Examination in School of Mathematical Sciences

Semester 1, 2016

003989 STATS 3001 Statistical Modelling III

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

NUMBER OF QUESTIONS: 5 TOTAL MARKS: 80

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 without remote communications or CAS capability are allowed.

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 a fixed $n \times p$ matrix of rank p, $E(\boldsymbol{\mathcal{E}}) = \mathbf{0}$, and $\text{var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I$ where I is the $n \times n$ identity matrix.

(a) State the formula for the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}$.

(b) Prove that $\hat{\boldsymbol{\beta}}$ is unbiased.

(c) Prove that $\text{var}(\hat{\boldsymbol{\beta}})$ is $\sigma^2(X^TX)^{-1}$.

(d) If $\boldsymbol{W} \sim N_r(\boldsymbol{\mu}, \Sigma)$ and A is a fixed $k \times r$ matrix and $\boldsymbol{b}$ is a fixed $k \times 1$ column vector, state the distribution of $A\boldsymbol{W} + \boldsymbol{b}$.

(e) Hence, if $\boldsymbol{x}_0^T$ is a fixed $1{\times}p$ vector and

$$\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}\right),$$

what is the distribution of $\hat{\eta}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$?

(f) Show that the random vector of residuals, $\hat{\boldsymbol{\mathcal{E}}}$, can be expressed as $(I - H)\boldsymbol{Y}$, where $H = X(X^TX)^{-1}X^T$.

(g) Hence, show that

$$\text{var}(\hat{\mathcal{E}}_i) = \sigma^2(1 - h_{ii}),$$

where h_{ii} is the ith diagonal element of H. You may use the fact $H = H^T = H^2$.

(h) If the observed ordinary residuals are $\hat{e}_i$, give the definition of standardized residuals $\hat{e}'_i$.

[20 marks]

2. (a) State the Gauss-Markov Theorem (without proof).

(b) Let $\boldsymbol{Y}$ be an $n{\times}1$ random vector with $E(\boldsymbol{Y}) = \boldsymbol{\eta} = X\boldsymbol{\beta}$ and $\text{Var}(\boldsymbol{Y}) = \sigma^2 V$, where V is a symmetric, positive-definite $n{\times}n$ matrix.

(i) Using $\hat{\boldsymbol{\beta}}_{OLS} = (X^TX)^{-1}X^T\boldsymbol{y}$, find $\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}\right)$.

(ii) Is $\hat{\boldsymbol{\eta}}_{OLS} = X\hat{\boldsymbol{\beta}}_{OLS}$ the best linear unbiased estimator (BLUE) for $\boldsymbol{\eta} = X\boldsymbol{\beta}$? Justify your answer.

[8 marks]

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

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

   (b) Based on the diagnostic plots for `A0`, which, if any, of these assumptions do not appear reasonable? Justify your answer.

   (c) Based on the Box-Cox output for `A0`, 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 the present context?

      (ii) What are the units of `y` in the present context?

   (e) Based on the diagnostic plots for `A1`, 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.

[17 marks]

4. (a) Consider $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{\eta} \in \mathcal{L} \subset \mathbb{R}^n$. Let P be the orthogonal projection on $\mathcal{L}$ and $(I - P)$ be the orthogonal projection on $\mathcal{L}^\perp$.

   (i) By considering what subspaces $\boldsymbol{y} - P\boldsymbol{y}$ and $P\boldsymbol{y} - \boldsymbol{\eta}$ are elements of, show that

$$\langle \boldsymbol{y} - P\boldsymbol{y}, P\boldsymbol{y} - \boldsymbol{\eta} \rangle = 0.$$

   (ii) Hence, show that

$$\|\boldsymbol{y} - \boldsymbol{\eta}\|^2 \geq \|\boldsymbol{y} - P\boldsymbol{y}\|^2$$

   with equality if and only if

$$P\boldsymbol{y} = \boldsymbol{\eta}.$$

(b) Suppose P_0 and P are the orthogonal projections on the linear subspaces $\mathcal{L}_0$ and $\mathcal{L}$, respectively, where $\mathcal{L}_0 \subset \mathcal{L}$.

