Exam 2017, questions and answers

Linear Statistical Models (University of Melbourne)

THE UNIVERSITY OF MELBOURNE

Student ID

Semester 1 Assessment, 2017

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)

SOLUTIONS

Authorised materials:

- Scientific calculators are permitted, but not graphical calculators.
- Two A4 double-sided handwritten sheets 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) Show that if A is a symmetric and idempotent matrix, then $r(A) = tr(A)$.

Solution [3 marks]: Diagonalise A as $A = PDP^T$. We know that A has eigenvalues which are only 0 or 1, hence D is a diagonal matrix with only 0 or 1s on the diagonal and therefore $r(D) = tr(D)$. Then

$$r(A) = r(PDP^T) = r(D) = tr(D) = tr(DP^TP) = tr(PDP^T) = tr(A).$$

(b) Show, without using the above result, that if X is an $n \times p$ matrix with $n > p$, then $r(X(X^TX)^cX^T) = r(X)$.

Solution [3 marks]:

$$r(X) \geq r(X(X^TX)^cX^T) \geq r(X(X^TX)^cX^TX) = r(X).$$

(c) Let $B\mathbf{x} = \mathbf{g}$ be a consistent linear system. Show that $\mathbf{x} = B^c\mathbf{g}$ is a solution to this system.

Solution [3 marks]: Since the system is consistent, it has a solution $\mathbf{x}_0$. Then

$$B(B^c\mathbf{g}) = BB^cB\mathbf{x}_0 = B\mathbf{x}_0 = \mathbf{g}.$$

Question 2 (13 marks) Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\right).$$

You are given the following R calculations.

```
> V <- matrix(c(2,1,1,3),2,2)
> P <- eigen(V)$vectors
> P%*%diag(sqrt(eigen(V)$values))%*%t(P)

          [,1]      [,2]
[1,] 1.3763819 0.3249197
[2,] 0.3249197 1.7013016

> P%*%diag(1/sqrt(eigen(V)$values))%*%t(P)

           [,1]       [,2]
[1,]  0.7608452 -0.1453085
[2,] -0.1453085  0.6155367
```

(a) Find two independent standard normal random variables which are linear combinations of $\mathbf{y}$ elements (and constants).

Solution [3 marks]: We know that

$$\mathbf{z} = V^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) = V^{-1/2}\mathbf{y} - V^{-1/2}\boldsymbol{\mu}$$

is multivariate normal with mean $\mathbf{0}$ and variance I, so its elements are the variables we require.

```
> mu <- c(0,-1)
> P%*%diag(1/sqrt(eigen(V)$values))%*%t(P)

           [,1]       [,2]
[1,]  0.7608452 -0.1453085
[2,] -0.1453085  0.6155367

> P%*%diag(1/sqrt(eigen(V)$values))%*%t(P)%*%mu

           [,1]
[1,]  0.1453085
[2,] -0.6155367
```

(b) Calculate $E[y_1^2 - 2y_1y_2]$.

Solution [3 marks]:

```
> A <- matrix(c(1,-1,-1,0),2,2)
> sum(diag(A%*%V)) + t(mu) %*% A %*% mu

     [,1]
[1,]    0
```

(c) Find all quadratic forms in $\mathbf{y}$ which have a non-central χ^2 distribution with 2 degrees of freedom.

Solution [4 marks]: For a quadratic form $\mathbf{y}^T A\mathbf{y}$ to have a non-central χ^2 distribution with 2 degrees of freedom, we require that AV be idempotent with rank 2. There is only one such 2×2 matrix, which is I_2. Hence the only such quadratic form is $\mathbf{y}^T V^{-1}\mathbf{y}$.

```
> solve(V)

     [,1] [,2]
[1,]  0.6 -0.2
[2,] -0.2  0.4
```

(d) Show that $4y_1^2 - 4y_1y_2 + y_2^2$ is independent of $y_1^2 + 6y_1y_2 + 9y_2^2$.

Solution [3 marks]:

```
> A <- matrix(c(4,-2,-2,1),2,2)
> B <- matrix(c(1,3,3,9),2,2)
> A%*%V%*%B

     [,1] [,2]
[1,]    0    0
[2,]    0    0
```

Question 3 (14 marks) The following data is a sample of 5 random countries. For each country we measure the following variables:

- `logPPgdp`: The logarithm of the 2001 gross domestic product per person in US dollars;

- `logFertility`: The logarithm of the birth rate per 1000 females in the year 2000;
- `Purban`: The percentage of the population which lives in an urban area.

