THE UNIVERSITY *of* ADELAIDE

Examination in School of Mathematical Sciences
Semester 2, 2018

104959 STATS 7107 Statistical Modelling and Inference PG

Official Reading Time:	10 mins
Writing Time:	180 mins
Total Duration:	190 mins

NUMBER OF QUESTIONS: 8 TOTAL MARKS: 100

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*, $\mathsf{MSE}_T(\theta)$, of T. [1 marks]

[1:0] Mark Scheme: 1 for definition

Solution:

$$\mathsf{MSE}_T(\theta) = \mathsf{E}[(T-\theta)^2].$$

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

[1:0] Mark Scheme: 1 for definition

Solution:

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

(c) Prove that

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

[3 marks]

[3:0] Mark Scheme: 3 for working

Solution:

$$\begin{aligned}
\mathsf{MSE}_T(\theta) &= \mathsf{E}[(T-\theta)^2] \\
&= \mathsf{E}[(T-\mathsf{E}[T]+\mathsf{E}[T]-\theta)^2] \\
&= \mathsf{E}[(T-\mathsf{E}[T])^2] + \mathsf{E}[(\mathsf{E}[T]-\theta)^2] + 2\mathsf{E}[(T-\mathsf{E}[T])(\mathsf{E}[T]-\theta)] \\
&= \mathsf{Var}(T) + \mathsf{E}[b_T(\theta)^2] + 2(\mathsf{E}[T]-\theta)\mathsf{E}[T-\mathsf{E}[T]] \\
&= \mathsf{Var}(T) + b_T(\theta)^2 + 2(\mathsf{E}[T]-\theta)0 \\
&= \mathsf{Var}(T) + b_T(\theta)^2.
\end{aligned}$$

(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 $\mathsf{MSE}_{\bar{Y}}(\mu)$.

[4 marks]

[0:4] Mark Scheme: 4 for working

Solution:

First calculate bias:

$$\begin{aligned} b_{\bar{Y}}(\mu) &= \mathsf{E}[\bar{Y}] - \mu \\ &= \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] - \mu \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[Y_i] - \mu \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu - \mu \\ &= \mu - \mu = 0. \end{aligned}$$

Next calculate $\mathsf{Var}(\bar{Y})$

$$\begin{aligned} \mathsf{Var}(\bar{Y}) &= \mathsf{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{Var}[Y_i] \quad \text{as independent} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Hence

$$\begin{aligned}\text{MSE}_{\bar{Y}}(\mu) &= \text{Var}(\bar{Y}) + b_{\bar{Y}}(\mu)^2 \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

[Total: 9]

Core: 5 Adv: 4

2\.

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

[3:0] Mark Scheme: 1 for Z, 1 for X, 1 for frac.

Solution:

Suppose $Z \sim N(0,1)$ and $X \sim \chi^2_k$ independently, and let

$$T = \frac{Z}{\sqrt{X/k}},$$

then T is said to have a t-distribution with k degrees of freedom.

(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]

[3:0] Mark Scheme: 3 for working

Solution:

$$
\begin{aligned}
\frac{\bar{Y}-\mu}{S/\sqrt{n}} &= \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\frac{\sigma/\sqrt{n}}{S/\sqrt{n}} \\
&= \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\Bigg/\sqrt{\frac{S^2}{\sigma^2}} \\
&= \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\Bigg/\sqrt{\frac{(n-1)S^2}{\sigma^2}(n-1)} \\
&= \frac{Z}{\sqrt{X/(n-1)}} \sim T_{n-1}
\end{aligned}
$$

(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]

[0:4] Mark Scheme: 4 for working

Solution:

$$
\begin{aligned}
M_{Z^2}(t) &= E[e^{tZ^2}] \\
&= \int_{-\infty}^{\infty} e^{tz^2}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{z^2(t-1/2)}dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-z^2(1-2t)/2}dz \\
&= \frac{1}{(1-2t)^{1/2}}\int_{-\infty}^{\infty} \frac{(1-2t)^{1/2}}{\sqrt{2\pi}}e^{-z^2(1-2t)/2}dz
\end{aligned}
$$

but the integral is one since it is the pdf of $N(0, \frac{1}{1-2t})$. So we have

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

(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]