   (i) Simplify $\mathcal{L}_0 + \mathcal{L} \cap \mathcal{L}_0^\perp$.

$$\text{(Hint: } \mathcal{L}_0 = \mathcal{L}_0 \cap \mathcal{L} \oplus \mathcal{L}_0 \cap \mathcal{L}^\perp\text{)}$$

   (ii) Noting that

$$\mathbb{R}^n = \mathcal{L}^\perp \oplus \mathcal{L} \cap \mathcal{L}_0^\perp \oplus \mathcal{L}_0,$$

   find the following:

   1. $(P - P_0)\boldsymbol{u}$, given $\boldsymbol{u} \in \mathcal{L}^\perp$,
   2. $(P - P_0)\boldsymbol{v}$, given $\boldsymbol{v} \in \mathcal{L} \cap \mathcal{L}_0^\perp$, and
   3. $(P - P_0)\boldsymbol{w}$, given $\boldsymbol{w} \in \mathcal{L}_0$.

   (iii) Hence, on what space does $P - P_0$ project?

[15 marks]

5. On the 28th January 1986, the NASA Space Shuttle Challenger disintegrated shortly after launch at Cape Canaveral, Florida. All seven crew members were killed.

The shuttle disintegrated as a result of failed O-rings, rubber components that create important seals between compartments of the shuttle.

Data on the failure for O-rings of $m = 23$ previous space shuttle flights, prior to the launch of the space shuttle Challenger, are analysed in Appendix B. The data contain the following variables for the 23 previous space shuttle flights, for which there were n_i O-rings per flight for $i = 1, 2, \ldots, 23$.

Variable	Description
`Temp`	Ambient launch temperature (in Fahrenheit)
`Damaged`	Number of O-rings damaged
`NotDamaged`	Number of O-rings not damaged

Note: Appendix C provides a range of critical values associated with different distributions that will be of help in answering some of the following questions.

(a) Give a careful interpretation of the parameter estimates in model `L1`.

(b) The forecast temperature for the day of launch of the Challenger space shuttle was 31 degrees Fahrenheit. Assuming extrapolation is justified, provide an estimate for the probability of a randomly selected O-ring on the Challenger space shuttle failing, π^*, for a launch temperature of 31 degrees Fahrenheit using model `L1`.

(c) Use the model output for `L1` and `L2` to calculate the log-likelihood ratio test statistic,

$$G^2 = 2(\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_0))$$

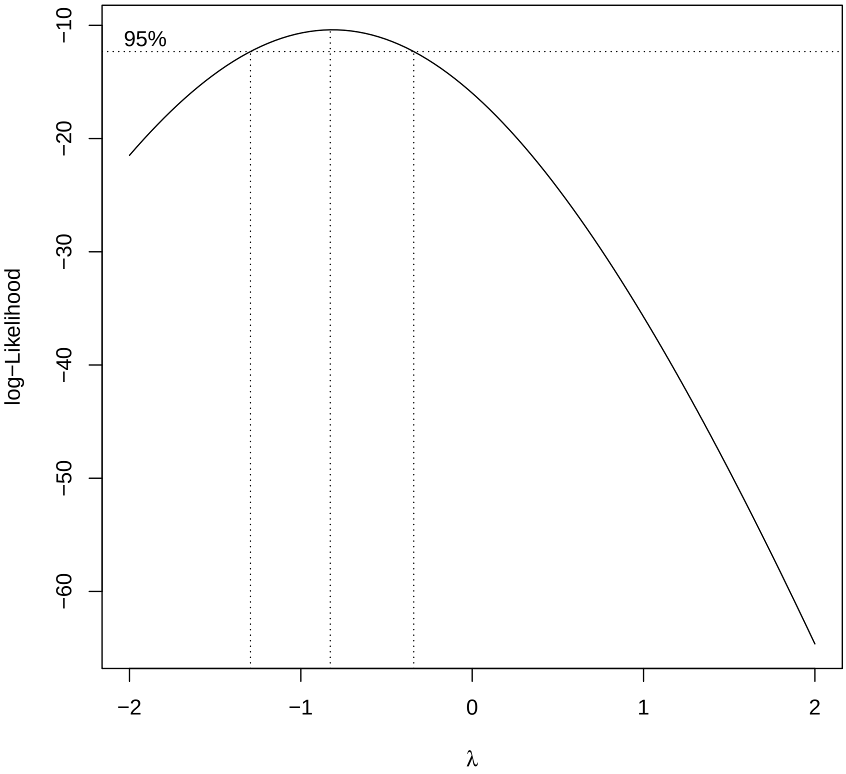
where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ are the maximum likelihood estimates under `L2` and `L1`, respectively. What are the associated null hypothesis and asymptotic null distribution (including degrees of freedom)? What do you conclude?

