COPY

MAST30025 Linear Statistical Models

Semester 1 Exam, 2013

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 8 questions. All questions should be attempted.
The number of marks for each question is indicated.
The total number of marks available is 60.

1. [6 marks]

   (a) Prove that a symmetric matrix is idempotent if and only if its eigenvalues are all 0 or 1.

   (b) Suppose that A is symmetric and $A^2 = A^3$. Show that A is idempotent.

2. [7 marks]

   (a) What does it mean for a matrix A to be (strictly) positive definite?

   (b) Show that if A is symmetric and positive definite then

      i. Its eigenvalues are all strictly positive

      ii. A^{-1} exists and is symmetric and positive definite

      iii. A has a square root which is symmetric and positive definite

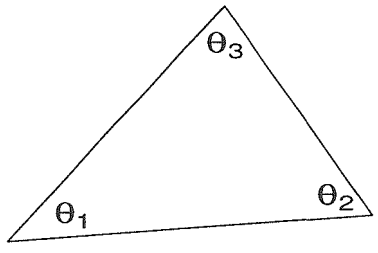
3. [5 marks] Suppose that X is an $n \times p$ matrix, $n > p$, of full rank. Using only X and the identity I, but not I alone, give examples of the following sorts of matrix

   (a) Symmetric and (strictly) positive definite

   (b) Symmetric and positive semidefinite, but not positive definite

   (c) Two symmetric and idempotent matrices that commute

   (d) Orthogonal



4. [7 marks] Let θ_1, θ_2 and θ_3 be the internal angles of a triangle (in radians). Each θ_i is measured, with independent errors $\varepsilon_i \sim N(0, \sigma^2)$.

   Let y_i be the measured value of θ_i. We consider a model for $\mathbf{y} = (y_1, y_2, y_3)^T$ of the form

$$\mathbf{y} = X\beta + \delta + \varepsilon$$

   where $\beta = (\theta_1, \theta_2)^T$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)^T$.

   (a) What are the design matrix X and the offset δ?

   (b) What is s^2, the sample variance, for this model?

   (c) How would you test H_0: $\theta_1 = \theta_2 = \theta_3$ against a general alternative? Give the test statistic and its distribution under H_0, assuming that the model is correct.

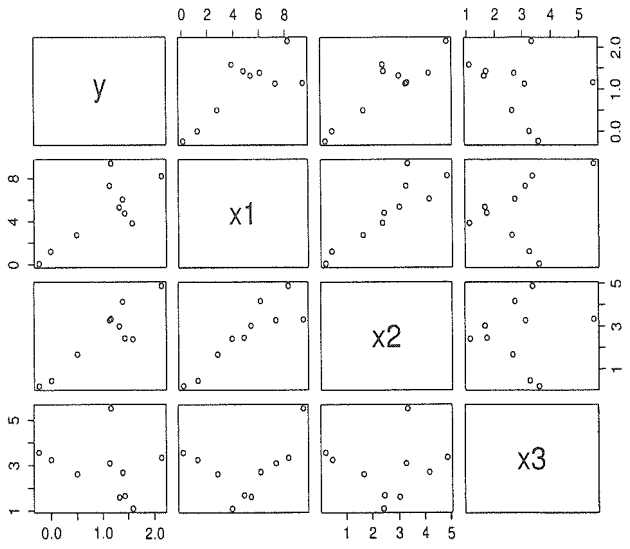
5. [5 marks] Define the following distributions

   (a) Multivariate normal

   (b) Noncentral χ^2

   (c) t

   (d) F



6. [14 marks] Consider the pairs plot above. A linear model was fitted for y in terms of x_1, x_2 and x_3. Here is a summary of the fitted model

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29462 -0.17353  0.02838  0.12575  0.27892

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.55829    0.28014   1.993   0.0933 .
x1           0.11624    0.08226   1.413   0.2073
x2           0.24512    0.15663   1.565   0.1686
x3          -0.25214    0.08718  -2.892   0.0276 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2457 on 6 degrees of freedom
Multiple R-squared: 0.9258,Adjusted R-squared: 0.8888
F-statistic: 24.97 on 3 and 6 DF,  p-value: 0.0008667
```

(a) The following values all appear in the model summary. In each case give formula for that value in terms of the response vector $\mathbf{y}$ and the design matrix X.

i. 0.55829 (intercept estimate)

ii. 0.2457 (residual standard error)

iii. -0.25214 (x3 estimate)

iv. 0.08718 (x3 standard error)

v. 0.0276 (x3 p-value)

vi. 24.97 (F-statistic)

(b) The F-statistic 24.97 is for comparing two models. What are they, and what do you conclude from this test?

(c) In this case we have the following value for $C = (X^T X)^{-1}$

```
       [,1]   [,2]   [,3]   [,4]
[1,]  1.300  0.191 -0.463 -0.329
[2,]  0.191  0.112 -0.200 -0.078
[3,] -0.463 -0.200  0.406  0.137
[4,] -0.329 -0.078  0.137  0.126
```

Give a 95% prediction interval for y when $x_1 = 10$, $x_2 = 0$ and $x_3 = 0$. You may use the fact that if $T \sim t_6$ then $\Pr(T < 2.447) = 0.975$.

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

(e) Do you think that removing x_1 from the model would increase/decrease/have no effect on the significance of x_2? Why?

7. [10 marks] The following "whiteness" ratings were obtained with specially designed equipment for 12 loads of washing, distributed over three different washing machines

	Machine 1	Machine 2	Machine 3	Totals
Detergent A	45	43	51	139
Detergent B	47	46	52	145
Detergent C	48	50	55	153
Detergent D	42	37	49	128
Totals	182	176	207	565

Consider the following R output

