Examination in School of Mathematical Sciences

Semester 2, 2017

104843 STATS 2107 Statistical Modelling & Inference II

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

NUMBER OF QUESTIONS: 5 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 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, $MSE_T(\theta)$, of T.

   Solution:

$$MSE_T(\theta) = E[(T - \theta)^2].$$

```
## 1 for definition
## Core: 1
```

   (b) Define the bias, $b_T(\theta)$, of T.

   Solution:

$$b_T(\theta) = E[T] - \theta.$$

```
## 1 for definition
## Core: 1
```

   (c) Prove that

$$MSE_T(\theta) = var(T) + b_T(\theta)^2.$$

   Solution:

$$\begin{aligned}
MSE_T(\theta) &= E[(T - \theta)^2] \\
&= E[(T - E[T] + E[T] - \theta)^2] \\
&= E[(T - E[T])^2] + E[(E[T] - \theta)^2] + 2E[(T - E[T])(E[T] - \theta)] \\
&= var(T) + E[b_T(\theta)^2] + 2(E[T] - \theta)E[T - E[T]] \\
&= var(T) + b_T(\theta)^2 + 2(E[T] - \theta)0 \\
&= var(T) + b_T(\theta)^2.
\end{aligned}$$

```
## 4 for working
## Core: 4
```

   (d) Suppose $Y_1, Y_2, \ldots, Y_n$ are $i.i.d.$ Bernoulli random variables with probability of success $0 \le p \le 1$, and that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is to be used as an estimator for p.

      (i) Show that $\bar{Y}$ is an unbiased estimator for p.

      Solution:

$$
\begin{aligned}
E[\bar{Y}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[Y_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} p \qquad \text{Bernoulli is binomial with n = 1 and using formulae sheet.} \\
&= p.
\end{aligned}
$$

```
## 2 for working
## Core: 2
```

(ii) Calculate $MSE_{\bar{Y}}(p)$.
Solution:

$$
\begin{aligned}
MSE_{\bar{Y}}(p) &= var(\bar{Y}) + b_{\bar{Y}}(p)^2 \\
&= var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) + 0 \qquad \text{as unbiased} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} var(Y_i) \qquad \text{as independent} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) \\
&= \frac{p(1-p)}{n}.
\end{aligned}
$$

```
## 3 for working
## Core: 3
```

[11 marks]

2. 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.

   (a) Define a $100(1-\alpha)\%$ confidence interval for the parameter θ.
   Solution:
   A random interval (L, U) such that

$$P(L \leq \theta \leq U) = 1 - \alpha.$$

```
## 1 for definition
## Core: 1
```

(b) Suppose that $Y_1, Y_2, \ldots, Y_n$ are *i.i.d.* $N(\mu, \sigma^2)$. Let c_1, c_2 be such that

$$P(c_1 < X < c_2) = 1 - \alpha,$$

where

$$X \sim \chi^2_{n-1}.$$

(i) Prove that the interval

$$\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1} \right),$$

where S^2 is the sample variance, is a $100(1-\alpha)\%$ confidence interval for σ^2. You may assume that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Solution:

$$\begin{aligned} & P\left(\frac{(n-1)S^2}{c_2} < \sigma^2 < \frac{(n-1)S^2}{c_1} \right) \\ = \; & P\left(\frac{1}{c_2} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{c_1} \right) \\ = \; & P\left(c_1 < \frac{(n-1)S^2}{\sigma^2} < c_2 \right) \\ = \; & 1 - \alpha. \end{aligned}$$

since,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

```
## 2 for working. Advanced as not seen before
## Adv: 2
```

(ii) In an experiment, 20 observations were made and the sample standard deviation was measured as 4. Calculate a 95% confidence interval for σ^2 of the form

$$(0, upper).$$

You may assume that the observations were randomly sampled from a normal distribution. The following R commands and output may be used.

```
qchisq(0.05, 19)
## [1] 10.11701
qchisq(0.05, 19, lower.tail = FALSE)
## [1] 30.14353
qchisq(0.025, 19)
## [1] 8.906516
qchisq(0.025, 19, lower.tail = FALSE)
## [1] 32.85233
```

Solution:

