THE UNIVERSITY of ADELAIDE

Examination in School of Mathematical Sciences

Semester 1, 2019

003989 STATS 3001 Statistical Modelling III

Official Reading Time:	10 mins
Writing Time:	120 mins
Total Duration:	130 mins

NUMBER OF QUESTIONS: 4 TOTAL MARKS: 70

Instructions

- Attempt all questions.
- Begin each answer on a new page.
- Examination materials must not be removed from the examination room.

Materials

- 1 Blue book is provided.
- Calculators are permitted.
- English and foreign language dictionaries are permitted.

DO NOT COMMENCE WRITING UNTIL INSTRUCTED TO DO SO.

1. Consider the multiple regression model,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}},$$

where X is an $n \times p$ matrix whose columns are linearly independent, $E(\boldsymbol{\mathcal{E}}) = \mathbf{0}$, and $\mathrm{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I$ where I denotes the $n \times n$ identity matrix.

(a) Show that the least squares estimates are given by $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$.

(b) Show that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

(c) Show that $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$.

(d) Suppose $\boldsymbol{\lambda} \in \mathbb{R}^p$ is a fixed vector.

(i) Show that $\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\lambda}^T \boldsymbol{\beta}$.

(ii) Show that $\mathrm{var}(\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{\lambda}^T (X^T X)^{-1} \boldsymbol{\lambda}$.

Typo in e)(ii): last term on RHS should be || I - X(X^TX)^{-1} \lambda||^2

(e) Suppose $\boldsymbol{\ell} \in \mathbb{R}^n$ is a fixed vector such that $E(\boldsymbol{\ell}^T \boldsymbol{Y}) = \boldsymbol{\lambda}^T \boldsymbol{\beta}$.

(i) Show that $(\boldsymbol{\ell} - X(X^T X)^{-1} \boldsymbol{\lambda})^T X = \mathbf{0}^T$.

(ii) Hence, show that $\boldsymbol{\ell}^T \boldsymbol{\ell} = \boldsymbol{\lambda}^T (X^T X)^{-1} \boldsymbol{\lambda} + \|\ell - (X^T X)^{-1} X \boldsymbol{\lambda}\|^2$.

(iii) Conclude that $\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator for $\boldsymbol{\lambda}^T \boldsymbol{\beta}$.

[20 marks]

2. Suppose $y_1, y_2, \ldots, y_n$ are independent observations such that $Y_i \sim \mathrm{Po}(\mu_i)$ for $i = 1, 2, \ldots, n$ and consider the log-linear model,

$$\boldsymbol{\eta} = X\boldsymbol{\beta},$$

where $\eta_i = \log \mu_i$ for $i = 1, 2, \ldots, n$.

(a) Show that the log-likelihood can be expressed in terms of $\boldsymbol{\eta}$ as

$$\ell(\boldsymbol{\eta}) = \sum_{i=1}^{n} \{y_i \eta_i - e^{\eta_i}\} - \log\left(\prod_{i=1}^{n} y_i!\right).$$

(b) Show that

$$\frac{\partial \ell}{\partial \eta_j} = y_j - \mu_j.$$

(c) Show that the score vector is given by

$$\mathcal{S}(\boldsymbol{\beta}) = \left(\frac{\partial \ell}{\partial \beta_k}\right) = X^T (\boldsymbol{y} - \boldsymbol{\mu}).$$

(d) Derive the Fisher information matrix, $\mathcal{I}(\boldsymbol{\beta})$.

[20 marks]

Please turn over for page 3

3. In a two-factor experiment, animals were exposed to one of four treatments labelled `A`, `B`, `C`, `D` and one of three poisons labelled `I`, `II`, `III` and the survival time in hours was recorded. Analyses performed in `R` are given in Appendix A.

   (a) State the assumptions of the linear model, `A0`.

   (b) Based on the relevant diagnostic plots shown in Figure 1, do these assumptions appear reasonable? Justify your answer.

   (c) Based on the Box Cox output in Figure 2, explain why the reciprocal transformation is indicated.

   (d) Consider the transformed response `y=1/Time`.

      (i) How would you interpret a large value of `y` in this context.

      (ii) What are the units for `y` in this context.

   (e) Based on the relevant diagnostic plots shown in Figure 3, do the linear model assumptions appear reasonable for the transformed model `A1`? Justify your answer.

   (f) Based on a suitable test of statistical significance, can the model `A1` be simplified to the additive model `A2`? Justify your answer.

   (g) Consider the estimated parameter `PoisonIII` in the model `A2`.

      (i) Interpret the numerical value of the estimated parameter in context.

      (ii) Using the fact that `qt(0.975,df=42)=2.018`, construct a 95% confidence interval for the true value of this parameter.

