STA 6208 – Spring 2002 – Exam 1

Print Name: **SSN:**

All questions are based on the following two regression models, where SIMPLE REGRESSION refers to the case where $p = 1$, and X is of full column rank (no linear dependencies among the predictor variables)

$$\textbf{Model 1:} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad i = 1, \ldots, n \quad \varepsilon_i \sim NID(0, \sigma^2)$$

$$\textbf{Model 2:} \quad \mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad \mathbf{X} \equiv n \times p' \quad \beta \equiv p' \times 1 \quad \varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

1) For **Model 1**:

a) Write out the the least squares estimate $\hat{\beta}_1$ as a linear function of Y_i $(i = 1, \ldots, n)$.

b) Derive $E[\hat{\beta}_1]$.

c) Derive $Var[\hat{\beta}_1]$.

2) A foam beverage insulator (beer hugger) manufacturer produces their product for firms that want their logo on beer huggers for marketing purposes. The firm's cost analyst wants to estimate their cost function. She interprets β_0 as the fixed cost of a production run, and β_1 as the unit variable cost (or marginal cost). Based on $n = 5$ production runs she observes the following pairs (X_i, Y_i) where X_i is the number of beer huggers produced in the i^{th} production run (in 1000s), and Y_i was the total cost of the run (in \$1000).

vse

i	X_i	Y_i
1	3.00	4.00
2	5.00	6.50
3	4.00	5.00
4	6.00	7.00
5	7.00	7.50

Obtain the following matrices and vectors: $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{X'X}$, $\mathbf{X'Y}$, $(\mathbf{X'X})^{-1}$, $\hat{\beta}$, $\hat{\mathbf{Y}}$, and $\mathbf{e}$

3) A regression model is fit relating selling price of homes (Y, in \$100000s) in a large subdivision to four factors: area (X_1 in 1000s of ft^2), number of bathrooms (X_2), age (X_3, in years), and an indicator of whether the house has a swimming pool ($X_4 = 1$ if yes, 0 if no). The realtor samples $n = 30$ homes that have sold in the division over the past 15 months.

a) The total of the squared prices is 250.0 (recall the units of Y), of which 60% is attributable to the mean, 32% is attributable to the regression, and 8% is attributable to residual (error). Give the Analysis of Variance (including sources of variation, degrees of freedom, sums of squares, and when appropriate mean squares).

b) Give the coefficient of multiple determination (R^2) for this model:

4) Consider the vector $\mathbf{D} = \hat{\mathbf{Y}} - \overline{Y}\mathbf{1}$ where $\mathbf{1}$ is a $n \times 1$ vector of 1^s.

a) Write $\mathbf{D}$ as a linear function of the data vector $\mathbf{Y}$. (That is, $\mathbf{D} = \mathbf{AY}$ for what matrix $\mathbf{A}$?)

b) Give the mean vector and variance-covariance matrix for $\mathbf{D}$ under Model 2.

