Venue

Student Number

Research School of Finance, Actuarial Studies & Statistics

EXAMINATION
Semester 2 - Final, 2020

STAT3015/STAT4030/STAT7030
GENERALISED LINEAR MODELLING

Writing Time: 2 hours
Reading Time: 15 minutes
Submission Time: 15 minutes
Total Time: 2 hours and 30 minutes

Exam Conditions:
This is an Open Book Exam, so any materials are permitted.
For the duration of the exam, no communication with other people is allowed. Any such communication will constitute a breach of ANU Academic Regulations.

Materials Permitted In The Exam Venue:
This is an Open Book Exam, so any materials are permitted.

Materials to Be Supplied To Students:
The exam paper will be available on Wattle in the assessment section.
It is recommended that you download the paper as soon as it becomes available.

Instructions to Students for the exam:

Attempt ALL questions

Each of the three questions caries equal marks

Start your solution to each question on a new page

To be a candidate for full marks show all steps in working out your solution and where appropriate briefly explain your reasoning. Marks may be deducted for failure to show appropriate calculations or formulae or for not providing any reasoning.

Instructions to Students for submission of scripts:

It is recommended that you write your solutions to the exam questions by hand.

At the end of the exam, you should transfer your solutions into a single pdf file in one of two ways:

either

scan your solutions into a single pdf file, if you have access to a scanner;

or

photograph your solutions, e.g. using a mobile phone, and save into a single pdf file.

Then upload this pdf file to Wattle.

You may also email the pdf file with your solutions directly to me as "insurance", but please be sure also to upload this file to Wattle.

The deadline for uploading the pdf file to Wattle containing your solutions is 150 minutes (two and a half hours) from the start of the exam.

1. We return to the productivity improvement data that was considered during the course. The dataset consisted of a measure of the productivity improvement of 27 business firms, where each firm was classifed according to whether their average expenditure for research and development in the past three years was high, moderate or low. Some R output is presented below.

```
> out1=lm(prodscre~RandD)
> summary(out1)

Call:
lm(formula = prodscre ~ RandD)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43333 -0.50556  0.06222  0.53333  1.32222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2000     0.3466  26.547  < 2e-16 ***
RandDLow     -2.3222     0.4217  -5.507 1.16e-05 ***
RandDMod     -1.0667     0.4000  -2.666   0.0135 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8001 on 24 degrees of freedom
Multiple R-squared:  0.5671,    Adjusted R-squared:  0.531
F-statistic: 15.72 on 2 and 24 DF,  p-value: 4.331e-05
```


(a) What type of model is being fitted here? Be brief but as precise as possible.

(b) Write down, in mathematical form, the model under consideration, using the "reference" parametrisation (i.e. the same parametrisation that is used by R). Briefly state the modelling assumptions that are being made.

(c) What are the degrees of freedom in each of the t-tests performed in the R summary output given above?

(d) Explain why, in this model, the (theoretical) variance of the estimator `(Intercept)` is equal to the (theoretical) covariance between the estimators `RandDLow` and `RandDMod`.

(e) There is a suspicion that there is no difference between low investment and moderate investment in the effect on productivity improvement. State the corresponding null hypothesis and then perform a suitable test of this hypothesis, expressing your result in the form of a p-value.

(f) In view of the p-value obtained in 1(e), what action if any would you consider taking concerning RandD? Briefly explain your thinking.

HINT FOR PART (d): *Consider carefully the implied R definitions of RandDLow and RandDMod, and hence decide which covariance to calculate.*

Solution to 1(a): *One-way analysis of variance.*

Solution to 1(b): *The fitted model may be written*

$$y_{ij} = \mu + \tau_j + \epsilon_{ij} = \mu_j + \epsilon_{ij}, \tag{1}$$

where $\tau_1 = 0$, and $j = 1$ corresponds to high investment (RandDHig), $j = 2$ corresponds to low investment (RandDLow) and $j = 3$ corresponds to moderate investment (RandDMod). The modelling assumptions are that the ϵ_{ij} are IID and that μ, τ_2 and τ_3 are fixed (i.e. non-random).

