## Problem Set 3

*Problem 1* If a vector $x$ of random variables has a normal distribution with mean (vector) $\mu$ and variance (matrix) $\Sigma$, then $S = (x-\mu)'\Sigma^{-1}(x-\mu)$ has a chi-squared ($\chi^2$) distribution with parameter (degrees of freedom) equal to the number of variables in $X$. If you know eigen-vectors and -values you can show this yourself. Use this fact to find an asymptotic test of the equality of the male and female regression coefficients in the wage regressions estimated separately on the male and female samples. We do not make the assumption that the variances of the errors in the two regressions are equal as was done in the Oaxaca paper.

(i) Derive the variance of the difference of the male and female regression coefficients. Do we have to worry about their covariance? Why (not)?

(ii) Assume that the CLR assumptions hold in the male and female populations and that the variance of the errors in the two populations is known. Suggest a test statistic that under the null hypothesis of equal coefficients has a $\chi^2$ distribution. What is the parameter, i.e. the degrees of freedom of this distribution? Hint: Use the fact that the sum or difference of (vectors of) random variables that have the normal distribution also have the normal distribution.

(iii) If the variances of the errors are unknown and potentially unequal, suggest a test that in large samples has a $\chi^2$ distribution if the null hypothesis is correct. Show that the test statistic has the same distribution under the null if the CLR assumptions do not hold, but we assume that the independent variables and the errors are uncorrelated. Hint: you can cite results in the lecture notes.

*Problem 2* The data set jivesh.asc contains the Angrist-Krueger data (329507 observations) in random order. The columns of the data set are BIRTHYR (year of birth), QOB (quarter of birth), EDUC (years of education) and LNWAGE (log of wage). Redo the analysis at the end of lecture 8 for the linear regression model with LNWAGE as dependent and BIRTHYR, the square of BIRTHYR and EDUC as independent variables. The instruments are the QOB dummies, i.e. the indicators of the second, third and fourth quarter (you have to make these variables from QOB). Consider sample sizes of 5000, 10000, and 20000. Because the data are in random order you can just take the first 5000, second 5000 etc. EDUC is endogenous and BIRTHYR is exogenous.

Use the samples to estimate the return to education using 2SLS and obtain the sampling variance of this estimator. Produce a table as in lecture 8 with one change. Instead of the t-test report the F-test for the joint significance of the QOB dummies in the first stage and the overall F-test for the significance of all coefficients in the first stage except the intercept. Use your result to recommend a procedure that would make it likely that the 2SLS estimator is approximately unbiased and has computed standard errors that are close to the standard deviation of the sampling distribution.