THE UNIVERSITY of ADELAIDE

Examination in School of Mathematical Sciences
Semester 1, 2017

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: 90

Instructions

- Attempt all questions.
- You are welcome to separate the appendices from the main booklet.
- 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.
- Bilingual dictionaries may be used.

DO NOT COMMENCE WRITING UNTIL INSTRUCTED TO DO SO.

1. Consider the multiple linear regression model

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

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

(a) State the formula for the ordinary least-squares estimate $\hat{\boldsymbol{\beta}}$ and prove that it uniquely minimises the sum of squares $Q(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2$.

(b) State, without proof, $E(\hat{\boldsymbol{\beta}})$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}})$.

(c) Define the vector $\hat{\boldsymbol{\mathcal{E}}}$ of ordinary residuals and derive expressions for $E(\hat{\boldsymbol{\mathcal{E}}})$ and $\mathrm{Var}(\hat{\boldsymbol{\mathcal{E}}})$. You may assume $H^T = H = H^2$ where $H = X(X^TX)^{-1}X^T$.

(d) Define the vector of standardized residuals $\hat{\mathbf{e}}'$ and explain the motivation for this definition.

(e) Suppose now that $E(\boldsymbol{\mathcal{E}}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 V$ where V is a known, symmetric, positive-definite $n \times n$ matrix.

(i) Prove that the ordinary least-squares estimate $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

(ii) Derive an expression for $\mathrm{Var}(\hat{\boldsymbol{\beta}})$.

(iii) State whether the ordinary least squares estimate, $\hat{\boldsymbol{\beta}}$, or the generalised least-squares estimate, $\hat{\boldsymbol{\beta}}_{GLS} = (X^TV^{-1}X)^{-1}X^TV^{-1}\boldsymbol{y}$, is preferable in the given context. Give a brief reason for your answer.

[23 marks]

2. Consider the multiple linear regression model,

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

where $\boldsymbol{Y}, \boldsymbol{\mathcal{E}} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and X is an $n \times p$ matrix with linearly independent columns. Suppose that $E(\boldsymbol{\mathcal{E}}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I$.

(a) The residual variance is defined as

$$s_e^2 = \frac{1}{n-p} \|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}\|^2.$$

In what follows, you may assume that $P^T = P = P^2$ where $P = X(X^TX)^{-1}X^T$.

(i) Show that

$$(n-p)s_e^2 = (\boldsymbol{Y} - \boldsymbol{\eta})^T(I - P)(\boldsymbol{Y} - \boldsymbol{\eta})$$

where $\boldsymbol{\eta} = X\boldsymbol{\beta}$.

(ii) Hence show that

$$E\left((n-p)s_e^2\right) = \sigma^2 \,\text{tr}(I - P).$$

(iii) Therefore show s_e^2 is an unbiased estimator for σ^2.

(b) State, without proof, the distribution of $(n-p)s_e^2/\sigma^2$ when it is assumed further that $\mathcal{E}_i \sim N(0, \sigma^2)$ independently for $i = 1, 2, \ldots, n$.

[10 marks]

Please turn over for page 4

3. Data were collected on 45 occupations in the U.S. in 1950. A summary of the variables recorded for each of the 45 occupations is below.

Variable	Description
`type`	Type of occupation with three levels: `prof`, professional and managerial; `wc`, white-collar (office job); and `bc`, blue-collar (manual labour).
`income`	Percent of males in occupation earning \$3500 or more per annum in 1950.
`education`	Percent of males in occupation in 1950 who were high-school graduates.
`prestige`	Percent of people surveyed rating the occupation as prestigious.

Excerpts from an analysis using `prestige` as a response variable are given in Appendix A.

(a) Provide the interpretation for the estimated `income` coefficient in the simple linear regression model `M0` in context. How does this interpretation differ for the estimated coefficient of `income` in the multiple linear regression model `M1`? Do the differences in the interpretation of the coefficients explain the difference in the estimated values?

(b) State the assumptions of the linear model, `M1`.

(c) Based on the diagnostic plots for `M1`, do the linear model assumptions appear reasonable? Make reference to the diagnostic plots (a)-(f) in Appendix A in justifying your answer.