Solution to 1(c): *The degrees of freedom in each t-test is the the degrees of freedom in the residual sum of squares, which (in the notation of the lecture notes) is $n - g = 27 - 3 = 24$.*

Solution to 1(d): *Bearing in mind that R uses the reference parametrisation for the one-way ANOVA, the parameter estimate (Intercept) equals $\widehat{\mu}_1$, the sample mean of group 1 (high investment); the parameter estimate RandDLow equals $\widehat{\mu}_2 - \widehat{\mu}_1$, where $\widehat{\mu}_2$ equals the sample mean of group 2 (low investment); and the parameter estimate RandDMod equals $\widehat{\mu}_3 - \widehat{\mu}_1$, where $\widehat{\mu}_3$ equals the sample mean of group 3 (moderate investment). A key point is that under the one-way ANOVA model, $\widehat{\mu}_1$,*

$\widehat{\mu}_2$ and $\widehat{\mu}_3$ are all independent. Therefore

$$\begin{aligned}\text{Cov}[\text{RandDLow}, \text{RandDMod}] &= \text{Cov}[\widehat{\mu}_2 - \widehat{\mu}_1, \widehat{\mu}_3 - \widehat{\mu}_1] \\ &= \text{Cov}[\widehat{\mu}_2, \widehat{\mu}_3] - \text{Cov}[\widehat{\mu}_2, \widehat{\mu}_1] \\ &\qquad - \text{Cov}[\widehat{\mu}_1, \widehat{\mu}_3] + \text{Cov}[\widehat{\mu}_1, \widehat{\mu}_1] \\ &= 0 - 0 - 0 + \text{Var}[\widehat{\mu}_1] \\ &= \text{Var}[(\text{Intercept})],\end{aligned}$$

as required.

Solution to 1(e): *The null hypothesis is $H_0 : \mu_2 = \mu_3$ or, equivalently, $H_0 : \mu_2 - \mu_1 = \mu_3 - \mu_1$. Moreover, from the R summary output, and using the identity $Var[aX + bY] = a^2\, Var[X] + b^2\, Var[Y] + 2ab\, Cov[X, Y]$,*

$$\begin{aligned}\text{Var}[\text{RandDLow} - \text{RandDMod}] &= 0.4217^2 + 0.4000^2 - 2 \times 0.3266^2 \\ &= 0.1245,\end{aligned}$$

so the standard error of RandDLow − RandDMod is

$$\sqrt{0.1245} = 0.3528.$$

So we should refer

$$\frac{\text{RandDLow} - \text{RandDMod}}{se(\text{RandDLow} - \text{RandDMod})} = \frac{-2.3222 + 1.0667}{0.3528} = -3.559$$

to the t-dsitrubtion with 24 degrees of freedom. The correponding 2-sided p-value is 0.0016. This provides fairly strong evidence for rejecting H_0, in that it is comfortably significant at the 0.01 level.

Solution to 1(f): *We might consider combining these two categories but, unless there is expert support for combining the low investment and moderate investment categories, might well be best to do nothing.*
Note: *Any sensible comments should be marked sympathetically.*

Mark scheme for Question 1. (a)=2. (b)=4. (c)=4. (d)=6. (e)= 6. (f)=3. **Total**: 25.

2. Consider a random variable Y with a discrete distribution whose probability mass function is given by

PMF :

$$\text{Prob}[Y = y] = f(y; p) = \begin{cases} (1-p)^2(y+1)p^y & y = 0, 1, 2, ... \\ 0 & y < 0. \end{cases} \quad (2)$$

(a) Demonstrate that $f(y; p)$ may be written in the form of a Generalised Linear Model (GLM) distribution, i.e. show that

$$f(y; p) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right\},$$

where θ, ϕ, $b(\theta)$ and $c(y, \phi)$ should be determined.

(b) Find the mean $\mu = E[Y]$ in terms of θ.

(c) What is the canonical link function for this GLM?

(d) Find the variance function $V(\mu)$ for the GLM associated with (2).

