Applied Linear Statistical Models
Spring 2004

***Name:*______________________________ *Cornell ID:*_______________**

Instructions: You must show your work where relevant to receive credit. There are three problems on three pages. The last page is your formula sheet, which you may remove.

All of the questions concern a simple linear regression model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently, for $i = 1, \ldots, n$.

1. Consider the least squares criterion,

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 .$$

(a) Write down the partial derivatives of Q with respect to the regression coefficients (2 pts)

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \\ \frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) \end{aligned}$$

(b) Write down the normal equations. Simplify the equations, and hence show that the least squares estimates, b_0 and b_1, are the solution of a pair of linear equations. (4 pts)

The least squares estimates satisfy the normal equations,

$$\begin{aligned} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) &= 0 \\ \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) &= 0 . \end{aligned}$$

These equations can be rewritten in the form

$$\begin{aligned} n b_0 + (\sum X_i) b_1 &= \sum Y_i & (1) \\ (\sum X_i) b_0 + (\sum X_i^2) b_1 &= \sum X_i Y_i ; & (2) \end{aligned}$$

i.e. as a pair of linear equations in b_0 and b_1.

(c) Show that $b_0 = \bar{Y} - b_1 \bar{X}$, and hence derive a formula for b_1. (4 pts)
Dividing equation (1) by n results in

$$b_0 + b_1 \bar{X} = \bar{Y} ,$$

or equivalently

$$b_0 = \bar{Y} - b_1 \bar{X} .$$

Substituting into equation (2) gives

$$(\sum X_i)(\bar{Y} - b_1 \bar{X}) + (\sum X_i^2) b_1 = \sum X_i Y_i .$$

Combining terms results in

$$\left(\sum X_i^2 - \frac{1}{n} (\sum X_i)^2 \right) b_1 = \sum X_i Y_i - \frac{1}{n} (\sum X_i)(\sum Y_i) .$$

Hence (by facts 2 and 3)

$$b_1 = \frac{SXY}{SXX} .$$

2. The least squares estimate of the slope can be written as a linear combination of the responses; i.e. $b_1 = \sum k_i Y_i$, where the constants, k_i, do not depend on the $Y's$. Using the facts on the formula sheet, show the following. (**Show the steps in the derivation, and state which facts you are using in each part.**)

   (a) $k_i = (X_i - \bar{X})/SXX$ (3 pts)
   Using the formula for b_1 derived in the previous question and fact #3

$$b_1 = \frac{SXY}{SXY} = \frac{\Sigma(X_i - \bar{X})Y_i}{SXX} = \sum k_i Y_i$$

   (b) $E(b_1) = \beta_1$ (3 pts)

$$E(b_1) = E(\sum k_i Y_i) = \sum k_i E(Y_i) = \sum k_i(\beta_0 + \beta_1 X_i) .$$

   But facts #1 and #2 respectively imply that $\sum k_i = 0$ and $\sum k_i X_i = 1$. Hence

$$E(b_1) = \beta_0(\sum k_i) + \beta_1(\sum k_i X_i) = \beta_1 .$$

   (c) $\text{var}(b_1) = \sigma^2/SXX$ (4 pts)
   Using fact #4 and the identity $\sum k_i^2 = 1/SXX$ we have

$$\text{var}(b_1) = \text{var}(\sum k_i Y_i) = \sum k_i^2 \text{var}(Y_i) .$$

   Finally, since var$(Y_i) = \sigma^2$,

$$\text{var}(b_1) = \frac{\sigma^2}{SXX} .$$

3. A simple linear regression using $n = 32$ observations resulted the following ANOVA table:

Source	SS	df	MS	F-ratio
Regression	(144)	(1)	(144)	(9)
Error	480	(30)	(16)	
Total	624			

(a) Fill in the six missing values in the table. (4 pts)

(b) What is the absolute value of the t-statistic for testing $H_0 : \beta_1 = 0$? (2 pts)

$$|t| = \sqrt{F} = 3 \text{ or } |t| = |b_1/s\{b_1\}| = 3/\sqrt{16/25} = 3.75$$

(NOTE: These answers should be the same! The value of b_1 in the next part should be 2.4.)

(c) Suppose that $b_1 = 3$ and $SXX = 25$. Find a 95% confidence interval for the true slope, β_1. (HINT: Use the fact that $t(.975, 30) = 2$.) (2 pts)
A 95% confidence interval is given by

$$b_1 \pm t \times s\{b_1\} = 3 \pm 2 \times \sqrt{16/25} = 3 \pm 1.6$$

(d) What proportion of the variation in the responses is explained by regression on the X-variable? You may give the answer as a fraction. (2 pts)

$$r^2 = \frac{SSR}{SSTO} = \frac{144}{624} = \frac{3}{13}$$

FACTS and FORMULAS

- FACT 1:

$$\sum(X_i - \bar{X}) = 0$$

- FACT 2:

$$SXX = \sum(X_i - \bar{X})^2 = \sum(X_i - \bar{X})X_i = \sum X_i^2 - \frac{1}{n}(\sum X_i)^2$$

- FACT 3:

$$SXY = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i - \bar{X})Y_i = \sum X_i Y_i - \frac{1}{n}(\sum X_i)(\sum Y_i)$$

- FACT 4: Suppose that $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$. Then, if $a_1, \ldots, a_n$ are constants,

$$\sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i\,, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

- STANDARD ERRORS

$$\begin{aligned} \sigma\{b_1\} &= \frac{\sigma}{\sqrt{SXX}} \\ \sigma\{\hat{Y}_h\} &= \sigma\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{SXX}} \\ \sigma\{pred\} &= \sigma\sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{SXX}} \end{aligned}$$

- R-SQUARED

$$r^2 = \frac{SSR}{SSTO}$$