CPH 931 — Fall 2008 — Dr. Charnigo

Written Assignment 1 Solutions

1a. The requested table is shown below.

Model, estimation method	Slope estimate	p-value
(i), ordinary least squares	4.776	0.0048
(ii), ordinary least squares	-19.84	0.5545
(i), M estimation	4.860	0.0027
(ii), M estimation	5.088	0.0017

1b. The two sets of results for model (i) are similar because there are no gross outlying values on Mortality to disrupt estimation by ordinary least squares. The two sets of results for model (ii) are not similar because there is a gross outlying value on MortalityALT. This outlier disrupts estimation by ordinary least squares but has little effect on M estimation, since M estimation places a finite upper bound on the contribution each observation can make to the criterion whose minimization determines the estimates. The two sets of results using M estimation are similar because Mortality and MortalityALT differ only on one observation. Changing the value of the response for one observation has little effect on M estimation, for the reason noted above.

2a. [Show your LOESS plot.]

2b. The estimated expected value of Mortality does not change much from 63 to 68 degrees, increases (although not monotonically) from 68 to 77 degrees, does not change much from 77 to 79 degrees, and decreases from 79 to 85 degrees. The increase from 68 to 77 degrees may reflect that people with chronic health problems are more vulnerable at higher temperatures, yielding more deaths in areas with higher temperatures even if those deaths are not officially attributed to heat related illnesses. However, such an explanation is difficult to reconcile with the decrease from 79 to 85 degrees. Closer inspection of the data reveals that the three areas with July temperatures above 80 degrees *and* mortalities less than 900 were the only three areas to have values less than 5 on HCPot, NOxPot, and SO2Pot. So reduced pollution potential — either because pollutants disperse differently in higher temperatures than in lower temperatures or because areas with higher temperatures do not have the same kinds of manufacturing activities as areas with lower temperatures — may help explain the decrease from 79 to 85 degrees. In any event, the LOESS smoothing shows that the bivariate relationship between JulyTemp and Mortality is not linear.

3a. The ordinary least squares coefficient estimates, standard errors, and p-values are reported below. A plot of ordinary residuals against fitted values shows that the residuals become more spread out as the fitted values increase. This suggests that the error variance becomes larger as the mean response increases. A plot of studentized residuals against fitted values conveys the same impression.

Variable	Coefficient estimate	Standard error	p-value
Intercept	466.0	224.4	0.0430
JanTemp	-0.524	0.956	0.5857
JulyTemp	4.084	2.353	0.0888
RelHum	1.605	1.582	0.3152
Rain	1.615	0.742	0.0342
PopDensity	0.00451	0.00517	0.3877
HCPot	-0.914	0.629	0.1523
NOxPot	1.898	1.272	0.1421
SO2Pot	0.143	0.184	0.4409

3b. The weighted least squares results, using weights inversely proportional to the squares of the smoothed absolute values of the ordinary residuals from part a, are reported below.

Variable	Coefficient estimate	Standard error	p-value
Intercept	577.5	159.3	0.0007
JanTemp	-0.254	0.642	0.6943
JulyTemp	2.814	1.846	0.1337
RelHum	0.750	0.885	0.4006
Rain	1.866	0.540	0.0011
PopDensity	0.00821	0.00521	0.1216
HCPot	-1.468	0.407	0.0007
NOxPot	2.980	0.836	0.0008
SO2Pot	0.083	0.171	0.6307

3c. The coefficient estimates have changed noticeably, although there is no systematic pattern to the changes (such as most being closer to 0 or most being further away from 0). The most striking changes in coefficient estimates are perhaps those for HCPot (from -0.914 to -1.468) and NOxPot (from 1.898 to 2.890), as HCPot (from $p = 0.1523$ to $p = 0.0007$) and NOxPot (from $p = 0.1421$ to $p = 0.0008$) are now identified as statistically significant predictors of Mortality, notwithstanding their role in the multicollinearity (to be formally diagnosed in exercise 4). The standard errors have also changed noticeably, although here there is a systematic pattern. Except for PopDensity and SO2Pot, all of the standard errors have been reduced by 22% to 44% compared to ordinary least squares.

4a. The variance inflation factors are as reported below. The explanatory variables causing the multicollinearity are HCPot and NOxPot.

Variable	Variance inflation factor
Intercept	NA
JanTemp	3.05
JulyTemp	3.30
RelHum	3.60
Rain	3.31
PopDensity	1.73
HCPot	182
NOxPot	186
SO2Pot	3.26

4b. [Show your ridge trace.]

4c. Choosing λ based on the ridge trace entails some subjectivity. I chose $\lambda = 0.010$, although a choice as large as 0.020 or as small as 0.005 is defensible. The results are shown below; I calculated approximate p-values using the Wald method. The coefficient estimates for HCPot and NOxPot are 77% and 78% smaller, but the standard errors are 76% and 76% smaller, so the conclusions about statistical significance persist. The coefficient estimate for SO2Pot has more than quadrupled, and the standard error has declined by 13%, so now SO2Pot is also a statistically significant predictor of Mortality. The coefficient estimate for RelHum has changed sign (from positive to negative), although the conclusion about lack of statistical significance is unaltered.

Variable	Coefficient estimate	Standard error	Approximate p-value
Intercept	610.9	160.2	< 0.001
JanTemp	-0.485	0.648	0.454
JulyTemp	2.765	1.861	0.137
RelHum	-0.133	0.836	0.874
Rain	2.491	0.500	< 0.001
PopDensity	0.00776	0.00546	0.155
HCPot	-0.331	0.098	< 0.001
NOxPot	0.645	0.200	0.001
SO2Pot	0.351	0.148	0.018