THE UNIVERSITY OF MELBOURNE

Student ID

Semester 1 Assessment, 2019

School of Mathematics and Statistics

MAST30025 Linear Statistical Models

Writing time: 3 hours

Reading time: 15 minutes

This is NOT an open book exam

This paper consists of 14 pages (including this page)

SOLUTIONS

Authorised materials:

- The only permitted scientific calculator is the Casio FX82.
- Two A4 double-sided handwritten sheets of notes.

Instructions to Students

- You must NOT remove this question paper at the conclusion of the examination.
- You should attempt all questions. Marks for individual questions are shown.
- The total number of marks available is 90.

Instructions to Invigilators

- Students must NOT remove this question paper at the conclusion of the examination.

Question 1 (10 marks)

(a) **[3 marks]** Let $A_1, \ldots, A_m$ be a set of symmetric idempotent matrices whose sum is also idempotent. Show directly that

$$r\left(\sum_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} r(A_i).$$

Solution:

$$r\left(\sum_{i=1}^{m} A_i\right) = tr\left(\sum_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} tr(A_i) = \sum_{i=1}^{m} r(A_i).$$

(b) **[3 marks]** Let $\mathbf{y}$ be a random vector with var $\mathbf{y} = V$. Show that V is positive semidefinite.

Solution: var $\mathbf{c}^T\mathbf{y} = \mathbf{c}^T V \mathbf{c} \geq 0$ for any $\mathbf{c}$. Hence V is positive semidefinite by definition.

(c) **[4 marks]** Show directly that for any matrix A, we have $A = A(A^TA)^cA^TA$. (*Hint: If $M^TM = 0$, then $M = 0$.*)

Solution:

$$\begin{aligned}\left(A - A(A^TA)^cA^TA\right)^T \left(A - A(A^TA)^cA^TA\right) &= A^TA - 2A^TA(A^TA)^cA^TA + A^TA(A^TA)^cA^TA(A^TA)^cA^TA \\ &= A^TA - 2A^TA + A^TA \\ &= 0.\end{aligned}$$

Hence $A - A(A^TA)^cA^TA = 0$.

Question 2 (12 marks) Let

$$\mathbf{y} \sim MVN\left(\begin{bmatrix} 4 \\ -1 \end{bmatrix}, \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}\right), \qquad A = \frac{1}{2}\begin{bmatrix} c & 1-3c \\ 1-3c & c \end{bmatrix}.$$

(a) **[3 marks]** Let $c = 1$. Calculate $E[\mathbf{y}^T A\mathbf{y}]$.

Solution:

```
> c <- 1
> A <- matrix(1/2*c(c,1-3*c,1-3*c,c),2,2)
> V <- matrix(c(3,1,1,3),2,2)
> mu <- c(4,-1)
> sum(diag(A%*%V)) + t(mu)%*%A%*%mu

     [,1]
[1,] 17.5
```

(b) **[3 marks]** Let $c = -1$. Describe the distribution of $A\mathbf{y}$.

Solution: $A\mathbf{y} \sim MVN(A\boldsymbol{\mu}, AVA^T)$.

```
> c <- -1
> A <- matrix(1/2*c(c,1-3*c,1-3*c,c),2,2)
> A %*% mu

     [,1]
[1,] -4.0
[2,]  8.5

> A %*% V %*% t(A)

      [,1]  [,2]
[1,] 10.75 -1.75
[2,] -1.75 10.75
```

(c) **[4 marks]** Find all values of c for which $\mathbf{y}^T A\mathbf{y}$ has a non-central χ^2 distribution.

Solution: We require

$$AV = \frac{1}{2}\begin{bmatrix} c & 1-3c \\ 1-3c & c \end{bmatrix}\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 3-8c \\ 3-8c & 1 \end{bmatrix}$$

to be idempotent. This requires

$$\begin{aligned}(AV)^2 &= \frac{1}{4}\begin{bmatrix} 1 & 3-8c \\ 3-8c & 1 \end{bmatrix}\begin{bmatrix} 1 & 3-8c \\ 3-8c & 1 \end{bmatrix} \\ &= \frac{1}{4}\begin{bmatrix} 1+(3-8c)^2 & 2(3-8c) \\ 2(3-8c) & 1+(3-8c)^2 \end{bmatrix} \\ &= AV \\ &= \frac{1}{2}\begin{bmatrix} 1 & 3-8c \\ 3-8c & 1 \end{bmatrix}.\end{aligned}$$

