CPH 931 — Fall 2008 — Dr. Charnigo

Written Assignment 1

Written Assignment 1 is due on Tuesday 23 September at the end of lecture.

The spreadsheet {PollutionMortality.xls} contains fictionalized data for 59 Standard Metropolitan Statistical Areas. The variable names and meanings are as follows.

Name	Meaning
city	City name
JanTemp	Mean January temperature (degrees Fahrenheit)
JulyTemp	Mean July temperature (degrees Fahrenheit)
RelHum	Relative Humidity
Rain	Annual rainfall (inches)
Mortality	Age adjusted mortality
MortalityALT	Age adjusted mortality with "mistake" in row 15
PopDensity	Population density
HCPot	HC pollution potential
NOxPot	Nitrous Oxide pollution potential
SO2Pot	Sulfur Dioxide pollution potential

[20] 1. Consider two linear regression models: (i) Mortality is the response variable and JulyTemp is the sole explanatory variable; (ii) MortalityALT is the response variable and JulyTemp is the sole explanatory variable.

[10] a. Fit models (i) and (ii) using ordinary least squares. Then fit models (i) and (ii) using M estimation. Construct a table that compares the four fitted models with respect to the estimated slope and its corresponding p-value.

[10] b. Explain why the two sets of results for model (i) are similar, why the two sets of results for model (ii) are not similar, and why the two sets of results using M estimation are similar.

[20] 2. As in exercise 1, let Mortality be the response variable and JulyTemp be the sole explanatory variable. However, we will not assume that the expected value of Mortality is a linear function of JulyTemp.

[10] a. Apply LOESS smoothing to obtain a plot of the estimated expected value of Mortality as a (nonlinear) function of JulyTemp.

[10] b. Describe the pattern revealed by LOESS smoothing. Is there a plausible public health or medical explanation for this pattern?

[30] 3. Consider a linear regression model in which Mortality is the response variable and JanTemp, JulyTemp, RelHum, Rain, PopDensity, HCPot, NOxPot, and S02Pot are explanatory variables.

[10] a. Fit the model using ordinary least squares. Report the coefficient estimates, standard errors, and p-values. Comment on the nature of the heteroscedasticity, if any.

[10] b. Refit the model using weighted least squares. Report the coefficient estimates, standard errors, and p-values.

[10] c. Identify the major differences, if any, between weighted least squares results and ordinary least squares results.

[30] 4. Continue from exercise 3.

[10] a. Report variance inflation factors for the model fitted using weighted least squares. Which explanatory variables appear to be causing multicollinearity?

[10] b. Still using the weights from part a, construct a ridge trace that displays coefficient estimates for values of the ridge parameter between 0 and 0.02 by increments of 0.001.

[10] c. Refit the model using ridge regression (or, more precisely, a hybrid of ridge regression and weighted least squares) with a value of the ridge parameter suggested by part b and the weights from part a. Identify the major differences, if any, between ridge regression results and weighted least squares results.