5) An agronomist is interested in the effects of concentration levels of fertilizer (X_1) and density of plants (plant stand, X_2). She believes that the effects of each factor will be linear in her range of levels of X_1 and X_2, and that the effect of each factor will be independent of the level of the other. She obtains the following matrices based on the design of her experiment:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 10 \\ 1 & 0 & 20 \\ 1 & 2 & 10 \\ 1 & 2 & 20 \\ 1 & 4 & 10 \\ 1 & 4 & 20 \\ 1 & 6 & 10 \\ 1 & 6 & 20 \\ 1 & 8 & 10 \\ 1 & 8 & 20 \end{bmatrix} \qquad \mathbf{X'X} = \begin{bmatrix} 10 & 40 & 150 \\ 40 & 240 & 600 \\ 150 & 600 & 2500 \end{bmatrix} \qquad (\mathbf{X'X})^{-1} = \begin{bmatrix} 1.2000 & -0.0500 & -0.0600 \\ -0.0500 & 0.0125 & 0.0000 \\ -0.0600 & 0.0000 & 0.0040 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{0.40} & 0.20 & 0.30 & 0.10 & 0.20 & 0.00 & 0.10 & -0.10 & 0.00 & -0.20 \\ 0.20 & \mathbf{0.40} & 0.10 & 0.30 & 0.00 & 0.20 & -0.10 & 0.10 & -0.20 & 0.00 \\ 0.30 & 0.10 & \mathbf{0.25} & 0.05 & 0.20 & 0.00 & 0.15 & -0.05 & 0.10 & -0.10 \\ 0.10 & 0.30 & 0.05 & \mathbf{0.25} & 0.00 & 0.20 & -0.05 & 0.15 & -0.10 & 0.10 \\ 0.20 & 0.00 & 0.20 & 0.00 & \mathbf{0.20} & 0.00 & 0.20 & 0.00 & 0.20 & 0.00 \\ 0.00 & 0.20 & 0.00 & 0.20 & 0.00 & \mathbf{0.20} & 0.00 & 0.20 & 0.00 & 0.20 \\ 0.10 & -0.10 & 0.15 & -0.05 & 0.20 & 0.00 & \mathbf{0.25} & 0.05 & 0.30 & 0.10 \\ -0.10 & 0.10 & -0.05 & 0.15 & 0.00 & 0.20 & 0.05 & \mathbf{0.25} & 0.10 & 0.30 \\ 0.00 & -0.20 & 0.10 & -0.10 & 0.20 & 0.00 & 0.30 & 0.10 & \mathbf{0.40} & 0.20 \\ -0.20 & 0.00 & -0.10 & 0.10 & 0.00 & 0.20 & 0.10 & 0.30 & 0.20 & \mathbf{0.40} \end{bmatrix}$$

Suppose we had observed data $Y_1, \ldots, Y_{10}$ in this design and obtained $s^2 = MS(\text{RESIDUAL})$.

a) Write out $\hat{Y}_1$ as a linear function of $Y_1, \ldots, Y_{10}$. What do the coefficients of $Y_1, \ldots, Y_{10}$ sum to?

b) What is $s^2(\hat{Y}_3)$? What is $s^2(\hat{Y}_9)$? Which design points (combination of X_1 and X_2) has the fitted values with the smallest variance? The largest variance?

c) What is $s^2(e_1)$? What is $s^2(e_7)$? Which design points (combination of X_1 and X_2) has the residuals with the smallest variance? The largest variance?

6) An electrical contractor has fit a regression model, relating costs of wiring a new home (Y, in dollars) to size of the home (X, in square feet). She has data on $n = 16$ homes, and obtained the following estimates:**(20 points)**

$$\hat{Y} = 50.00 + 0.22X \qquad s^2 = 1600.00 \qquad \sum(X - \overline{X})^2 = 4,000,000 \qquad \overline{X} = 2000.0$$

a) Give the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_0$.

b) Give a 95% confidence interval for β_1.

b) Give a 95% confidence interval for the mean cost of all homes with $X_0 = 2000$.

c) Give a 95% prediction interval for the cost of her brother–in–law's house with $X_0 = 2000$.

7) For the Analysis of Variance in Model 2, give the expected values of:

$$SS(\text{MODEL}) = \mathbf{Y'PY} \quad \text{and} \quad SS(\text{ERROR}) = \mathbf{Y'(I - P)Y}$$

Hint: If $\mathbf{Y}$ is a random vector with the following mean vector and variance-covariance matrix:

$$\mathbf{E[Y]} = \boldsymbol{\mu} \qquad \mathbf{Var[Y]} = \mathbf{V_Y} = \mathbf{V}\sigma^2$$

Then, the expectation of a quadratic form $\mathbf{Y'AY}$ is:

$$\mathbf{E[Y'AY]} = tr(\mathbf{AV_Y}) + \boldsymbol{\mu}'\mathbf{A}\,\boldsymbol{\mu} = \sigma^2 tr(\mathbf{AV}) + \boldsymbol{\mu}'\mathbf{A}\,\boldsymbol{\mu}$$