mu$

(1 mark)

Question 1A continued

(e) Rank the seven varieties in order from largest to smallest average **Yield**. Which of the varieties had average **Yield** that differed significantly from the overall average?

*The summary table includes coefficients for six of the seven varieties. As the model was fitted using a sum constraint ($\eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5 + \eta_6 + \eta_7 = 0$), the coefficient for the missing **Variety** (η_7) can be found using:*

$$\eta_7 = -(\eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5 + \eta_6)$$
$$= -(-0.530571 + 1.067429 + \cdots - 4.420571) = 0.6534286$$

So, varieties 4, 2, 7, 5 *and* 3 *(in descending order) all produce average yields above the overall mean, whilst varieties* 1 *and* 6 *(again in descending order) produce average yields below the overall mean. Judging by the p-values in the table of coefficients, only varieties* 4 *and* 6 *differ significantly from the overall mean – variety* 7 *lies between varieties* 2 *and* 5 *and as all the standard errors associated the η coefficients are the same, variety* 7 *will have a p-value somewhere between* 0.3199 *and* 0.7581*, which is definitely greater than* $\alpha = 0.05$.

Assuming the aim of the experiment is to determine the variety that produces the best yield, then it appears to be variety 4*; the only variety to have an average yield significantly above the overall average ($\eta_4 = \mu_4 - \mu = 2.6$, $t_{24} = 2.47$, $p = 0.0212$).*

(3 marks)

(f) Has blocking been effect in this instance? What proportion of the overall variability is due to variation between **Block**s?

From the random effects section of the summary output:

$\hat{\sigma}_\lambda = 0.5886959$ *is the estimated standard deviation between **Block**s, and*

$\hat{\sigma}_\varepsilon = 2.538402$ *is the estimated residual (within **Block**s) standard deviation.*

So $\hat{\sigma}_\lambda^2 \big/ \left(\hat{\sigma}_\lambda^2 + \hat{\sigma}_\varepsilon^2\right) = 0.5886959^2 \big/ \left(0.5886959^2 + 2.538402^2\right) \approx 5.1\%$

*So, variation between **Block**s is only a small proportion (*5%*) of the overall variability, but this is still enough to make accounting for differences between **Block**s an important part of the experimental design.*

(2 marks)

Question 2 **(13 marks)**

A survey of attitudes towards the introduction of a non means-tested age pension resulted in the following data:

	Age of Respondent				
	20	30	40	50	60
Sample Size	12	12	12	12	12
Number in favour	4	6	9	11	12

Pages 5 to 7 of the R output some the results of fitting a series of logisitic regression models to the above data, with the proportion in favour of a non means-tested pension as the response variable and treating age as a continuous explanatory variable.

(a) Use the model pension.glm, summary output for which is shown on page 8 of the R output, to estimate the age at which 50% of respondents are in favour of a non means-tested age pension.

The fitted model is: E[logit(prptn)] = –3.1576993 + 0.1112453 *age*

When the proportion in favour is 50%, logit(prptn) = 0.

Solving the above equation for age:

age = –(–3.1576993) / 0.1112453 = 28.385 *or just over* 28 *years.*

(1 mark)

(b) Is age a significant predictor of the response in the model pension.glm? Which of the above statistics in the R output are important in answering this question?

*The t statistic associated with the coefficient of age in the table of coefficients (*3.596698*) is larger than the appropriate critical value for a t with* 3 *degrees of freedom (the residual degrees of freedom),* $t_3(0.975) = 3.1824$ *indicating that we should reject the null hypothesis that the coefficient is equal to* 0 *and conclude that the term involving age is a significant part of the model, i.e. age is a significant predictor of the response variable and the proportion in favour of a non-means tested age pension (as logit is an order preserving transformation).*

*The same conclusion could also be based on the change in deviance associated with age (*19.98251*) in the analysis of deviance table. This change in deviance has an approximate chi-square distribution with* 1 *degree of freedom, with associated p-value* 0.000008 *(less than our usual significance of* 0.05*), again indicating the significance of the term involving age.*

Note that we can interpret the unscaled change of deviance in this fashion as we are fitting a binomial model where the dispersion factor is assumed to be 1 *and the scaled deviance will be the same as the unscaled deviance divided by* 1.

(2 marks)

Question 2 continued

Residual diagnostic plots for the model pension.glm are given on page 9 of the R output.