Name	logFertility	logPPgdp	Purban
Moldova	0.23	5.8	41
Netherlands	0.38	10.1	90
Estonia	0.14	8.3	69
Uganda	1.36	5.5	15
Hungary	0.13	8.6	65

We wish to predict `logFertility` using `logPPgdp` and `Purban`, using a linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *with no constant (intercept) term*.

You are given the following R calculations.

```
> qt(0.975,1:5)

[1] 12.706205  4.302653  3.182446  2.776445  2.570582

> qf(0.95,1,1:5)

[1] 161.447639  18.512821  10.127964   7.708647   6.607891

> qf(0.95,2,1:5)

[1] 199.500000  19.000000   9.552094   6.944272   5.786135
```

(a) Calculate the least squares estimates of $\boldsymbol{\beta}$.

Solution [4 marks]:

```
> X <- matrix(c(5.8,10.1,8.3,5.5,8.6,41,90,69,15,65),5,2)
> y <- c(0.23,0.38,0.14,1.36,0.13)
> (b <- solve(t(X)%*%X,t(X)%*%y))

             [,1]
[1,]  0.30973645
[2,] -0.03418006
```

(b) Calculate and interpret a 95% confidence interval for the parameter associated with `Purban`. You are given the sample variance $s^2 = 0.0887$.

Solution [3 marks]:

```
> n <- 5
> p <- 2
> (s2 <- sum((y-X%*%b)^2)/(n-p))

[1] 0.08865523

> b[2] + c(-1,1) * qt(0.975,df=n-p) * sqrt(s2 * solve(t(X)%*%X)[2,2])

[1] -0.065127396 -0.003232731
```

For every increase of 1% of population living in an urban area, the log-fertility goes down by an amount between 0.003 and 0.065.

(c) Test for model relevance at a 5% significance level.

Solution [3 marks]:

```
> SSReg <- t(y)%*%X%*%b
> SSReg/2/s2

         [,1]
[1,] 10.25001
```

As $10.25 > 9.55$, we conclude that the model is relevant at the 5% level.

(d) Croatia has a gross domestic product of \$4500 per person, and 58% of its population lives in an urban area. Calculate a 95% prediction interval for its birth rate.

Solution [4 marks]:

```
> tt <- c(log(4500),58)
> ci <- tt%*%b + c(-1,1) * qt(0.975,df=n-p) * sqrt(s2) *
+         sqrt(1 + t(tt) %*% solve(t(X)%*%X) %*% tt)
> exp(ci)

[1] 0.6404194 5.4284146
```

Question 4 (13 marks) Consider a full rank linear model $\mathbf{y} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$. We wish to derive the formula for a prediction interval for the sum of the responses of two independent future observations with predictors $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. This uses the unbiased point estimator $(\mathbf{x}_1 + \mathbf{x}_2)^T\mathbf{b}$, where $\mathbf{b}$ is the least squares estimator of $\boldsymbol{\beta}$.

(a) Calculate the variance of the prediction error of this estimator.

Solution [3 marks]: If the responses are y_1 and y_2, then the variance of the prediction error is

$$\begin{aligned}\text{var } (y_1 + y_2 - (\mathbf{x}_1 + \mathbf{x}_2)^T\mathbf{b}) &= \text{var } y_1 + \text{var } y_2 + \text{var } (\mathbf{x}_1 + \mathbf{x}_2)^T\mathbf{b} \\ &= 2\sigma^2 + (\mathbf{x}_1 + \mathbf{x}_2)^T(X^TX)^{-1}\sigma^2(\mathbf{x}_1 + \mathbf{x}_2) \\ &= \sigma^2\left[2 + (\mathbf{x}_1 + \mathbf{x}_2)^T(X^TX)^{-1}(\mathbf{x}_1 + \mathbf{x}_2)\right].\end{aligned}$$

(b) Show that this estimator is normally distributed.

Solution [2 marks]: This follows from the fact that $\mathbf{b}$ (or $\mathbf{y}$) is multivariate normal and any linear combination of multivariate normals is normal.

(c) Show that this estimator is independent of SS_{Res}.

Solution [2 marks]: This follows from the independence of $\mathbf{b}$ and SS_{Res}.