   [15 marks]

4. A large company employs approximately 4,000 staff. Each year, about 15% of staff leave the company for a variety of reasons. To understand the factors that lead to a staff member leaving, the variables summarised in Table 1 of Appendix B were recorded. Various logistic regression models were fit and the output is shown in Appendix B.

   (a) Based on model `B1`, find the estimated log-odds of `Attrition` for a female employee, aged 30 years, who has been with the company for 10 years, travels frequently and is divorced. Hence find the estimated probability of `Attrition`.

   (b) Test the hypothesis that marital `Status` has no effect on `Attrition` in the model `B2`, using the likelihood ratio test. Give the value of the likelihood ratio test statistic and the degrees of freedom for the χ^2 distribution. State your conclusion in context.
   Note: The critical value for the test with significance level 5% is given below.

```
qchisq(0.95,df)
## [1] 5.991465
```

(c) Based on model `B3`, the human resources manager notices that the probability of `Attrition` is higher for single employees than those who are married or divorced. She wonders whether this could be because single employees would be younger and hence more likely to move to another job. Comment on whether this explanation is supported by the data.

(d) Consider two employees, one male and one female who have both worked for the company for 10 years and have identical values on all other predictor variables. Based on the model `B3`, which of these employees is more likely to leave the company?

(e) Explain why the two models `B2` and `B3` are equivalent and illustrate the equivalence by comparing the parameters involving `Gender` and `Year`.

[15 marks]

Appendix A - R output for Question 3

```
#Model A0
A0=lm(Time~Treatment*Poison,data=poison)
A0.res=studres(A0)
A0.fit=fitted(A0)
A0.residuals=data.frame(fitted=A0.fit,residuals=A0.res)
ggplot(A0.residuals,aes(fitted,residuals))+geom_point()+
  geom_hline(yintercept=0)+ggtitle("Residuals vs Fitted for Model A0")
ggplot(A0.residuals,aes(sample=residuals))+geom_qq()+geom_qq_line()+
  ggtitle("Normal Quantile Plot for Model A0")
```