```
> y <- c(45,47,48,42,43,46,50,37,51,52,55,49)
> detergent <- rep(c("A","B","C","D"), 3)
> machine <- c(rep(1,4), rep(2,4), rep(3,4))
> X <- matrix(0, 12, 8)
> X[,1] <- 1
> X[detergent=="A",2] <- 1
> X[detergent=="B",3] <- 1
> X[detergent=="C",4] <- 1
> X[detergent=="D",5] <- 1
> X[machine==1,6] <- 1
> X[machine==2,7] <- 1
> X[machine==3,8] <- 1
```

```
> X
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
 [1,]    1    1    0    0    0    1    0    0
 [2,]    1    0    1    0    0    1    0    0
 [3,]    1    0    0    1    0    1    0    0
 [4,]    1    0    0    0    1    1    0    0
 [5,]    1    1    0    0    0    0    1    0
 [6,]    1    0    1    0    0    0    1    0
 [7,]    1    0    0    1    0    0    1    0
 [8,]    1    0    0    0    1    0    1    0
 [9,]    1    1    0    0    0    0    0    1
[10,]    1    0    1    0    0    0    0    1
[11,]    1    0    0    1    0    0    0    1
[12,]    1    0    0    0    1    0    0    1
> XtX <- t(X) %*% X
> XtXc <- matrix(0, 8, 8)
> XtXc[3:8,3:8] <- solve(XtX[3:8,3:8])
> 12*XtXc
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    0    0    0    0    0    0    0    0
[2,]    0    0    0    0    0    0    0    0
[3,]    0    0    8    4    4   -4   -4   -4
[4,]    0    0    4    8    4   -4   -4   -4
[5,]    0    0    4    4    8   -4   -4   -4
[6,]    0    0   -4   -4   -4    6    3    3
[7,]    0    0   -4   -4   -4    3    6    3
[8,]    0    0   -4   -4   -4    3    3    6
> b <- XtXc %*% t(X) %*% y
> b
           [,1]
[1,]  0.0000000
[2,]  0.0000000
[3,]  2.0000000
[4,]  4.6666667
[5,] -3.6666667
[6,] 44.7500000
[7,] 43.2500000
[8,] 51.0000000
> sum((y - X %*% b)^2)
[1] 18.833333
> C <- matrix(c(0,  1,-1, 0, 0,  0, 0, 0,
+               0,  1, 0,-1, 0,  0, 0, 0,
+               0,  1, 0, 0,-1,  0, 0, 0,
+               0,  0, 0, 0, 0,  1,-1, 0,
+               0,  0, 0, 0, 0,  1, 0,-1), 5, 8, byrow=T)
> C %*% b
        [,1]
[1,] -2.0000
[2,] -4.6667
[3,]  3.6667
[4,]  1.5000
[5,] -6.2500
```

```
> solve(C %*% XtXc %*% t(C))
      [,1]  [,2]  [,3]    [,4]    [,5]
[1,]  2.25 -0.75 -0.75  0.0000  0.0000
[2,] -0.75  2.25 -0.75  0.0000  0.0000
[3,] -0.75 -0.75  2.25  0.0000  0.0000
[4,]  0.00  0.00  0.00  2.6667 -1.3333
[5,]  0.00  0.00  0.00 -1.3333  2.6667
> solve(C %*% XtXc %*% t(C), C %*% b)
         [,1]
[1,]  -3.7500
[2,] -11.7500
[3,]  13.2500
[4,]  12.3333
[5,] -18.6667
> qt(.975, 6)
[1] 2.4469119
> qf(.95, 2, 6)
[1] 5.1432528
> qf(.95, 3, 6)
[1] 4.7570627
> qf(.95, 5, 6)
[1] 4.3873742
```

(a) Estimate the mean whiteness using detergent type C with machine type 2.

(b) Give a 95% confidence interval for your estimate

(c) Does the type of detergent or type of machine make a difference to the whiteness? Justify your answers using appropriate hypothesis tests (at the 95% level).

Some calculations will be required. You may assume that there is no interaction between the type of washing machine and the type of detergent.

8. [6 marks] A study is to be conducted to evaluate the effect of two types of contact lens on a person's peripheral vision. The evaluation will consist of measuring the reaction time to a given stimulation, in a situation where the reaction time is known to depend on a person's peripheral vision. The study will take a total of 40 observations, and the following suggestions have been made concerning their disposition.

(a) Amy says that she is willing to be "the subject" for the study. That is, to wear each type of lens and have her response time measured as often as needed. Give one point in favour and one point against this proposal.

(b) Ben suggests that it would be better to have 40 different subjects, and to allocate them at random to the different lens types. Give two reasons why this design might be better than the one suggested by Amy.

(c) Cat claims that it would be better to use 20 subjects, with each subject using, on separate occasions, each of the lens types. Give one point in favour and one point against this approach.

(d) Don proposes using 10 subjects, with each using each lens type twice. Give two points either for or against this proposal.

Which design do you prefer, and why?

End of examination