(d) Derive a t-distributed quantity based on this estimator and state its degrees of freedom.

Solution [3 marks]: A t-distributed quantity is

$$\frac{y_1 + y_2 - (\mathbf{x}_1 + \mathbf{x}_2)^T\mathbf{b}}{\sigma\sqrt{2 + (\mathbf{x}_1 + \mathbf{x}_2)^T(X^TX)^{-1}(\mathbf{x}_1 + \mathbf{x}_2)}} \Big/ \sqrt{\frac{SS_{Res}}{(n-p)\sigma^2}} = \frac{y_1 + y_2 - (\mathbf{x}_1 + \mathbf{x}_2)^T\mathbf{b}}{s\sqrt{2 + (\mathbf{x}_1 + \mathbf{x}_2)^T(X^TX)^{-1}(\mathbf{x}_1 + \mathbf{x}_2)}}.$$

This has $n-p$ degrees of freedom.

(e) Thus write down a formula for a $100(1-\alpha)\%$ prediction interval for the sum of two responses.

Solution [3 marks]: The prediction interval formula is

$$(\mathbf{x}_1 + \mathbf{x}_2)^T \mathbf{b} \pm t_{\alpha/2} s \sqrt{2 + (\mathbf{x}_1 + \mathbf{x}_2)^T (X^T X)^{-1} (\mathbf{x}_1 + \mathbf{x}_2)}.$$

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

(a) State two methods that can be used to fit this model to data and compare them.

Solution [2 marks]: We can use the method of least squares, or maximum likelihood estimation. This gives identical $\boldsymbol{\beta}$ parameters but ML has a biased variance estimator while LS does not.

(b) Explain the difference between a confidence interval and a prediction interval.

Solution [2 marks]: A confidence interval is an interval for the expected response, given specified explanatory variables; a prediction interval is an interval for the response from a single sample with those explanatory variables.

(c) Define Akaike's information criterion and explain why it is useful as a goodness-of-fit measure for model selection.

Solution [2 marks]: Akaike's information criterion is defined as

$$AIC = -2 \ln(\text{likelihood}) + 2p.$$

It balances the fit of the model to the data (expressed in the likelihood) against the model complexity (p).

(d) Give an advantage and a disadvantage of stepwise selection over forward selection.

Solution [2 marks]: Advantage: stepwise selection is more flexible as it can remove a variable in the model. Disadvantage: stepwise selection uses goodness-of-fit measures rather than F tests of significance, potentially resulting in the inclusion of insignificant variables.

(e) Define interaction between two categorical predictors.

Solution [2 marks]: Interaction occurs when the effect of one predictor depends on the level of the other predictor.

(f) List two principles of experimental design.

Solution [2 marks]: Two of: control, blocking, randomisation, blinding.

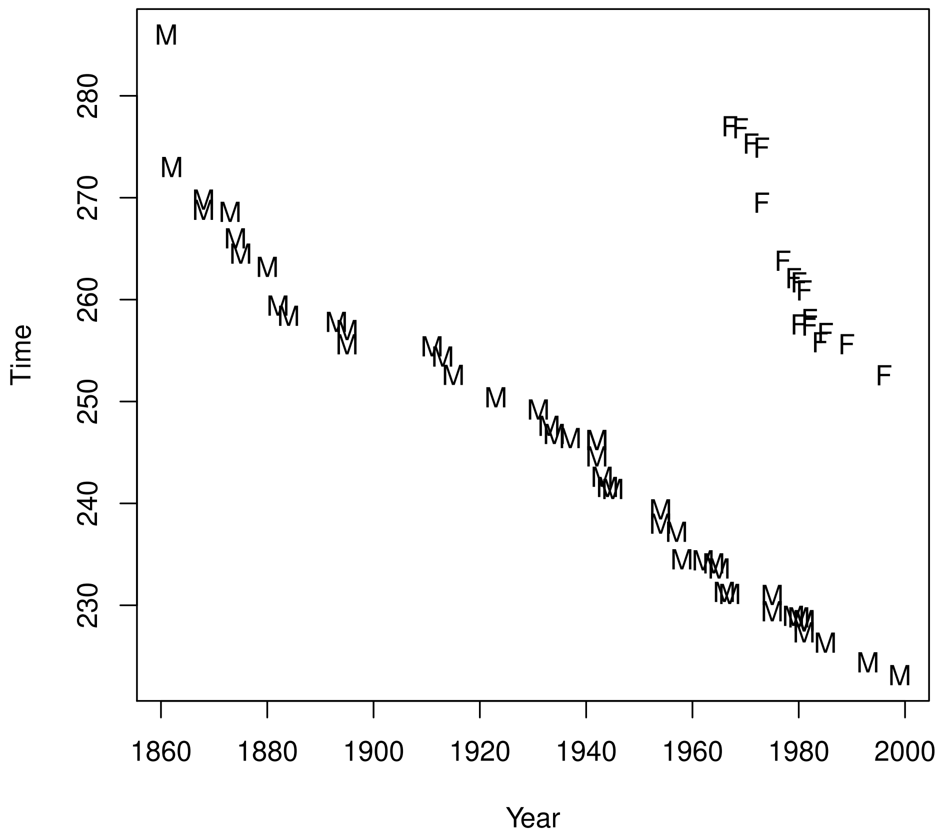
(g) Explain what a Latin square is and its use in experimental design.