Set $c_2 = \infty$ and c_1 such that $P(X < c_1) = 0.05$, i.e., $c_1 = 10.11701$.
This gives

```
s  <- 4
c1  <- qchisq(0.05, 19)
(20 - 1) * s^2 / c1
## [1] 30.04839
```

Final interval is

$$(0, 30.0484)$$

```
## 1 for getting c1 and c2; 1 for working
## Core: 2
```

(c) Prove that, even though S^2 is unbiased for σ, S is not unbiased for σ.
Hint: You may assume that $var(S) > 0$.
Solution:

$$\begin{aligned} 0 < var(S) &= E[S^2] - E[S]^2 \\ &= \sigma^2 - E[S]^2 \\ \Rightarrow E[S]^2 &< \sigma^2 \\ \Rightarrow E[S] &< \sigma \end{aligned}$$

Hence $E[S] \neq \sigma$ and so S is biased.

```
## 4 for working
## Adv: 4
```

[9 marks]

3. Consider the multiple regression model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response random variables, X is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors with $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$.

(a) Prove that if $X_{n\times p}$ is a matrix with linearly independent columns then the symmetric, $p \times p$ matrix $X^T X$ is invertible.

Solution:

Since the determinant of a matrix is the product of its eigenvalues, it is sufficient to check that 0 is not an eigenvalue of $X^T X$. That is, we must show that there exists no $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $X^T X \boldsymbol{\alpha} = \mathbf{0}$. Now observe,

$$
\begin{aligned}
& (X^T X)\boldsymbol{\alpha} = \mathbf{0} \\
\Rightarrow\quad & \boldsymbol{\alpha}^T (X^T X)\boldsymbol{\alpha} = 0 \\
\Rightarrow\quad & (X\boldsymbol{\alpha})^T (X\boldsymbol{\alpha}) = 0 \\
\Rightarrow\quad & \|X\boldsymbol{\alpha}\|^2 = 0 \\
\Rightarrow\quad & X\boldsymbol{\alpha} = \mathbf{0} \\
\Rightarrow\quad & \boldsymbol{\alpha} = \mathbf{0} \text{ since the columns of } X \text{ are linearly independent.}
\end{aligned}
$$

```
## 4 for working
## Core: 4
```

(b) Prove that

$$(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T (X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}) = 0,$$

where

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

You may assume that the columns of X are linearly independent.

Solution:

$$
\begin{aligned}
(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T (X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}) &= (\boldsymbol{y} - X(X^T X)^{-1} X^T \boldsymbol{y})^T X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \left((I - X(X^T X)^{-1} X^T)\boldsymbol{y}\right)^T X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \boldsymbol{y}^T (I - X(X^T X)^{-1} X^T) X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= 0.
\end{aligned}
$$

```
## 5 for working
## Adv: 5
```

(c) Hence, prove that

$$\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 = \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 + \|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|^2$$

Solution:

$$\begin{aligned}
\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 &= \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|^2 \\
&= \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 + \|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|^2 + 2(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}) \\
&= \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 + \|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|^2 && \text{from (a)}
\end{aligned}$$

```
## 3 for working
## Adv: 3
```

(d) Hence, show that least squares estimates are given by

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Solution:

Since $\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|^2 \geq 0$, and using the result from part (b), we have

$$\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \geq \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2$$

Hence, the minimal value of the sum of squares is $\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2$, and this is only achieved when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

```
## 4 for working
## Adv: 4
```

[16 marks]

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

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

(a) Write down the likelihood.

Solution:

$$\prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}$$

```
## Need to simplify for mark
## Core: 1
```

(b) Write down the log-likelihood.

Solution:

$$n \log(\lambda) - \lambda \sum_{i=1}^{n} y_i$$

```
## 1 for working
## Core: 1
```

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

Solution:

Differentiate the log-likelihood w.r.t. λ

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i$$

Set equal to zero and solve:

$$\begin{aligned} \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} y_i = 0 \\ \Rightarrow \frac{n}{\hat{\lambda}} = \sum_{i=1}^{n} y_i \\ \Rightarrow \hat{\lambda} = \frac{1}{\bar{y}}. \end{aligned}$$

```
## 4 for working
## Core: 4
```

(d) Find the Fisher information.

