PUBPOL 639: ASSIGNMENT 3
Winter 2011

Due: Monday, March 14th at the start of class

Service-Sector Unionization

In your new job as policy analyst for SEIU, you are helping to craft a memo that will be used to promote the benefits of unionization to potential new service-sector members. You will need to use multivariable regression in order to estimate the relationship between unionization and wages, holding other factors constant.

Data

The dataset "union.dta" contains data from the 2009 Current Population Survey. It contains a random subsample of 1,000 currently employed participants in the "Earner Study" supplement to the CPS who worked last year in a service industry (retail, health care, education, personal services, and government). The dataset contains five variables:

1. **hrwage** - Hourly wage last year in dollars. This was estimated by dividing wage and salary income by the approximate number of hours worked last year (weeks worked * usual hours worked per week). Observations with hourly wages less than $3 and more than $40 were excluded.

2. **union** - A dummy variable indicating whether the worker was a union member or covered by some other collective bargaining agreement.

3. **age** - Age in years.

4. **empsize** - The size of the firm the person works for. This was originally a categorical variable with ranges (e.g. 10-24, 25-99, etc) for which I have imputed the midpoint of the ranges.

5. **manager** - A dummy variable indicating the person worked in a managerial occupation.

Presentation

You should answer the questions below, but also present the results from all your regressions in a single table that is easy-to-read with each column corresponding to a separate regression. You should use Table 7.1 in the textbook as an example (though you do not have to report the SER). Everything should be labeled, variables should have names, and notes should provide enough information to understand what you have done. If you are interested, the Stata command "outreg2" will spit out regression output into a text file, that you can then paste into an excel table. Since this command is executed as an ado file (basically a do-file that executes a specific routine), you may need to install it first. Try "net search outreg2" or "net install outreg2". You would use it like this:

```
regress y x1, robust
outreg2 using assignment4.txt, replace
regress y x1 x2, robust
outreg2 using assignment4.txt, append
```

Questions

1. Estimate the relationship between hourly wages and union status using bivariate regression and report your results in a table.

    a. What is the population regression function (or equation) you have estimated? [do not use Y and X for variable names]

    b. Interpret the coefficient on union status.

    c. Interpret the constant.

    d. Test the null hypotheses that the coefficient on union status in the population regression function is zero.

    e. How much of the variation in wages can be accounted for by union status?

2. After your training in PubPol 639, you are reluctant to interpret the bivariate relationship between wages and union status as the *causal effect* of unionization given all the possible confounders. One variable you are particularly worried about is age.

    a. Now regress hourly wage on union status and age and report the results in the table.

    b. What is the population regression function (or equation) you have estimated? [do not use Y and X for variable names]

    c. What happens to the coefficient on union status once age is included in the regression? Why? Explain in terms of omitted variable bias.

    d. Given your answer to (c), what is the sign of the correlation between hourly wage and age? What about age and union status?

3. You recall from your pre-Ford days as a union organizer that you would specifically target employers that had a lot of workers so that you could organize the most people in one place. You also recall hearing that wages tend to vary with firm size, though you cannot remember the direction. You figure you should control for firm size too, just in case.

    a. Now regress hourly wage on union status, age, and empsize, reporting your results in the table.

    b. What happens to the coefficient on union status once age and employer size are included in the regression?

    c. Do larger firms pay more or less, conditional on union status and age?

4. The data contains workers from many different occupations in the main service industries, including managers.

   a. Do managers tend to be more or less unionized than non-managers?

   b. Do managers tend to have higher or lower wages than non-managers?

   c. Given your answers to (a) and (b), what is the consequence of not accounting for management status when regressing hourly wages on union status?

   d. Regress hourly wage on union status, age, empsize, and management, reporting your results in the table.

   e. Interpret the coefficient on union status.

5. When you sit down with your boss to share your analysis, she voices two concerns:

   "This can't be the whole story since you haven't controlled for sex. Women make less than men, so you have to account for sex."

   "What about industry? Much of our growth has been in the public sector, which tends to pay more."

Since you do not have variables on sex and industry in your dataset, you can't directly address these concerns by adding them to the regression. However, you are able to find some information about wages and unionization rates by sex and industry from another source (shown in table below).

	Men	**Women**	**Gov't**	**Non-Gov't**
Average Hourly Wage	17.65	15.36	21.24	15.49
Fraction Unionized	0.1747	0.1736	0.484	0.129

   a. What is the consequence of failing to control for sex in your regressions above?

   b. What is the consequence of failing to control for industry (particularly, government or non-government sector) in your regressions above?

   c. List at least two other factors that you think may be important to control for. For each, state what you anticipate the sign of the correlation between the variable and hourly wages and between the variable and union status to be. How does omitting each of these variables from the bivariate regression of question 1 affect the coefficient on union status?