# Exam 2016, questions and answers

Linear Statistical Models (University of Melbourne)

THE UNIVERSITY OF MELBOURNE

Student ID

Semester 1 Assessment, 2016

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 11 pages (including this page)

**Authorised materials**:

- Scientific calculators are premitted, but not graphical calculators.
- One A4 double-sided handwritten sheet 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.

This paper may be held in the Baillieu Library

**This paper must not be removed from the examination room**

**Question 1 (9 marks)**

(a) Let $A$ be a square matrix and suppose that $A^k = A^{k+1}$ for some $k \geq 1$. Show that $A$ is idempotent.

**Solution [3 marks]:** Let $\lambda$ be an eigenvalue of $A$ with corresponding eigenvector $\mathbf{x}$. Then

$$\lambda^k \mathbf{x} = A^k \mathbf{x} = A^{k+1}\mathbf{x} = \lambda^{k+1}\mathbf{x}.$$

Since $\mathbf{x}$ is nonzero, we have $\lambda^k(1-\lambda) = 0$, which gives $\lambda = 0$ or 1. Therefore $A$ is idempotent.

(b) Let $X$ be an $n \times p$ matrix of full rank, where $n > p$. Show that $H = X(X^TX)^{-1}X^T$ is idempotent, and find its rank. (You may assume that $H$ is symmetric.)

**Solution [3 marks]:**

$$H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H.$$

Since $H$ is symmetric and idempotent, we have

$$r(H) = tr(H) = tr((X^TX)^{-1}X^TX) = tr(I_p) = p.$$

(c) Show that if a square matrix $A$ is positive semidefinite, then its eigenvalues are non-negative.

**Solution [3 marks]:** Let $\lambda$ be an eigenvalue of $A$ with corresponding eigenvector $\mathbf{x}$. Then

$$0 \leq \mathbf{x}^T A\mathbf{x} = \lambda \mathbf{x}^T\mathbf{x},$$

and since $\mathbf{x}^T\mathbf{x} > 0$, we have $\lambda \geq 0$.

**Question 2 (10 marks)** Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \sim MVN\left( \begin{bmatrix} a \\ -a \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \right),$$

where $a$ is a constant.

(a) What is the distribution of $y_1 + y_2$?

**Solution [3 marks]:** $y_1 + y_2$ has a normal distribution with mean 0 and variance

$$\begin{aligned} \text{var}\,(y_1+y_2) &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \\ 0 \end{bmatrix} \\ &= 6. \end{aligned}$$

(b) What is the distribution of $\frac{1}{2}\left(y_1^2 - 2y_1y_2 + y_2^2 + y_3^2\right)$?

**Solution [4 marks]:** We have

$$\begin{aligned}
\frac{1}{2}\left(y_1^2 - 2y_1y_2 + y_2^2 + y_3^2\right) &= \mathbf{y}^T \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\mathbf{y} \\
\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} &= \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.
\end{aligned}$$

It is easy to check that this last matrix is idempotent with rank 2, so $\frac{1}{2}\left(y_1^2 - 2y_1y_2 + y_2^2 + y_3^2\right)$ has a noncentral $\chi^2$ distribution with 2 d.f. and noncentrality parameter

$$\begin{bmatrix} a & -a & 0 \end{bmatrix}\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} a \\ -a \\ 0 \end{bmatrix} = \begin{bmatrix} a & -a & 0 \end{bmatrix}\begin{bmatrix} a \\ -a \\ 0 \end{bmatrix} = 2a^2.$$

(c) Suppose $a = 0$. For what values of $c$ does

$$c\frac{y_1^2 - 2y_1y_2 + y_2^2 + y_3^2}{y_1^2 + 2y_1y_2 + y_2^2}$$

have an $F$ distribution?