[0:3] Mark Scheme: 3 for working

Solution:

$$\begin{aligned}
M_X(t) &= E\left\{\exp(tX)\right\} \\
&= E\left\{\exp\left(t\sum_{i=1}^{k} Z_i^2\right)\right\} \\
&= E\left\{\exp(tZ_1^2)\exp(tZ_2^2)\cdots\exp(tZ_k^2)\right\} \\
&= E\{\exp(tZ_1^2)\}E\{\exp(tZ_2^2)\}\cdots E\{\exp(tZ_k^2)\} \text{ (by independence)} \\
&= (1-2t)^{-1/2}(1-2t)^{-1/2}\cdots(1-2t)^{-1/2} \\
&= \frac{1}{(1-2t)^{k/2}}.
\end{aligned}$$

X has the chi-squared distribution with k degrees of freedom.

(e) Hence, or otherwise, show that if

$$X \sim \chi^2_k,$$

then

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

[5 marks]

[0:5] Mark Scheme: 5 for working

Solution:

$$\begin{aligned}
E[X] &= \left.\frac{d}{dt}M_x(t)\right|_{t=0} \\
&= \left.\frac{d}{dt}(1-2t)^{-k/2}\right|_{t=0} \\
&= \left.k(1-2t)^{-k/2-1}\right|_{t=0} \\
&= k.
\end{aligned}$$

$$\begin{aligned} E[X^2] &= \frac{d^2}{dt^2} M_x(t) \bigg|_{t=0} \\ &= \frac{d}{dt} k(1-2t)^{-k/2-1} \bigg|_{t=0} \\ &= k(k+2)(1-2t)^{-k/2-2} \big|_{t=0} \\ &= k(k+2). \end{aligned}$$

Now

$$\begin{aligned} \text{Var}(X) &= E[X^2] - E[X]^2 \\ &= k(k+2) - k^2 \\ &= 2k. \end{aligned}$$

[Total: 18]

Core: 6 Adv: 12

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]

[0:3] Mark Scheme: 3 for working

Solution:

$$\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \\
&= \sum_{i=1}^{n} y_i(x_i - \bar{x}) - \sum_{i=1}^{n} \bar{y}(x_i - \bar{x}) \\
&= \sum_{i=1}^{n} y_i(x_i - \bar{x}) - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \sum_{i=1}^{n} y_i(x_i - \bar{x}) \qquad\qquad \text{as } \sum_{i=1}^{n}(x_i - \bar{x}) = 0.
\end{aligned}$$

(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]

[0:2] Mark Scheme: 2 for working

Solution:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{S_{xx}} \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})Y_i}{S_{xx}} \\
&= \sum_{i=1}^{n} a_i Y_i, \quad \text{where } a_i = \frac{(x_i - \bar{x})}{S_{xx}}
\end{aligned}$$

(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]

[0:6] Mark Scheme: 3 for E; 3 for Var

Solution:

$$\begin{aligned}
\mathsf{E}[\hat{\beta}_1] &= \mathsf{E}\left[\sum_{i=1}^{n} a_i Y_i\right] \\
&= \sum_{i=1}^{n} a_i \mathsf{E}[Y_i] \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} \mathsf{E}[Y_i] \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} (\beta_0 + \beta_1 x_i) \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} \beta_0 + \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} \beta_1 x_i \\
&= \beta_1 \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} x_i \qquad\qquad \text{as } \sum_{i=1}^{n} (x_i - \bar{x}) = 0. \\
&= \beta_1 \frac{S_{xx}}{S_{xx}} = \beta_1.
\end{aligned}$$

$$
\begin{aligned}
\text{Var}[\hat{\beta}_1] &= \text{Var}\left[\sum_{i=1}^{n} a_i Y_i\right] \\
&= \sum_{i=1}^{n} a_i^2 \text{Var}[Y_i] \qquad\qquad \text{as independent} \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{S_{xx}^2} \text{Var}[Y_i] \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{S_{xx}^2} \sigma^2 \\
&= \frac{S_{xx}}{S_{xx}^2} \sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}
$$

[Total: 11]

Core: 0 Adv: 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]

[3:0] Mark Scheme: direction - 1, strength - 1, three lines - 1

Solution:

There is a weak negative non-linear relationship between cty and displ. The three lines do not look parallel.