(c) Do the residuals plots suggest any problems with the underlying assumptions?

Are there any problem(s) shown on the "Residuals vs Fitted" plot on page 9?
If so describe the problem(s):

The plot displays definite curvature, i.e. a violation of the assumption of independent errors. We could try adding a quadratic term in age to the model (see the later parts of the question) or possibly some "linearising" transformation.

Are there any problem(s) shown on the "Normal Q-Q" plot on page 9?
If so describe the problem(s):

No obvious problems, the residuals look about as "normal" as you could reasonably expect with just 5 *observations.*

Are there any problem(s) shown on the "Cook's distance" plot on page 9?
If so describe the problem(s):

Observation 1 *has a relatively high Cook's D value compared to the other observations, but the real problem here is the lack of independence noted in conjunction with the main residual plot.*

I probably would not have produced and examined any outlier plot, until I had done something to try and fix the obvious problem with the main residual plot.

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

- ☒ Residuals are not independent (obvious pattern)
- ☐ Residuals do not display constant deviance/dispersion
- ☐ 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)

(d) Is there evidence of under-dispersion or over-dispersion with the model pension.glm?

In fitting the model, R *assumed a value of* 1 *for the dispersion parameter. We can estimate the dispersion separately by dividing the residual deviance* (1.0292) *by the residual degrees of freedom* (3) *to get* 0.34*, which is considerably less than* 1*, suggesting under-dispersion and hence a potential problem with the model.*

Note: The question didn't ask if the under or over-dispersion was ***significant****.*
If it had, then applying the rule of thumb wouldn't work in this instance: under that rule, we have under-dispersion if 0.34 *is less than* 1 – 3 * *sqrt*(2/3) = –1.45*,*
which is not possible, as the estimated dispersion cannot be less than 0*.*

As discussed in lectures, a much better approach is always to use the scaled residual deviance (scaled by the assumed dispersion, which is 1 *in this instance), which should have an approximate chi-square distribution with the residual degrees of freedom. In this instance* 1.0292 *lies within the interval:*
$\left(\chi^2_3(0.025) = 0.2158, \chi^2_3(0.975) = 9.3484\right)$

So, there is no evidence of significant under-dispersion, but the problems with the main residual plot alone are enough to lead us to questions if this model is really an appropriate model for this small sample of data.

(2 marks)

(e) Find a 95% confidence interval for the coefficient of age in the above model. What does this confidence interval imply about the relationship between the proportion in favour and age?

The coefficient of age in the table of coefficients is positive, suggesting that the proportion in favour increases as age increases.

95% *confidence interval:*
coefficient of age ± $t_3(0.975)$ × *standard error(coefficient of age)*
= 0.1112453 ± (3.1824)(0.03092984) = 0.11125 ± 0.09843
= (0.01282, 0.20968)

The above confidence interval is for the expected increase in the logit (log odds) of the proportion in favour of a non-means tested age pension as age increases by 1 year. As the entire confidence interval is greater than 0, this indicates that the odds of being in favour increases with age, i.e. the older the age of the person, the more likely they are to be in favour of a non means-tested age pension.

However, the problems with the model discussed in the earlier parts of this question cast considerable doubt on the reliability of this inference.

(2 marks)

Question 2 continued

(f) One suggestion for improving the model pension.glm is to try including a quadratic term in age in the model. This results in the model pension.glm1, summary output for which is shown on page 10 of the R output. Does this output suggest that this model is a significant improvement on the earlier model?

The two statistics associated with the additional term in the model both show that the term is not a significant improvement. The t (z) statistic in the table of coefficients is only 0.8020842*, which is much smaller than* $t_2(0.975) = 4.3027$ *and the change in deviance is only* 0.75209 *with a p-value* (0.3858143) *greater than the significance level of* 0.05.

Note you could also do this comparison using the standard "normal" p-value associated with the z statistic (on the grounds that the key deviance parameter here, the dispersion, is a known constant = 1*). This is why the current version of R produces output labelled z, rather the older versions of S-Plus, which used an ambiguous label and didn't show p-values (leaving the decision of whether to use z or t tables up to the reader). I think there are other sources of uncertainty at play here and that making comparisons using a more conservative Student's t cut-off is justified, especially in a small sample situation.*

Given the added quadratic term is not significant and we still have qualms about the residual plots (see the next part of the question), I do not feel this model is a significant improvement on the earlier model. We need to investigate other possible solutions (changing the link function, transformations to the explanatory variables), but it is distinctly possible that none of these will result in a better model, especially given the small number of observations.

