COPY

# MAST30025 (620-328) Linear Statistical Models

## Semester 1 Exam, 2011

Department of Mathematics and Statistics
The University of Melbourne

**Exam duration: 3 hours**
**Reading time: 15 minutes**
**This exam has 7 pages, including this page.**

| |
|---|
| *Authorised materials:* |
| Scientific calculators are permitted, but not graphical calculators. |
| One A4 double-sided handwritten sheet of notes. |

| |
|---|
| *Instructions to invigilators:* |
| The exam paper may be taken out of the examination room. |

| |
|---|
| *Instructions to students:* |
| There are 6 questions. All questions should be attempted. |
| The number of marks for each question is indicated. |
| The total number of marks available is 80. |

1. [13 marks] Consider the column vectors

$$\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}.$$

   (a) [1 mark] Show that these vectors are mutually orthogonal.

   (b) [1 mark] What constant $c$ makes the following matrix orthogonal?

$$P = c \begin{bmatrix} 1 & 2 & 2 \\ 2 & -2 & 1 \\ 2 & 1 & -2 \end{bmatrix}$$

   (c) [2 marks] Let $A = P\Lambda P^T$, where

$$\Lambda = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

   What are the rank $r(A)$ and trace $tr(A)$?

   (d) [3 marks] Show that

$$A^c = P \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} P^T$$

   is a conditional inverse of $A$.

   (e) [3 marks] Find a conditional inverse of $A^T A$.

   (f) [3 marks] Show that for your choice of $(A^T A)^c$ above, $A(A^T A)^c A^T$ is symmetric and idempotent (by direct calculation or otherwise).

2. [10 marks] Consider the following ANCOVA model, with a single factor and a single regression variable $x$,

$$y_{ij} = \mu + \tau_i + \gamma x_{ij} + \epsilon_{ij}$$

   Suppose that the factor has two levels, and that for each level there are two observations. Also suppose that $\sum_j x_{1j} = 0$, $\sum_j x_{2j} = 1$, and $\sum_{i,j} x_{ij}^2 = 3$.

   (a) [1 mark] What are the parameter vector $\beta$ and the design matrix $X$ for this model?

   (b) [3 marks] Write down $X^T X$ and hence show that it has rank $r(X^T X) = 3$ (provided the $x_{ij}$ are not pathological).

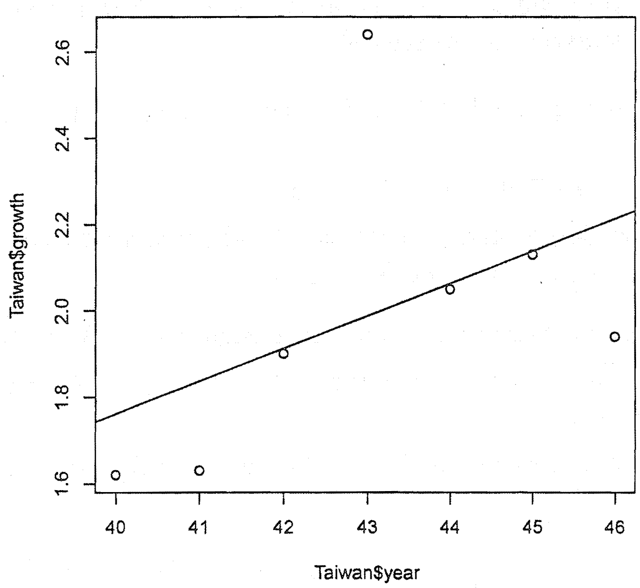
   (c) [3 marks] Give a conditional inverse for $X^T X$. You may use the fact that, when it exists,

$$\begin{bmatrix} x & 0 & a \\ 0 & y & b \\ a & b & c \end{bmatrix}^{-1} = \frac{1}{cxy - a^2 y - b^2 x} \begin{bmatrix} cy - b^2 & ab & -ay \\ ab & cx - a^2 & -bx \\ -ay & -bx & xy \end{bmatrix}.$$