(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]

[2:0] Mark Scheme: 1 for intercept; 1 for slope

Solution:

For rear-wheel drive cars, we have

$$cty = 22.5914 - 3.0124 + (-2.0663 + 1.0039) \times displ$$

(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]

[4:0] Mark Scheme: 1 for hypotheses; 1 for F; 1 for P; 1 for conclusion

Solution:

$$H_0 : \beta_4 = \beta_5 = 0$$
$$H_a : \text{at least one of } \beta_4, \beta_5 \neq 0$$

where β_4 and β_5 are the coefficients associated with the interactions term.

The value of the test statistic is $F = 9.3963$.

The P-value is 0.0001199.

We reject the null hypothesis at the 5

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

[2:0] Mark Scheme: 1 for separate; 1 for justification.

Solution:

The best model appears to be the separate regression model as this has the smallest AIC with a value of 1051.857.

(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]

[4:0] Mark Scheme: 1 for each assumption.

Solution:

Linearity: Residual versus fitted (top left) shows random scatter so reasonable. There are possible outliers (numbers 222, 213, 223).

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 except for the points at the far right.

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

[Total: 15]

Core: 15 Adv: 0

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]

[1:0] Mark Scheme: 1 for expanded likelihood

Solution:

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

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

[1:0] Mark Scheme: 1 for formula

Solution:

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

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

[0:3] Mark Scheme: 1 for diff; 1 for set to zero; 1 for solve.

Solution:

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

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

Set equal to zero and solve:

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

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

[0:3] Mark Scheme: 1 for diff; 1 for E; 1 for Fisher

Solution:

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \lambda^2} &= -\frac{\sum_{i=1}^n y_i}{\lambda^2}\\ I_\lambda &= E\left[-\frac{\partial^2 \ell}{\partial \lambda^2}\right]\\ &= E\left[\frac{\sum_{i=1}^n y_i}{\lambda^2}\right]\\ &= \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}.\end{aligned}$$

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

[1:0] Mark Scheme: 1 for giving answer.

Solution:

We know that

$$\begin{aligned}\hat{\phi} &= \log(\hat{\lambda})\\ &= \log(\bar{y})\end{aligned}$$

[Total: 9]

Core: 3 Adv: 6

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]

[0:5] Mark Scheme: 2 for setup; 3 for working

Solution:

Let S be the number of sons with haemophilia. If the woman is not a carrier, then

$$P(S=0|\theta=0) = 1,$$

as the sons will only obtain a haemophilia carrying X chromosome X^H from their mother, and if does not have it, she cannot pass it on. If the woman is a carrier she has a probability of 1/2 of passing it on, assuming independence, we have

$$P(S=0|\theta=1) = \frac{1}{4}.$$

Putting this together with Bayes' rule gives

$$\begin{aligned} P(\theta=1|S=0) &= \frac{P(S=0|\theta=1)P(\theta=1)}{P(S=0|\theta=1)P(\theta=1)+P(S=0|\theta=0)P(\theta=0)} \\ &= \frac{1/4 \times 1/2}{1/4 \times 1/2 + 1 \times 1/2} \\ &= \frac{1}{5}. \end{aligned}$$

(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]

[0:3] Mark Scheme: 1 for setup; 2 for working.

Solution:

Let

$$S_0 = \begin{cases} 1 & \text{if Son 3 is haemophilliac,} \\ 0 & \text{if Son 3 is not haemophilliac.} \end{cases}$$

$$\begin{aligned} P(S_0 = 1|S = 0) &= \sum_\theta P(S_0 = 1|\theta)P(\theta|S = 0) \\ &= P(S_0 = 1|\theta = 1)P(\theta = 1|S = 0) + P(S_0 = 1|\theta = 0)P(\theta = 0|S = 0) \\ &= \frac{1}{2} \times \frac{1}{5} + 0 \times \frac{4}{5} = \frac{1}{10}. \end{aligned}$$

[Total: 8]

Core: 0 Adv: 8

7. Consider the multiple regression model

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

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 $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors with $\varepsilon_i \sim N(0, \sigma^2), i = 1, ..., n$.

(a) State the necessary and sufficient condition on X for the least squares estimate

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

to be uniquely identified. [1 marks]

[1:0] Mark Scheme: 1 for statement.

Solution:

The columns of X are linearly independent.