**Solution [3 marks]:** Since $a = 0$, the numerator divided by 2 has a $\chi^2$ distribution with 2 d.f.. From (a), the denominator divided by 6 has a $\chi^2$ distribution with 1 d.f.. We also need the two $\chi^2$ distributions to be independent:

$$\begin{aligned}
\begin{bmatrix} 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}.
\end{aligned}$$

Thus for this quantity to have an $F$ distribution, we need to divide the numerator by 2 and then 2 again (for the d.f.), and the denominator by 6. Thus we require $c = \frac{3}{2}$.

**Question 3 (14 marks)** Consider the full rank linear model, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

(a) State the assumptions involved in fitting this model.

**Solution [2 marks]:** We assume that the linear relationship holds and that $\boldsymbol{\varepsilon} \sim MVN(0, \sigma^2 I)$.

(b) Define the term BLUE (best linear unbiased estimator).

**Solution [2 marks]:** The BLUE of a quantity $\mathbf{t}^T\boldsymbol{\beta}$ is a linear estimator $L\mathbf{y}$ such that $E[L\mathbf{y}] = \mathbf{t}^T\boldsymbol{\beta}$ and the variance of the elements of $L\mathbf{y}$ is minimised over all linear estimators of $\mathbf{t}^T\boldsymbol{\beta}$.

(c) Is it better to fit this model using the method of least squares or maximum likelihood estimation? Justify your answer.

**Solution [2 marks]:** Fitting parameters using least squares or MLE produces the same result. It is better to use least squares to estimate $\sigma^2$ as the MLE is biased.

(d) Define and explain the purpose of the leverage of a point.

**Solution [2 marks]:** The leverage of a point, defined as the corresponding diagonal element in the hat matrix $X(X^TX)^{-1}X^T$, is a measure of how much influence the particular point has on the fit of the linear model.

(e) Explain the difference between a model relevance test and a model relevance test using a corrected sum of squares.

**Solution [2 marks]:** A model relevance test tests if all parameters are 0, including the intercept; using a corrected sum of squares tests if all parameters except the intercept are 0, in the presence of the intercept.

(f) When is a model with fewer explanatory variables more desirable or less desirable than a model with more explanatory variables?

**Solution [2 marks]:** A model with fewer explanatory variables is more desirable if and only if the variables left out are not relevant. Reducing the number of variables prevents overfitting, but removing relevant variables hurts the fit.

(g) Explain why the residual sum of squares $SS_{Res}$ is not an appropriate goodness-of-fit measure for model selection.

