STAT 8004 Midterm Exam I Optional Makeup

Nooreen Dabbish

March 15, 2015

1 Suppose that we have observable random variables y_1, y_2, y_3 and y_4 satisfying $E(y_1) = 2\beta_1 - \beta_2 + \beta_3 - \beta_4$, $E(y_2) = 2\beta_1 + \beta_3$, $E(y_3) = \beta_2$, and $E(y_4) = 2\beta_1 + \beta_2 + \beta_3$.Let $\mathbf{Y} = (y_1,y_2,y_3,y_4)^T$, and $\beta = (\beta_1,\beta_2, \beta_3, \beta_4)^T$. Answer parts (a)-(e) in this scenario.

1.1 a

Find $\mathbf{X}$ and ϵ such that a model for $\mathbf{Y}$ can be expressed in the form of $\mathbf{Y}=\mathbf{X}\beta+\epsilon$.

We write $\mathrm{E}(\epsilon_i) = 0$ and $\mathbf{X} = \begin{pmatrix} 2 & -1 & 1 & -1 \\ 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 \end{pmatrix}$.

To give the model:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 1 & -1 \\ 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

1.2 b

Is $\mathbf{X}$ in your model full rank? Why or Why not?

$\mathbf{X}$ is not full rank. We can easily tell because the fourth row is equal to the sum of the second row and the third row. Another way to verify that $\mathbf{X}$ is not full rank is to show that it has a determinant of zero. This is verified in R with `det(X1)`, as well as calculating the determinant by hand. Calculating along the third row we have

$$detX = (-1)^{3+2} \begin{vmatrix} 2 & 1 & -1 \\ 2 & 1 & 0 \\ 2 & 1 & 0 \end{vmatrix} = 0.$$

Where the determinant of the submatrix is obviously zero because column 1 = 2 x column 2.

1.3 c

Clearly and precisely state the minimal conditions under which your model in part (a) is a Gauss-Markov model.

1. $E(\epsilon) = \mathbf{0}$
2. $Var(\epsilon) = \sigma^2\mathbf{I}$

1.4 d

Is β_4 estimable? If yes, find a linear unbiased estimator for β_4. If no, why?

Yes. β_4 is estimable because there is a vector c in the rowspace of $\mathbf{X}$ such that $\beta_4 = c^t\beta$. Specifically c = Row 4 - Row 1 - 2xRow 3:

$$\begin{aligned}\beta_4 &= c^T\beta \\ &= \begin{pmatrix}0 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} \\ &= \begin{pmatrix}-1 & 0 & -2 & 1\end{pmatrix}\begin{pmatrix}2 & -1 & 1 & -1\\ 2 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 2 & 1 & 1 & 0\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}\end{aligned}$$

Find a linear unbiased estimate for β_4:

$$\hat{\beta}_4 = \widehat{c^T\beta} = c^T(X^TX)^-X^TY$$

Evaluating gives the following (see scanned scratchwork or verification in R code appendix for additional details).

$$(X^TX)^- = \begin{pmatrix}0 & 0 & 0 & 0\\ 0 & 2/3 & -1/3 & -1\\ 0 & -1/3 & 2/3 & 1\\ 0 & -1 & 1 & 3\end{pmatrix},\ (X^TX)^-X^T = \begin{pmatrix}0 & 0 & 0 & 0\\ 0 & -1/3 & 2/3 & 1/3\\ 0 & 2/3 & -1/3 & 1/3\\ -1 & 1 & -1 & 0\end{pmatrix}$$

$$c^t(X^TX)^-X^T = (-1, 1, -1, 0)$$

$$\widehat{c^t\beta} = c^t(X^TX)^-X^TY = -y_1 + y_2 - y_3$$

R code for verification:

```
ct <- c(0,0,0,1)
ct %*% ginv(t(X1)%*%X1)%*%t(X1)
qr(ginv(t(X1)%*%X1))$rank
qr((t(X1)%*%X1)[2:4,2:4])$rank
A22inv <- solve((t(X1)%*%X1)[2:4,2:4])
G <- rbind(c(0,0,0,0), (cbind(c(0,0,0), A22inv)))
t(X1)%*%X1
G%*%t(X1)
c(0,0,0,1)%*%G%*%t(X1)
```

2 Consider an experiment with two factors: A (with levels A_1 and A_2) and B (with levels B_1 and B_2). Let y_{ijk} be the outcome of the kth unit at the level of A_i factor and B_j factor (i,j = 1,2). Data are collected as in the following table . . .

