COPY

620-328 Linear Statistical Models

Semester 1 Exam — June 18, 2010

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:
Calculators are permitted, but you may be requested to erase the memory of programmable calculators.
Any printed or handwritten material is permitted, including textbooks.
Computers are NOT permitted.

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 approximate number of marks for each question is indicated.
The total number of marks available is 100.

1. [16 marks]

   (a) Show that the columns of an orthogonal matrix form an orthonormal set.

   (b) Show that the hat matrix, $H = X(X^TX)^{-1}X^T$, is idempotent.

   (c) Find the rank of $\begin{bmatrix} 3 & -5 & 2 & 0 \\ 6 & -9 & -2 & -3 \\ 0 & -1 & 6 & 3 \end{bmatrix}$.

   (d) Let A be an $n \times n$ symmetric matrix with all eigenvalues either 0 or 1. Show that A is idempotent. *(Hint: The eigenvectors of A form a basis of $\mathcal{R}^n$.)*

   (e) Show from first principles that if A is a symmetric matrix, $\frac{\partial}{\partial \mathbf{y}}\mathbf{y}^T A\mathbf{y} = 2A\mathbf{y}$.

2. [13 marks] Let

$$A = \frac{1}{3}\begin{bmatrix} -1 & 0 & 2 \\ 0 & 3 & 0 \\ 2 & 0 & -1 \end{bmatrix}, V = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix},$$

and let $\mathbf{y}$ be a normal random vector with mean 0 and variance V.

   (a) Find $E[\mathbf{y}^T A\mathbf{y}]$.

   (b) Find $E\left[\frac{\partial}{\partial \mathbf{y}}\mathbf{y}^T A\mathbf{y}\right]$.

   (c) Describe the distribution of $\mathbf{y}^T A\mathbf{y}$.

   (d) Let Z be a standard normal random variable that is independent from $\mathbf{y}$. Describe the distribution of $\frac{Z^2}{\mathbf{y}^T A\mathbf{y}}$.

   (e) Suppose that C and D are matrices such that $\mathbf{y}^T C\mathbf{y}$ and $\mathbf{y}^T D\mathbf{y}$ have noncentral χ^2 distributions with k_1 and k_2 degrees of freedom respectively, and suppose that $CD = 0$ and $CV = VC$. Show that $\mathbf{y}^T C\mathbf{y} + \mathbf{y}^T D\mathbf{y}$ has a noncentral χ^2 disribution.

   (f) What is the degrees of freedom of $\mathbf{y}^T C\mathbf{y} + \mathbf{y}^T D\mathbf{y}$?

3. [18 marks] We wish to fit a linear model to explain the length of time that a motor will run for, based on the amount of fuel left in the tank. A study is performed and the following data collected:

Amount of fuel (litres)	Time motor runs (hours)
4	1.1
7	1.8
14	1.9
15	2.2
18	2.6
22	2.4

We fit the linear model $\mathbf{y} = X\boldsymbol{\beta} + \varepsilon$, where $\mathbf{y}$ is the vector of response values, X is the design matrix, ε is the vector of errors, and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ is the parameter vector. Here β_0 is an intercept term and β_1 is a parameter associated with the amount of fuel.

(a) Calculate the normal equations for this model.

(b) Calculate the least square estimators for this model.

(c) Name one disadvantage to using the maximum likelihood estimator as an estimator for σ^2.

(d) What is the difference between a BLUE and an UMVUE estimator?

(e) Calculate the leverage of the first data point.

(f) Consider the following R output for the model:

```
> cars <- data.frame(fuel = c(4, 7, 14, 15, 18, 22), time = c(1.1,
+     1.8, 1.9, 2.2, 2.6, 2.4))
> model <- lm(time ~ fuel, data = cars)
> newcar <- data.frame(fuel = 5)
> predict(model, newcar, interval = "confidence", level = 0.95)

       fit       lwr      upr
1 1.406158 0.9145878 1.897729

> predict(model, newcar, interval = "prediction", level = 0.95)

       fit       lwr      upr
1 1.406158 0.5381483 2.274168
```

Interpret this output.

(g) Find a joint 95% confidence region for the parameters. Express this as a single scalar inequality (you do not need to simplify it, however). You may take the critical value of an F distribution with 2 and 4 degrees of freedom as $f_{0.05} = 6.944$ and the residual sum of squares as $SS_{Res} = 0.266$.

(h) Under what circumstances would a logarithmic transformation of the response variable be necessary?

4. [19 marks] Consider the study in question 3.

   (a) Test for model adequacy at the 99% level using a corrected sums of squares approach. You may take the critical value of an F distribution with 1 and 4 degrees of freedom as $f_{0.01} = 21.20$, and recall that $SS_{Res} = 0.266$.

   (b) Explain the difference between testing $\beta_1 = 0$ in the presence of β_0 and in the presence of no other parameters.

