MAST30025 Linear Statistical Models

Semester 1 Exam, 2012

Department of Mathematics and Statistics
The University of Melbourne

Exam duration: 3 hours
Reading time: 15 minutes
This exam has 6 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 9 questions. All questions should be attempted.
The number of marks for each question is indicated.
The total number of marks available is 85.

1. [17 marks] Let $\mathbf{y} \sim N(\boldsymbol{\mu}, I_n)$.

   (a) If A is a symmetric idempotent $n \times n$ matrix, then what are the distributions of $A\mathbf{y}$ and $\mathbf{y}^T A\mathbf{y}$?

   (b) Given that A and B are $n \times n$ symmetric matrices such that $AB = 0$, prove that $\mathbf{y}^T A\mathbf{y}$ and $\mathbf{y}^T B\mathbf{y}$ are independent. State clearly any results you appeal to.

   Suppose now that $\mathbf{y} \sim N(\boldsymbol{\mu}, V)$ for some $n \times n$ matrix V.

   (c) For an $n \times n$ symmetric matrix A and an $m \times n$ matrix B, state (do not prove) conditions for $\mathbf{y}^T A\mathbf{y}$ and $B\mathbf{y}$ to be independent.

   (d) Suppose that

$$V = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

   Show that $\bar{y} = \sum_i y_i/n$ and $s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$ are independent.

   (e) Show that if A is symmetric and idempotent then it is also positive semidefinite.

   (f) Show that if A is idempotent then A and I are both conditional inverses of A.

2. [3 marks] State the Gauss-Markov theorem for a linear model of full rank.

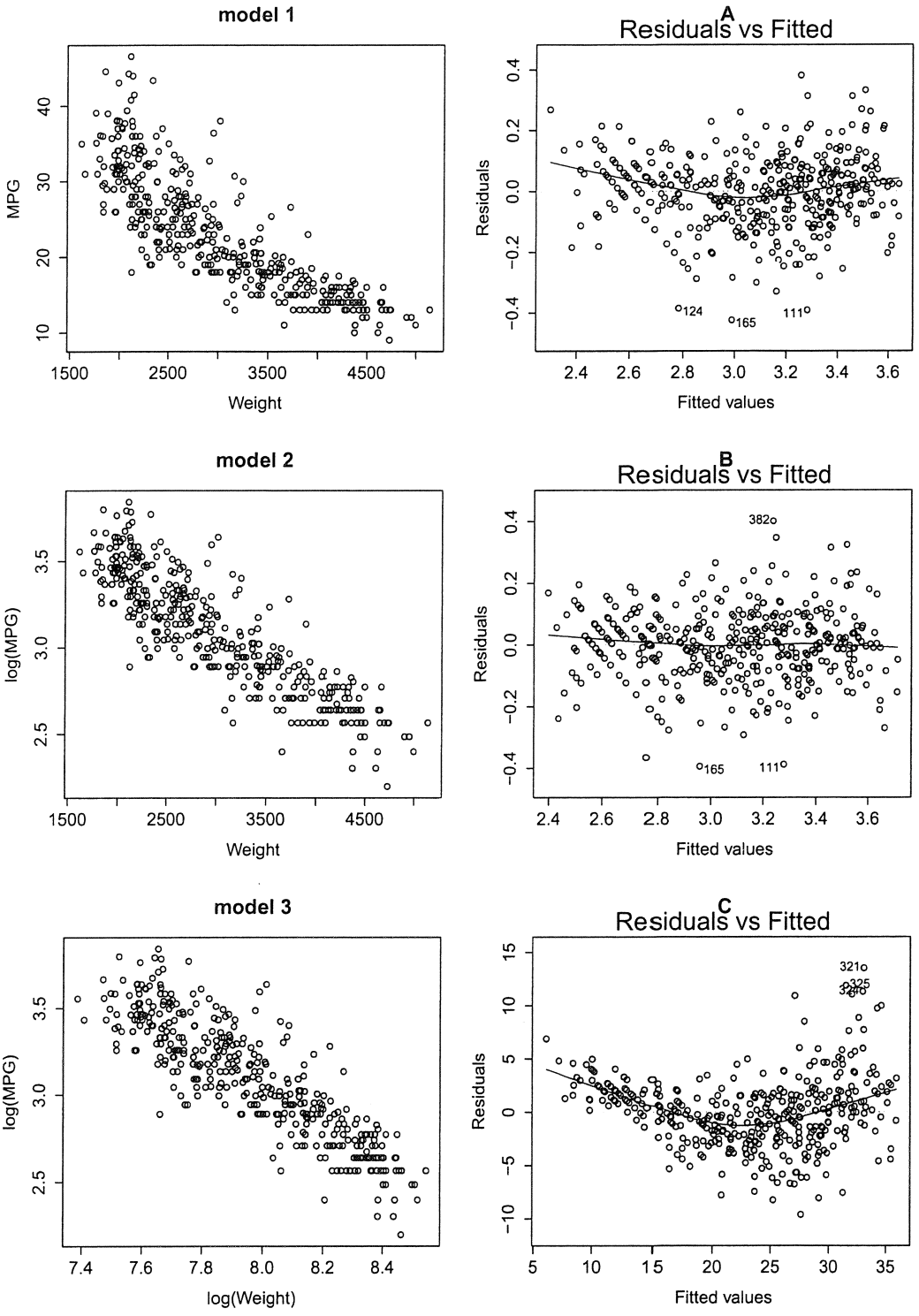
3. [5 marks] Why are symmetric idempotent matrices important in the theory of linear models? Be specific in your answer, and use no more than one page.

