ADVANCED ECONOMETRICS

**EMPIRICAL PROBLEM SET #1**

*Due September 24*

This exercise examines regression analysis. Feel free to work in groups of two or three, but each one of you must hand in your own problem set with your own interpretation of the results. Please include a concise summary of your empirical results/answers. We will analyze the following STATA data sets:

Data 1: german92.dta

The data set comes from a 1992 survey of workers in Germany (see DiNardo and Pischke (1997), “The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?”). The unit of observation is the individual. There are 20,042 observations on 17 variables. The included variables are:

Variable Description

lnw Natural log of the hourly wage reported by the individual

ed Years of schooling

exp/exp2 Work experience of person and its square (exp2)

female 1 if female, 0 otherwise

mar 1 if married, 0 otherwise

femmar female\*mar

part 1 if part-time worker, 0 otherwise

city 1 if lives in city, 0 otherwise

beamter 1 if civil servant, 0 otherwise

blue 1 if blue color worker, 0 if white color worker

occ 4-digit occupation codes

computer 1 if person uses a computer at work, 0 otherwise

calc 1 if person uses a calculator at work, 0 otherwise

telefon 1 if person uses a telephone at work, 0 otherwise

pencil 1 if person uses a pencil at work, 0 otherwise

hammer 1 if person uses a hammer at work, 0 otherwise

Data 2: NSW\_ladonde.dta

The data set comes from the National Supported Work Demonstration (NSW), a randomized job training program (see LaLonde (1986), “Evaluating the Econometric Evaluations of Training Programs with Experimental Data”). The unit of observation is the individual. There are 16,714 observations on 10 variables. The included variables are:

Variable Description

experimt 1 if experimental data, 0 if non-experimental (CPS) data

treat 1 for treatment group, 0 for control group

age Age

educ Years of schooling

black 1 if black, 0 otherwise

hispanic 1 if Hispanic, 0 otherwise

married 1 if married, 0 otherwise

nodegree 1 if no high school degree, 0 otherwise

re75 Real earnings in 1975 (before the NSW Program)

re78 Real earnings in 1978 (after the NSW Program)

1. **Estimating the returns to education (using German data)**
2. Run the bivariate regression of log-wages on a constant and years of schooling. Briefly interpret the “economic meaning” of slope coefficient. Show the scatter plot of log-wages on the y-axis and years of schooling on the x-axis. Based on the plot, is there any evidence of heteroskedasticity? Now run the multivariate regression of log-wages on a constant, years of schooling, experience, experience-squared, the gender indicator, and the marital status indicators, while adjusting for heteroskedasticity (use “robust” option in STATA). Explain briefly how these estimated standard errors are corrected for heteroskedasticity.
3. Now apply the partialling-out method (a.k.a. Frisch-Waugh-Lovell Theorem) to the multivariate regression – (i) run a regression of log-wages on a constant, experience, experience-squared, the gender indicator, and the marital status indicators; (ii) save the residuals using STATA command [predict *variable name*, residual]; (iii) run a regression of years of schooling on a constant, experience, experience-squared, the gender indicator, and the marital status indicators; (iv) save the residuals; (v) regress the residuals from (ii) on the residuals from (iv), while adjusting for heteroskedasticity. Compare this regression coefficient to the one (the coefficient on years of schooling) from the multivariate regression in part (a). Are they the same? What about standard errors and *t*-statistics?
4. **Estimating the returns to “computer use” (using German data)**
5. Write down the OLS estimator when the explanatory variable is binary (in a bivariate regression). Run a regression of log-wages on a constant and the computer indicator. Do the two sample t-test for the mean difference in log-wages between workers who use computers on the job and non-users (“ttest” command in STATA). Compare the mean difference, standard error and *t*-statistic to the ones from the bivariate regression. Are they the same?
6. Now regress log-wages on a constant, the computer indicator, years of schooling, experience, experience-squared, the gender indicator, the marital status indicator, female\*married, a dummy for part-time worker, living in city, and civil servants, while adjusting for heteroskedasticity. How does the regression-adjusted estimate of the return to computer use compare to the one from part (a). What does this imply about the similarity in observables between computer users and non-users?
7. Now compare the mean characteristics – e.g., age, gender, marital status, education (years of schooling), part-time worker, etc. – of computer users to non-users. Are they statistically different? Why might this cast doubt on the causal interpretation of the return to computer use? Can you think of (unobserved) variables that we have not controlled for that may be related to both computer use and log-wages? Explain how this could lead to omitted variables bias in the OLS estimate of the return to computer use.
8. Now add the indicators for calculator, telephone, and pencil use to the regression you ran in part (b). Compare the estimated return to computer use to the one from part (b). Now run a regression that also controls for the individual’s occupation category as “fixed effects” [areg y x, absorb(occ) robust]. Interpret the implications of your findings for the role of potential omitted variables bias in the OLS estimate of the effect of computer use on log-wages (see DiNardo and Pischke (1997) for their interpretation).
9. Now run the same regression as in part (d), while using the “cluster” option in STATA to correct the estimated standard errors for clustering at the occupation-level [areg y x, absorb(occ) cluster(occ)]. Explain why the standard error on the estimated return to computer use is higher (and lower *t*-statistic) than when clustering is not corrected for.
10. **Estimating the effect of the NSW program on earnings (using NSW data)**
11. Use experimental data [experimt==1]. Test the mean differences in characteristics (e.g., age, race, ethnicity, marital status, and years of schooling) between treatment and control group. What does the result of this test imply about the identification of the causal effect of the NSW program? Estimate the average causal effect of the NSW program on earnings (re78) using a regression: i) without control variables; and ii) with control variables (age, educ, black, hispanic, married, nodegree, re75). Interpret the results. Are these two estimates (with and without controls) statistically same? Why would this be the case?
12. Now use the non-experimental control group from the CPS data, along with the experimental treatment group [(experimt==0 & treat==0)|treat==1]. As you did in part (a), test the mean differences in characteristics between treatment and control group. Interpret the results. Again, as you did in part (a), run a regression of earnings (re78) on treatment dummy (treat): i) without control variables; and ii) with control variables. Interpret the results. How do these results compare to the causal estimates using experimental data in part (a)?