(b) Prove that $\hat{\boldsymbol{\beta}}$ uniquely minimises the sum of squares $Q(\boldsymbol{\beta}) = ||\boldsymbol{y} - X\boldsymbol{\beta}||^2$. [14 marks]

[0:14] Mark Scheme: 14 for working

Solution:

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

Now,

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

So we have that

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

Since $\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\|^2 \geq 0$ 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 $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

[Total: 15]

Core: 1 Adv: 14

8. To investigate the effect of Vitamin C on tooth growth in Guinea Pigs, 60 Guinea Pigs were given doses of Vitamin C. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day), denoted by `dose`, by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC): denoted by `supp`. The response is the length of odontoblasts (cells responsible for tooth growth) denoted `len`.

   An analysis was performed in R and the output is given in Appendix C.

(a) Describe what the code in the section **Clean data** is doing. [2 marks]

[0:2] Mark Scheme: 1 for recode; 1 for factor order.

Solution:

It is reclassifying the dose as low (0.5), medium (1), and high (2).

Also converting to a factor with order low, medium, then high.

(b) Is the experiment balanced? Justify your answer. [2 marks]

[0:2] Mark Scheme: 1 for yes, 1 for justification.

Solution:

Yes, there are 10 subjects in each treatment group.

(c) Using the interaction plot in Figure 3, describe the relationship between length of odontoblasts and dose; and between length of odontoblasts and delivery method. Does an interaction between dose and supplementary appear to be present? [3 marks]

[3:0] Mark Scheme: 1 for dose; 1 for supp; 1 for parallel

Solution:

As dose increases, then mean length increases.

Length is on average greater for OJ compared to VC.

Lines not parallel so interaction appears necessary.

(d) From the output, is an interaction term necessary? Justify your conclusion with reference to the output. [2 marks]

[2:0] Mark Scheme: 1 for yes; 1 for justification

Solution:

From the ANOVA, we have a p-value of 0.02186, and so an interaction term is necessary at the 5% level.

(e) Using the interaction model, calculate the predicted mean length of odontoblasts for Guinea Pigs on a low dose of Vitamin C given as orange juice. [2 marks]

[0:2] Mark Scheme: 1 for answer; 1 for working

Solution:

So the predicted value is

$$13.23$$

as this is the reference level for both dose and supp.

(f) Using the interaction model, calculate the predicted mean length of odontoblasts for Guinea Pigs on a high dose of Vitamin C given as ascorbic acid. [2 marks]

[0:2] Mark Scheme: 1 for answer; 1 for working

Solution:

So the predicted value is

$$13.23 + 12.830 - 5.25 + 5.33 = 26.14.$$

(g) Using the normal QQ-plots given in Figure 4, is the assumption of normality of length for each treatment reasonable? Justify your conclusion. [2 marks]

[0:2] Mark Scheme: 1 for answer; 1 for working

Solution:

All roughly linear so normality is reasonable.

[Total: 15]

Core: 5 Adv: 10

Q	core	adv
1	5	4
2	6	12
3	0	11
4	15	0
5	3	6
6	0	8
7	1	14
8	5	10
total	35	65
	NA	100