(e) Let $\ell(\mu; y)$ denote the log likelihood of this GLM for a single observation from (2), but parametrised with μ rather than p. Show that $\ell(\mu; y)$ has the form

$$\ell(\mu; y) = y \log(\mu) - (y+2) \log(\mu + 2) + a(y), \quad (3)$$

where the function $a(y)$ should be determined. What is the maximum likelihood estimator of μ in the model (3)? You may state the result without proof.

(f) Suppose now that we have response data $y_1, \ldots, y_n$. After fitting a particular GLM with response distribution of the form (2), the fitted values corresponding to $y_1, \ldots, y_n$, were found to be $\widehat{\mu}_1, \ldots, \widehat{\mu}_n$, respctively. Find an expression for the deviance residual for observation i.

(g) Consider a plot of fitted values versus deviance residuals, with the fitted values on the horizontal axis. Provide a rough sketch of what you might expect to see in the plot if the variance function is approximately correct for smaller fitted values and approximately correct for fitted values in the middle of the range, but the variance function tends to over-estimate the variance of the response variable for larger fitted values.

Solution to 2(a): *Taking logs, and using the fact that if* $\theta = \log(p)$ *then* $p = e^\theta$*, we have*

$$\begin{aligned}\log\{f(y;p)\} &= 2\log(1-p) + \log(y+1) + y\log(p)\\ &= y\theta + 2\log(1-e^\theta) + \log(y+1)\\ &= \frac{y\theta - b(\theta)}{\phi} + c(y,\phi),\end{aligned}$$

where $\theta = \log(p)$*,* $b(\theta) = -2\log(1-e^\theta)$*,* $\phi = 1$ *and* $c(y,1) = \log(y+1)$*.*

Solution to 2(b): *From the theory of GLMs,*

$$\mu = E[Y] = b'(\theta) = \frac{db}{d\theta}(\theta) = \frac{2e^\theta}{1-e^\theta}. \tag{4}$$

Solution to 2(c): *The canonical link function is characterised by the requirement that* $\eta = \theta$*, where* η *is the linear predictor and* θ *is the natural parameter. Making* e^θ *the subject of (4), we find that*

$$e^\theta = \frac{\mu}{\mu+2} \tag{5}$$

and so

$$\theta = \log(\mu) - \log(\mu+2) = g(\mu)$$

is the canonical link function.

Solution to 1(d): *The variance function is given by* $b''(\theta)$*, expressed as a function of* μ*. Now*

$$\begin{aligned}b''(\theta) &= \frac{d}{d\theta}\frac{2e^\theta}{1-e^\theta}\\ &= \frac{2e^\theta}{1-e^\theta} + \frac{2e^\theta e^\theta}{(1-e^\theta)^2}\\ &= \frac{2e^\theta}{(1-e^\theta)^2}\\ &= \mu\frac{1}{1-e^\theta}.\end{aligned}$$

From (5) it follows that $1-e^\theta = 1-\mu/(\mu+2) = 2/(\mu+2)$, and therefore $(1-e^\theta)^{-1} = 1+(\mu/2)$. Consequently, the variance function is given by

$$V(\mu) = \frac{\mu}{1-e^\theta} = \mu\left(1+\frac{\mu}{2}\right) = \mu + \frac{1}{2}\mu^2.$$

Solution to 2(e): *Since* $p = e^\theta = \mu/(\mu+2)$, *after substituting* $p = \mu/(\mu+2)$ *into (2), using* $1-p = 2/(\mu+2)$, *we find*

$$f(y; p(\mu)) = \left(\frac{2}{\mu+2}\right)^2 (y+1)\left(\frac{\mu}{\mu+2}\right)^y,$$

so taking logs we obtain

$$\begin{aligned}\ell(\mu; y) &= \log\{f(y; p(\mu))\} \\ &= 2\log(2) - 2\log(\mu+2) + \log(y+1) + y\log(\mu) - y\log(\mu+2) \\ &= y\log(\mu) - (y+2)\log(\mu+2) + a(y),\end{aligned}$$

as required, where $a(y) = 2\log(2) + \log(y+1)$. The maximum likelihood estimator of μ is y.