(2 marks)

Question 2 continued

Residual diagnostic plots for the model pension.glm1 are given on page 11 of the R output.

(g) Do the residuals 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):

The plot no longer displays definite curvature, but there are still some signs of a pattern (though it is very hard to tell with so few observations).

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

Even given the relatively small number of observations, this new normal quantile plot looks distinctly "non-normal"!

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

Observation 1 *still has a smaller, but still relatively high Cook's D value compared to the other observations.*

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

- ☐ Residuals are not independent (obvious pattern)
- ☐ Residuals do not display constant deviance/dispersion
- ☐ 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 3 **(14 marks)**

Myers, R.H., Mongomery, D.C. and Vining, G.G. (2002) *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley: New York, which was one of the texts in the recommended reading list for this course, includes as an exercise (4.10, p.154) some data collected for a student project. The student was examining the impact of popping temperature, amount of oil and the popping time on counts of the number of inedible kernels of popcorn.

The data for this experiment is shown on pages 12 of the R output, followed by summary output and graphs for a series of models.

(a) There are some obvious problems with the residual plot for the model popcorn.glm shown on page 13 of the R output. Identify the data points that are causing these problems. What can be done to further investigate these problems?

The two points that cause most concern are one in the far right of the residual plot that is probably highly influential and a possible vertical outlier towards the top of the residual plot. The influential point is observation #7 with a linear predictor value of 4.719304 *and a standardised residual of* 3.26162695. *The potential outlier is observation #2 with a linear predictor of* 2.564289 *and a standardised residual of* 4.36158572.

We could further investigate how influential these points were in determining the fit of the model by calculating their leverage values, examining the deletion residuals, calculating the value of Cooks' D for all the observations and examining the change in the model coefficients if we excluding one or both of the suspect observations.

If the suspect points:

have relatively high leverage (i.e. greater than $2p/n = (2 \times 7)/15 = 0.933$*);*
cause marked changes between the usual residual plot and a deleted residual plot;
have Cooks' D values considerably larger than those of the other observations; or
cause a substantial change in the estimated model coefficients when excluded,

then we would consider them to be discordant observations and treat them accordingly, by identifying, where possible what causes them to behave differently from the rest of the data and either correcting them (if they are in error), or by making an appropriate modification to the model.

However, these two points are not the only observations with large standardised residuals and the fact that we have 4/15 *or* 27% *of the observations lying more than* 2 *standard deviations away from the model, suggests we may not yet have the right model. We could try switching to a stronger link function, before we go down the line of declaring observations to be outliers.*

(3 marks)

Question 3 continued

(b) Summary output for the model popcorn.glm is shown on pages 13 and 14 of the R output. Apart from the 0 degrees of freedom associated with the three-way interaction term suggesting that there is insufficient data to sensibly fit such a term, what does the ANOVA table and the table of coefficients suggest are the significant terms in this initial model?

As it does not appear possible to fit the three-way interaction term, we examine the two-way interaction terms and find that the interaction between temp and time is significant with a change in deviance of 77.16164 *with a p-value less than* 0.05 *in the analysis of deviance table and a t value of* 8.874033 *in the table of coefficients, which is considerably greater than* $t_8(0.975) = 2.3060$.

The interaction between oil and time is also significant (change in deviance = 16.44048, *t value =* –4.064890, *p-value <* 0.05*), however, the interaction between temp and oil is not significant (change in deviance =* 3.06750, *t value =* –1.810598, *p-value >* 0.05*).*

(3 marks)

(c) Is there evidence of significant over-dispersion? What does this suggest about the adequacy of this model?

Divide the residual deviance (36.71496) *by the residual degrees of freedom* (8) *to estimate the dispersion parameter. This gives* 4.58937 > 1 + 3 * sqrt(2/8) = 2.5, *indicating over-dispersion.*

Better yet, instead of using the "rule of thumb", conduct a formal hypothesis test:

$H_0 : \phi = 1 \quad H_A : \phi \neq 1$

$\left(\chi^2_8(0.025) = 2.1797, \chi^2_8(0.975) = 17.5345 \right)$

The scaled residual deviance (36.715) *is not in the above interval, indicating significant over-dispersion and questioning the adequacy of the model.*

(3 marks)

Question 3 continued

(d) Summary output for another model popcorn.glm1 is shown at the bottom of page 14 of the R output. This is an attempt to refine the initial model. Is this model new an adequate one for the data? Would you suggest any further refinements to the model?