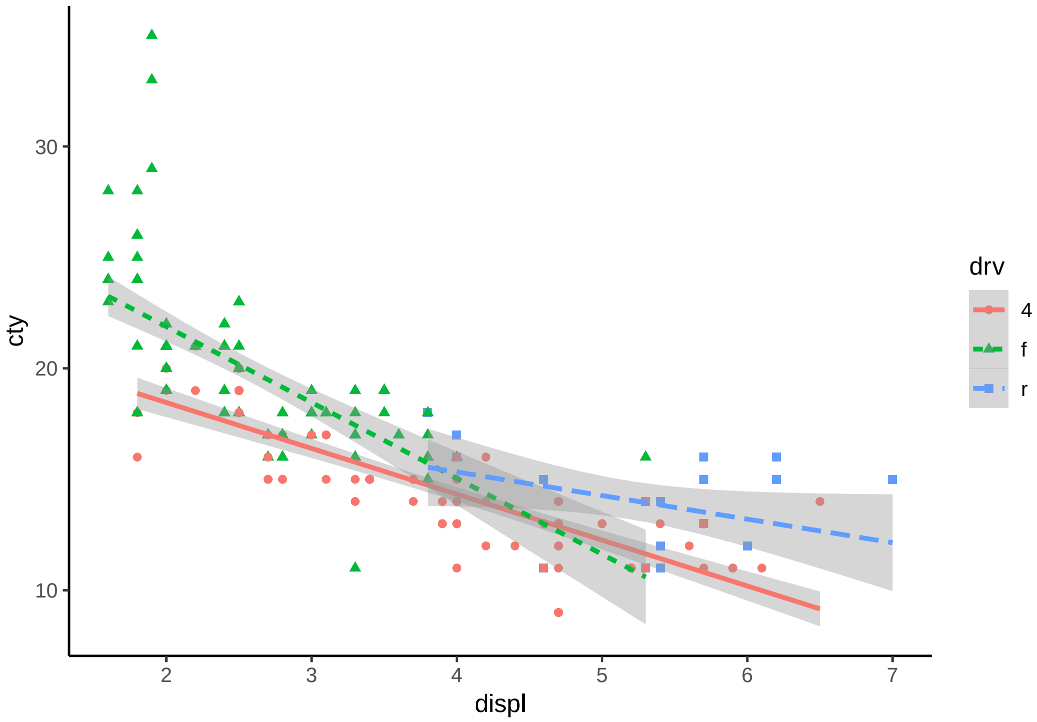

Figure 1: Scatterplot of Fuel efficiency against displacement for the MPG dataset. Colour and shape of points indicates drive (type).

Appendix A

Load the data

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

Visualise data

```r
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 21

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

Model Coefficients

```
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
```

```
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 22

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

```
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

```
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 23

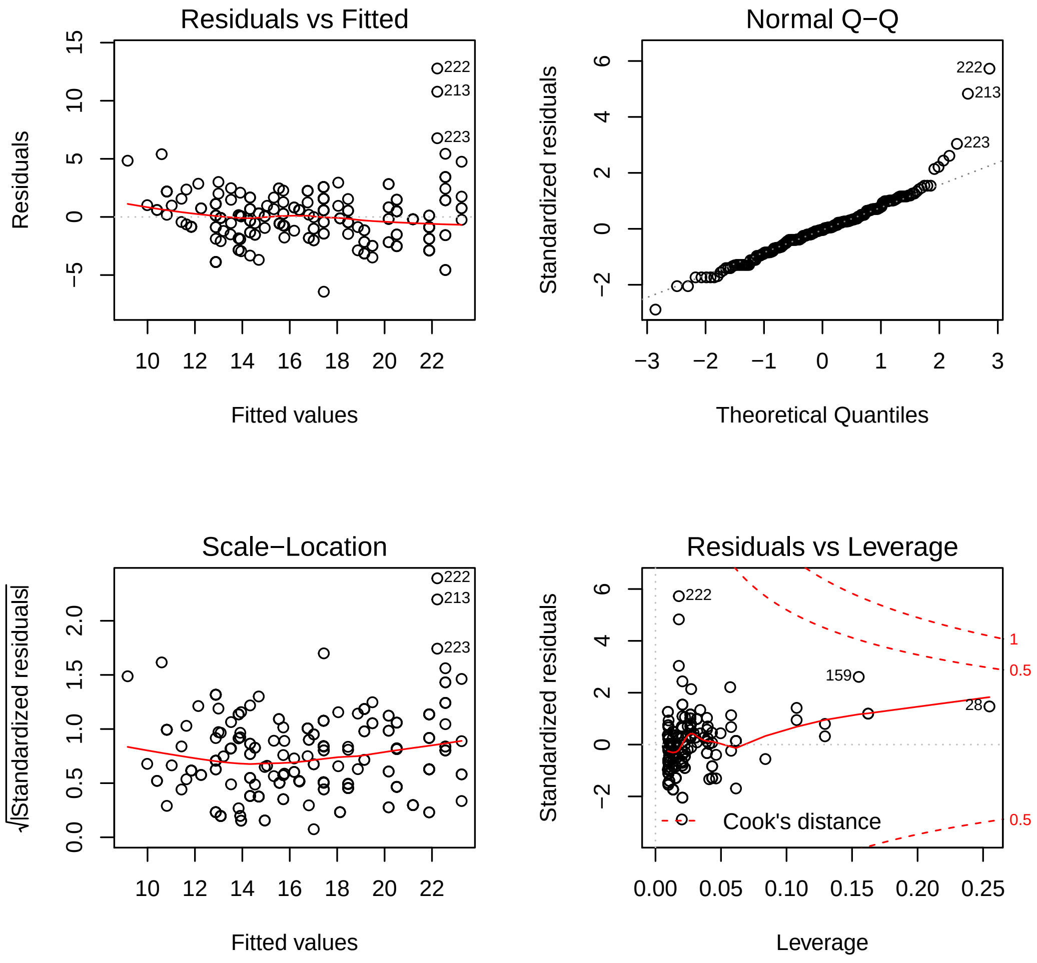

Figure 2: Assumption plots of the separate model for the MPG dataset.

