Stat 511 Exam 1

February 23, 2004 Prof. Vardeman

1. Consider an instance of the linear model for $n = 5$ observations,

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

a) This is a full rank model. One way to easily see this is to argue that rows 1, 2, and 3 of the model matrix are linearly independent (and so the rank is at least 3). Put these three rows into a 3×3 matrix $\mathbf{M}$ and show this matrix is non-singular by arguing that $\mathbf{Mc} = \mathbf{0}$ implies that $\mathbf{c} = \mathbf{0}$.

Notice that the columns of $\mathbf{X}$ are perpendicular, so that $\mathbf{X}'\mathbf{X}$ is diagonal.

b) In a Gauss-Markov version of this model, which of the parameters $\beta_1, \beta_2,$ or β_3 can be estimated with the greatest precision? Explain carefully.

c) Compute a matrix $\mathbf{P}_{\mathbf{X}}$ that projects any element of $\Re^5$ onto $C(\mathbf{X})$ (in a perpendicular fashion).

d) In a Gauss-Markov version of this model, which row of the $\mathbf{X}$ matrix represents a set of conditions under which $\mathrm{E}y$ can be estimated with the best precision? Explain carefully.

For the next two parts of this question (parts **e)** and **f)),** suppose that $\mathbf{Y}$ is such that $SSE = 3$ and $\mathbf{b}'_{\mathrm{OLS}} = (5, 6, 2)$. Consider an analysis under the normal version of the Gauss Markov model.

e) In the future, two new observations, $y_{\mathrm{new}1}$ and $y_{\mathrm{new}2}$ are going to be observed under the conditions described respectively by the 1st and 2nd rows of the $\mathbf{X}$ matrix. Give 95% two-sided prediction limits for $y_{\mathrm{new}1} - y_{\mathrm{new}2}$. (Plug correct numbers into correct formulas, but do not take time to do arithmetic.)

f) Write the hypothesis $\mathrm{H}_0 : \mathrm{E}\, y_1 = \mathrm{E}\, y_2$ and $\mathrm{E}\, y_1 = \mathrm{E}\, y_3$ in testable form $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ for an appropriate matrix $\mathbf{C}$ (write out such a matrix) and compute an F statistic for testing this (you need not do the arithmetic, but plug correct numbers into a correct formula).

g) Consider an Aitken version of the model on page 1, where $\mathbf{V} = \text{diag}(1, \delta, 1, 1, \delta)$ for δ small. Generalized least squares estimation of $\boldsymbol{\beta}$ under these circumstances will essentially force $\hat{y}_2 = \beta_1 - \beta_2 - \beta_3 \approx y_2$ and $\hat{y}_5 = \beta_1 + \beta_2 + \beta_3 \approx y_5$. This is $\beta_1 \approx (y_2 + y_5)/2$ and $\beta_2 + \beta_3 \approx (y_5 - y_2)/2$. Take these approximations as given and find estimates of β_2 and β_3 if $\mathbf{Y}' = (15, 4, 6, 8, 10)$.

2. Attached to this exam is a printout of an R session for a time series analysis (via an ordinary linear model) of 6 years worth of quarterly retail sales data (for the JC Penney Company). For consecutive 3-month periods that we will simply label as $t = 1, 2, \ldots, 24$ we'll model

$$y_t = \text{sales in period } t$$

as roughly linearly increasing in t, but with different "effects" for the 4 quarters of the year. That is, with

$$q_i(t) = \begin{cases} 1 & \text{if period } t \text{ is from the } i\text{th quarter of the year} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, 2, 3, 4$ we consider a model

$$y_t = \beta_0 + \beta_1 t + \gamma_1 q_1(t) + \gamma_2 q_2(t) + \gamma_3 q_3(t) + \gamma_4 q_4(t) + \varepsilon_t$$

for $t = 1, 2, \ldots, 24$ the values $\beta_0, \beta_1, \gamma_1, \gamma_2, \gamma_3, \gamma_4$ unknown constants and the $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{24}$ iid normal $(0, \sigma^2)$. (Period $t = 1$ is a first quarter period.) You may use the printout to answer the following questions. Refer very carefully to where you find anything you take from the printout (give page and location on the page).

a) Is the parametric function $\gamma_1 - \gamma_2$ estimable in this model? Argue this very carefully. (Write the $\mathbf{X}$ matrix for the first 5 periods below and use it in your argument.)