Solution:

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \lambda^2} &= -\frac{n}{\lambda^2} \\ I_\lambda &= E\left[-\frac{\partial^2 \ell}{\partial \lambda^2}\right] \\ &= E\left[\frac{n}{\lambda^2}\right] \\ &= \frac{n}{\lambda^2}. \end{aligned}$$

```
## 3 for working
## Core: 3
```

(e) The following observations were made from an exponential distribution:

$$0.1, 0.2, 0.3, 0.5, 0.6, 0.1, 0.2, 0.2, 0.3, 0.2.$$

Calculate a 95% confidence interval for λ. You may assume that $P(Z < 1.96) = 0.975$, where $Z \sim N(0, 1)$.

Solution:

```
y <- c(0.1, 0.2, 0.3, 0.5, 0.6, 0.1, 0.2, 0.2, 0.3, 0.2)

n  <- length(y)

lambda  <- 1 / mean(y)

I <- n / lambda^2

lwr <- lambda - 1.96 * sqrt(1 / I)
upr <- lambda + 1.96 * sqrt(1 / I)
c(lwr, upr)
## [1] 1.408124 5.999283
```

```
## 4 for working
## Adv: 4
```

(f) An alternative form of the exponential distribution is

$$f(y_i;\beta) = \frac{1}{\beta}e^{-y_i/\beta}, y_i > 0, \beta > 0.$$

Give an expression for the maximum likelihood estimate of β.

Solution:

$$\beta = \frac{1}{\lambda}$$
$$\Rightarrow \hat{\beta} = \frac{1}{\hat{\lambda}} = \bar{y}.$$

```
## 1 for answer; 1 for justification
## Adv: 2
```

[15 marks]

5. An analysis of the effect of displacement (`displ`) and class (`class`) on the highway fuel efficiency (`hwy`) 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 volume of the cylinders, while the class is the type of car, in this case, we have just two levels - midsize and SUV. Of note, is the fact that the minimum displacement of SUV's is 2.5 litres.

   (a) Consider the scatterplot of highway fuel efficiency against displacement given in Figure 1. Describe the relationship.

   Solution:

   There is a weak negative linear relationship between hwy and displ. The two lines look parallel with midsize having a larger intercept that suv.

```
## 1 for weak; 1 for negative; 1 for linear; 1 for comparison.
## Core: 4
```

(b) Consider the separate regression model. Write down the two lines of best fit for the relationship between displacement and highway fuel efficiency: one for midsize cars and one for SUV cars.

Solution:

For midsize, we have

$$hty = 31.8013 - 1.5430 \times displ$$

For SUV, we have

$$hty = (31.8013 - 5.5397) + (-1.5430 - 0.2819) \times displ = 26.2616 - 1.8249 \times displ$$

```
## 1 for each equation; 1 for working.
## Core: 3
```

(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.

Solution:

$$H_0 : \beta_3 = 0$$
$$H_0 : \beta_3 \neq 0,$$

where β_3 is the coefficient associated with the interaction term.

The value of the test statistic is -0.531.

The P-value is 0.59646.

We retain the null hypothesis at the 5% significance level and conclude that there is no evidence that the interaction term is necessary for the model.

```
## 1 for hypo; 1 for test; 1 for pv; 2 for conclusion.
## Core: 5
```

(d) Calculate a 95% confidence interval for the slope in the identical model. The following R command may be useful. Interpret the confidence interval in context.

```r
qt(0.975, 101)
## [1] 1.983731
```

Solution:

```r
b1 <- -3.4132
se  <- 0.2676
t  <- qt(0.975, 101)
lwr  <- b1 - t * se
upr  <- b1 + t * se
c(lwr, upr)
```

Please turn over for page 11

```
## [1] -3.944046 -2.882354
```

We are 95% confident that if displacement increases by 1 litre, then we expect the highway fuel efficiency to decrease by 2.882354 to 3.944046 miles per gallon.

```
## 1 for each endpoint; 1 for interpretation.
## Core: 3
```

(e) Assess the assumptions of the linear model used in the parallel model. The plots given in Figure 2 may be used where appropriate.

Solution:

Linearity: Residual versus fitted (top left) shows random scatter so reasonable. Although of note is the five points in the SUV group with the minimum of 2.5 litres.

Homoscedascity: Standardised residual versus fitted (bottom left) shows equal spread as move from left to right so reasonable.

