EXERCISE ON NON-NORMAL RESIDUALS, DUE THURSDAY 11/14

(1) Suppose

$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 3 \\ 10 \end{bmatrix} \qquad y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ .9 \\ 2.1 \\ 2.9 \\ 1.1 \\ 1.9 \\ 3.1 \\ 4 \end{bmatrix} .$$

It is easy to see that the first 9 observations lie very close to a straight line, while the 10'th is far off that line. We will investigate how various approaches to "robust" estimation of a straight-line fit behave in this situation.

(a) Estimate a linear regression of y on x and a constant by ordinary least squares. Form the standardized residuals (residuals divided by their estimated standard errors). Do they show extreme outliers (values larger than, say, 3 or 4 standard error units)?

(b) Compare the coefficient standard errors generated by the standard $\text{Cov}(Y \mid X) = \sigma^2 I$ regression theory to those generated from an HCCM (heteroskedasticity-consistent covariance matrix) calculation.

(c) Generate a sample of 10,000 or more Monte Carlo draws from the posterior distribution of the coefficients, assuming the residuals are i.i.d. conditional on x and distributed as σ times a t distribution with degrees of freedom 3.5. Do this also with the degrees of freedom 7. Show an effective sample size computation. (There is more than one way to do this. One is to use the `effectiveSize` function in the `coda` package of R.) Also show a trace plot for each coefficient. Do this for both degrees of freedom values, and comment on whether there is any indication of lack of convergence. An R program that generates such draws is available in the file `tshock.R` on the course web site.

(d) Show the mean and standard deviation of the sample from the posterior for each degrees of freedom case and compare them to the results from OLS. Display plots of the posterior densities for the two coefficients under each degrees of freedom case. Probably you will see multiple peaks in at lest some of these. Discuss how conclusions might differ with the t-distribution model from what is implied by sticking to OLS or OLS with HCCM errors.

(2) David Card and Alan Krueger wrote a classic paper on the minimum wage, observing what happened to employment in fast food restaurants when New Jersey passed a minimum wage law that affected fast food workers, while neighboring Pennsylvania did not. They found a statistically significant *positive* effect of the minimum wage on firm-level employment. The original article, a description of the data, and two versions of the data set are available on the course web site. The `ckdata.RData` file has all the data as an R data frame that can be brought in to R with `load("ckdata.RData")`. The original data file used by Card and Krueger is also available. It has the data in fixed-width columns in `public.dat` and a description of what's in each column in `codebook`. There is a SAS program to read in the data in `check.sas`. We will be using (unless you want to do more on your own for fun (?)) just the data in the second-period employment, first-period employment, and starting-wage series. You need to follow the Card and Krueger article's description of how to use the full-time and part-time employment data to construct total employment and how to construct their `gap` variable from the data on starting wage and the minimum wage value (5.05).

(a) Estimate by OLS the regression of change in employment on `gap` and a constant, reproducing the Card-Krueger result. Check whether using HCCM standard errors alters any conclusions.

(b) Estimate by OLS the regression of period-2 employment on a constant, `gap`, and period-1 employment. Again compare results with HCCM standard errors. This regression probably shows an "insignificant", rather than significantly positive, effect of `gap` on employment.

(c) Repeat your estimates of the two equations above, this time using MCMC posterior simulation under the assumption of t-distributed errors with 3.5 and 7 degrees of freedom.

(d) Briefly discuss what you might conclude from the pattern of results you find.