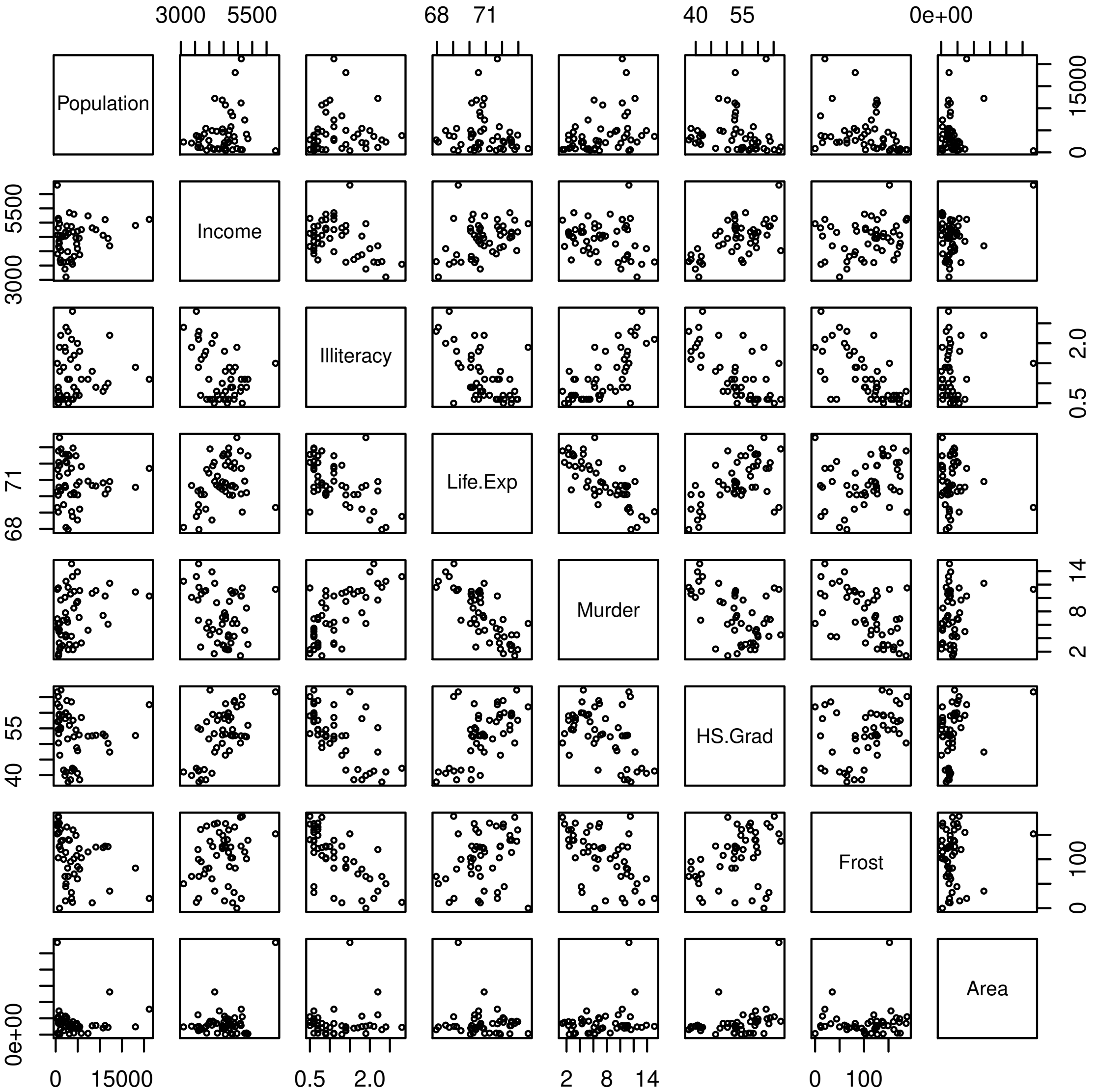
**Solution [2 marks]:** $SS_{Res}$ always decreases when a variable is added, so it does not discourage overfitting.

**Question 4 (17 marks)** In this question, we study a dataset of 50 US states. This dataset contains the variables:

- `Population`: population estimate as of July 1, 1975
- `Income`: per capita income (1974)
- `Illiteracy`: illiteracy (1970, percent of population)
- `Life.Exp`: life expectancy in years (1969–71)
- `Murder`: murder and non-negligent manslaughter rate per 100,000 population (1976)
- `HS.Grad`: percentage of high-school graduates (1970)
- `Frost`: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- `Area`: land area in square miles