Equating gives

$$\begin{aligned} 1 + (3 - 8c)^2 &= 2 \\ 8 - 48c + 64c^2 &= 0 \\ 8(1 - 2c)(1 - 4c) &= 0 \\ c &= \frac{1}{2}, \frac{1}{4}. \end{aligned}$$

(d) **[2 marks]** Let $c = \frac{1}{2}$. Determine if $\mathbf{y}^T A \mathbf{y}$ is independent of $y_1 + y_2$.

```
> B <- c(1,1)
> c <- 1/2
> A <- matrix(1/2*c(c,1-3*c,1-3*c,c),2,2)
> B %*% V %*% A

     [,1] [,2]
[1,]    0    0
```

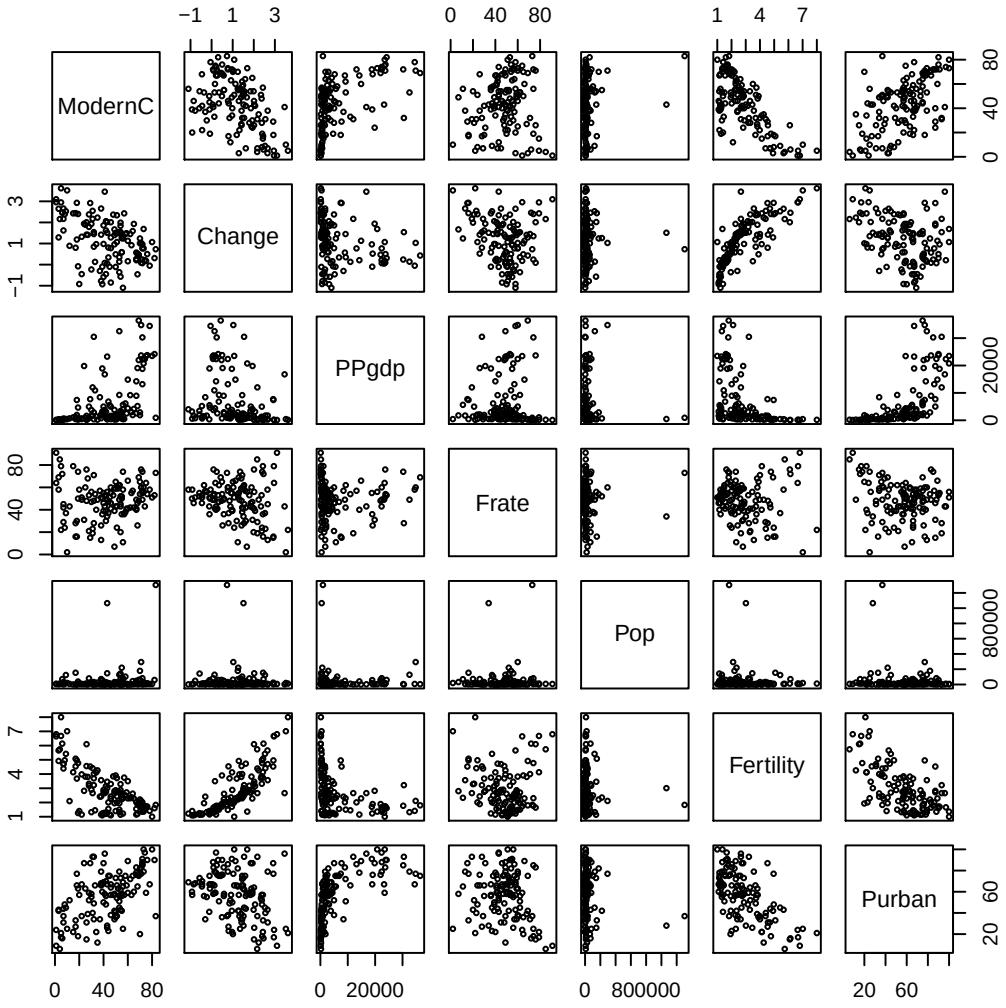
Solution: $\mathbf{y}^T A \mathbf{y}$ is independent of $y_1 + y_2$ for $c = \frac{1}{2}$.