Normality: Residual QQ-plot (top right) is roughly linear so reasonable.

Independence: The fuel efficiency of one car should not affect the fuel efficiency of the other cars so this is reasonable.

```
## 1 for each assumption.
## Core: 4
```

[19 marks]

Solution:

Q	adv	core	total	p	type
1	0	11	11	1.00	MSE
2	6	3	9	0.33	CI
3	12	4	16	0.25	Linear models
4	6	9	15	0.60	Likelihood
5	0	19	19	1.00	MLR

```
## [1] 46
## [1] 24
## [1] 0.6571429
## [1] 70
```

Appendix A

```
## load libraries ----
library(tidyverse)

## Switch off significant stars - sorry folks - said I would ----
options(show.signif.stars=FALSE)

## Load MPG datasets ----
data(mpg)

## Filter for just midsize and SUV cars ----
mpg  <- mpg %>%
  filter(class %in% c("midsize", "suv"))

## Look at relationship between fuel efficiency and displacement
ggplot(mpg, aes(x = displ, hwy, col = class)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Displacement (litres)",
       y = "Highway fuel efficiency (miles per gallon)") +
  theme(legend.position = "top")
```

```
## Identical regression model ----
identical.model  <- lm(hwy ~ displ, data = mpg)
summary(identical.model)

##
## Call:
## lm(formula = hwy ~ displ, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8605 -1.8725  0.1395  2.3221  8.1874
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.9028     1.0780   32.38   <2e-16
## displ        -3.4132     0.2676  -12.75   <2e-16
##
## Residual standard error: 3.256 on 101 degrees of freedom
## Multiple R-squared:  0.6169,Adjusted R-squared:  0.6131
## F-statistic: 162.7 on 1 and 101 DF,  p-value: < 2.2e-16
```

