THE UNIVERSITY of ADELAIDE

Examination in School of Mathematical Sciences

Semester 2, 2018

104843 STATS 2107 Statistical Modelling and Inference II

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

NUMBER OF QUESTIONS: 6 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.
- Formulae sheets are provided.
- Calculators without remote communications capability are allowed.
- English and foreign-language dictionaries may be used.

DO NOT COMMENCE WRITING UNTIL INSTRUCTED TO DO SO.

1. Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed (*i.i.d.*) random variables with probability density function $f(y;\theta)$ for a real scalar parameter $\theta \in \Theta$, where Θ denotes the parameter space.

   Let $T = T(Y_1, Y_2, \ldots, Y_n)$ be an estimator for θ.

(a) Define the *mean squared error*, $\text{MSE}_T(\theta)$, of T. [1 marks]

(b) Define the *bias*, $b_T(\theta)$, of T. [1 marks]

(c) Prove that

$$\text{MSE}_T(\theta) = \text{Var}(T) + b_T(\theta)^2.$$

[3 marks]

(d) Suppose $Y_1, Y_2, \ldots, Y_n$ are independent identically distributed (i.i.d.) $N(\mu, \sigma^2)$ random variables and let

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

be an estimator for μ. Calculate $\text{MSE}_{\bar{Y}}(\mu)$.

[4 marks]

[Total: 9]

2.

(a) Carefully define the t-distribution with k degrees of freedom. [3 marks]

(b) Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, then prove that

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

You may assume that $\bar{Y}$ and S^2 are independent. [3 marks]

(c) Let $Z \sim N(0,1)$. Show that the moment generating function of Z^2 is

$$M_{Z^2}(t) = (1-2t)^{-\frac{1}{2}}, \quad t < 1/2.$$

[4 marks]

Please turn over for page 3

(d) Suppose $Z_1, Z_2, \ldots, Z_k$ are independent and identically distributed $N(0,1)$ random variable and let

$$X = \sum_{i=1}^{k} Z_i^2.$$

Show that the moment generating function of X is

$$M_X(t) = (1-2t)^{-\frac{k}{2}}, \quad t < 1/2.$$

[3 marks]

(e) Hence, or otherwise, show that if

$$X \sim \chi_k^2,$$

then

$$\mathsf{E}[X] = k \text{ and } \mathsf{Var}(X) = 2k.$$

[5 marks]

[Total: 18]

3. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent with $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$.

(a) Consider

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

Show that S_{xy} can be written as

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) y_i.$$

[3 marks]

(b) Given that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

find the constants $a_1, a_2, \ldots, a_n$, such that

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i.$$

[2 marks]

(c) Prove that

$$\mathsf{E}[\hat{\beta}_1] = \beta_1 \text{ and } \mathsf{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}},$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x}) x_i.$$

[6 marks]

[Total: 11]

4. An analysis of the effect of displacement (`displ`) and drive type (`drv`) on the city fuel efficiency (`cty`) for 38 popular models of car was performed in R. The commands and output are given in Appendix A.

   The displacement of a car is the volume of the cylinders, while the drive is the type of drive, in this case we have just three levels - front-wheel drive, rear-wheel drive and four-wheel drive.

   Three models are fitted:

- `cty` on `displ` (Model 1 - identical regression)
- `cty` on `displ` and `drv` (Model 2 - parallel regression)
- `cty` on `displ` and `drv` with interaction (Model 3 - separate regression)

(a) Consider the scatterplot of city fuel efficiency against displacement given in Figure 1. Describe the relationship. [3 marks]

(b) Consider the separate regression model. Write down the line of best fit for the relationship between displacement and city fuel efficiency for rear-wheel drive cars. [2 marks]

(c) Test for a statistically significant interaction term in the separate regression model at the 5% significance level. Remember to include the null and alternative hypotheses, the value of the test statistic, the P-value and your conclusion. [4 marks]

(d) Using the Akaike's Information Criterion, which model fits the data the best? Justify your answer. [2 marks]

Please turn over for page 5

(e) Assess the assumptions of the linear model used in the separate regression model. The plots given in Figure 2 may be used where appropriate. [4 marks]

[Total: 15]

5. Suppose $y_1, y_2, \ldots, y_n$ are independent Poisson observations with parameter λ, $\lambda > 0$. That is, for $i = 1, 2, \ldots, n$,

$$f(y_i; \lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}, \quad y_i > 0.$$

(a) Write down the likelihood. [1 marks]

(b) Write down the log-likelihood. [1 marks]

(c) Find the maximum likelihood estimate of λ, $\hat{\lambda}$. [3 marks]

(d) Find the Fisher information. [3 marks]

(e) Let $\phi = \log(\lambda)$. Write down the maximum likelihood estimate, $\hat{\phi}$. [1 marks]