(d) Assuming that $n_i\pi_i(1 - \pi_i)$ are all large for $i = 1, 2, \ldots, 23$, test the adequacy of the model fit of `L1`.

[20 marks]

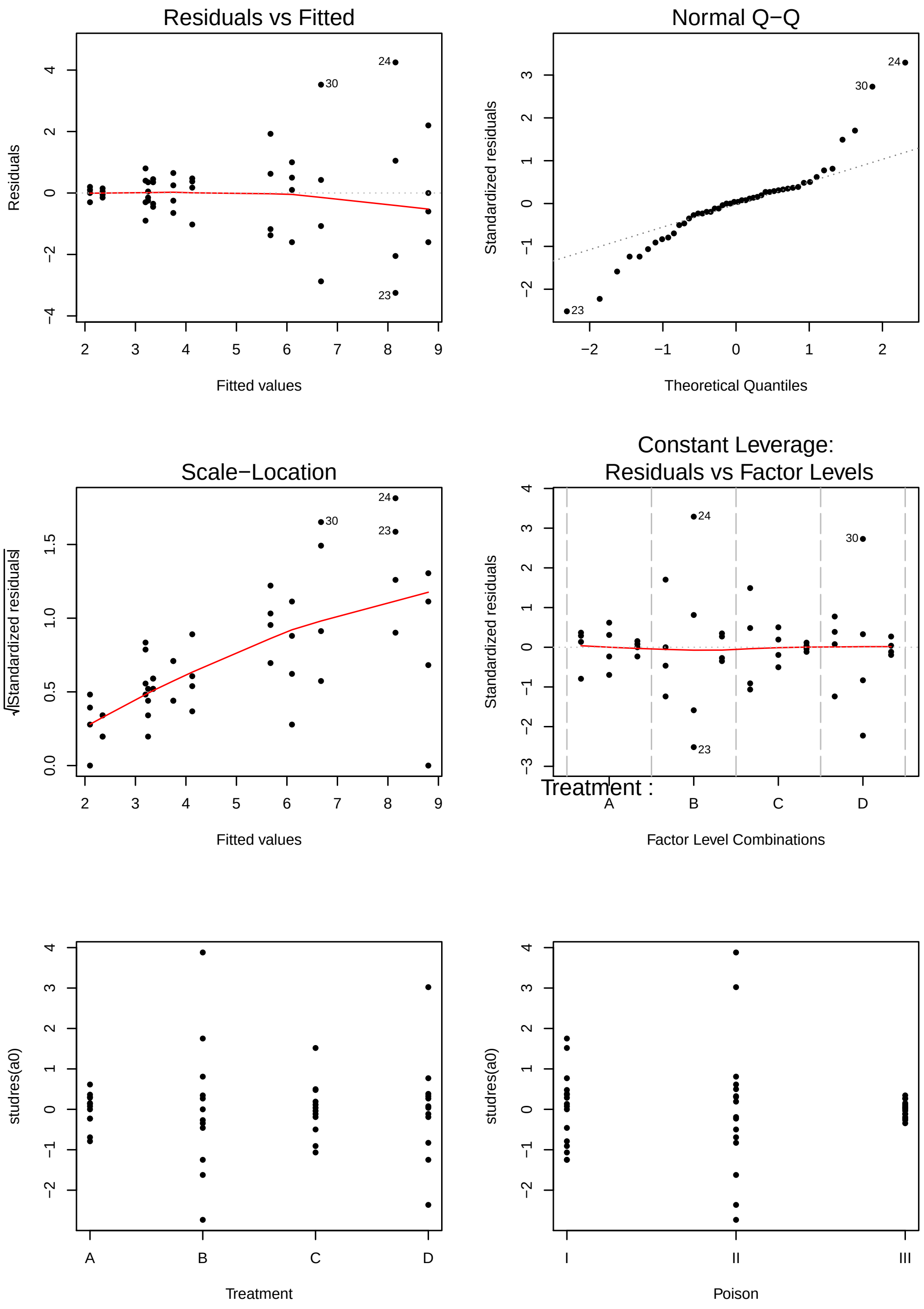
Appendix A: R output for Question 3

```
library(MASS)
poison<-read.csv("poison.csv",header=TRUE)
head(poison)

##   Poison Treatment Time
## 1      I         A  3.1
## 2      I         A  4.5
## 3      I         A  4.6
## 4      I         A  4.3
## 5      I         B  8.2
## 6      I         B 11.0

Time<-poison$Time
Treatment<-factor(poison$Treatment)
Poison<-factor(poison$Poison)

# model A0
a0<-lm(Time~Treatment*Poison)
boxcox(a0)
```