Please turn over for page 13

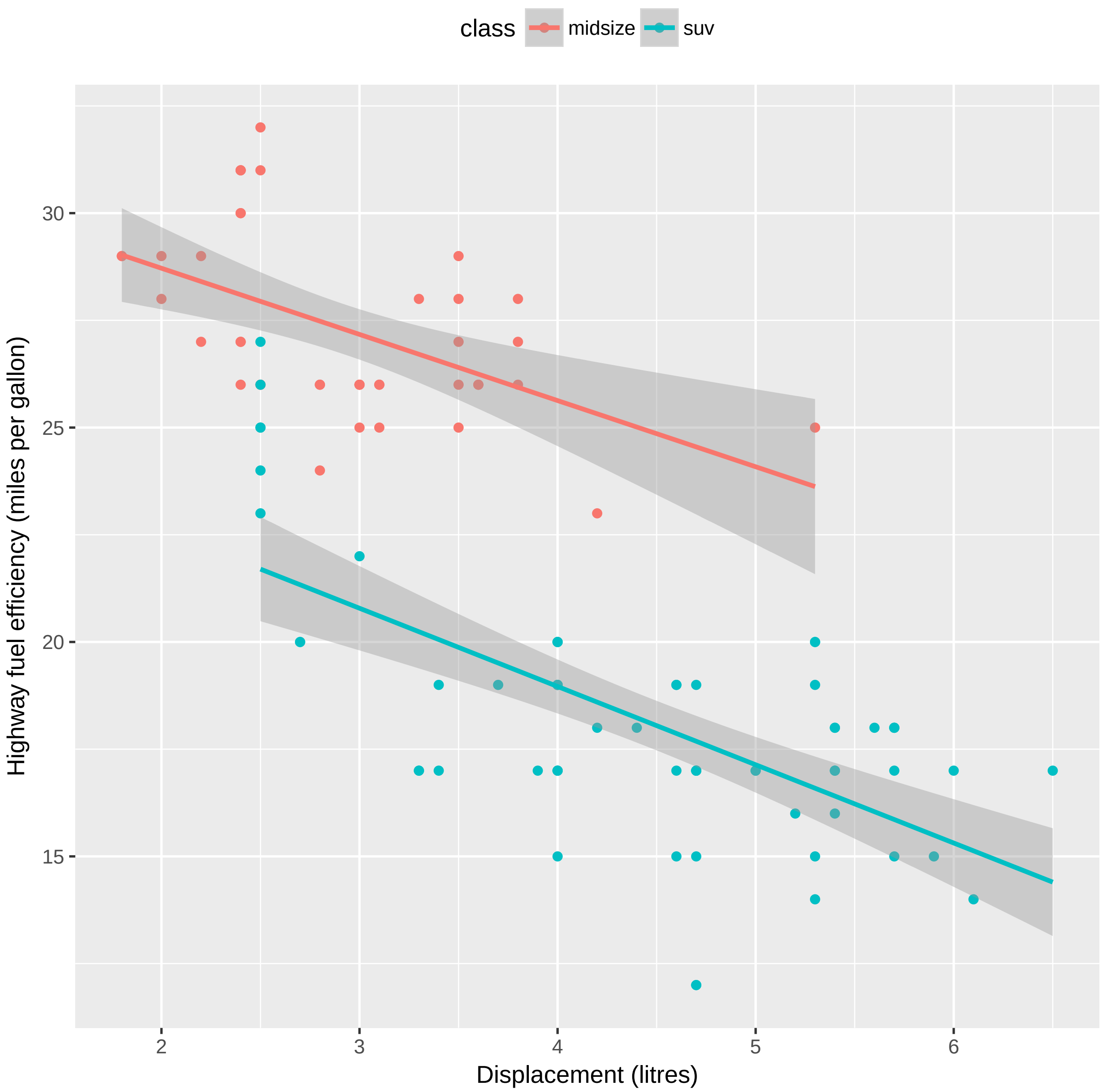

Figure 1: Scatterplot of highway fuel efficiency against displacement for midsize and SUV cars in MPG dataset.

Please turn over for page 14

```r
## Parallel regression model ----
parallel.model  <- lm(hwy ~ displ + class, data = mpg)
summary(parallel.model)

##
## Call:
## lm(formula = hwy ~ displ + class, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7003 -1.2284 -0.2284  1.5318  5.4273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.4358     0.7285  44.524  < 2e-16
## displ        -1.7602     0.2224  -7.915 3.46e-12
## classsuv     -6.4627     0.5447 -11.865  < 2e-16
##
## Residual standard error: 2.109 on 100 degrees of freedom
## Multiple R-squared:  0.8409,Adjusted R-squared:  0.8377
## F-statistic: 264.3 on 2 and 100 DF,  p-value: < 2.2e-16
```

```r
## Separate regression model ----
separate.model  <- lm(hwy ~ displ * class, data = mpg)
summary(separate.model)

##
## Call:
## lm(formula = hwy ~ displ * class, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6846 -1.3344 -0.2321  1.4848  5.3006
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      31.8013     1.4006  22.706  < 2e-16
## displ            -1.5430     0.4658  -3.313  0.00129
## classsuv         -5.5397     1.8215  -3.041  0.00302
## displ:classsuv  -0.2819     0.5307  -0.531  0.59646
##
## Residual standard error: 2.117 on 99 degrees of freedom
## Multiple R-squared:  0.8414,Adjusted R-squared:  0.8366
```

Please turn over for page 15

```
## F-statistic:   175 on 3 and 99 DF,  p-value: < 2.2e-16
```

```
## Plots for assumption checking ----
tmp  <- par(mfrow = c(2,2))
plot(parallel.model)
par(tmp)
```