[Total: 9]

6. Haemophilia is a X-chromosome linked, recessive disorder. Suppose a woman has a haemophiliac brother, her father is normal, and her mother is a carrier. Let

$$\theta = \begin{cases} 1 & \text{if the woman is a carrier,} \\ 0 & \text{otherwise.} \end{cases}$$

It follows from genetic considerations that the prior distribution is

$$p(\theta) = \begin{cases} \frac{1}{2} & \text{if } \theta = 1, \\ \frac{1}{2} & \text{if } \theta = 0. \end{cases}$$

(a) Suppose the woman has two sons, of which neither have haemophilia. Find the probability the woman is a carrier. [5 marks]

(b) Suppose the woman has a third son. Given that the first two sons are not haemophiliacs, what is the probability that the third son is not a haemophiliac? [3 marks]

[Total: 8]

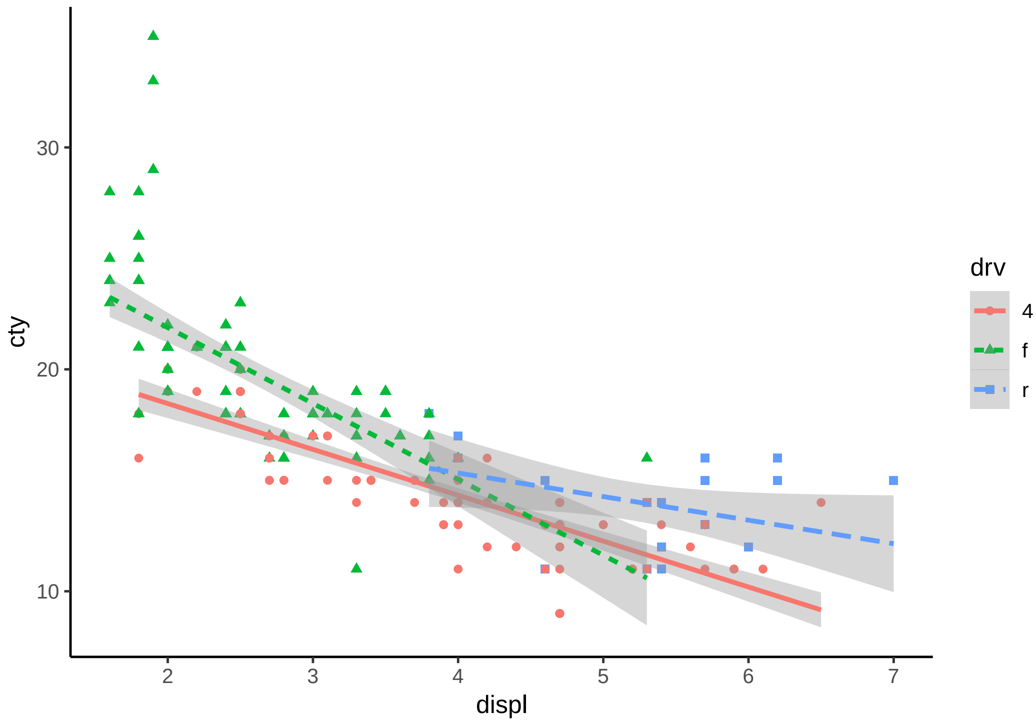

Figure 1: Scatterplot of Fuel efficiency against displacement for the MPG dataset. Colour and shape of points indicates drive (type).

Appendix A

Load the data

```
library(tidyverse)
data(mpg)
theme_set(theme_classic())
```

Visualise data

```
mpg %>%
  ggplot(aes(displ, cty, col = drv, shape = drv)) +
  geom_point() +
  geom_smooth(method = "lm", aes(linetype = drv))
```

Fit models

Please turn over for page 7

```r
identical  <- lm(cty ~ displ, data = mpg)
parallel  <- lm(cty ~ displ + drv, data = mpg)
separate  <- lm(cty ~ displ * drv, data = mpg)
```

Model Coefficients

```r
summary(separate)
```

```
##
## Call:
## lm(formula = cty ~ displ * drv, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4363 -1.2957 -0.0863  1.1203 12.7768
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.5914     0.8136  27.768  < 2e-16 ***
## displ        -2.0663     0.1958 -10.554  < 2e-16 ***
## drvf          6.1284     1.1632   5.269 3.18e-07 ***
## drvr         -3.0124     3.1043  -0.970 0.332872
## displ:drvf   -1.3529     0.3696  -3.661 0.000313 ***
## displ:drvr    1.0039     0.6048   1.660 0.098285 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.252 on 228 degrees of freedom
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7201
## F-statistic: 120.9 on 5 and 228 DF,  p-value: < 2.2e-16
```

```r
anova(separate)
```

```
## Analysis of Variance Table
##
## Response: cty
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## displ       1 2691.06 2691.06 530.7574 < 2.2e-16 ***
## drv         2  277.99  138.99  27.4136 2.144e-11 ***
## displ:drv   2   95.28   47.64   9.3963 0.0001199 ***
```