   (c) Suppose we have two models, M_1 and M_2, where M_1 contains all the variables in M_2 (as well as some other variables). Explain why R^2 for the M_1 model is larger than R^2 for the M_2 model.

   (d) Explain why stepwise selection using Akaike's information criterion does not necessarily find the model with the overall lowest AIC.

   (e) Calculate Mallows' C_p statistic for the model with only an intercept term.

   (f) Suppose we want to test $\beta_1 \neq 2\beta_0 - 1$ using a general linear hypothesis. Write down the null and alternative hypotheses, in matrix form.

   (g) Is this model mutually orthogonal? Why or why not?

5. [21 marks] A study of three major coal seams in a region is conducted, to compare the sulfur content of coal drawn from each seam. The following data is collected (in terms of percentage of sulfur):

	Seam	
1	2	3
1.8	2.3	1.8
1.6	2	1.5

The linear model that we use is

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

where τ_i is the effect on the sulfur content associated with seam i, and ε_{ij} is the error for the jth sample from seam i.

(a) Calculate the normal equations for this model.

(b) Show that if $\mathbf{t}^T\boldsymbol{\beta}$ is estimable (where $\boldsymbol{\beta}$ is the parameter vector), then $\mathbf{t}^T(X^TX)^c\mathbf{t}$ is unique (i.e. invariant to the choice of conditional inverse). *(Hint: Consider two different conditional inverses and show that the expression is the same for both of them.)*

(c) Find a general form for all solutions of the normal equations.

(d) Estimate $\mu + \tau_2$.

(e) Prove that the difference between a treatment effect and the average of the other treatment effects is estimable.

(f) Calculate s^2, the estimator for the variance of the errors.

(g) Calculate a 95% confidence interval for the average sulfur content of coal seam 3. You may take the critical value for a t distribution with 3 degrees of freedom to be $t_{0.025} = 3.182$.

6. [13 marks] We study the efficiency of a chemical reactor, based on the percentage of material that is reacted with a given set of inputs. There are 4 factors involved: feed rate, catalyst amount, agitation rate, and temperature. Each factor has two different levels, as shown in the following table:

	Factor	1	2
1	Feed rate (l/min)	10	15
2	Catalyst (%)	1	2
3	Agitation rate (rpm)	100	120
4	Temperature (°C)	140	180

Two samples from each combination of factor levels (i.e. 32 data points in total) are collected and the following R output is obtained.

```
> reactor <- read.csv("reactor.csv")
> options(contrasts = c("contr.treatment", "contr.poly"))
> model <- lm(reacted ~ (feed + catalyst + agitation + temperature)^2,
+     data = reactor)
> summary(model)

Call:
lm(formula = reacted ~ (feed + catalyst + agitation + temperature)^2,
    data = reactor)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9687  -6.3750  -0.1563   6.5000  12.0313

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            104.094     34.015   3.060  0.00594 **
feed                    -3.250     15.741  -0.206  0.83841
catalyst               -35.125     15.741  -2.231  0.03668 *
agitation              -12.125     15.741  -0.770  0.44971
temperature            -31.125     15.741  -1.977  0.06128 .
feed:catalyst            5.375      5.950   0.903  0.37654
feed:agitation           1.375      5.950   0.231  0.81946
feed:temperature        -4.625      5.950  -0.777  0.44561
catalyst:agitation       1.625      5.950   0.273  0.78742
catalyst:temperature    28.625      5.950   4.811 9.37e-05 ***
agitation:temperature    4.625      5.950   0.777  0.44561
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.414 on 21 degrees of freedom
Multiple R-squared:  0.79,       Adjusted R-squared: 0.6901
F-statistic: 7.902 on 10 and 21 DF,  p-value: 3.828e-05
```

```
> model2 <- lm(reacted ~ feed + catalyst + agitation + temperature,
+     data = reactor)
> linear.hypothesis(model2, C, dst)

Linear hypothesis test

Hypothesis:
feed = 0
catalyst - agitation = 0

Model 1: reacted ~ feed + catalyst + agitation + temperature
Model 2: restricted model

  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     27  3277.8
2     29  4721.9 -2   -1444.0 5.9473 0.007243 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) Estimate the percentage of reacted material if the feed rate is 10 l/min, the agitation rate is 120 rpm, the temperature is 140° C and there is 2% catalyst.

(b) Let μ be the overall mean and τ_{ij} be the effect on the mean of factor i being in level j. Express the hypothesis that is being tested in the `linear.hypothesis` command, in the matrix form of a generalised linear hypothesis. Define the parameter vector ($\boldsymbol{\beta}$) that you use.

(c) Show that this hypothesis is testable.

(d) Does feed rate have a significant effect on reacted material?

(e) What significant interactions are there?

(f) Give R commands for constructing an `lm` model with only the significant interaction terms.

(g) Give R commands for testing the presence of interaction between any two factors (as a single test, ignoring possible three- or more-factor interactions).

End of examination