(d) The provided diagnostic plots for `M1` do not include a Standardised Residuals vs Leverage plot. If you were told point 9, which has a standardised residual of 3.31, was *not* a point of concern (for undue influence on the regression estimates), what does this imply about the leverage of point 9, namely h_{99}?

(e) Derive the value for the quantity $\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0$ under model `M1` using the `R` output provided where $\boldsymbol{x}_0$ is the vector of predictor values corresponding to `education=60`, `income=80` and `type='prof'`. You may use the fact $t_{40}(0.025) = 2.021$.

(f) Do you have any concerns about the prediction interval using model `M1` with predictor values of `education=60`, `income=80` and `type='prof'`?

(g) Based on a suitable test of statistical significance, can the model `M1` be simplified to the reduced model `M2`? Justify your answer and specify the hyptothesis being tested.

(h) The `income` and `prestige` variables were calculated from the sample in different ways. The `income` variable is the percentage of men *within* each occupation earning \$3500 or more per annum in 1950. In contrast, all respondents rated all occupations as prestigious or not prestigious. The `prestige` variable is the percentage of respondents who rated an occupation as prestigious. If you were to fit a model with `income` as the response variable instead of `prestige`, why might weighted least squares (WLS) be more appropriate than multiple linear regression? What additional information would you need to fit the WLS model?

[24 marks]

4. Consider the vector space $\mathbb{R}^n$ and suppose P_0, P_1 and P_2 are the orthogonal projections on the linear subspaces $\mathcal{L}_0$, $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively, where $\mathcal{L}_0 \subset \mathcal{L}_1 \subset \mathcal{L}_2$.

   (a) Noting that
   $$\mathbb{R}^n = \mathcal{L}_2^\perp \oplus \mathcal{L}_2 \cap \mathcal{L}_1^\perp \oplus \mathcal{L}_1 \cap \mathcal{L}_0^\perp \oplus \mathcal{L}_0,$$
   simplify the following:

   (i) $(P_2 - P_0)\boldsymbol{w}$, given $\boldsymbol{w} \in \mathcal{L}_2^\perp$,

   (ii) $(P_2 - P_0)\boldsymbol{v}_2$, given $\boldsymbol{v}_2 \in \mathcal{L}_2 \cap \mathcal{L}_1^\perp$,

   (iii) $(P_2 - P_0)\boldsymbol{v}_1$, given $\boldsymbol{v}_1 \in \mathcal{L}_1 \cap \mathcal{L}_0^\perp$, and

   (iv) $(P_2 - P_0)\boldsymbol{v}_0$, given $\boldsymbol{v}_0 \in \mathcal{L}_0$.

   (b) Simplify $\mathcal{L}_2 \cap \mathcal{L}_1^\perp \oplus \mathcal{L}_1 \cap \mathcal{L}_0^\perp$.

   (Hint: $\mathcal{L}_1^\perp = \mathcal{L}_1^\perp \cap \mathcal{L}_0^\perp$ and $\mathcal{L}_1 = \mathcal{L}_2 \cap \mathcal{L}_1$.)

   (c) Therefore, on what space does $P_2 - P_0$ project?

[12 marks]

Please turn over for page 6

5. Madsen (1976) reported the results of a survey of residents' satisfaction with housing conditions in Copenhagen. Residents of rented accommodation were questioned about their satisfaction with their accommodation as well as their degree of contact with the other residents. The accommodation was classified as a house or tower block. The degree of contact with other residents was classified as low or high. The proportion of residents expressing high satisfaction within each of the four categories are:

	Low contact	High contact
Tower block	100/219	100/181
Houses	62/177	104/339

Note: Appendix B includes logistic regression analyses without an interaction (model `L0`) and with an interaction (model `L1`). Appendix C also provides a range of critical values associated with different distributions that will be of assistance in answering some of the following questions.

(a) Write down the fitted model `L0` defining any terms you use.

(b) Hence, use model `L0` to calculate the estimated probability of a house renting resident, with low contact with other residents, expressing high satisfaction with their accomodation.

(c) Give a careful interpretation of the estimated `(Intercept)` coefficient in model `L1`.