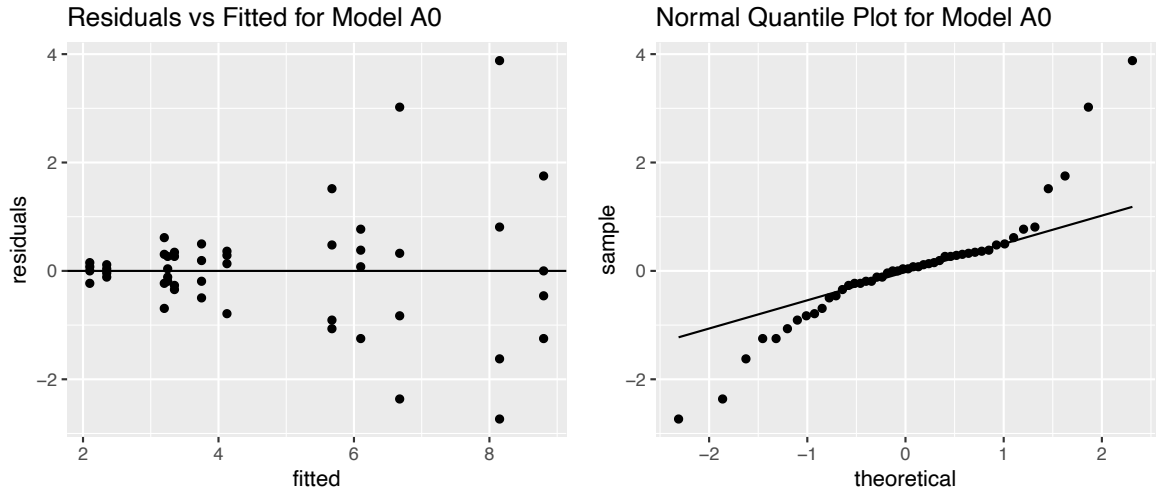

Figure 1: Diagnostic plots for Model A0

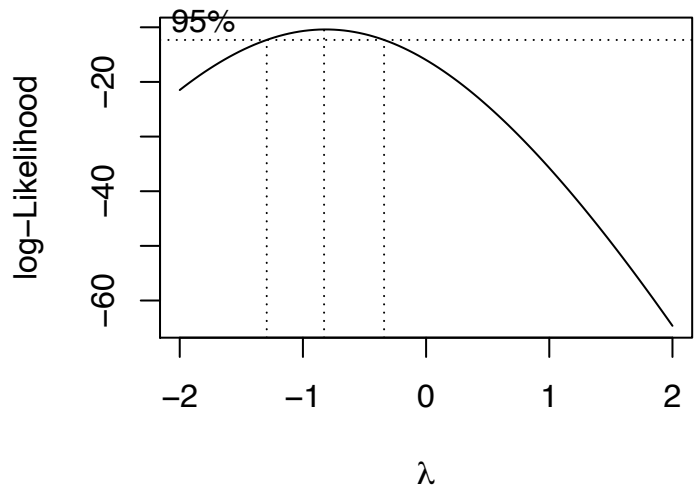

Figure 2: Box-Cox Analysis

```
boxcox(A0)
```

```
#Model A1
A1=lm(1/Time~Treatment*Poison,data=poison)
A1.res=studres(A1)
A1.fit=fitted(A1)
A1.residuals=data.frame(fitted=A1.fit,residuals=A1.res)
```

```
ggplot(A1.residuals,aes(fitted,residuals))+geom_point()+
  geom_hline(yintercept=0)+ggtitle("Residuals vs Fitted for Model A1")
ggplot(A1.residuals,aes(sample=residuals))+geom_qq()+geom_qq_line()+
  ggtitle("Normal Quantile Plot for Model A1")
```

Please turn over for page 7

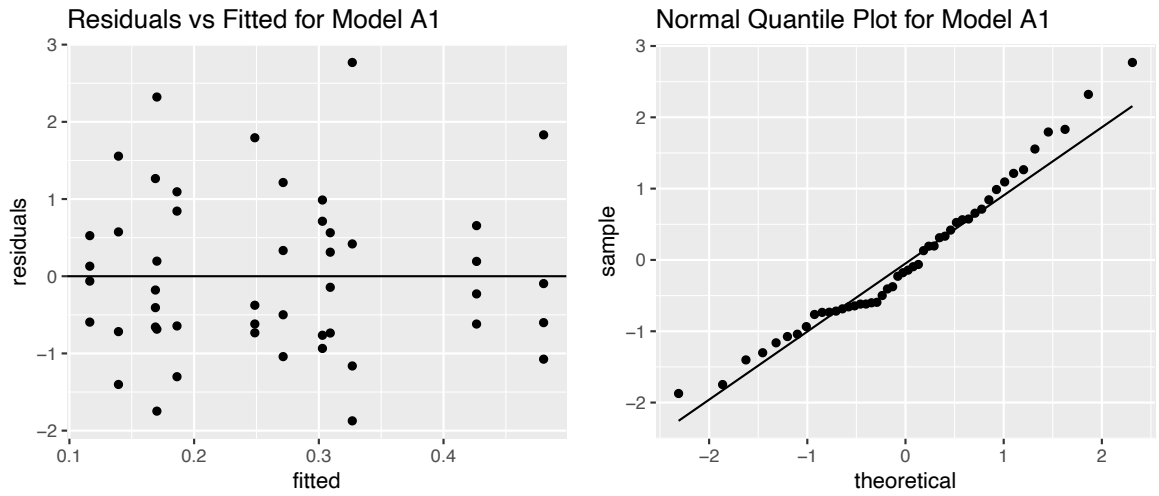

Figure 3: Diagnostic plots for Model A1

```
#Model A2
A2=lm(1/Time~Treatment+Poison,data=poison)
summary(A2)

##
## Call:
## lm(formula = 1/Time ~ Treatment + Poison, data = poison)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.082757 -0.037619  0.002116  0.027568  0.118153
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.26977    0.01744  15.473  < 2e-16 ***
## TreatmentB  -0.16574    0.02013  -8.233 2.66e-10 ***
## TreatmentC  -0.05721    0.02013  -2.842  0.00689 **
## TreatmentD  -0.13583    0.02013  -6.747 3.35e-08 ***
## PoisonII     0.04686    0.01744   2.688  0.01026 *
## PoisonIII    0.19964    0.01744  11.451 1.69e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04931 on 42 degrees of freedom
## Multiple R-squared:  0.8441,Adjusted R-squared:  0.8255
## F-statistic: 45.47 on 5 and 42 DF,  p-value: 6.974e-16
```

Please turn over for page 8

```
anova(A2,A1)

## Analysis of Variance Table
##
## Model 1: 1/Time ~ Treatment + Poison
## Model 2: 1/Time ~ Treatment * Poison
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     42 0.102139
## 2     36 0.086431  6  0.015708 1.0904 0.3867
```

Appendix B - R output for Question 4

Table 1: Variables in the `Staff` data set

Variable	Description	Values
`Attrition`	employee left in the last year	`Yes` `No`
`Gender`	sex of the employee	`Female` `Male`
`Age`	age of the employee in years	18-60
`Years`	number of years employed at company	0-40