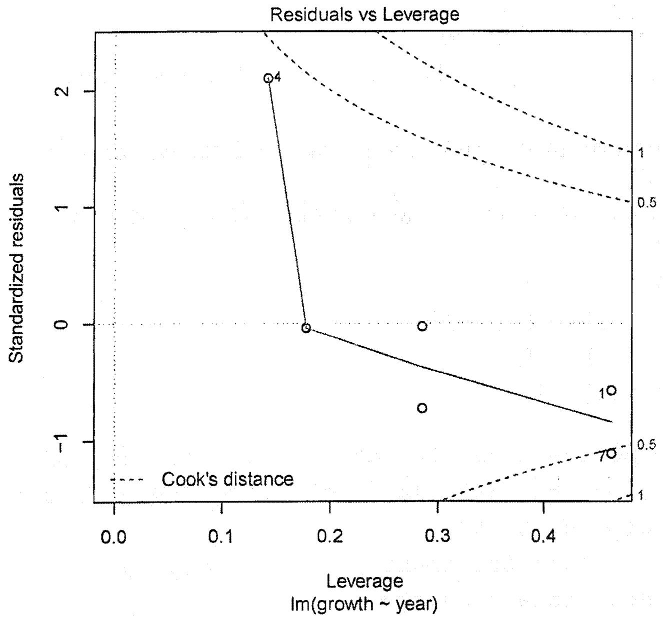
   (d) [3 marks] Give a solution to the normal equations.

3. [13 marks] The following data concerns population growth in Taiwan.

```
> Taiwan <- data.frame(year = 40:46, growth = c(1.62, 1.63, 1.9,
+     2.64, 2.05, 2.13, 1.94))
> model <- lm(growth ~ year, data = Taiwan)
> summary(model)

Call:
lm(formula = growth ~ year, data = Taiwan)

Residuals:
        1         2         3         4         5         6         7
-0.141071 -0.206429 -0.011786  0.652857 -0.012500 -0.007857 -0.273214

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.25321    2.73157  -0.459    0.666
year         0.07536    0.06346   1.188    0.288

Residual standard error: 0.3358 on 5 degrees of freedom
Multiple R-squared:  0.22,      Adjusted R-squared: 0.064
F-statistic:  1.41 on 1 and 5 DF,  p-value: 0.2883

> plot(Taiwan$year, Taiwan$growth)
> abline(coef = model$coefficients)
```



```
> plot(model, which = 5)
```




(a) [6 marks] The second plot indicates there may be some outliers. Define standardised residuals, leverage and Cook's distance, and briefly explain how they are used for regression diagnostics.

(b) [4 marks] Using the output above and below, calculate the Cook's distance for observation 7. (Do not just read it from the graph.)

```
> X <- matrix(nrow = 7, ncol = 2)
> X[, 1] <- 1
> X[, 2] <- Taiwan$year
> (H <- X %*% solve(t(X) %*% X) %*% t(X))
            [,1]          [,2]       [,3]      [,4]       [,5]          [,6]
[1,]  0.46428571  3.571429e-01 0.25000000 0.1428571 0.03571429 -7.142857e-02
[2,]  0.35714286  2.857143e-01 0.21428571 0.1428571 0.07142857 -6.938894e-16
[3,]  0.25000000  2.142857e-01 0.17857143 0.1428571 0.10714286  7.142857e-02
[4,]  0.14285714  1.428571e-01 0.14285714 0.1428571 0.14285714  1.428571e-01
[5,]  0.03571429  7.142857e-02 0.10714286 0.1428571 0.17857143  2.142857e-01
[6,] -0.07142857 -7.216450e-16 0.07142857 0.1428571 0.21428571  2.857143e-01
[7,] -0.17857143 -7.142857e-02 0.03571429 0.1428571 0.25000000  3.571429e-01
            [,7]
[1,] -0.17857143
[2,] -0.07142857
[3,]  0.03571429
[4,]  0.14285714
[5,]  0.25000000
[6,]  0.35714286
[7,]  0.46428571
```

(c) [3 marks] Give a joint 95% confidence region for the intercept and slope. You may express your region as an inequality in matrix form, and note that the upper 5% point for an $F_{2,5}$ distribution is 5.79.

4. [22 marks] Suppose that $\mathbf{y} \sim MVN(\mu, I_n)$, and that $A$ is an $n \times n$ symmetric idempotent matrix of rank $k$.

   (a) [6 marks] Show that the rank equals the trace $r(A) = tr(A)$.

   (b) [6 marks] Show that $\mathbf{y}^T A \mathbf{y} \sim \chi^2_{k,\mu^T A\mu/2}$.

   (c) [10 marks] Suppose that $B$ is symmetric and $AB = 0$. Show that $\mathbf{y}^T A \mathbf{y}$ and $\mathbf{y}^T B \mathbf{y}$ are independent.
   If in addition $B$ is idempotent of rank $h$, what is the distribution of $\mathbf{y}^T A \mathbf{y} + \mathbf{y}^T B \mathbf{y}$?