Solution to 2(f): *The deviance contribution from observation* i *is*

$$\begin{aligned}d_i^2 &= 2[\ell(y_i; y_i) - \ell(\widehat{\mu}_i; y_i)] \\ &= 2[y_i\log(y_i) - (y_i+2)\log(y_i+2) + a(y_i) \\ &\quad - \{y_i\log(\widehat{\mu}_i) - (y_i+2)\log(\widehat{\mu}_i+2) + a(y_i)\}] \\ &= 2\left[y_i\log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i+2)\log\left(\frac{y_i+2}{\widehat{\mu}_i+2}\right)\right],\end{aligned}$$

and so the deviance residual for observation i *is given by* $\text{sign}(y_i - \widehat{\mu}_i)d_i$, *where* $\text{sign}(.)$ *is the sign function defined in the lectures.*

Solution to 2(g): *Assuming for simplicity that the horizontal density of the points is not too non-uniform, one would expect to see more or less constant vertical spread for fitted values below the upper region; and for fitted values in the upper region we would expect to see reduced vertical spread compared to the lower and middle regions. Any reasonable attempt should be marked sympathetically.*

Mark scheme for Question 2. (a)=4. (b)=3. (c)=3. (d)=4. (e)=4. (f)=4. (g)=3. **Total**: 25.

3(a) Different doses of two chemicals, $\mathcal{A}$ and $\mathcal{B}$, were used in a trial whose purpose was to reduce cockroach numbers. The variable $x1$ gives the dose of chemical $\mathcal{A}$ and the variable $x2$ gives the dose of chemical $\mathcal{B}$. In the R code, the first column of c gives the number of cockroaches killed and the second column of c gives the number of cockroaches that survived. The following R outputs were obtained:

```
>  out=glm(c~x1+x2,family=binomial)
> summary(out)

Call:
glm(formula = c ~ x1 + x2, family = binomial)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.7922   0.5388  -0.3190  -1.2973   0.4378  -0.7025   0.4556   2.1441

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -97.8769    22.8731   A       1.68e-05 ***
x1           56.4856    13.6157   B       3.35e-05 ***
x2           -0.5368     0.3122   C       D .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.2024  on 7  degrees of freedom
Residual deviance:   8.1925  on 5  degrees of freedom
AIC: 40.391

Number of Fisher Scoring iterations: 5

> anova(out)
Analysis of Deviance Table

Model: binomial, link: E

Response: c

Terms added sequentially (first to last)
```

	Df	Deviance	Resid.Df	Resid.Dev
NULL			7	F
x1	1	G	H	11.232
x2	J	3.04	5	K

(i) Determine the missing information indicated by the letters A,B,C,D, E,F,G,H,J and K. Note that for E you are required to specify the link function.

(ii) Write down the relevant model in mathematical form, focusing on the contribution of observation i to the likelihood.

(iii) Briefly indicate your impressions of the results of the the statistical analysis so far.

(iv) What are the next questions you would investigate in the statistical analysis? State what your next two steps would be.

Solution to 3(a)(i): *The missing information is*

- $A = 4.280$.
- $B = 4.148$.
- $C = -1.719$.
- $D = 0.086$.
- E is the logit (or logistic) link.
- $F = 284.2024$.
- $G = 272.9704$.
- $H = 6$.
- $J = 1$.
- $K = 8.1925$.

Solution to 3(a)(ii): *The model states that y_i, $i = 1, \ldots, n$, where here $n = 8$, are independent binomial random variables, with $y_i \sim$ Binomial(m_i, p_i) where m_i is the number of binomial trials for observation i and p_i is the probability of killing a cockroach in trial i. The*

mathematical form of p_i *is*

$$p_i = \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)} = \frac{\exp(\beta_0+\beta_1 x1+\beta_2 x2)}{1+\exp(\beta_0+\beta_1 x1+\beta_2 x2)},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ *and* $\boldsymbol{x}_i = (1, x1, x2)^\top$. *As usual the binomial probabilities are given by*

$$\text{Prob}[Y_i = y_i] = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i-y_i}, \quad y_i = 0, 1, \ldots, m_i.$$