Question 3 (18 marks) In this question, we study a dataset of 125 countries, collected by the United Nations. This dataset contains the variables:

- `ModernC`: Percent of unmarried women using a modern method of contraception
- `Change`: Annual population growth rate, percent
- `PPgdp`: Per capita gross national product, US dollars
- `Frate`: Percent of females over age 15 economically active
- `Pop`: Total 2001 population, 1000s
- `Fertility`: Expected number of live births per female, 2000
- `Purban`: Percent of population that is urban, 2001

We wish to model the birth rate (`Fertility`) in terms of the other variables. The following R calculations are produced (with some removed):

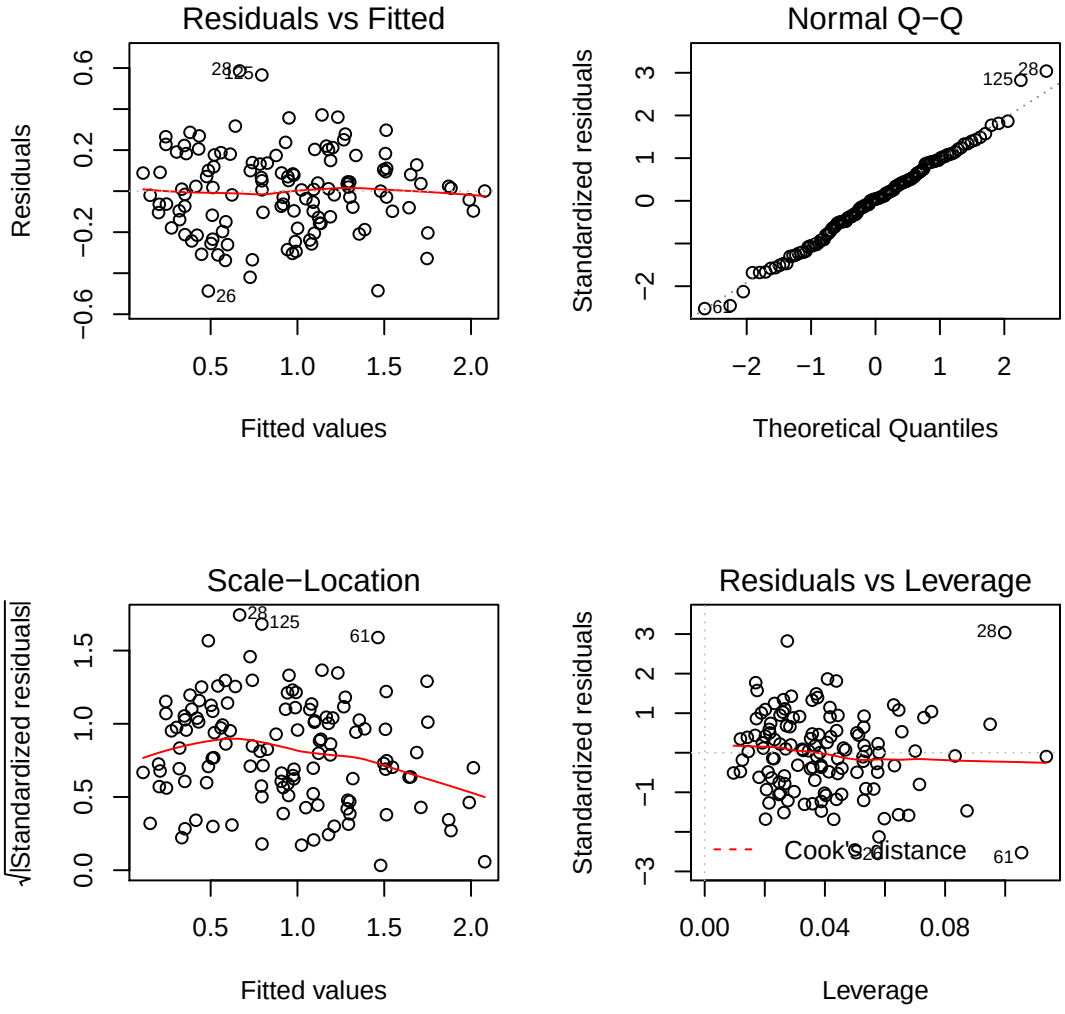
```
> UN <- read.csv('UN3.csv', header=T)
> pairs(UN, cex=0.5)
```



```
> UN$Fertility <- log(UN$Fertility)
> UN$PPgdp <- log(UN$PPgdp)
> UN$Pop <- log(UN$Pop)
> fullmodel <- lm(Fertility ~ ., data=UN)
> deviance(fullmodel)

[1] 4.929815

> # Input removed

Start:  AIC=-390.13
Fertility ~ ModernC + Change + PPgdp + Frate + Pop + Purban

          Df Sum of Sq     RSS     AIC
- Frate    1    0.0003  4.9301 -392.12
- PPgdp    1    0.0277  4.9575 -391.43
<none>                  4.9298 -390.13
- Pop      1    0.3084  5.2382 -384.54
- ModernC  1    0.3461  5.2759 -383.65
- Purban   1    0.5826  5.5124 -378.16
- Change   1   10.2407 15.1705 -251.62

Step:  AIC=-392.12
Fertility ~ ModernC + Change + PPgdp + Pop + Purban

          Df Sum of Sq     RSS     AIC
- PPgdp    1    0.0285  4.9586 -393.40
<none>                  4.9301 -392.12
+ Frate    1    0.0003  4.9298 -390.13
- Pop      1    0.3102  5.2403 -386.49
- ModernC  1    0.3559  5.2860 -385.41
- Purban   1    0.6135  5.5436 -379.46
- Change   1   10.9186 15.8487 -248.15

Step:  AIC=-393.4
Fertility ~ ModernC + Change + Pop + Purban

> summary(model)

Call:
lm(formula = Fertility ~ ModernC + Change + Pop + Purban, data = UN)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4866 -0.1282  0.0084  0.1321  0.5862

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.285452   0.102858  12.497  < 2e-16 ***
```

```
ModernC     -0.003971   0.001121  -3.541 0.000567 ***
Change       0.323302   0.019679  16.429  < 2e-16 ***
Pop         -0.024369   0.009324  -2.613 0.010110 *
Purban      -0.006111   0.000985  -6.205 8.09e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2033 on 120 degrees of freedom
Multiple R-squared:  0.8462,        Adjusted R-squared:  0.841
F-statistic:   165 on 4 and 120 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(model)
```



```
> library(car)
> linearHypothesis(model, c(0,0,0,log(2),0), log(0.98))

Linear hypothesis test

Hypothesis:
0.693147180559945 Pop = - 0.0202027073175195

Model 1: restricted model
Model 2: Fertility ~ ModernC + Change + Pop + Purban

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    121 4.9694
2    120 4.9586  1  0.010847 0.2625 0.6093

> qt(0.975,118:120)

[1] 1.980272 1.980100 1.979930

> qf(0.975,1,118:120)

[1] 5.154550 5.153431 5.152331

> qf(0.975,2,118:120)

[1] 3.806642 3.805631 3.804638
```