We use linear models to model life expectancy in terms of the other variables. The following R output is produced.

```
> data(state)
> statedata <- data.frame(state.x77, row.names=state.abb, check.names=TRUE)
> pairs(statedata, cex=0.5)
```



```
> statedata$Population <- log(statedata$Population)
> statedata$Area <- log(statedata$Area)
> nullmodel <- lm(Life.Exp ~ 1, data = statedata)
> fullmodel <- lm(Life.Exp ~ ., data = statedata)
> model <- step(fullmodel, scope = ~ .)

Start:  AIC=-23.6
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost + Area

             Df Sum of Sq    RSS      AIC
- Income      1    0.0018 22.650 -25.5934
- Illiteracy  1    0.0556 22.704 -25.4746
- Area        1    0.2106 22.859 -25.1344
<none>                    22.648 -23.5973
- Frost       1    1.2374 23.886 -22.9374
- Population  1    1.8854 24.533 -21.5992
- HS.Grad     1    2.4375 25.086 -20.4864
- Murder      1   23.2760 45.924   9.7483

Step:  AIC=-25.59
Life.Exp ~ Population + Illiteracy + Murder + HS.Grad + Frost +
    Area

             Df Sum of Sq    RSS      AIC
- Illiteracy  1    0.0556 22.705 -27.4708
- Area        1    0.2197 22.870 -27.1107
<none>                    22.650 -25.5934
- Frost       1    1.2602 23.910 -24.8862
+ Income      1    0.0018 22.648 -23.5973
- Population  1    2.1909 24.841 -22.9768
- HS.Grad     1    4.0374 26.687 -19.3918
- Murder      1   24.2130 46.863   8.7601

Step:  AIC=-27.47
Life.Exp ~ Population + Murder + HS.Grad + Frost + Area

             Df Sum of Sq    RSS     AIC
- Area        1    0.2157 22.921 -28.998
<none>                    22.705 -27.471
+ Illiteracy  1    0.0556 22.650 -25.593
+ Income      1    0.0017 22.704 -25.475
- Population  1    2.2792 24.985 -24.688
- Frost       1    2.3760 25.082 -24.495
- HS.Grad     1    4.9491 27.655 -19.612
- Murder      1   29.2296 51.935  11.899

Step:  AIC=-29
Life.Exp ~ Population + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
<none>                    22.921 -28.998
+ Area        1     0.216 22.705 -27.471
+ Illiteracy  1     0.052 22.870 -27.111
+ Income      1     0.011 22.911 -27.021
- Frost       1     2.214 25.135 -26.387
- Population  1     2.450 25.372 -25.920
- HS.Grad     1     6.959 29.881 -17.741
- Murder      1    34.109 57.031  14.578

> summary(model)
```

```
Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = statedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.41760 -0.43880  0.02539  0.52066  1.63048

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.720810   1.416828  48.503  < 2e-16 ***
Population   0.246836   0.112539   2.193 0.033491 *
Murder      -0.290016   0.035440  -8.183 1.87e-10 ***
HS.Grad      0.054550   0.014758   3.696 0.000591 ***
Frost       -0.005174   0.002482  -2.085 0.042779 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7137 on 45 degrees of freedom
Multiple R-squared:  0.7404,        Adjusted R-squared:  0.7173
F-statistic: 32.09 on 4 and 45 DF,  p-value: 1.17e-12
```

```
> anova(nullmodel, model, fullmodel)
```

```
Analysis of Variance Table

Model 1: Life.Exp ~ 1
Model 2: Life.Exp ~ Population + Murder + HS.Grad + Frost
Model 3: Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost + Area
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1     49 88.299
2     45 22.921  4    65.378 30.3101 6.901e-12 ***
3     42 22.648  3     0.273  0.1688    0.9168
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> signif(vcov(model), 6)
```

```
            (Intercept)   Population       Murder      HS.Grad        Frost
(Intercept)  2.00740000 -1.18811e-01 -1.98357e-02 -1.44506e-02 -1.42795e-03
Population  -0.11881100  1.26650e-02 -3.56651e-04  2.36109e-04  8.91432e-05
Murder      -0.01983570 -3.56651e-04  1.25601e-03  1.84375e-04  3.42863e-05
HS.Grad     -0.01445060  2.36109e-04  1.84375e-04  2.17798e-04 -3.18945e-06
Frost       -0.00142795  8.91432e-05  3.42863e-05 -3.18945e-06  6.15931e-06
```

(a) Why do we take a logarithmic transformation of population and area?

**Solution [2 marks]:** The relationship between life expectancy and population/area does not appear linear, and the variables are very right-skewed; taking logarithmic transformations might help.

(b) Find the Akaike's Information Criterion for the model with variables `Population`, `Murder`, `Frost`, and `Area`.

**Solution [1 mark]:** -19.612.

(c) Write down the final fitted model (including any variable transforms used).

**Solution [2 marks]:**

$$\texttt{Life.Exp} = 68.72 + 0.25\ln(\texttt{Population}) - 0.29\,\texttt{Murder} + 0.055\,\texttt{HS.Grad} - 0.0052\,\texttt{Frost}.$$

(d) Calculate the sample variance $s^2$ for the final model.

