

Applied Linear Statistical Models
Spring 2004

Name: _____ **Cornell netID:** _____

Instructions: You must show your work where relevant to receive credit. There are five problems worth a total of 40 points. The last page is your formula sheet, which you may remove.

All of the questions concern a general linear model of the form

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

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

1. Consider the simple linear regression model,

$$E(Y_i) = \beta_0 + \beta_1 X_i, \quad i = 1, \dots, n$$

- (a) Write down the form of the matrix $\mathbf{X}'\mathbf{X}$ for this model (2 pts)

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

- (b) Derive the form of $(\mathbf{X}'\mathbf{X})^{-1}$, (3 pts)

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum X_i^2 - (\sum X_i)^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} = \frac{1}{SXX} \begin{bmatrix} \frac{1}{n} \sum X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix}$$

- (c) Using the “fact” on the formula sheet, show that the variance of the least squares estimate, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, is equal to $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. (3 pts)

Note that $\mathbf{b} = \mathbf{A}\mathbf{Y}$, where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Hence

$$\text{var}(\mathbf{b}) = \sigma^2 \mathbf{A}\mathbf{A}' = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- (d) Write down a formula for the covariance between the intercept and slope estimates. Hence show that they are uncorrelated if $\bar{X} = 0$. (2 pts)

From part (b)

$$\text{cov}(b_0, b_1) = -\sigma^2 \frac{\bar{X}}{SXX}$$

Hence, $\text{cov}(b_0, b_1) = 0$ if $\bar{X} = 0$

2. A study is conducted to analyze relationships between the variables Y = percentage of vote for Democratic candidate, X_1 = percentage of registered voters who are Democrats, and X_2 = percentage of registered voters who vote in the election. The following prediction equation is based on data from several congressional elections.

$$\hat{Y} = 20 + 0.3X_1 + 0.05X_2 + 0.005X_1X_2$$

- (a) Find the prediction equations relating Y to X_1 for each of the values: $X_2 = 30$, 60. (2 pts)

$$\hat{Y}(30) = (20 + 0.05 \times 30) + (0.3 + 0.005 \times 30)X_1 = 21.5 + 0.45X_1$$

$$\hat{Y}(60) = (20 + 0.05 \times 60) + (0.3 + 0.005 \times 60)X_1 = 23.0 + 0.60X_1$$

- (b) For each of the X_2 levels in (a), determine the X_1 value required for the predicted percentage of vote for a Democrat to be 50%. (2 pts)

$$\text{When } X_2 = 30, X_1 = \frac{50 - 21.5}{0.45} = 63.3$$

$$\text{When } X_2 = 60, X_1 = \frac{50 - 23.0}{0.60} = 45.0$$

- (c) Does a high voter turnout favor the Democrats? Yes or No. Explain briefly. (2 pts)

Yes. The probability of a Democrat winning increases with turnout, regardless of the percentage of registered Democrats.

3. Consider the single-factor ANOVA framework with $r = 3$ factor levels and equal sample sizes ($n_i = 2, i = 1, 2, 3$).

(a) Write down the form of the \mathbf{X} and $\boldsymbol{\beta}$ for the *cell means model*, with factor level means, μ_1, μ_2 and μ_3 . (2 pts)

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

(b) Write down the form of \mathbf{X} and $\boldsymbol{\beta}$ for the *sum constraint* model, $\mu_i = \mu. + \tau_i$, in which $\sum \tau_i$ is constrained to be zero. (2 pts)

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu. \\ \tau_1 \\ \tau_2 \end{bmatrix}$$

4. The following data comes from a study of computer use among highschool students. Students were asked to report the number of hours they spent using a computer each week, and to specify an income category for their parents.

| Income | Sample Size | Sample Mean | Sample Std.Dev. |
|--------|-------------|-------------|-----------------|
| Low | 10 | 10.0 | 3.0 |
| Middle | 15 | 15.0 | 3.0 |
| High | 5 | 13.0 | 4.0 |

- (a) Calculate the overall mean, SSTR and SSE (3 pts)

$$\bar{Y}_{..} = \frac{10 \times 10 + 15 \times 15 + 5 \times 13}{30} = \frac{390}{30} = 13$$

$$SSTR = 10(10 - 13)^2 + 15(15 - 13)^2 + 5(13 - 13)^2 = 150$$

$$SSE = 9 \times 3^2 + 14 \times 3^2 + 4 \times 4^2 = 81 + 126 + 64 = 271$$

- (b) Fill in the values in the ANOVA table assuming that SSTR = 150 and SSE = 270. (4 pts)

| Source | SS | df | MS | F-ratio |
|--------|-----|----|----|---------|
| Income | 150 | 2 | 75 | 7.5 |
| Error | 270 | 27 | 10 | |
| Total | 420 | | | |