2.1 a

Is this data set a balanced one?

No. This data set is not balanced because there are two observations of A_1 B_1, A_1 B_2, and A_2 B_2, but only one observation of A_2 B_1.

2.2 b

Express the mean outcomes $\mu_{ij} = E(y_{ijk})$ (i,j = 1,2) corresponding to all possible combinations of the factors A nd B as functions of β_1, β_2, β_3, and β_4.

$$\mu_{11} = \beta_1 + \beta_2 + \beta_3 + \beta_4$$
$$\mu_{12} = \beta_1 + \beta_2 - \beta_3 - \beta_4$$
$$\mu_{21} = \beta_1 - \beta_2 + \beta_3 - \beta_4$$
$$\mu_{22} = \beta_1 - \beta_2 - \beta_3 + \beta_4$$

2.3 c

Express the overall mean of the outcomes as a funtion of β_1, β_2, β_3,β_4

$$\mu_{..} = 1/4(\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}) = \beta_1$$

2.4 d

Find a 95% confidence interval for μ_{21}- μ_{11}

2.5 e

Find the F statistic for H_0: $\mu_{12} = \mu_{22} = 35$ vs H_a: not H_0, and give its degrees of freedom.

F statistic is given by

$$F = \frac{\frac{1}{\sigma^2}\frac{(\widehat{C\beta}-d)^T(C(X^TX)^-C^T)^{-1}(\widehat{C\beta}-d)}{l}}{\frac{1}{\sigma^2}\frac{SSE}{n-rank(X)}}$$

where n = number of observations, here n = 7, rank X = 4 , so n- rank(X) = 3. The matrix $\mathbf{C}$ and vector $\mathbf{d}$ are given by the null hypothesis, here $\mathbf{C} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$ and $\mathbf{d} = \begin{pmatrix} 35 \\ 35 \end{pmatrix}$. l is determined by the rank of the $\mathbf{C}$ matrix, here l=2.

Evaluating gives F = 1/4 with 2, 3 degrees of freedom, please see included scratchwork and/or R code for verification.

```
C2 <- matrix(c(1,1,-1,-1,1,-1,-1,1),nrow=2,ncol=4,byrow=TRUE)
d <- c(35,35)
betahat <- c(36.25,-8.75,2.5,-7.5)
C2%*%betahat
```

```
XtXinv <- matrix((1/32)*c(5,-1,1,-1,-1,5,-1,1,1,-1,5,-1,-1,1,-1,5),nrow=4,ncol=4,byrow=TRUE)
C2%*%XtXinv%*%t(C2)
solve(C2%*%XtXinv%*%t(C2))
((t(C2%*%betahat - d)%*%solve(C2%*%XtXinv%*%t(C2))%*%(C2%*%betahat - d))/2)/(75/3)
```

2.6 f

Suppose three new outcomes are observed at the condition with respectively level A_1 and B_2, find a 95% prediction interval for the average of the three new observations.

2.7 g

Suppose that the fifth row (outcome = 55) in the data table is now removed. Does it change the estimiability of any of the parameters β_j ($j = 1, \ldots, 4$) in the model? Why or why not?

2.8 h

Does the change in the previous part (g) have any impact on the least square estimation of $\mu_{11}$and μ_{12}? Explain your answer.

If y_{211}, the fifth outcome, is removed, it will not have any impact on the least squares estimation of μ_{11} or μ_{12}. To see this we can look at the least squares estimate of these quantities with the fifth outcome in place and removed and it becomes obvious that only the first two outcomes matter for the least squares estimate of μ_{11} and only the third and fouth outcomes are part of the the least square estimate of μ_{12}. That is, only the observations from the conditions of the respective means. To show this explicitly note that:

$$
\begin{aligned}
\widehat{\mu_{11}} &= (1,1,1,1)(X^TX)^-X^TY \\
&= (1/2,1/2,0,0,0,0,0,0)\begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{221} \\ y_{222} \end{pmatrix} = \frac{1}{2}(y_{111}+y_{112}) \\
\widehat{\mu^\star_{11}} &= (1,1,1,1)(X^{\star T}X^\star)^-X^{\star T}Y^\star \\
&= (1/2,1/2,0,0,0,0,0,0)\begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{221} \\ y_{222} \end{pmatrix} = \frac{1}{2}(y_{111}+y_{112}) \\
\widehat{\mu_{12}} &= (1,1,-1,-1)(X^TX)^-X^TY \\
&= (0,0,1/2,1/2,0,0,0,0)Y = \frac{1}{2}(y_{121}+y_{122}) \\
\widehat{\mu^\star_{12}} &= (1,1,-1,-1)(X^{\star T}X^\star)^-X^{\star T}Y^\star \\
&= (0,0,1/2,1/2,0,0,0,0)\mathbf{Y}^\star = \frac{1}{2}(y_{121}+y_{122})
\end{aligned}
$$