**Solution [2 marks]:**

```
> (s2 <- 22.921/45)

[1] 0.5093556
```

(e) Calculate a 95% confidence interval for $\beta_{Population} - \beta_{Murder}$. (The 97.5% critical value for a $t$ distribution with 45 d.f. is 2.014.)

**Solution [4 marks]:**

```
> 0.246836+0.290016+c(-1,1)*2.014*sqrt(0.0126650+2*0.000356651+0.00125601)

[1] 0.2932137 0.7804903
```

(f) What conclusions do you draw from the tests in the ANOVA table?

**Solution [2 marks]:** At least one of the variables in the final model is relevant, while the variables that have been removed are clearly irrelevant.

(g) If you were to perform an $F$ test of $H_0 : \beta_{Frost} = 0$ in the final model, what would your $F$ statistic and $p$-value be?

**Solution [2 marks]:** The $F$ statistic would be $-2.085^2 = 4.347$, and the $p$-value would be 0.043.

(h) Explain the $F$-statistic for the final model (last line of the `summary` call). Why is it different to the $F$-value in line 2 of the ANOVA table?

**Solution [2 marks]:** The $F$-statistic for the final model is a test of model relevance (with the corrected sum of squares) using $s^2$ for the final (selected) model, but the $F$-value in the ANOVA table uses the $s^2$ for the full model.

**Question 5 (14 marks)** Consider the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, which may be of full or less than full rank.

(a) Define the term estimable.

**Solution [2 marks]:** A quantity $\mathbf{t}^T\boldsymbol{\beta}$ is estimable if there exists an unbiased linear estimator, i.e. a matrix $L$ such that $E[L\mathbf{y}] = \mathbf{t}^T\boldsymbol{\beta}$.

(b) Show that if $\mathbf{t}^T = \mathbf{t}^T(X^TX)^cX^TX$, then $\mathbf{t}^T\boldsymbol{\beta}$ is estimable.

**Solution [3 marks]:**

$$E[\mathbf{t}^T(X^TX)^cX^T\mathbf{y}] = \mathbf{t}^T(X^TX)^cX^TX\boldsymbol{\beta} = \mathbf{t}^T\boldsymbol{\beta}.$$

Thus by definition, $\mathbf{t}^T\boldsymbol{\beta}$ is estimable.

(c) Show that in a one-factor model, all treatment contrasts are estimable.

**Solution [3 marks]:** A treatment contrast is defined as $\sum_i a_i\tau_i$ where $\sum_i a_i = 0$. But this is equal to $\sum_i a_i(\mu + \tau_i)$, which is a linear combination of elements of $X\boldsymbol{\beta}$. Therefore it is estimable.

(d) If $\mathbf{t}^T\boldsymbol{\beta}$ is estimable, derive the distribution of $\mathbf{t}^T\mathbf{b}$, where $\mathbf{b}$ is the least squares estimator of $\boldsymbol{\beta}$.

**Solution [3 marks]:** $\mathbf{t}^T\mathbf{b}$ is multivariate normal with mean $\mathbf{t}^T\boldsymbol{\beta}$ and variance

$$\begin{aligned} \text{var } \mathbf{t}^T(X^TX)^cX^T\mathbf{y} &= \mathbf{t}^T(X^TX)^cX^T\sigma^2IX(X^TX)^c\mathbf{t} \\ &= \mathbf{t}^T(X^TX)^c\mathbf{t}\sigma^2. \end{aligned}$$

(e) If $\mathbf{t}^T\boldsymbol{\beta}$ is estimable, show that $\mathbf{t}^T\mathbf{b}$ is independent of the sample variance $s^2$.

**Solution [3 marks]:**

$$\begin{aligned} BVA &= \mathbf{t}^T(X^TX)^cX^T\sigma^2I(I - X(X^TX)^cX^T) \\ &= \left[\mathbf{t}^T(X^TX)^cX^T - \mathbf{t}^T(X^TX)^cX^T\right]\sigma^2 \\ &= 0. \end{aligned}$$

Therefore $\mathbf{t}^T\mathbf{b}$ is independent of $SS_{Res}$, and by extension $s^2$.