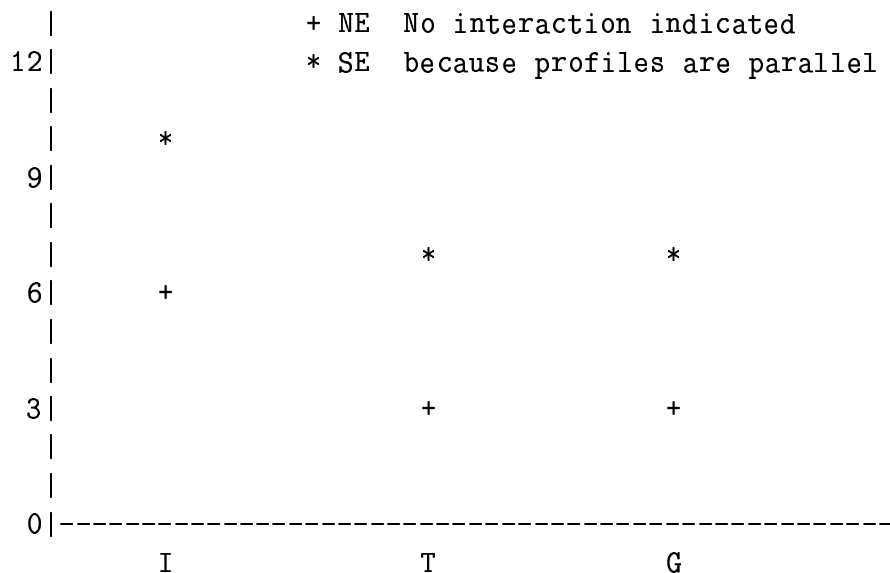
- (c) Calculate the t -statistic for testing for a difference between the “Low” and “Middle” parental income groups. Set up the calculation (3 pts)

$$t^* = \frac{10 - 15}{\sqrt{10 \left(\frac{1}{10} + \frac{1}{15} \right)}} = \frac{-5}{1.29} = -3.87$$

5. The table below gives sample means and variances of murder rates for US cities cross-classified by TYPE (Industrial, Trade or Governmental) and REGION (Northeast or Southeast). Each mean (and variance) is based on a sample of $n = 4$ cities.

| | IND. | TRADE | GOV. |
|-----------|--------|-------|-------|
| NORTHEAST | 6 (2) | 3 (1) | 3 (1) |
| SOUTHEAST | 10 (2) | 7 (2) | 7 (2) |

- (a) Construct a plot of the three sample means for the two regions. Does the plot indicate the presence of interaction? Explain. (2 pts)



- (b) Calculate the REGION and ERROR sums of squares. (2 pts)
 Note that $n = 4$, $a = 2$ and $b = 3$. The NE and SE overall means are $\bar{Y}_{1..} = 4$ and $\bar{Y}_{2..} = 8$ respectively. So $\bar{Y}_{...} = 6$

$$SS_{Region} = 12(4 - 6)^2 + 12(8 - 6)^2 = 96$$

$$SSE = \sum_i \sum_j (n - 1) s_{ij}^2 = 3(2 + 1 + 1 + 2 + 2 + 2) = 30$$

- (c) Assuming that $SSTR = 222$ complete the ANOVA table below. (Your answers do not have to agree with those in part (b).) (2 pts)

| SOURCE | SS | DF |
|-------------|-----|----|
| Region | 144 | 1 |
| Type | 72 | 2 |
| Region*Type | 6 | 2 |
| Error | 54 | 18 |
| Total | 276 | |

- (d) What linear contrast among the sample means could be used to compare Industrial cities with Trade and Governmental cities? Specify your answer in terms of the sample means, \bar{Y}_j , $j = 1, 2, 3$ (2 pts)

$$\hat{C} = \bar{Y}_{.1} - \frac{1}{2}(\bar{Y}_{.2} + \bar{Y}_{.3})$$

- (e) Using the summary data and the ANOVA table, how would you estimate the variance of the linear contrast in part (d)? Setup the calculation. (2 pts)

$$MSE \times \left[\frac{1}{8} + \left(\frac{1}{2}\right)^2 \frac{1}{8} + \left(\frac{1}{2}\right)^2 \frac{1}{8} \right] = 3 \left[\frac{1}{8} + \frac{1}{32} + \frac{1}{32} \right]$$