Math 644, Fall 2012

Homework 6 Due: Friday, 11/30/2012

1. In regression analysis of on-the-job head injuries of warehouse laborers caused by falling objects. Y is a measure of severity of the injury, X_1 is an index reflecting both the weight of the object and the distance it fell, and D_1 and D_2 are indicator variables for nature of head protection worn at the time of the accident, coded as follows

type of protection	D_1	D_2
Hard hat	1	0
Bump cap	0	1
None	0	0

the response function to be used in the study is $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2$.

   (a) Develop the response function for each type of the protection category.

   (b) For each of the following questions, specify H_0 and H_a for each appropriate test: (1) with X_1 fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection? (2) with X_1 fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap?

2. Refer to the tool wear example (in the lecture notes, part 5 of Chapter 2), consider model $Y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \beta_5 X_1 D_1 + \beta_6 X_1 D_2 + \beta_7 X_1 D_3 + \varepsilon$. Indicating the meaning of (1) β_3, (2) $\beta_4 - \beta_3$, (3) β_1, (4) $\beta_7 = 0$, (5) $\beta_5 - \beta_6$.

3. **Steroid level.** An endocrinologist was interested in exploring the relationship between the level of a steroid (Y) and age (X) in healthy female subjects whose ages ranged from 8 to 25 years. She collected a sample of 27 healthy females in this age range.

   (a) Fit regression model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$. Plot the fitted regression function and the data. Does the quadratic regression function appear to be a good fit here? Find R^2.

   (b) Test whether or not there is a regression relation; use $\alpha = 0.01$. State the alternatives, decision rule, and conclusion. What is the p-value of the test?

(c) Predict the averaged steroid levels of females aged 15 using a 99 percent prediction interval. Interpret your interval.

(d) Test whether the quadratic term can be dropped from the model; use $\alpha = 0.01$. State the alternatives, decision rule, and conclusion.

4. Refer to **Commercial properties**, the vacancy rate predictor (X_3) does not appear to be needed when property age (X_1), operating expenses and taxes (X_2), and total square footage (X_4) are included in the model as predictors of rental rates (Y).

    (a) The age of the property (X_1) appears to exhibit some curvature when plotted against the rental rates (Y). Fit a polynomial regression model with property age (X_1), the square of property age (X_1^2), operating expenses and taxes (X_2), and total square footage (X_4). Plot the Y observations against the fitted values. Does the response function provide a good fit?

    (b) Calculate R_a^2. What information does this measure provide?

    (c) Test whether or not the square of property age (X_1^2) can be dropped from the model; use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What is the p-value of the test?

    (d) Estimate the mean rental rate when $X_1 = 8, X_2 = 16$, and $X_4 = 250,000$; use a 95 percent confidence interval. Interpret your interval.

5. Refer to regression model $Y_i = \beta_0 + \beta_2 X_{i2} + \varepsilon_i$, where $X_2 = 1$ if firm incorporated and 0 otherwise, and exclude variable X_1.

    (a) Obtain the $\mathbf{X}'\mathbf{X}$ matrix for this special case of a single qualitative predictor variable, for $i = 1, \ldots, n$ when n_1 firms are not incorporated.

    (b) Find $\mathbf{b}$.

    (c) Find SSE and SSR.

6. The **CDI** provides selected country demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data ware deleted from the data set. The information generally pertains to the years 1990 and 1992. The 17 variables are:

    1 Identification number: 1-440;

2 County: County name;

3 State: Two-letter state abbreviation;

4 Land area: Land area (square miles);

5 Total population: Estimated 1990 population;

6 Percent of population aged 18-34: Percent of 1990 CDI population aged 18-34;

7 Percent of population 65 or older: Percent of 1990 CDI population aged 65 years old or older;

8 Number of active physicians: Number of professionally active nonfederal physicians during 1990;

9 Number of hospital beds: Total number of beds, cribs, and bassinets during 1990;

10 Total serious crimes: Total number of serious in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies;

11 Percent high school graduates: Percent of adult population (persons 25 years old or older) who completed 12 or more years of school;

12 Percent bachelors degrees: Percent of adult population (persons 25 years old or older) with bachelors degrees;

13 Percent below poverty level: Percent of 1990 CDI population with income below poverty level 3;

14 Percent unemployment: Percent of 1990 CDI labor force that is unemployed;

15 Per capita income: Per capita income of 1990 CDI population (dollars);

16 Total personal income: Total personal income of 1990 CDI population (in millions of dollars);

17 Geographic region: Geographic region classification is that used by the U.S. Bureau of the Census, where: 1=NE, 2=NC, 3=S, 4=W.

The number of active physicians (Y) is to be regressed against total population (X_1), total personal income (X_2), and geographic region (X_3, X_4, X_5).

(a) Fit a first-order regression model. Let $X_3 = 1$ if NE and 0 otherwise, $X_4 = 1$ if NC and 0 otherwise, and $X_5 = 1$ if S and 0 otherwise.

(b) Examine whether the effect for the northeastern region on number of active physicians differs form the effect for the north central region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate.

(c) Test whether any geographic effects are present; use $\alpha = 0.01$. State the alternatives, decision rule, and conclusion. What is the p-value of the test?