Please turn over for page 8

```
## Residuals 228 1156.01    5.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(separate, parallel)
```

```
## Analysis of Variance Table
##
## Model 1: cty ~ displ * drv
## Model 2: cty ~ displ + drv
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    228 1156.0
## 2    230 1251.3 -2   -95.283 9.3963 0.0001199 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Selection

```r
AIC(identical, parallel, separate)
```

```
##           df      AIC
## identical  3 1109.336
## parallel   5 1066.391
## separate   7 1051.857
```

Assumption checking

Please turn over for page 9

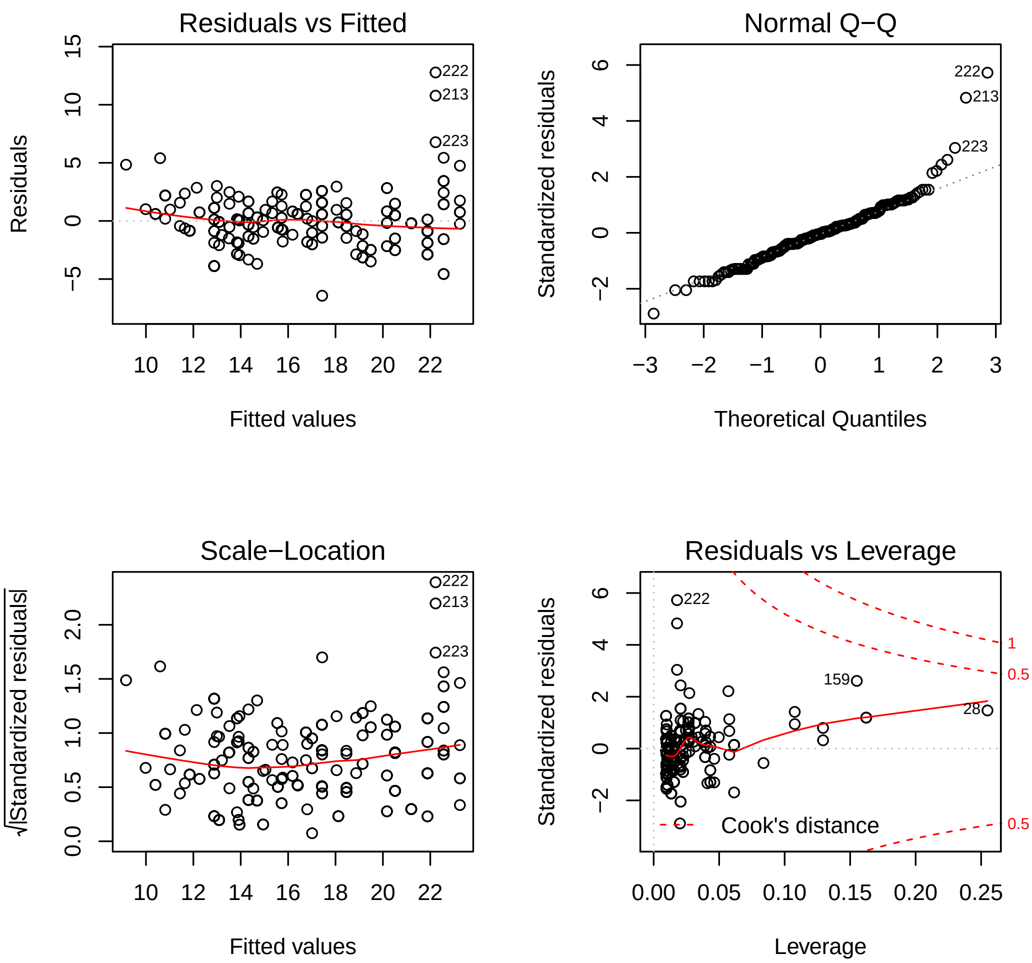

Figure 2: Assumption plots of the separate model for the MPG dataset.

Please turn over for page 10

Appendix B

Distribution	Probability mass function / probability density function	Expectation	Variance
Binomial	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$	np	$np(1-p)$
Geometric	$p(x) = p(1-p)^{x-1}$ for $x = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$	λ	λ
Uniform	$f(x) = \frac{1}{b-a}$ for $a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$f(x) = \lambda e^{-\lambda x}$ for $x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x > 0$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Normal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)(x-\mu)^2}$ for $-\infty < x < \infty$	μ	σ^2
Beta	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma\beta} \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $0 < \theta < 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$