Solution [2 marks]: A Latin square is an $n \times n$ square which contains the numbers 1 to n such that every number occurs once in each row and column. It is used to design experiments where there are too many factor combinations for a complete block design.

Question 6 (15 marks) Data was collected on the world record times for the one-mile run. For males, the records are from the period 1861–2003, and for females, from the period 1967–2003. This data is analysed below.

```
> mile <- read.csv('mile.csv', header=T)
> mile$Gender <- factor(mile$Gender)
> plot(Time ~ Year, data = mile, pch=as.character(Gender))
> imodel <- lm(Time ~ (Year + Gender)^2, data = mile)
> summary(imodel)

Call:
lm(formula = Time ~ (Year + Gender)^2, data = mile)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4512 -1.6160 -0.1137  1.1784 13.7265

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2309.4247   202.0583  11.429  < 2e-16 ***
Year               -1.0337     0.1021 -10.126 1.95e-14 ***
GenderMale      -1355.6778   203.1441  -6.673 1.03e-08 ***
Year:GenderMale     0.6675     0.1027   6.502 2.00e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.989 on 58 degrees of freedom
Multiple R-squared:  0.9663,        Adjusted R-squared:  0.9645
F-statistic: 553.8 on 3 and 58 DF,  p-value: < 2.2e-16

> amodel <- lm(Time ~ Year + Gender, data = mile)
> summary(amodel)

Call:
lm(formula = Time ~ Year + Gender, data = mile)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9071 -2.0988 -0.1141  1.2002 13.1863

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1003.00334   27.84691   36.02   <2e-16 ***
Year          -0.37364    0.01406  -26.57   <2e-16 ***
GenderMale   -34.85078    1.30099  -26.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.896 on 59 degrees of freedom
Multiple R-squared:  0.9417,        Adjusted R-squared:  0.9397
F-statistic: 476.3 on 2 and 59 DF,  p-value: < 2.2e-16

> anova(amodel, imodel)
```

```
Analysis of Variance Table

Model 1: Time ~ Year + Gender
Model 2: Time ~ (Year + Gender)^2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     59 895.62
2     58 518.03  1    377.59 42.276 2.001e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> linearHypothesis(imodel, c(0,1,0,1), -0.3)

Linear hypothesis test

Hypothesis:
Year  + Year:GenderMale = - 0.3

Model 1: restricted model
Model 2: Time ~ (Year + Gender)^2

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     59 850.63
2     58 518.03  1     332.6 37.238 9.236e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> qt(c(0.95,0.975,0.99,0.995), df=58)

[1] 1.671553 2.001717 2.392377 2.663287

> qt(c(0.95,0.975,0.99,0.995), df=59)

[1] 1.671093 2.000995 2.391229 2.661759
```



(a) Do you think that this data satisfies the linear model assumptions? Explain.

Solution [2 marks]: While the data look quite linear, there is a big problem: the record can only decrease. So the data are not independent and cannot satisfy the linear model assumptions.

(b) What are the covariates used in the `imodel` object (in the context of the study)?

Solution [2 marks]: The covariates used are year, gender, and interaction between year and gender.

(c) What is being tested in the `anova` function call? What do you conclude from the results?

Solution [2 marks]: The `anova` call tests if there is interaction, i.e., if there is a significant difference between the improvements in male and female records. We conclude that there is significant interaction and we should use the model with interaction.