`Travel`	amount of business travel required	`Frequently`
		`Rarely`
		`Never`

`Status`	marital status of employee	`Divorced`
		`Married`
		`Single`

```
#Model B1
B1=glm(Attrition~Age+Gender+Years+Travel+Status,data=Staff,family="binomial")
summary(B1)

##
## Call:
## glm(formula = Attrition ~ Age + Gender + Years + Travel + Status,
##     family = "binomial", data = Staff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1567  -0.6245  -0.4726  -0.3286   3.1366
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.009226   0.228905   0.040   0.9678
## Age           -0.035624   0.005171  -6.889 5.63e-12 ***
## GenderMale     0.083918   0.087970   0.954   0.3401
## Years         -0.061124   0.009705  -6.298 3.02e-10 ***
## TravelNever   -1.380863   0.196469  -7.028 2.09e-12 ***
## TravelRarely  -0.651159   0.098558  -6.607 3.93e-11 ***
## StatusMarried  0.241183   0.127839   1.887   0.0592 .
## StatusSingle   0.997848   0.125271   7.965 1.65e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3895.7  on 4409  degrees of freedom
## Residual deviance: 3559.0  on 4402  degrees of freedom
## AIC: 3575
##
## Number of Fisher Scoring iterations: 5
```

Please turn over for page 10

```r
#Model B2
B2=glm(Attrition~Age+Gender*Years+Travel+Status,data=Staff,family="binomial")
summary(B2)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + Gender * Years + Travel + Status,
##     family = "binomial", data = Staff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2252  -0.6161  -0.4747  -0.3281   3.1950
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.262857   0.238425  -1.102 0.270256
## Age              -0.035060   0.005176  -6.774 1.25e-11 ***
## GenderMale        0.489196   0.136962   3.572 0.000355 ***
## Years            -0.020166   0.013413  -1.503 0.132726
## TravelNever      -1.397506   0.197037  -7.093 1.32e-12 ***
## TravelRarely     -0.639463   0.098784  -6.473 9.58e-11 ***
## StatusMarried     0.249183   0.128061   1.946 0.051677 .
## StatusSingle      1.007399   0.125576   8.022 1.04e-15 ***
## GenderMale:Years -0.073040   0.018712  -3.903 9.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3895.7  on 4409  degrees of freedom
## Residual deviance: 3544.0  on 4401  degrees of freedom
## AIC: 3562
##
## Number of Fisher Scoring iterations: 5
```

Please turn over for page 11

```r
#Model B3
B3=glm(Attrition~Age+Gender/Years+Travel+Status,data=Staff,family="binomial")
summary(B3)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + Gender/Years + Travel + Status,
##     family = "binomial", data = Staff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2252  -0.6161  -0.4747  -0.3281   3.1950
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.262857   0.238425  -1.102 0.270256
## Age                -0.035060   0.005176  -6.774 1.25e-11 ***
## GenderMale          0.489196   0.136962   3.572 0.000355 ***
## TravelNever        -1.397506   0.197037  -7.093 1.32e-12 ***
## TravelRarely       -0.639463   0.098784  -6.473 9.58e-11 ***
## StatusMarried       0.249183   0.128061   1.946 0.051677 .
## StatusSingle        1.007399   0.125576   8.022 1.04e-15 ***
## GenderFemale:Years -0.020166   0.013413  -1.503 0.132726
## GenderMale:Years   -0.093206   0.013495  -6.907 4.96e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3895.7  on 4409  degrees of freedom
## Residual deviance: 3544.0  on 4401  degrees of freedom
## AIC: 3562
##
## Number of Fisher Scoring iterations: 5
```

Please turn over for page 12

```r
#Model B4
B4=glm(Attrition~Age+Gender*Years+Travel,data=Staff,family="binomial")
summary(B4)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + Gender * Years + Travel, family = "binomial",
##     data = Staff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0847  -0.6411  -0.5127  -0.3479   3.1822
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.478192   0.211769   2.258 0.023941 *
## Age               -0.040752   0.005156  -7.904 2.71e-15 ***
## GenderMale         0.471764   0.135808   3.474 0.000513 ***
## Years             -0.025113   0.013458  -1.866 0.062043 .
## TravelNever       -1.408898   0.194619  -7.239 4.51e-13 ***
## TravelRarely      -0.648960   0.097178  -6.678 2.42e-11 ***
## GenderMale:Years -0.071794   0.018796  -3.820 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3895.7  on 4409  degrees of freedom
## Residual deviance: 3639.9  on 4403  degrees of freedom
## AIC: 3653.9
##
## Number of Fisher Scoring iterations: 5
```