(a) **[2 marks]** Give two reasons in favour of the logarithmic transformation of the `PPgdp` variable.

Solution: From the `pairs` plot, the `PPgdp` variable has a clear non-linear relationship with the `Fertility` response variable, and also appears to be right-skewed.

(b) **[4 marks]** Test the hypothesis $H_0 : \beta_{\texttt{Pop}} = 0$ for the full model (with all variables included) at the 5% significance level. Clearly state your F-statistic and critical value, and interpret your conclusion in the context of the study.

```
> 0.3084/(4.929815/118)

[1] 7.381859
```

Solution: Since $7.38 > 5.154$, we reject the hypothesis; population has a significant (power-law) effect on birth rate.

(c) **[2 marks]** Identify the variable selection procedure that has been used.

Solution: Stepwise selection using AIC as a goodness-of-fit measure, starting from the model with all variables. (Sharp observers will notice that the procedure has not terminated.)

(d) **[2 marks]** Perform one step of backwards elimination on `model`.

Solution: All variables in the model are significant, so backwards elimination will stop at this model.

(e) **[2 marks]** Write down the final fitted model, including any variable transformations used.

Solution:

$$\log \texttt{Fertility} = 1.285 - 0.00397\,\texttt{ModernC} + 0.323\,\texttt{Change} - 0.024\,\log \texttt{Pop} - 0.00611\,\texttt{Purban}.$$

(f) **[2 marks]** Calculate a 95% confidence interval for the parameter corresponding to the `Change` variable.

Solution:

```
> 0.323302 + c(-1,1)*1.97993*0.019679

[1] 0.284339 0.362265
```

(g) **[2 marks]** Interpret the results of the `linearHypothesis` function in the context of the study.

Solution: The function tests whether birth rate goes down by 2% for every doubling of the population. With a p-value of 0.6093, we cannot reject this hypothesis.

(h) **[2 marks]** From the diagnostic plots, comment on the suitability of the linear model for this data.

Solution: All the diagnostic plots show that the model is a good fit to this data; distributional assumptions appear to be satisfied and there are no outliers.

Question 4 (13 marks) Consider the full rank linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with p parameters.