Appendix B

Distribution	Probability mass function / probability density function	Expectation	Variance
Binomial	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \dots, n$	np	$np(1-p)$
Geometric	$p(x) = p(1-p)^{x-1}$ for $x = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \dots$	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 $0 < \theta < 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Please turn over for page 25

Appendix C

Load data

```
library(tidyverse)
data("ToothGrowth")
ToothGrowth  <- as_tibble(ToothGrowth)
ToothGrowth
```

```
## # A tibble: 60 x 3
##       len supp   dose
##     <dbl> <fct> <dbl>
##   1   4.2 VC      0.5
##   2  11.5 VC      0.5
##   3   7.3 VC      0.5
##   4   5.8 VC      0.5
##   5   6.4 VC      0.5
##   6  10   VC      0.5
##   7  11.2 VC      0.5
##   8  11.2 VC      0.5
##   9   5.2 VC      0.5
## 10   7   VC      0.5
## # ... with 50 more rows
```

Clean data

```
ToothGrowth  <-
  ToothGrowth %>%
  mutate(dose = case_when(
    dose == 0.5 ~ "low",
    dose == 1 ~ "medium",
    TRUE ~ "high"
  ))
ToothGrowth$dose  <- factor(ToothGrowth$dose,
                            levels = c("low", "medium", "high"))
```

Descriptive analysis

```
ToothGrowth %>%
  count(dose, supp) %>%
  spread(supp, n)
```

Please turn over for page 26

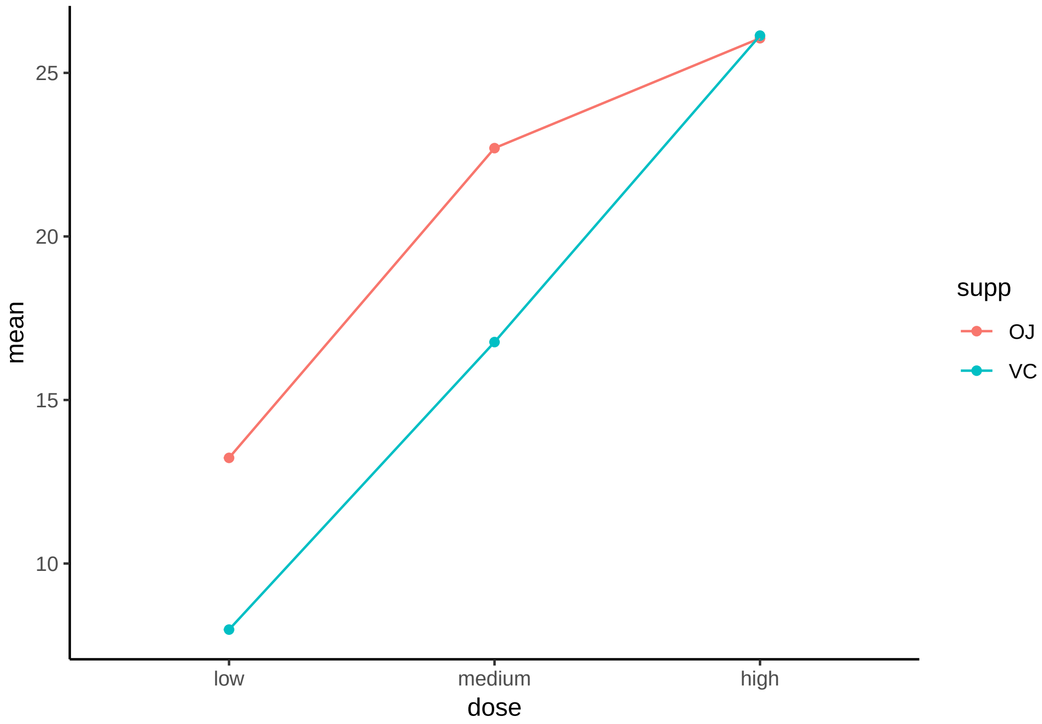

Figure 3: Interaction plot of mean length for each dose and type of supplement.

```
## # A tibble: 3 x 3
##   dose      OJ    VC
##   <fct>  <int> <int>
## 1 low       10    10
## 2 medium    10    10
## 3 high      10    10
```