**Question 6 (12 marks)** The nursing director at a private hospital wishes to compare the weekly number of complaints received against the nursing staff during the 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 | |
|---|---|---|---|
| | number of observations | mean | sample variance |
| shift 1 | 5 | 10 | 2 |
| shift 2 | 6 | 9 | 4.8 |
| shift 3 | 6 | 12 | 4.4 |

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

(a) What kind of experimental design is this?

**Solution [1 mark]:** It is a completely randomised design.

(b) Calculate the sample variance $s^2$ for the linear model.

**Solution [2 marks]:**

```
> ns <- c(5,6,6)
> n <- sum(ns)
> r <- 3
> means <- c(10,9,12)
> vars <- c(2,4.8,4.4)
> (s2 <- sum((ns-1)*vars)/(n-r))

[1] 3.857143
```

(c) Calculate a 95% prediction interval for the total number of complaints received in a day. (The 97.5% critical value of a $t$ distribution with 14 d.f. is 2.145.) (*Hint:* You will need to modify the formula for a prediction interval.)

**Solution [4 marks]:**

```
> XtXc <- diag(c(0,1/ns))
> b <- c(0,means)
> tt <- c(3,1,1,1)
> tt%*%b + c(-1,1)*2.145*sqrt(s2)*sqrt(3+t(tt)%*%XtXc%*%tt)

[1] 23.08133 38.91867
```

(d) Test the hypothesis that shift has no effect on the number of complaints. (The 95% critical value of an $F$ distribution with 2 and 14 d.f. is 3.739.)

**Solution [5 marks]:**

```
> C <- matrix(c(0,1,-1,0,0,0,1,-1),2,4,byrow=T)
> (Fstat <- t(C%*%b) %*% solve(C%*%XtXc%*%t(C)) %*% C%*%b / 2 / s2)

         [,1]
[1,] 3.614379
```

The effect of shift is not significant at a 5% level.

### Question 7 (14 marks)

(a) Discuss when it is best to use a completely randomised design, complete block design, or Latin square design.

**Solution [3 marks]:** A CRD is always useful, but may not perform as well as the other two if there are confounding factors (which are relevant). A CBD is best used when there is one confounding factor which must be dealt with. A Latin square design is used for two or more confounding factors, when a CBD is impractical.

(b) For a complete block design, why do we fit an additive model and not an interaction model?

**Solution [2 marks]:** An interaction model makes little sense, as the purpose of blocking is to isolate and remove the block effects.

(c) Write down a design matrix and parameter vector for a balanced incomplete block design for a model with 3 treatments and 3 blocks, each of size 2.

**Solution [2 marks]:** One possible design matrix and parameter vector is

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}.$$

(d) Calculate the reduced design matrix $X_{2|1}$ for this model.

**Solution [5 marks]:**

```
> X1 <- matrix(0,6,4)
> X1[,1] <- 1
> X1[cbind(1:6,rep(2:4,each=2))] <- 1
> X2 <- matrix(0,6,3)
> X2[cbind(1:6,c(1,2,1,3,2,3))] <- 1
> X1tX1c <- matrix(0,4,4)
> X1tX1c[2:4,2:4] <- solve((t(X1)%*%X1)[2:4,2:4])
> (diag(6) - X1%*%X1tX1c%*%t(X1))%*%X2

     [,1] [,2] [,3]
[1,]  0.5 -0.5  0.0
[2,] -0.5  0.5  0.0
[3,]  0.5  0.0 -0.5
[4,] -0.5  0.0  0.5
[5,]  0.0  0.5 -0.5
[6,]  0.0 -0.5  0.5
```

(e) Do you expect the reduced normal equations for this model to have the same solution as the normal equations for a completely randomised design of 6 experimental units over 3 treatments?

**Solution [2 marks]:** In general, no; the blocks are not orthogonal to the treatments, so the BIBD does not have the same solution as a CRD.

**End of Exam—Total Available Marks = 90.**