The three-way interaction term and the non-significant interaction between temp and oil, identified in part (b) have been deleted. It is unlikely that these modifications will have solved the problems with the residual plot from part (a) and the over-dispersion is still significant, the residual deviance (39.9847) *is still above the interval:*

$\left(\chi^2_9(0.025) = 2.7004, \chi^2_9(0.975) = 19.0228\right)$

The analysis of deviance table confirms that the remaining interactions (temp:time and oil:time) are still significant, and as these significant interactions involve all three of the main effects, all three should generally be included in the model, regardless of their significance (note that the table of coefficients and the analysis of deviance table for the original model also make conflicting suggestions about the significance of the main effect for oil).

So, making "further refinements" by deleting more terms from the model is not the solution in this instance and we need to investigate other approaches (transformations etc.) for correcting the problems with the model.

(4 marks)

Question 4 **(13 marks)**

A paper by R.E. Chapman, titled "Degradation study of a Photographic Develop to Determine Shelf Life" (*Quality Engineering* (10), 1997-98, pp.137-140) presents the results of an experiment designed to study the relationship between the shelf life and the density of a photographic developer. Density is considered a good indicator of overall developer performance.

In the experiment, the shelf life of 21 batches of the photographic developer were accelerated by subjecting each batch to a set number of hours at a high temperature and the resulting maximum density of each batch was recorded:

								temperature
life time (hours)	72	144	216	288	360	432	504	72°C
maximum density	3.55	3.27	2.89	2.55	2.34	2.14	1.77	
life time (hours)	48	96	144	192	240	288	336	82°C
maximum density	3.52	3.35	2.50	2.10	1.90	1.47	1.19	
life time (hours)	24	48	72	96	120	144	168	92°C
maximum density	3.46	2.91	2.27	1.49	1.20	1.04	0.65	

When the experiment was designed, it was suggested that the life times might follow an exponential distribution, as might the resulting maximum densities. Based on this suggestion R was used to fit the model photo.glm. Summary output for this model is shown on page 15 of the R output; page 16 shows a plot of the data with the fitted values from the model superimposed; and page 17 shows a residual plot for the model.

(a) For a Gamma GLM, there are two simple ways to estimate the dispersion factor. Use the summary output for the model photo.glm to find these two estimates and compare them. Is there any evidence of under or over-dispersion?

R *uses the Coefficient of Variation (CV) estimate of the dispersion parameter for a Gamma GLM – this is the value of* 0.00379838 *given in the output on page* 15.

The alternative approach is to divide the residual deviance (0.056144) *by the residual degrees of freedom* (15)*, giving an estimate of* 0.003742933. *As the two estimates are close, there is no evidence of either under or over-dispersion.*

Again, a better approach is to do a formal hypothesis test:

$H_0 : \phi = \phi_{CV} \quad H_A : \phi \neq \phi_{CV}$

$\left(\chi^2_{15}(0.025) = 6.2621, \quad \chi^2_{15}(0.975) = 27.4884 \right)$

The scaled residual deviance (0.056144/0.00379838 = 14.78) *is in the above interval, indicating no significant evidence of either under or over-dispersion.*

(2 marks)

(b) How does fitting a Gamma GLM help to implement the suggestion that the life times and the maximum densities follow an exponential distribution? Do the estimated dispersion factors from part (a) support the suggestion that the error distribution is exponential?

An exponential distribution is the special case of the Gamma (α, β) distribution, with the first parameter α equal to 1.

The dispersion parameter for a GLM with an error distribution from the Gamma family should be equal to $1/\alpha$*, so in the case of the exponential distribution the dispersion should be* 1. *The two estimates of the dispersion parameter in part (a) consistently suggest that* α *is considerably larger than* 1 *(either* $1/0.0037984 = 263$ *or* $1/0.0037429 = 267$*), suggesting that the error distribution is better modelled by a more general Gamma distribution rather than the exponential distribution.*

(2 marks)

(c) Does the analysis in parts (a) and (b) and the plots shown on pages 16 and 17 of the R output suggest that the model is an appropriate one for the data?

