Economics 312 Spring 2020
Project #5 Assignment Due: 11:59pm, Monday, March 2

Partner assignments

Shisham Adhikari	Amrita Sawhney
Blaise Albis-Burdige	Lauren Rabe
Isabelle Caldwell	Grisha Post
Anne Cao	Evian Oosthuizen
Aditya Gadkari	Max Nobel
Catherine Gong	Paul Nguyen
Lirui Jiao	Peter Mulgrew
Jonathan Li	Charlie Lyu
Roy Shannon	George Zhao
Matt Wan	

Data and background

The data set wage2_class.dta is extracted from a data set in Wooldridge's collection. It is drawn from the National Longitudinal Study of Young Men (NLSYM), a panel study that began in 1966. The NLSYM tracked over time, through repeated surveys, a sample of (initially) 5000+ men beginning at ages 14–24. The data include numerous standard variables relating to each young man. The paper from which the data are taken has information for both 1980 and from varying dates in the early 1970s. It is unclear which data are in wage2_class.dta and why there are only 935 observations.

In addition to the standard and self-explanatory variables, the dataset includes two test scores. The IQ variable is a standard intelligence-quotient test, which is constructed to have a population mean of 100. According to a paper using this dataset, "the [Knowledge of the World of Work] test examines respondents' knowledge about the labor market, covering the duties, educational attainment, and relative earnings of ten occupations." (Note that both of these variables are capital letters in the dataset and remember that Stata names *are* case-sensitive.)

Variables in the data set are shown below:

Variable	Description
wage	average hourly earnings
IQ	IQ score
KWW	knowledge of world of work score
educ	years of education
exper	years of work experience
tenure	years with current employer
age	age in years
married	=1 if married
black	=1 if black
south	=1 if live in south
urban	=1 if live in SMSA
sibs	number of siblings
brthord	birth order
meduc	mother's education
feduc	father's education

1. Exploring the data

a. Examine the summary statistics of the variables in the data set. Are there any surprises? Discuss very briefly anything notable that we learn about the variables.

b. Are there variables for which observations are missing? What is the most likely explanation for this? Short of the imputation models we will study at the end of the course, what would be the options for dealing with this situation? Which will you use?

c. Are there sets of variables in the data set that you would expect to be highly correlated with each other? Explore the correlations among groups of explanatory variables (*i.e.,* not the wage) that would potentially cause multicollinearity concerns.

d. In many labor data sets, the experience variable is really "potential experience," constructed artificially as age minus education minus 6. Does this appear to be the case for NLSYM data? Is this good or bad?

e. The variable descriptions of *exper* and *tenure* suggest that the former is total work experience and the latter is experience with current employer. The difference *exper – tenure* should then be experience prior to joining the current employer. Create and examine this variable. Does your analysis raise any red flags? If so, what would be the reasonable options for dealing with the problem?

2. Basic regression analysis

a. The labor-economics literature has settled on using the natural log of wage as the dependent variable in its preferred specification. This will be a problem if were are observations with a zero wage (representing zero earnings), which will then be omitted from the sample because the log of zero is not defined. Is this a problem in this data set?

b. OLS regression will lead to biased and inconsistent estimates if one or more of the regressors is endogenous. A regressor is endogenous if a random shock to the wage (*i.e.,* the error term of the equation) induces "reverse causality" by causing a change in the regressor. For example, we can safely assume that a change in one's wage would not cause one to be older or younger, so the age variable would not be subject to reverse causality and could be treated as exogenous. Is it reasonable to treat all of the other variables in the data set as exogenous with respect to wage or are there any that raise concerns?

c. There are a lot of potential regressors in the data set. If we were to include quadratic or higher-power terms and/or interactions for all the variables we would have a mind-numbing mess of an equation. With the merits of parsimony clearly in mind, are there particular variables for which you think a quadratic term or an interaction term would be useful? Explain your logic.

d. Start with a basic linear model and then explore the nonlinearities (quadratics and interactions) that you think might be useful. As a practical expedient, I suggest trying one variation at a time, and either adopting the variation (if appropriate) or abandoning it before going on to the next variation. Otherwise you will end up with an unmanageable set of models. Present your results in one or more outreg2 tables so that the reader can easily compare models. Identify your preferred specification and explain why you chose it.

3. Implications of your regression. Use your estimated model(s) to answer the following questions based on the evidence in the NLSYM sample, in each case controlling for the other characteristics.

a. Are aspects of the person's childhood family background important in determining wage?

b. Does the location of his residence matter for wage?

c. Does one seem to get a higher wage, other things being equal, by remaining with the same employer more than by switching jobs?

d. What are the effects of IQ and KWW test scores? Some have argued that all that these tests measure is the ability to take tests, in which case it would be redundant to keep both in the equation and questionable whether they would have any effect on earnings. Is there evidence for this based on their correlation coefficient and/or regression results?

4. Prediction of wage. The logged dependent variable is awkward and cannot be directly interpreted. If we were predicting the implications of our model for individuals, we would want to predict fitted values for wage, not log-wage.

a. An obvious way to create fitted values for wage would be to compute *lwage* from your regression and generate $elwage = \exp\left(lwage\right)$. Do this and compare the summary statistics of *elwage* with those of *wage*. Does this seem like a good predictor of *wage*?

b. Section 6-4c of Wooldridge's text discusses the problem of prediction with a log dependent variable. Implement his corrected prediction using equation [6.40]. Compare the summary statistics for these predictions with those of *wage*. Does this seem to have fixed the problem?

5. Conclusion. Briefly summarize what you think are the most important or surprising implications of your analysis. What are the strengths and weaknesses of your work and what could, in principle, be done to improve them?