(a) **[3 marks]** Show that $\mathbf{b}$, the least squares estimator of $\boldsymbol{\beta}$, and s^2, the sample variance, are independent.

Solution:

$$\begin{aligned} BVA &= (X^TX)^{-1}X^T\sigma^2 I\frac{1}{n-p}(I-H) \\ &= \frac{\sigma^2}{n-p}(X^TX)^{-1}(X^T - X^TH) \\ &= 0. \end{aligned}$$

(b) **[3 marks]** Calculate the expected value of the maximum likelihood estimator of σ^2 and thereby show that it is biased.

Solution:

$$\begin{aligned} E[\tilde{\sigma}^2] &= E[\frac{SS_{Res}}{n}] \\ &= \frac{n-p}{n}E[s^2] \\ &= \frac{n-p}{n}\sigma^2 \\ &\neq \sigma^2. \end{aligned}$$

(c) **[4 marks]** Given a (symmetric) variance matrix var $\mathbf{y} = V$, the method of generalised least squares estimates the parameters $\boldsymbol{\beta}$ by minimising $\mathbf{e}^TV^{-1}\mathbf{e}$. Derive an expression for the resulting estimators.

Solution:

$$\begin{aligned} \mathbf{e}^TV^{-1}\mathbf{e} &= \mathbf{y}^TV^{-1}\mathbf{y} - 2\mathbf{y}^TV^{-1}X\mathbf{b} + \mathbf{b}^TX^TV^{-1}X\mathbf{b} \\ \frac{d}{d\mathbf{b}}(\mathbf{e}^TV^{-1}\mathbf{e}) &= -2X^TV^{-1}\mathbf{y} + 2X^TV^{-1}X\mathbf{b} \\ &= 0 \\ \mathbf{b} &= (X^TV^{-1}X)^{-1}X^TV^{-1}\mathbf{y}. \end{aligned}$$

(d) **[3 marks]** Calculate the variance of the estimators derived in question 4c.

Solution:

$$\begin{aligned} \text{var } \mathbf{b} &= (X^TV^{-1}X)^{-1}X^TV^{-1}VV^{-1}X(X^TV^{-1}X)^{-1} \\ &= (X^TV^{-1}X)^{-1}. \end{aligned}$$

Question 5 (12 marks) Consider the general linear model, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. This model may be of full or less than full rank.

(a) **[2 marks]** State two methods that can be used to fit this model to data, and compare them.

Solution: We can use the method of least squares, or maximum likelihood estimation. This gives identical $\boldsymbol{\beta}$ estimates, but ML has a biased variance estimator while LS does not.

(b) **[2 marks]** Is a high leverage or low leverage more desirable? Explain.

Solution: Low leverage is generally more desirable as a point with high leverage has the potential to affect the fit drastically.

(c) **[2 marks]** What are two possible reasons to transform a response variable?

Solution: Two possible reasons might be a non-linear relationship with predictor variables, and non-normal distribution.

(d) **[2 marks]** Define overfitting and explain how it may be avoided.

Solution: Overfitting occurs when irrelevant predictor variables are fit to noise in the data rather than the underlying relationship. It can be avoided with careful variable selection.

(e) **[2 marks]** Define interaction between two continuous predictors and explain how to model it.

Solution: Interaction between two continuous predictors occurs when the value of one predictor linearly affects the parameter (effect) associated with the second. It can be modelled by including a product term $x_1 x_2$ as another predictor.

(f) **[2 marks]** Under what circumstances should one prefer a complete block design to a completely randomised design, and vice versa?

Solution: One should prefer a complete block design when the blocking variable has an actual effect; otherwise a completely randomised design is slightly better. (If one is unsure, a CBD is better.)

Question 6 (14 marks) The nursing director at a private hospital wishes to compare the weekly number of complaints received against the nursing staff during three daily shifts: first (7am–3pm), second (3pm–11pm), and third (11pm–7am). Her plan is to sample 17 weeks and select a shift at random from each week sampled, recording the number of complaints received during the selected shift.

The following data is collected:

	number of complaints		
	observations	mean	sample variance
shift 1	9, 9, 11, 9, 12	10	2
shift 2	8, 11, 6, 8, 9, 12	9	4.8
shift 3	15, 14, 10, 11, 10, 12	12	4.4

The data is analysed using a one-way classification model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, 2, 3, j = 1, \ldots, n_i.$$

You are also given the following R calculations:

```
> qt(0.975,14:17)

[1] 2.144787 2.131450 2.119905 2.109816

> qf(0.975,1,14:17)

[1] 6.297939 6.199501 6.115127 6.042013

> qf(0.975,2,14:17)

[1] 4.856698 4.765048 4.686665 4.618874
```

(a) [**2 marks**] Write down the normal equations for this model.