Two-way ANOVA

```
ToothGrowth_M1  <- lm(len ~ dose * supp, data = ToothGrowth)
ToothGrowth_M2  <- lm(len ~ dose + supp, data = ToothGrowth)
anova(ToothGrowth_M1)
```

```
## Analysis of Variance Table
##
## Response: len
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## dose       2 2426.43 1213.22  92.000 < 2.2e-16 ***
## supp       1  205.35  205.35  15.572 0.0002312 ***
## dose:supp  2  108.32   54.16   4.107 0.0218603 *
```

Please turn over for page 27

```
## Residuals 54  712.11   13.19
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(ToothGrowth_M1, ToothGrowth_M2)
```

```
## Analysis of Variance Table
##
## Model 1: len ~ dose * supp
## Model 2: len ~ dose + supp
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     54 712.11
## 2     56 820.43 -2   -108.32 4.107 0.02186 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(ToothGrowth_M1)
```

```
##
## Call:
## lm(formula = len ~ dose * supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         13.230      1.148  11.521 3.60e-16 ***
## dosemedium           9.470      1.624   5.831 3.18e-07 ***
## dosehigh            12.830      1.624   7.900 1.43e-10 ***
## suppVC              -5.250      1.624  -3.233  0.00209 **
## dosemedium:suppVC   -0.680      2.297  -0.296  0.76831
## dosehigh:suppVC      5.330      2.297   2.321  0.02411 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.7937, Adjusted R-squared:  0.7746
## F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

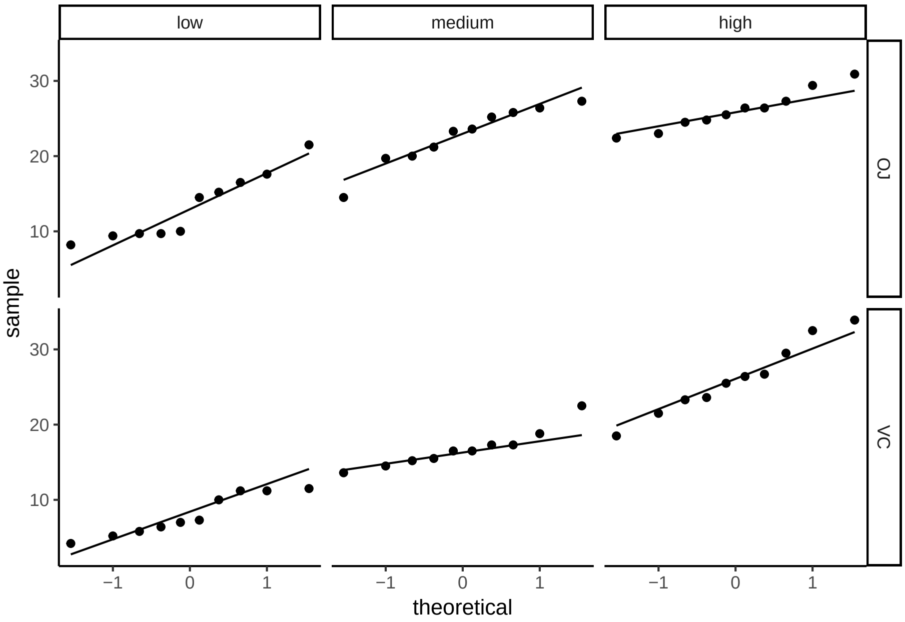

Figure 4: Normal QQ-plots of length for each treatment.

```
summary(ToothGrowth_M2)
```

```
##
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.085 -2.751 -0.800  2.446  9.650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4550     0.9883  12.603  < 2e-16 ***
## dosemedium    9.1300     1.2104   7.543 4.38e-10 ***
## dosehigh     15.4950     1.2104  12.802  < 2e-16 ***
## suppVC       -3.7000     0.9883  -3.744 0.000429 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 56 degrees of freedom
## Multiple R-squared:  0.7623, Adjusted R-squared:  0.7496
## F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

Final page