4. [13 marks] This question uses some data on cars collected in 1983 by Ernesto Ramos and David Donoho. For 392 cars we have the following variables: MPG (Miles Per Gallon), Cylinders, Horsepower, Weight, Acceleration, Year, Origin (1 for USA, 2 for Japan, or 3 for Europe). We wish to develop a model for MPG in terms of the other variables.

   (a) Three different models were fitted, using MPG or log(MPG) as the response and Weight or log(Weight) as a dependent variable. The residuals for these three models are given on the right in the figure below, possibly out of order.

      i. Indicate which residual plot (A, B, or C) should go with which model (1, 2, or 3).

      ii. What best describes the relationship between MPG and Weight: linear, polynomial, or exponential?



(b) Consider the following R output.

```
> model <- lm(log(MPG) ~ Cylinders + log(Horsepower) + log(Weight) +
              log(Acceleration) + Year + Origin, data = cars83)
> drop1(model, scope = ~ ., test = "F")
Single term deletions

Model:
log(MPG) ~ Cylinders + log(Horsepower) + log(Weight) + log(Acceleration) +
```

```
    Year + Origin
                  Df Sum of Sq    RSS     AIC  F value    Pr(>F)
<none>                         4.8848 -1703.0
Cylinders          1    0.0331 4.9179 -1702.3   2.6024  0.107521
log(Horsepower)    1    0.3056 5.1904 -1681.2  24.0207 1.408e-06 ***
log(Weight)        1    0.9084 5.7932 -1638.1  71.4078 6.091e-16 ***
log(Acceleration)  1    0.1214 5.0062 -1695.3   9.5420  0.002154 **
Year               1    3.7309 8.6157 -1482.5 293.2918 < 2.2e-16 ***
Origin             2    0.1140 4.9988 -1697.9   4.4821  0.011908 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model2 <- lm(log(MPG) ~ log(Horsepower) + log(Weight) + log(Acceleration) +
               Year + Origin, data = cars83)
> summary(model2)

Call:
lm(formula = log(MPG) ~ log(Horsepower) + log(Weight) + log(Acceleration) +
    Year + Origin, data = cars83)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38165 -0.07105  0.00479  0.06803  0.38237

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.50051    0.30838  24.323  < 2e-16 ***
log(Horsepower)   -0.28272    0.05791  -4.882 1.54e-06 ***
log(Weight)       -0.63061    0.05757 -10.953  < 2e-16 ***
log(Acceleration) -0.15805    0.05652  -2.796  0.00543 **
Year               0.03062    0.00174  17.595  < 2e-16 ***
Origin2            0.05405    0.01773   3.049  0.00246 **
Origin3            0.04716    0.01825   2.584  0.01014 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.113 on 385 degrees of freedom
Multiple R-squared: 0.8912,Adjusted R-squared: 0.8895
F-statistic: 525.7 on 6 and 385 DF,  p-value: < 2.2e-16
```

i. Using backwards elimination with the F-test, which variable should we remove first, and why?
   If we were to use the AIC instead of the F-test, would the outcome be the same?

ii. Give a 95% confidence interval for the log(Weight) coefficient in the second model.
   Note that for a t distribution with 385 degrees of freedom, $t_{0.025} = 1.9661$.