Solution:

```
> (XtX <- matrix(c(17,5,6,6,5,5,0,0,6,0,6,0,6,0,0,6),4,4))

     [,1] [,2] [,3] [,4]
[1,]   17    5    6    6
[2,]    5    5    0    0
[3,]    6    0    6    0
[4,]    6    0    0    6

> (Xty <- c(176,50,54,72))

[1] 176  50  54  72
```

(b) [**2 marks**] Write down two distinct estimates for the parameters of the model. One of these should estimate $\hat{\mu} = 10$.

Solution:

```
> means <- c(10,9,12)
> (b <- c(0,means))

[1]  0 10  9 12

> b + 10*c(1,-1,-1,-1)

[1] 10  0 -1  2
```

(c) **[3 marks]** The sample variance s^2 is calculated to be 3.857. Calculate a 95% prediction interval for the weekly number of complaints generated in the first shift.

Solution:

```
> s2 <- 3.857
> tt <- c(1,1,0,0)
> XtXc <- diag(c(0,1/5,1/6,1/6))
> tt %*% b + c(-1,1)*2.144787*sqrt(s2)*sqrt(1+t(tt)%*%XtXc%*%tt)

[1]  5.385766 14.614234
```

(d) **[3 marks]** The nursing director wishes to test if the third shift generates as many complaints as the average of the other two. Express this as a formal hypothesis and show that it is testable.

Solution: The hypothesis is $H_0 : \tau_3 - \frac{1}{2}(\tau_1 + \tau_2) = 0$.

```
> tt <- c(0,-1/2,-1/2,1)
> tt %*% XtXc %*% XtX

     [,1] [,2] [,3] [,4]
[1,]    0 -0.5 -0.5    1
```

(e) **[4 marks]** Test this hypothesis at the 5% significance level. Clearly state your F-statistic and critical value.

```
> C <- t(tt)
> (Fstat <- (C%*%b) %*% solve(C%*%XtXc%*%t(C)) %*% C%*%b / s2)

         [,1]
[1,] 6.272634
```

Solution: As $6.272 < 6.298$, we cannot reject the hypothesis at the 5% level.

Question 7 (11 marks)

(a) **[4 marks]** Sleepeze is a new drug that is claimed to reduce insomnia. You plan to conduct an experiment to test this claim. You are given resources to monitor a total of 120 person-nights (120 people for 1 night each, or 60 people for two nights each, etc.), and wish to perform an experiment with a complete block design, with gender as a blocking factor.

Briefly describe how control, blocking, randomisation, and blinding may be used in a design of this experiment.

Solution: Control: The experiment will include a control group who take placebos. Blocking: Two blocks, 60 males and 60 females with insomnia, will be chosen as the test subjects. Randomisation: Within each block, 30 people will be randomly chosen to take Sleepeze, and the rest will take the placebo. Blinding: Neither the subjects nor the experimenters will know who takes Sleepeze and who takes the placebos.

(b) **[2 marks]** Write down a design matrix and parameter vector for a model which has a complete block design with 2 treatments and 3 blocks, and one sample per treatment/block combination.

Solution: One possible answer is

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \tau_1 \\ \tau_2 \end{bmatrix}.$$

(c) **[2 marks]** Write down estimators for the model from question 7b for the treatment parameters only, as a linear function of the response vector $\mathbf{y}$.

Solution:

$$\mathbf{b}_2 = \frac{1}{6} \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix} \mathbf{y}.$$

(d) **[3 marks]** For the model from question 7b, directly calculate the variance of the estimator of the difference between the two treatment parameters. You may write it in terms of the error variance σ^2.

Solution:

$$\begin{aligned} \text{var} \begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{b}_2 &= \text{var} \frac{1}{6} \begin{bmatrix} 2 & -2 & 2 & -2 & 2 & -2 \end{bmatrix} \mathbf{y} \\ &= \sigma^2 \frac{1}{36} \begin{bmatrix} 2 & -2 & 2 & -2 & 2 & -2 \end{bmatrix} I_6 \begin{bmatrix} 2 \\ -2 \\ 2 \\ -2 \\ 2 \\ -2 \end{bmatrix} \\ &= \frac{2}{3}\sigma^2. \end{aligned}$$

End of Exam—Total Available Marks = 90.