(d) Use the model output for `L0` and `L1` 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 `L1` and `L0`, respectively. What is the associated null hypothesis and asymptotic null distribution (including degrees of freedom)? What do you conclude?

(e) Is your conclusion in (d) consistent with the follwing tests: (i) the deviance statistic of `L0`; and (ii) the significance of the interaction term in `L1` (Wald test)? Provide reasoning with your answer.

(f) What practical conclusions can you draw from this study?

[21 marks]

Appendix A: R output for Question 3

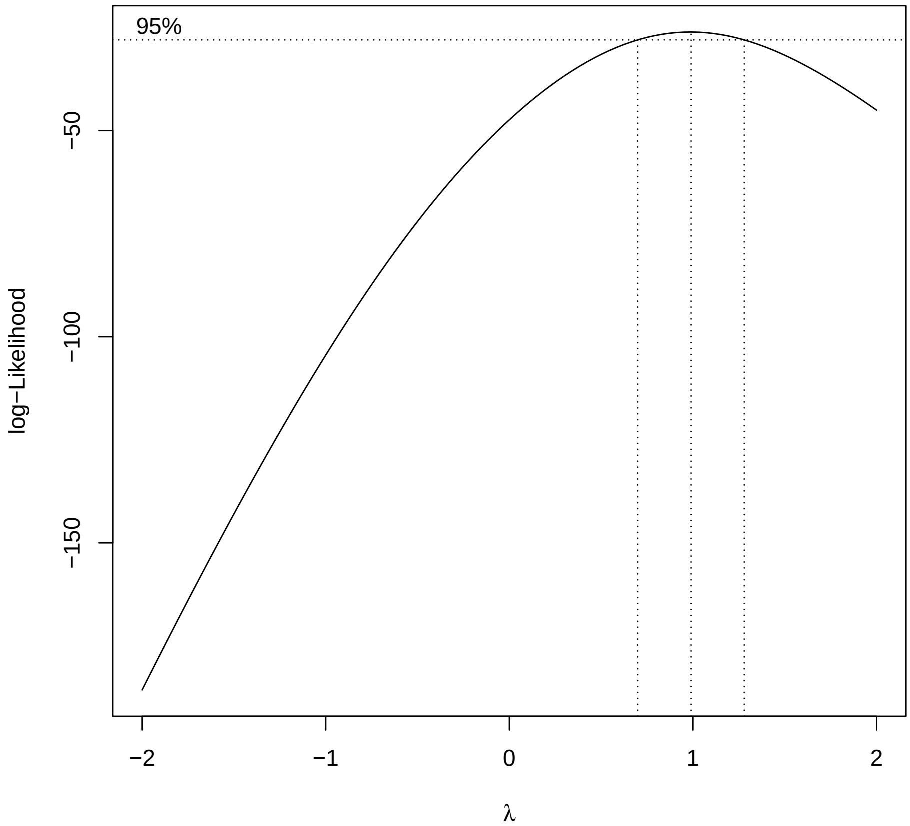
```
library(MASS)
occ<-read.csv("occ.csv")
head(occ)

##        occupation type income education prestige
## 1      accountant prof     62        86       82
## 2       architect prof     75        92       90
## 3 insurance.agent   wc     55        71       41
## 4     store.clerk   wc     29        50       16
## 5       carpenter   bc     21        23       33
## 6     electrician   bc     47        39       53
```

Please turn over for page 8

```
# model M0
M0<-lm(prestige ~ income, data=occ)
summary(M0)

##
## Call:
## lm(formula = prestige ~ income, data = occ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.566  -9.421   0.257   9.167  61.855
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4566     5.1901   0.473    0.638
## income        1.0804     0.1074  10.062 7.14e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.4 on 43 degrees of freedom
## Multiple R-squared:  0.7019,Adjusted R-squared:  0.695
## F-statistic: 101.3 on 1 and 43 DF,  p-value: 7.144e-13

# model M1
M1<-lm(prestige ~ income + education + type, data=occ)
summary(M1)

##
## Call:
## lm(formula = prestige ~ income + education + type, data = occ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.890  -5.740  -1.754   5.442  28.972
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.18503    3.71377  -0.050  0.96051
## income         0.59755    0.08936   6.687 5.12e-08 ***
## education      0.34532    0.11361   3.040  0.00416 **
## typeprof      16.65751    6.99301   2.382  0.02206 *
## typewc       -14.66113    6.10877  -2.400  0.02114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.744 on 40 degrees of freedom
## Multiple R-squared:  0.9131,Adjusted R-squared:  0.9044
## F-statistic:   105 on 4 and 40 DF,  p-value: < 2.2e-16
```