The first plot shows a reasonably good fit to the data and the residual plot shows no obvious problems with any of the usual underlying assumptions. This together with the consistency of the two estimates of the dispersion parameter in part (a) suggests this model is appropriate (but definitely not exponentially distributed).

(2 marks)

Question 4 continued

(d) Based on the table of coefficients for the model, which terms in the model are significant predictors of the maximum density (and therefore of the overall performance of the developer)?

In the model, temperature has been fitted as a discrete factor, rather than as a continuous covariate. As shown in the plot on page 16 *of the* R *output (which shows the data with the model superimposed), this produces a model similar to an ANCOVA model, consisting of three separate log regression models, one for each level of temperature, between the response variable (density) and the continuous covariate (lifetime). The model involves a log link function, so if the plot had shown log(density) rather than density, the three models would have appeared as three straight line regression models.*

The same questions apply to this model as do to a standard ANCOVA model: do we need a single model with a common intercept and common slope for the three levels of temperature; or do we need parallel models with the same slope, but different intercepts; or do we need three completely separate models with different intercepts and different slopes?

We examine the t values in the table of coefficients – a coefficient with a t value greater in absolute value than $t_{15}(0.975) = 2.1314$ *will indicate that the corresponding term is a significant part of the model.*

The coefficient of lifetime has a large t value (–9.64)*, suggesting lifetimes do have a significant effect on log(density). This base slope coefficient is negative and as can be seen from the plot, longer lifetimes are associated with lower maximum densities at all three levels of temperature.*

The t value for the second of the two temperature coefficients (2.439) *is larger than* 2.1314 *suggesting a significant difference in the intercepts for the highest temperature* (92°C) *and the base or reference temperature* (72°C)*, however, the difference in the intercepts between the middle temperature* (82°C) *and the base temperature is not significant* (t = 1.438).

The significant t values for the two interaction coefficients (–7.877 *and* –19.344) *suggest that we do need to modify the base slope for the different levels of temperature, in short, the chosen model should allow for different slopes and different intercepts for the different levels of temperature.*

(4 marks)

Question 4 continued

(e) Based on the analysis of deviance table, which of the explanatory variables are significant predictors of the maximum density? Are these conclusions consistent with your answers to part (d)?

As parts (a) and (b) have demonstrated that the dispersion parameter for this Gamma GLM is not equal to 1*, before we can correctly interpret the change in deviances shown in the analysis of deviance table, we need to scale them by dividing by the estimated dispersion. Using the CV estimate (though we could use either estimate of the dispersion from part (a) as it would make very little difference in this instance):*

For the term involving lifetime: 0.45327/0.00379838 = 119.3318 *is greater than* qchisq(0.95, 1) = 3.8415*, indicating that there is a significant linear relationship between log(density) and lifetime.*

For the term involving temperature: 1.81813/0.00379838 = 478.6593 *is greater than* qchisq(0.95, 2) = 5.9915*, indicating that this term is significant in the model (and that at least one of the associated coefficients is non-zero). This implies that there are significant differences between the intercepts (of the above linear relationship between log(density) and lifetime) for some of the three levels of temperature. In part (d), the fact that not all of the t statistics for the coefficients associated with this term were significant suggests that not all three of the intercepts differ significantly, but at least one of them differs significantly for the others.*

For the interaction term between lifetime and temperature: 1.45322/0.00379838 = 382.5894 *is greater than* qchisq(0.95, 2) = 5.9915*, indicating that this term is significant in the model. This implies that there are significant differences between some of the slopes for the three levels of temperature.*

The results are consistent with the results of part (d).

(3 marks)

Question 5 **(13 marks)**

568 male and female residents in Australian metropolitan and rural areas were surveyed about their preference for locally manufactured or imported motor vehicles, with the following results:

Car Preference	Males		Females	
	City Residents	Country Residents	City Residents	Country Residents
Imported	68	12	16	24
Local	168	32	84	164

Pages 18 and 19 of the R output show some analyses of these data.

(a) Use the observed and expected values shown on page 18 of the R output to find, for males and females separately, the Pearson χ^2 statistics for tests of independence between car preference and residence. Do these statistics suggest that there is a significant association between car preference and residence for either males or females?

