**COMPREHENSIVE WRITTEN EXAMINATION, PAPER III**

**FRIDAY AUGUST 15, 2008 9:00 A.M. – 1:00 P.M.**

**STOR 664 Questions**

**Part A:** (47 points)

Suppose that $Y$ has mean $\beta_0 + \beta_1 x + \beta_2 x^2$ and variance $\sigma^2$, for $-1 \leq x \leq 1$. Four independent observations, $Y_1, \ldots, Y_4$, are collected at the $x$ settings, $-1, -c, c, +1$, respectively, where $0 \leq c < 1$.

1. (10 points) A matrix notation for the problem is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

   Write down each term of the equation, $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$; write down two different common assumptions on $\boldsymbol{\epsilon}$.

2. (14 points) Derive the variance-covariance matrix of the least squares estimator of $\boldsymbol{\beta}$.

3. (11 points) Determine the value of $c$ to maximize the precision of the least squares estimator of $\beta_2$.

4. (12 points) Assuming that $c$ is fixed, show that $E(Y|x_0)$ is most accurately estimated at

$$x_0 = \pm\frac{1}{2}\sqrt{\frac{1 + 6c^2 + c^4}{1 + c^2}}.$$

**Part B:** (53 points)

Consider the following analysis of Asphalt shingles sales data consisting of the variables Sales, Promotional Accounts (PrAc), Active Accounts (AcAc), Competing Brands (CpBr), and Potential (Poten). For the model, Sales $\sim$ PrAc + AcAc + CpBr + Poten, the summary is as follows:

| | Estimate | Std. Error | $t$ values | $Pr(>\|t\|)$ |
|---|---|---|---|---|
| (Intercept) | 177.2286 | 8.7874 | 20.1685 | 0.0000 |
| PrAc | 2.1702 | 0.6737 | 3.2213 | 0.0092 |
| AcAc | 3.5380 | 0.1092 | 32.4136 | 0.0000 |
| CpBr | −22.1583 | 0.5454 | −40.6297 | 0.0000 |
| Poten | 0.2035 | 0.3189 | 0.6383 | 0.5376 |

Residual standard error: 5.119 on 10 degrees of freedom
Multiple R-squared: 0.9971
$F$-statistic: 851.7 on 4 and 10 degrees of freedom, the $p$-value is 1.285e-12

1. (10 points) How many observations on (Sales, PrAc, AcAc, CpBr, Poten) does this "sales" data set contain? Based on the analysis above, which variable, if any, is not useful for predicting the sales volume? Explain.

2. Suppose we are interested in two parameters, $\gamma_1 = \beta_{PrAc} - \beta_{AcAc}$ and $\gamma_2 = \beta_{CpBr} - \beta_{Poten}$:

   (a) (12 points) Explain how to construct simultaneous 95% confidence intervals for $(\gamma_1, \gamma_2)$.

   (b) (12 points) Derive a general $F$-test procedure in testing $H_0$: $(\gamma_1, \gamma_2) = (-1, -20)$ versus $H_0$: $(\gamma_1, \gamma_2) \neq (-1, -20)$.

3. (5 points) For the data above, the following R analysis provides information on four sets of linear models. Each set involves Sales and one of the explanatory variables above as dependent variables, and the remaining three as explanatory variables. Then a simple linear regression model of their residuals is fitted. The regression coefficient of the model for residuals is printed for each set.

```
e.y <- lm(Sales~AcAc + CpBr + Poten, data=sales)$resid
e.x.1<-lm(PrAc~AcAc + CpBr + Poten, data=sales)$resid
lm(e.y~e.x.1)$coeff[2]
e.x.1
2.17021

e.y <- lm(Sales~PrAc + CpBr + Poten, data=sales)$resid
e.x.2<-lm(AcAc~PrAc + CpBr + Poten, data=sales)$resid
```

```
lm(e.y~e.x.2)$coeff[2]
e.x.2
3.538014

e.y <- lm(Sales~PrAc + AcAc + Poten, data=sales)$resid
e.x.3<-lm(CpBr~PrAc + AcAc + Poten, data=sales)$resid
lm(e.y~e.x.3)$coeff[2]
e.x.3
-22.15834

e.y <- lm(Sales~PrAc + AcAc + CpBr, data=sales)$resid
e.x.4<-lm(Poten~PrAc + AcAc + CpBr, data=sales)$resid
lm(e.y~e.x.4)$coeff[2]
e.x.4
0.2035384
```

What, if any, similarities do you find among the regression coefficients in the multiple regression model in part 1 above and the regression coefficients of the models for residuals in this part? Explain.

4. (14 points) Let $Y$ denote "Sales", and $X_1, X_2, X_3, X_4$ denote the four explanatory variables, respectively, as introduced above (that is, $X_1$ represents PrAc, Etc.). Let $X_{(-i)}$ denote the design matrix for the model containing an intercept and all other variables except $X_i$. For $i = 1, 2, 3, 4$, the regressions in part 2 represent residuals of the regression of $Y$ on $X_{(-i)}$, and of the regression of $X_i$ on $X_{(-i)}$; then a simple linear regression model of these residuals is fitted. It turns out that the regression coefficient in the simple linear regression model for residuals is the same as the regression coefficient of $X_i$ in the multiple regression model.

   Prove that this relationship between the regression coefficients of the multiple regression and the regression coefficients of the simple linear regression model for appropriate residuals holds in general.

**Part C:** Reference Formula

Suppose all matrix inverses exist, then
I.

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

II. $(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$.