(d) Write down the final fitted models for the male and female records.

Solution [2 marks]:

```
> imodel$coef[c(1,2)] + imodel$coef[c(3,4)]
```

```
(Intercept)        Year
953.7469611  -0.3661867
```

For females,

$$\text{Time} = 2309 - 1.034\,\text{Year}.$$

For males,

$$\text{Time} = 954 - 0.366\,\text{Year}.$$

(e) Calculate a point estimate for the year when the female world record will equal the male world record. Do you expect this estimate to be accurate? Why or why not?

Solution [3 marks]: Equating the above lines,

```
> -imodel$coef[3]/imodel$coef[4]

GenderMale
   2030.95
```

We expect that the world records will be equal around the year 2031. However this is unlikely to be an accurate estimate as we are extrapolating well beyond the range of the data.

(f) Calculate a 95% confidence interval for the amount by which the gap between the male and female world records narrow every year.

Solution [2 marks]:

```
> se <- coef(summary(imodel))[,2]
> imodel$coef[4] + c(-1,1)*qt(0.975,df=58)*se[4]

[1] 0.4620087 0.8730100
```

(g) What is the hypothesis being tested in the `linearHypothesis` function call? What do you conclude from the output?

Solution [2 marks]: The function call tests if the slope of the male line (amount by which the male world record decreases each year) is equal to -0.3. We conclude that it is not; the record decreases faster than this rate.

Question 7 (12 marks)

(a) Randomisation eliminates confounding from both known and unknown factors. Explain why it is additionally advantageous to use blocking for some confounding factors.

Solution [2 marks]: Blocking reduces variation within the blocks, increasing statistical power.

(b) A study is to be conducted to evaluate the effect of a drug on brain function. The evaluation consists of measuring the response of a particular part of the brain using an MRI scan. The drug is prescribed in doses of 1, 2 and 5 milligrams. Funding allows only 24 observations to be taken in the current study.

Explain how control might be used in a design of this experiment.

Solution [2 marks]: We should have a control group which takes a placebo containing no dosage of the drug. This group will be compared against the groups taking the other three doses.

(c) For the scenario above, explain how replication might be used in a design of this experiment.

Solution [2 marks]: Each of the 4 groups contains 6 observations; having more observations in each group provides better statistical power.

(d) For a complete block design with b blocks and k treatments, the reduced design matrix $X_{2|1}$ satisfies

$$X_{2|1}^T X_{2|1} = b\left[I_k - \frac{1}{k}J_k\right],$$

where I_k is the $k \times k$ identity matrix and J_k is the $k \times k$ matrix with all elements equal to 1. Show that

$$(X_{2|1}^T X_{2|1})^c = \frac{1}{b}I_k.$$

Solution [3 marks]:

$$\begin{aligned} X_{2|1}^T X_{2|1}\frac{1}{b}I_k X_{2|1}^T X_{2|1} &= b\left[I_k - \frac{1}{k}J_k\right]\left[I_k - \frac{1}{k}J_k\right] \\ &= b\left[I_k - \frac{2}{k}J_k + \frac{k}{k^2}J_k\right] \\ &= b\left[I_k - \frac{1}{k}J_k\right] \\ &= X_{2|1}^T X_{2|1}. \end{aligned}$$

(e) For the above complete block design, show that if a quantity $\mathbf{t}^T\boldsymbol{\tau}$ involving only the treatment parameters is estimable, then it must be a treatment contrast. (That is, $\mathbf{t}^T\mathbf{1} = 0$.)

Solution [3 marks]: If the quantity is $\mathbf{t}^T\boldsymbol{\tau}$, then

$$\begin{aligned} \mathbf{t}^T(X_{2|1}^T X_{2|1})^c X_{2|1}^T X_{2|1} &= \mathbf{t}^T\frac{1}{b}I_k b\left[I_k - \frac{1}{k}J_k\right] \\ &= \mathbf{t}^T\left[I_k - \frac{1}{k}J_k\right] \\ &= \mathbf{t}^T \\ \mathbf{t}^T\mathbf{1} &= \mathbf{t}^T\left[I_k - \frac{1}{k}J_k\right]\mathbf{1} \\ &= \mathbf{t}^T\left[\mathbf{1} - \frac{1}{k}k\mathbf{1}\right] \\ &= 0. \end{aligned}$$

End of Exam—Total Available Marks = 90.