Since the model as originally posed is not full rank, a call to R's `lm()` function introduces a restriction in order to produce a full rank version. The restriction used by R in this case is to set to 0 the coefficient for the last column of the model matrix entered in the function call. That is, R fits the model

$$y_t = \beta_0^* + \beta_1^* t + \gamma_1^* q_1(t) + \gamma_2^* q_2(t) + \gamma_3^* q_3(t) + \varepsilon_t$$

b) In this model, find 90% two-sided confidence limits for σ. (No need to simplify after plugging in.)

c) Give 95% two-sided confidence limits for $\gamma_1^* - \gamma_2^*$. (No need to simplify after plugging in.)

d) Give 95% prediction limits for y_{28} (the retail sales in the 4th quarter of the year after the end of the data in hand) based on this model. (Plug correct numbers into a correct formula, but you need not do arithmetic.)

Stat 511 Exam I Spring 2004 Printout

```
> data
      sales  t q1 q2 q3 q4
 [1,]  4452  1  1  0  0  0
 [2,]  4507  2  0  1  0  0
 [3,]  5537  3  0  0  1  0
 [4,]  8157  4  0  0  0  1
 [5,]  6481  5  1  0  0  0
 [6,]  6420  6  0  1  0  0
 [7,]  7208  7  0  0  1  0
 [8,]  9509  8  0  0  0  1
 [9,]  6755  9  1  0  0  0
[10,]  6483 10  0  1  0  0
[11,]  7129 11  0  0  1  0
[12,]  9072 12  0  0  0  1
[13,]  7339 13  1  0  0  0
[14,]  7104 14  0  1  0  0
[15,]  7639 15  0  0  1  0
[16,]  9661 16  0  0  0  1
[17,]  7528 17  1  0  0  0
[18,]  7207 18  0  1  0  0
[19,]  7538 19  0  0  1  0
[20,]  9573 20  0  0  0  1
[21,]  7522 21  1  0  0  0
[22,]  7211 22  0  1  0  0
[23,]  7729 23  0  0  1  0
[24,]  9542 24  0  0  0  1
> JCPenney<-lm(sales~t+q1+q2+q3+q4)
> summary(JCPenney)

Call:
lm(formula = sales ~ t + q1 + q2 + q3 + q4)

Residuals:
    Min      1Q  Median      3Q     Max
-1232.1  -274.0   157.3   332.1   853.9

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7858.76     331.26  23.724 1.40e-15 ***
t              99.54      16.93   5.878 1.16e-05 ***
q1          -2274.21     331.12  -6.868 1.49e-06 ***
q2          -2564.58     328.94  -7.796 2.45e-07 ***
q3          -2022.79     327.63  -6.174 6.22e-06 ***
q4                NA         NA      NA       NA
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 566.7 on 19 degrees of freedom
Multiple R-Squared: 0.8683,     Adjusted R-squared: 0.8405
F-statistic:  31.3 on 4 and 19 DF,  p-value: 4.014e-08
```

```
> vcov(JCPenney)
            (Intercept)          t          q1          q2          q3
(Intercept)  109733.448 -4014.6383 -65572.4262 -61557.7879 -57543.1495
t             -4014.638   286.7599    860.2796    573.5198    286.7599
q1           -65572.426   860.2796 109637.8613  55249.0705  54388.7908
q2           -61557.788   573.5198  55249.0705 108204.0619  54102.0310
q3           -57543.150   286.7599  54388.7908  54102.0310 107343.7823
> predict(JCPenney)
        1         2         3         4         5         6         7         8
 5684.089  5493.256  6134.589  8256.923  6082.254  5891.420  6532.754  8655.087
        9        10        11        12        13        14        15        16
 6480.418  6289.585  6930.918  9053.251  6878.582  6687.749  7329.082  9451.415
       17        18        19        20        21        22        23        24
 7276.746  7085.913  7727.246  9849.580  7674.911  7484.077  8125.411 10247.744
> anova(JCPenney)
Analysis of Variance Table

Response: sales
          Df   Sum Sq  Mean Sq F value    Pr(>F)
t          1 16217509 16217509  50.495 9.296e-07 ***
q1         1  2282899  2282899   7.108   0.01526 *
q2         1  9473203  9473203  29.496 3.064e-05 ***
q3         1 12242274 12242274  38.118 6.219e-06 ***
Residuals 19  6102250   321171
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```