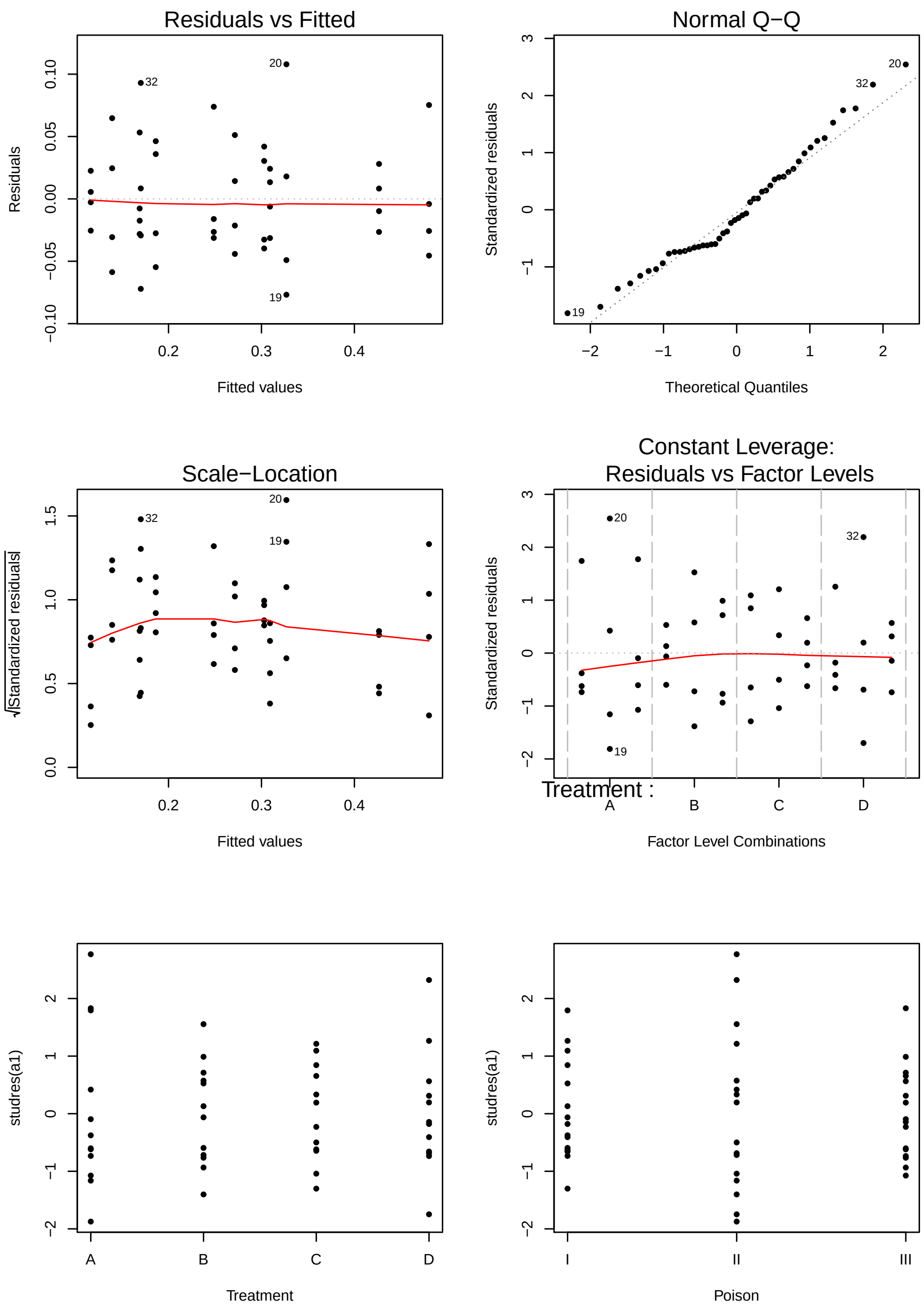


Please turn over for page 7

```
par(mfrow=c(3,2))
plot(a0)
plot(Treatment,studres(a0))
plot(Poison,studres(a0))
```



```r
y<-1/Time
# model A1
a1<-lm(y~Treatment*Poison)
par(mfrow=c(3,2))
plot(a1)
plot(Treatment,studres(a1))
plot(Poison,studres(a1))
```



Please turn over for page 9

```
# model A2
a2<-lm(y~Treatment+Poison,data=poison)
anova(a1,a2)

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

Please turn over for page 10

Appendix B: **R** output for Question 5

```
chal<-read.table("challenger.txt",header=TRUE)
# min and max Temps
c(min_temp=min(chal$Temp),max_temp=max(chal$Temp))
```

```
## min_temp max_temp
##       53       81
```

```
# log-odds depend on Temp as linear effect
L1<-glm(cbind(Damaged,NotDamaged)~Temp,family=binomial,data=chal)
# allow for non-linear Temp effect
L2<-glm(cbind(Damaged,NotDamaged)~Temp+I(Temp^2),family=binomial,data=chal)
# L1 fit
summary(L1)
```

```
##
## Call:
## glm(formula = cbind(Damaged, NotDamaged) ~ Temp, family = binomial,
##     data = chal)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.95227  -0.78299  -0.54117  -0.04379   2.65152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.08498    3.05247   1.666   0.0957 .
## Temp        -0.11560    0.04702  -2.458   0.0140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 18.086  on 21  degrees of freedom
## AIC: 35.647
##
## Number of Fisher Scoring iterations: 5
```