Formula for the Pearson χ^2: $\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Males:

$$\frac{(68-67.42857)^2}{67.42857} + \frac{(12-12.57143)^2}{12.57143} + \frac{(168-168.57143)^2}{168.57143} + \frac{(32-31.42857)^2}{31.42857}$$

Females:

$$\frac{(16-13.88889)^2}{13.88889} + \frac{(24-26.11111)^2}{26.11111} + \frac{(84-86.11111)^2}{86.11111} + \frac{(164-161.88889)^2}{161.88889}$$

The result for males is 0.043 *and for females* 0.571. *Both of these are based on* 2×2 *tables, so the degrees of freedom are* (*number of rows* – 1) × (*number of columns* – 1) = 1×1 = 1. *Both the results are less than* qchisq(0.95, 1) = 3.8415, *so we can't reject the null hypothesis of independence and therefore conclude there is no significant association between car preference and residence for either males or females.*

(3 marks)

Question 5 continued

(b) The top of page 19 of the R output shows the data table collapsed so that the effects of sex are ignored. Find the expected values for the collapsed data assuming that car preference and residence are independent and then find the deviance statistic for a test of independence (also called the likelihood ratio χ^2 statistic). Is there a significant association between car preference and residence ignoring the effects of sex?

Find row, column and overall totals and expected values for each of the cells:

	City Residents	*Country Residents*	*Total*
Imported	(120 × 336)/568 = 70.986	(120 × 232)/568 = 49.014	84 + 36 = 120
Local	(448 × 336)/568 = 265.014	(448 × 232)/568 = 182.986	252 + 196 = 448
Total	84 + 252 = 336	36 + 196 = 232	120 + 448 = 568

The Likelihood Ratio χ^2 for this 2×2 table is:

$$2\sum_i\sum_j O_{ij}\ln\left(\frac{O_{ij}}{E_{ij}}\right) =$$

$$2\left[84\ln\left(\frac{84}{70.986}\right)+36\ln\left(\frac{36}{49.014}\right)+252\ln\left(\frac{252}{265.014}\right)+196\ln\left(\frac{196}{182.986}\right)\right]$$

The result is 7.62 (*which is again approximately χ^2 with* 1 *degree of freedom*) *and as this is greater than* qchisq(0.95, 1) = 3.8415, *we reject the null hypothesis of independence and conclude that there is a significant association between car preference and residence, having ignored the effects of sex.*

(4 marks)

Question 5 continued

(c) Does the analysis on page 19 of the R output suggest that there are significant associations between sex and either car preference or residence? If so, what do these associations suggest about the relationships between sex and the other variables?

The Pearson χ^2 for the collapsed table of sex by car preference is large (18.36716) *and has a small p-value* (0.0000182171)*, which is less than the significance level of* 0.05*, so we conclude that there is a significant association between sex and car preference having ignored the effects of residence.*

Examining the discrepancies between the observed and expected counts, males are about twice as likely as females to prefer imported cars: 80 *out of* 280 *or* 28% *of males prefer imported cars, compared with only* 40 *out of* 288 *or* 14% *of females.*

Similarly, the Likelihood Ratio χ^2 for the collapsed table of sex by residence is also large (539.3575) *and has a small p-value* (0)*, again less than the significance level of 0.05, so we conclude that there is a significant association between sex and residence having ignored the effects of car preference.*

Again examining the discrepancies between the observed and expected counts, the sample is heavily skewed towards city males: 236 *out of* 280 *or* 84% *of males reside in the city, compared with only* 100 *out of* 288 *or* 35% *of females.*

(4 marks)

(d) The results of part (a) and (b) appear to contradict each other, though the results of part (c) may help to explain this apparent contradiction. Is this an example of Simpson's paradox and if so, how might the "paradox" be explained in this instance?

Yes, this is an example of Simpson's paradox, as at one level of aggregation, when we consider the two sexes separately as in part (a), there appears to be no association between car preference and residence, but if we aggregate over sex as in part (b), there appears to be an association between car preference and residence.

The results of part (c) help to explain this apparent contradiction, as they show that both car preference and residence are related to sex. So whilst there is no association between car preference and residence (accounting for sex), if we ignore sex, then the underlying associations between sex and both of these other two variables manifests as an apparent relationship between the two variables.

(2 marks)

END OF EXAMINATION