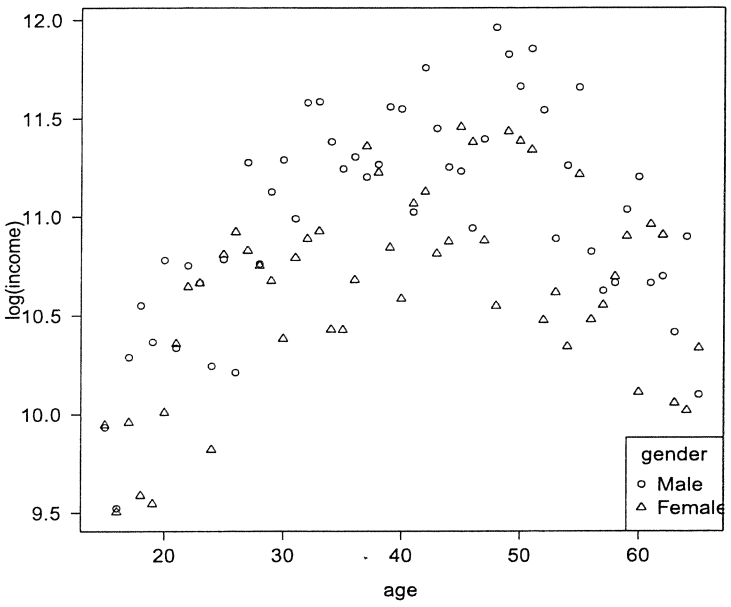
iii. On average, which cars are the most fuel efficient, American, Japanese, or European? Use the fitted model to justify your answer.

iv. What would be the p-value of the F-test used to compare the model without Cylinders or log(Acceleration), with the model without Cylinders?
   Also, what would the value of the test statistic be, and how many degrees of freedom would we use?

v. What are the residual sum of squares and the corrected total sum of squares for model two?

5. [11 marks]

(a) The figure below gives the log income of over 100 men and women, according to their age.
Suggest a linear model for these data.



(b) Consider the following two-way classification model with interaction:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{1}$$

for $1 \leq i \leq 2$, $1 \leq j \leq 2$ and $1 \leq k \leq 2$.

i. Write down the design matrix, and give its rank.

ii. Give a full rank version of this model. Write down its design matrix, and give its rank.

iii. Give a version of the model without interaction. Write down its design matrix, and give its rank.

iv. Explain how to use the model (1) to test for the presence of an interaction. Your explanation should include a precise description of an appropriate hypothesis test.

6. [10 marks] Consider a one-way classification problem, for which we have the following data

Factor level:	A	B	C
No. observations:	12	8	16
Mean response:	10.2	11.3	8.4

We are also given $s^2 = 4.9$.

Let μ_A, μ_B and μ_C be the true means for each factor level.

(a) Give a 95% CI for $\mu_A - \mu_B$.
Note that for a t distribution with 33 degrees of freedom, $t_{0.025} = 2.0345$.

(b) Calculate the test statistic used to test the hypothesis $\mu_A = \mu_B = \mu_C$.
What are the degrees of freedom?

7. [12 marks] Consider the two-way classification model

$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\theta}^T = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2)$ and μ is the overall mean, α_i the effect of factor 1 at level i, and β_j the effect of factor 2 at level j.

   (a) Which of the following quantities are estimable?

   i. μ
   ii. $\mu + \alpha_1$
   iii. $\alpha_1 - \alpha_2$
   iv. $\mu + \alpha_1 - \alpha_2$
   v. $\mu + 2\alpha_1 - \alpha_2$
   vi. $\alpha_1 - 2\alpha_2 + \alpha_3$
   vii. $\alpha_1 - \beta_2$

   (b) Define what it means for $\mathbf{t}^T\boldsymbol{\theta}$ to be estimable.

   (c) Show that $\mathbf{t}^T\boldsymbol{\theta}$ is estimable if

$$\mathbf{t}^T(X^TX)^cX^TX = \mathbf{t}^T.$$

   (Do *not* show the converse.)

8. [8 marks] Give examples of the following types of experimental designs. In each case suppose that we have three treatments and at least six experimental units.

   (a) A completely randomised design.

   (b) A complete block design.

   (c) A latin square design.

   (d) An incomplete block design.

9. [6 marks] You are asked to assess the effectiveness of two cholesterol reducing drugs, and given the resources to use thirty test subjects, all of whom will be recruited by a local hospital and known to have high cholesterol. You may specify what sort of test subjects you want (there is a large pool available), but each subject may only be tested once.

   Describe an appropriate experimental design. Take care to consider the human as well as the mathematical aspects of the design.

End of examination