Please turn over for page 16

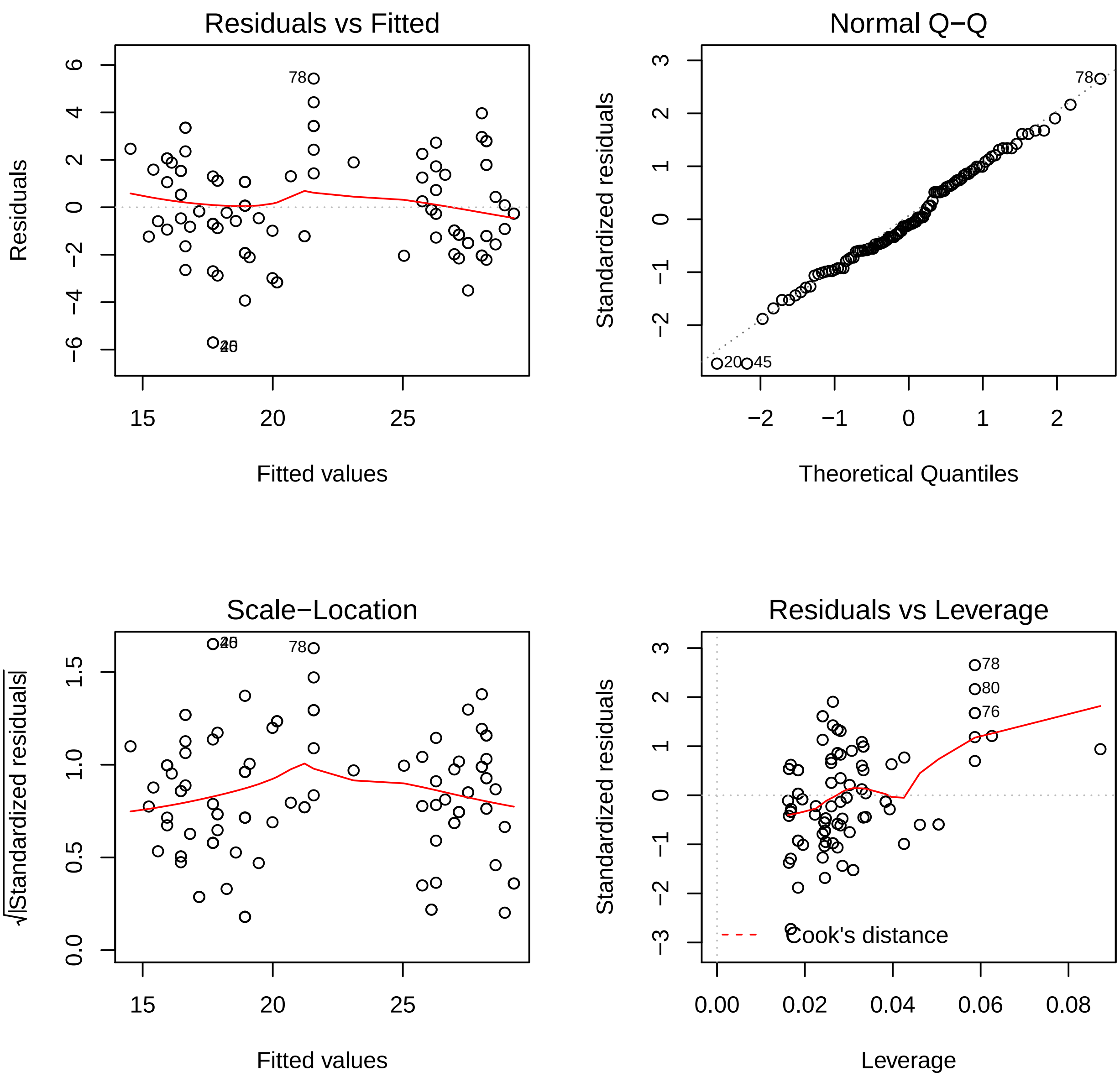

Figure 2: Plots to check assumptions for the parallel regression model

Please turn over for page 17

Appendix B

Binomial Distribution

- $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$
- $E(X) = np$
- $var(X) = np(1-p)$

Geometric Distribution

- $p(x) = p(1-p)^{x-1}$ for $x = 1, 2, \ldots$
- $E(X) = \frac{1}{p}$
- $var(X) = \frac{1-p}{p^2}$

Poisson Distribution

- $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$
- $E(X) = \lambda$
- $var(X) = \lambda$

Uniform Distribution

- $f(x) = \frac{1}{b-a}$ for $a < x < b$
- $E(X) = \frac{a+b}{2}$
- $var(X) = \frac{(b-a)^2}{12}$

Exponential Distribution

- $f(x) = \lambda e^{-\lambda x}$ for $x > 0$
- $E(X) = \frac{1}{\lambda}$
- $var(X) = \frac{1}{\lambda^2}$

Gamma Distribution

- $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x > 0$
- $E(X) = \frac{\alpha}{\lambda}$
- $var(X) = \frac{\alpha}{\lambda^2}$

Normal Distribution

- $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)(x-\mu)^2}$ for $-\infty < x < \infty$
- $E(X) = \mu$
- $var(X) = \sigma^2$