Solution to 3(a)(iii): *It appears that variable* $x1$*, the dose of chemical* $\mathcal{A}$*, is quite important as the p-value of the Wald test is very small and the associated reduction in deviance is large. The case for the importance of the variable* $x2$*, the dose of chemical* $\mathcal{B}$*, is somewhat less clear-cut, the p-value based on the Wald test is around 0.09 and the sign of the coefficient is negative, which perhaps is a bit strange.*

Solution to 3(a)(iv): *Few possibilities are: to fit the model with* $x1$ *included and* x_2 *excluded, to assess the importance of* $x2$ *through the change-in-deviance test; see if there is any interaction between* $x1$ *and* $x2$*; try replacing* $x1$ *and* $x2$ *by their logs, assuming these variables are positive, and perhaps try some other link functions. Any sensible suggestions shuld be marked sympathetically.*

Mark scheme for Question 3(a). (a)(i) = 5. (a)(ii)=3. (a)(iii)=3. (a)(iv)=3. **Total**: 14.

3(b) Blood groups of peptic ulcer and control patients in London, Manchester and Newcastle were recorded in a case-control study. Blood groups A and O were represented in a factor B with two levels; the cities mentioned were represented as a factor C with three levels, L, M and N; and U represented a factor with two levels, Control and Ulcer. In this case-control study it is appropriate to treat B as a response factor and C and U both as covariate factors. The data were entered into R as follows.

```
> B=c("A","A","A","A","A","A","O","O","O","O","O","O")
> C=c("L","L","M","M","N","N","L","L","M","M","N","N")
> U=c("C","U","C","U","C","U","C","U","C","U","C","U")
> count=c(4219,579,3775,246,5261,219,4578,911,4532,361,6598,396)
```

The following models were fitted, using Poisson regression with log link. The deviance (as defined in the lecture notes) and the degrees of freedom (df) are given in the third and fourth columns, respectively.

Model	Model Formula	Deviance	df
$\mathcal{M}_1$	`count~B*C+U`	754.47	7
$\mathcal{M}_2$	`count~B*C+U`	737.75	5
$\mathcal{M}_3$	`count~B*U+C`	700.97	6
$\mathcal{M}_4$	`count~B+C*U`	83.559	5
$\mathcal{M}_5$	`count~B*U+C*U`	30.106	4
$\mathcal{M}_6$	`count~B*C+C*U`	66.878	3
$\mathcal{M}_7$	`count~B*C+B*U`	684.25	4
$\mathcal{M}_8$	`count~B*C+B*U+C*U`	2.9655	2

(i) Construct the three-way contingency table from the data that has been input into R.

(ii) Provide a list of the models in the table that are relevant to the situation where B is a response factor and C and U are covariate factors. Briefly explain why the models on your list, and no others, are the relevant models.

(iii) What do you conclude from the results in the table? Which model is to be preferred? How should this model be interpreted? Discuss briefly.

Solution to 3(b)(i): *It should be an easy exercise though I have not asked them to do this before so it is possible one or two might struggle.*

Solution to 3(b)(ii): *If we wish to treat B as a response factor and C and U as covariate factors, then we need to fix the totals at every combination of levels of C and U. This is achieved by only considering those models which include the interaction term C^*U. In the list, the models which have the interaction term C^*U are: $\mathcal{M}_4$, $\mathcal{M}_5$, $\mathcal{M}_6$ and $\mathcal{M}_8$.*

Solution to 3(b)(iii): *Using the change of deviance test. we clearly reject all of the models in the table except for $\mathcal{M}_8$, which we do not reject at any reasonable level as the p-value is* 0.23. *This model is not particularly easy to interpret: the distribution of B depends on the level of both C and U, but this model is still simpler than the saturated model (corresponding to the 3-way interaction.*

Mark scheme for Question 3(b). (b)(i)=3. (b)(ii)=5. (b)(iii)=3.
Total: 11.

Total for Question 3: 14+11=25.