Please turn over for page 11

```
# L2 fit
summary(L2)

## 
## Call:
## glm(formula = cbind(Damaged, NotDamaged) ~ Temp + I(Temp^2), 
##     family = binomial, data = chal)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -0.84320  -0.72385  -0.61980  -0.01335   2.52101  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 22.126148  23.794426   0.930    0.352
## Temp        -0.650885   0.740756  -0.879    0.380
## I(Temp^2)    0.004141   0.005692   0.727    0.467
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 17.592  on 20  degrees of freedom
## AIC: 37.152
## 
## Number of Fisher Scoring iterations: 5
```

Please turn over for page 12

Appendix C

```
# critical values of the Chi-squared dist at the 5% level for various degrees of freedom
alpha<-0.05
deg_of_freedom<-c(1,2,3,4,5,20,21,22,23,24)
data.frame(
        deg_of_freedom
        ,crit_val=qchisq(1-alpha,df=deg_of_freedom)
)
```

```
##    deg_of_freedom  crit_val
## 1               1  3.841459
## 2               2  5.991465
## 3               3  7.814728
## 4               4  9.487729
## 5               5 11.070498
## 6              20 31.410433
## 7              21 32.670573
## 8              22 33.924438
## 9              23 35.172462
## 10             24 36.415029
```