5. [11 marks] Four tropical feeds were fed to baby chicks. The gains in weight (in grams) after two weeks were:

| | | | | | | |
|---|---|---|---|---|---|---|
| Feed A | 42 | 68 | 85 | | | |
| Feed B | 42 | 97 | 81 | 95 | 61 | 103 |
| Feed C | 61 | 112 | 30 | 89 | 63 | |
| Feed D | 169 | 137 | 169 | 111 | 154 | |

```
> chicks <- data.frame(gain = c(42, 68, 85, 42, 97, 81, 95, 61,
+     103, 61, 112, 30, 89, 63, 169, 137, 169, 111, 154), feed = rep(c("A",
+     "B", "C", "D"), c(3, 6, 5, 5)))
> options(contrasts = c("contr.treatment", "contr.poly"))
> model <- lm(gain ~ feed, data = chicks)
> summary(model)

Call:
lm(formula = gain ~ feed, data = chicks)

Residuals:
   Min     1Q Median     3Q    Max
-41.00 -14.92   3.00  19.00  41.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    65.00      14.94   4.351 0.000571 ***
feedB          14.83      18.30   0.811 0.430247
feedC           6.00      18.90   0.317 0.755251
feedD          83.00      18.90   4.392 0.000525 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.88 on 15 degrees of freedom
Multiple R-squared: 0.6758,     Adjusted R-squared: 0.6109
F-statistic: 10.42 on 3 and 15 DF,  p-value: 0.0005872
```

   (a) [6 marks] Give formulae for the $F$-statistic and $p$-value given on the last line of output above. Take care to define all the terms you use. What do you conclude from this $p$-value?

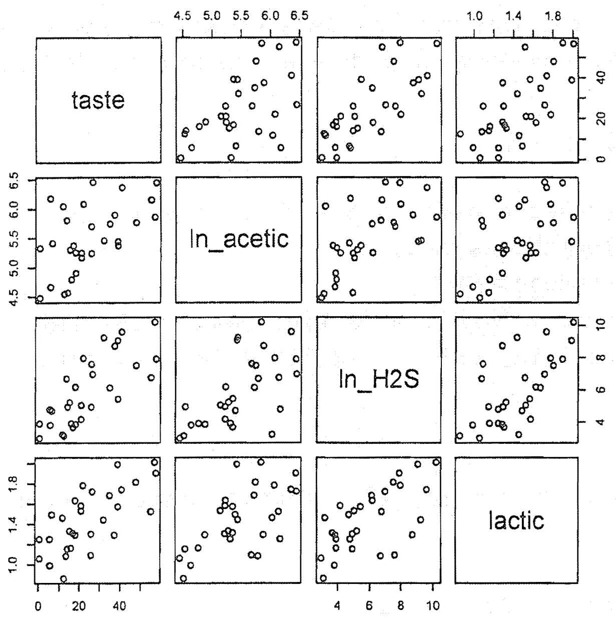
   (b) [2 marks] Estimate the mean weight gain for a chick on feed D.

   (c) [3 marks] Can you estimate the difference between the mean weight gain for a chick on feed D, and the mean weight gain for a chick on feed A, B or C? (That is, the difference between $\tau_4$ and $(\tau_1 + \tau_2 + \tau_3)/3$.) If so, explain why, and then give the estimate. If not, explain why not.

6. [11 marks] In an experiment to understand what makes a good cheese, a variety of cheeses were selected and subjected to a taste test by a panel of experts, who gave each cheese a numerical rating. The levels of acetic acid, hydrogen sulphide, and lactic acid were then measured for each cheese.

   Here is an analysis of the data in R.

```
> cheese <- read.table("cheese.csv", sep = ",", header = T)
> cheese$ln_acetic <- log(cheese$acetic)
> cheese$ln_H2S <- log(cheese$H2S)
> pairs(~taste + ln_acetic + ln_H2S + lactic, data = cheese)
> full_model <- lm(taste ~ ln_acetic + ln_H2S + lactic, data = cheese)
> summary(full_model)

Call:
lm(formula = taste ~ ln_acetic + ln_H2S + lactic, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.611  -1.008   4.907  25.448

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8725    19.7428  -1.462  0.15561
ln_acetic     0.3268     4.4612   0.073  0.94217
ln_H2S        3.9121     1.2486   3.133  0.00425 **
lactic       19.6701     8.6287   2.280  0.03108 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518,      Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```




(a) [1 mark] Why were the variables `acetic` and `H2S` transformed?

(b) [1 mark] Write down the fitted model.

(c) [7 marks] Suppose that a new cheese has the following measured values

$$\texttt{acetic} = 200, \texttt{H2S} = 2000, \texttt{lactic} = 1.5$$

For these values a 95% confidence interval for the mean taste is $(25.75, 38.45)$. Give a 95% prediction interval for the `taste` of this cheese. (Note that the upper 2.5% point for a $t_{26}$ distribution is 2.056.)

(d) [2 marks] If you were to perform one step of backward elimination, which variable would you remove, if any, and why?

End of examination