R code to verify results above:

```
X2 <- matrix(c(rep(c(1,1,1,1),2),rep(c(1,1,-1,-1),2),1,-1,1,-1,rep(c(1,-1,-1,1),2)),nrow=7,ncol=4,byr
Ch <- c(1,1,-1,-1)
Ch%*%ginv(t(X2)%*%X2)%*%t(X2)

X2 <- X2[-5,]
```

3 Consider the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x^2{}_i + \epsilon_i$ . . .

3.1 a

Find the maximum likelihood estimator for x_0.

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$\frac{\partial E(y)}{\partial x} = \beta_1 + 2\beta_2 x$$

Set derivative to zero to find maxima/minima.

$$\hat{x_0} = \frac{-\hat{\beta_1}}{2\hat{\beta_2}}$$

3.2 b

Find a (1 - α) level confidence interval for x_0. You may assume n large here. If you solve this part without assuming n large, there will be 3 extra points.

We write $\hat{x_0} = \frac{-\hat{\beta_1}}{2\hat{\beta_2}}$. We know the OLS estimator for our parameter vector $\hat{\beta}$ is multivariate normally distributed with expectation β and variance $\sigma^2(X^TX)^-$. Therefore $\hat{\beta_1} \sim N(\beta_1, \sigma^2(X^TX)^-_{2,2})$ and $\hat{\beta_2} \sim N(\beta_2, \sigma^2(X^TX)^-_{3,3})$. It should therefore be possible to use partial fraction decomoption to re-write x_0 interms of standardized normal distributions. The square-root of the square of the denominator will give a t-distribution with 1 degree of freedom.

Our (1 - α) confidence interval will be $\hat{x_0} \pm st_{\alpha/2}/\sqrt{n}$ where s is the standard deviation in our estimator $\hat{x_0} = \frac{-\hat{\beta_1}}{2\hat{\beta_2}}$.

4 Appendix: Tangled R code

```
library(MASS); library(xtable)
      lvector <- function(x, dig = 2, dsply=rep("f",ncol(x)+1)) {
       x <- xtable(x, align=rep("",ncol(x)+1),display=dsply,digits=dig) # We repeat empty
       print(x, floating=FALSE, tabular.environment="pmatrix",
         hline.after=NULL, include.rownames=FALSE, include.colnames=FALSE)
       }

X1 <- matrix(c(2,-1,1,-1,2,0,1,0,0,1,0,0,2,1,1,0), nrow=4, ncol=4, byrow=TRUE)
lvector(X1)

ct <- c(0,0,0,1)
ct %*% ginv(t(X1)%*%X1)%*%t(X1)
qr(ginv(t(X1)%*%X1))$rank
```

```
qr((t(X1)%*%X1)[2:4,2:4])$rank
A22inv <- solve((t(X1)%*%X1)[2:4,2:4])
G <- rbind(c(0,0,0,0), (cbind(c(0,0,0), A22inv)))
t(X1)%*%X1
G%*%t(X1)
c(0,0,0,1)%*%G%*%t(X1)

C2 <- matrix(c(1,1,-1,-1,1,-1,-1,1),nrow=2,ncol=4,byrow=TRUE)
d <- c(35,35)
betahat <- c(36.25,-8.75,2.5,-7.5)
C2%*%betahat


XtXinv <- matrix((1/32)*c(5,-1,1,-1,-1,5,-1,1,1,-1,5,-1,-1,1,-1,5),nrow=4,ncol=4,byrow=T
C2%*%XtXinv%*%t(C2)
solve(C2%*%XtXinv%*%t(C2))
((t(C2%*%betahat - d)%*%solve(C2%*%XtXinv%*%t(C2))%*%(C2%*%betahat - d))/2)/(75/3)

X2 <- matrix(c(rep(c(1,1,1,1),2),rep(c(1,1,-1,-1),2),1,-1,1,-1,rep(c(1,-1,-1,1),2)),nrow
Ch <- c(1,1,-1,-1)
Ch%*%ginv(t(X2)%*%X2)%*%t(X2)

X2 <- X2[-5,]
```