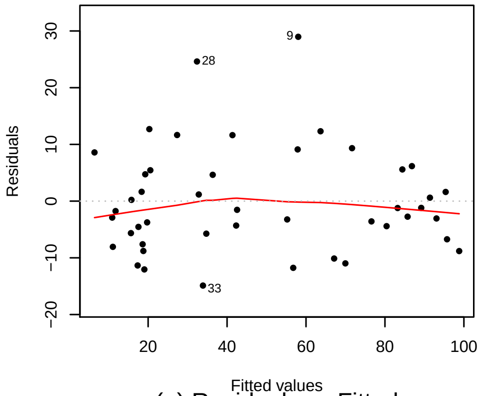
Please turn over for page 9

```r
boxcox(M1)
```

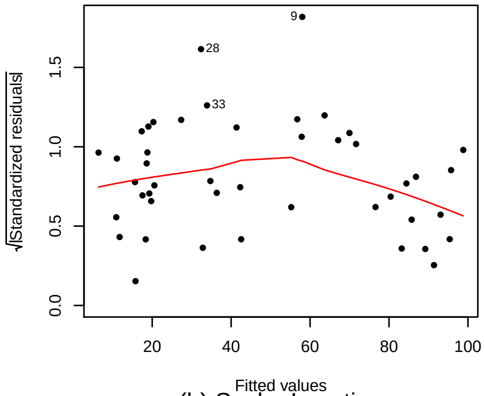


Please turn over for page 10

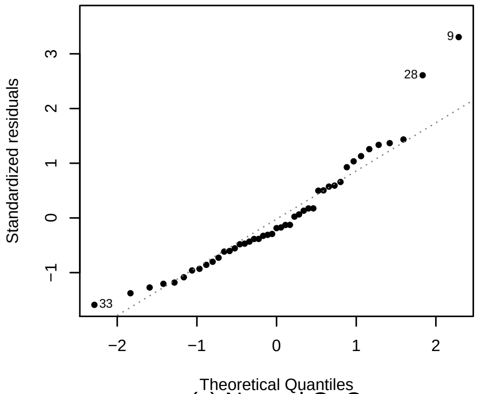
```r
plot(M1,which=c(1,3,2))
plot(stdres(M1) ~ occ$education)
plot(stdres(M1) ~ occ$income   )
plot(stdres(M1) ~ occ$type     )
```



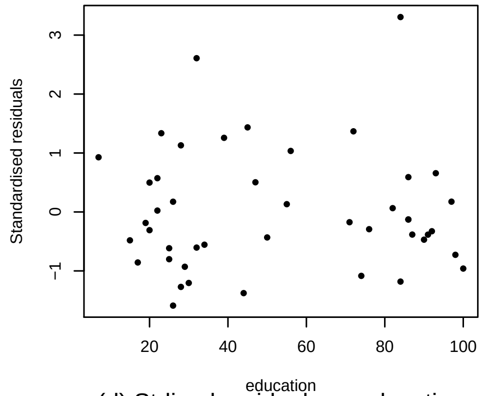
(a) Residuals vs Fitted



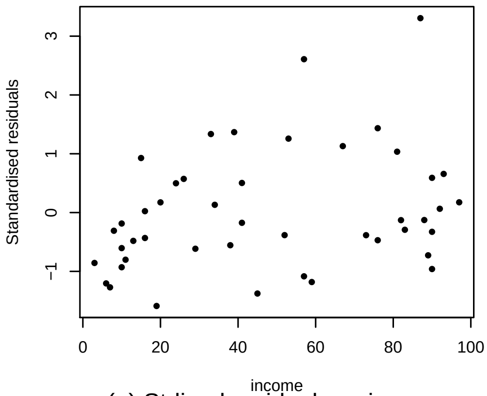
(b) Scale−Location



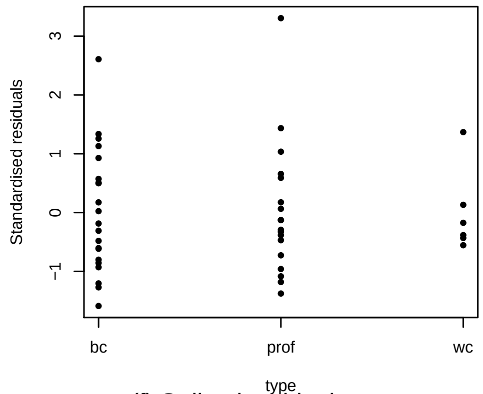
(c) Normal Q−Q



(d) Stdised residuals vs education



(e) Stdised residuals vs income



(f) Stdised residuals vs type

Please turn over for page 11

```
predict(M1
    ,newdata=data.frame(education=60,income=80,type="prof")
    ,interval="prediction")

##        fit    lwr      upr
## 1 84.99536 63.629 106.3617
```

```
# model M2
M2<-lm(prestige ~ education + income, data=occ)
anova(M1,M2)

## Analysis of Variance Table
##
## Model 1: prestige ~ income + education + type
## Model 2: prestige ~ education + income
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     40 3798.0
## 2     42 7506.7 -2   -3708.7 19.53 1.208e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Please turn over for page 12

Appendix B: R output for Question 5

```
# dataset
satis

##     x   n accom contact
## 1 100 219 tower      lo
## 2 100 181 tower      hi
## 3  62 177 house      lo
## 4 104 339 house      hi

# outcome: 2 column matrix with rows (x, n-x)
y <- cbind(satis$x, satis$n-satis$x)
```

```
# model L0
L0 <- glm(y ~ accom + contact, data=satis, family="binomial")
summary(L0)

##
## Call:
## glm(formula = y ~ accom + contact, family = "binomial", data = satis)
##
## Deviance Residuals:
##       1        2        3        4
## -0.9876   1.0868   1.1740  -0.8492
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.71580    0.10557  -6.780 1.20e-11 ***
## accomtower   0.76443    0.14067   5.434 5.51e-08 ***
## contactlo   -0.08882    0.14161  -0.627    0.531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.516  on 3  degrees of freedom
## Residual deviance:  4.256  on 1  degrees of freedom
## AIC: 33.388
##
## Number of Fisher Scoring iterations: 3
```

```
# model L1
L1 <- glm(y ~ accom * contact, data=satis, family="binomial")
summary(L1)
```

```
##
## Call:
## glm(formula = y ~ accom * contact, family = "binomial", data = satis)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.8152     0.1178  -6.922 4.46e-12 ***
## accomtower                1.0259     0.1903   5.391 7.01e-08 ***
## contactlo                 0.1974     0.1967   1.003   0.3156
## accomtower:contactlo     -0.5821     0.2819  -2.065   0.0389 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  3.4516e+01  on 3  degrees of freedom
## Residual deviance: -3.8636e-14  on 0  degrees of freedom
## AIC: 31.132
##
## Number of Fisher Scoring iterations: 2
```

Please turn over for page 14

Appendix C

```
# critical values (crit_val) of the Chi-squared dist
# for various degrees of freedom (df) and levels of significance (alpha)
df <- 1:4
alpha <- c(0.025,0.05)
chi_tab <- expand.grid(df=df, alpha=alpha)
chi_tab$crit_val <- qchisq(1-chi_tab$alpha, df=chi_tab$df)
chi_tab

##   df alpha  crit_val
## 1  1 0.025  5.023886
## 2  2 0.025  7.377759
## 3  3 0.025  9.348404
## 4  4 0.025 11.143287
## 5  1 0.050  3.841459
## 6  2 0.050  5.991465
## 7  3 0.050  7.814728
## 8  4 0.050  9.487729
```