PUBPOL 639: ASSIGNMENT 3 - Solutions
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
```

Table 1. Results of OLS Regression of Hourly Wage on Union Status and Various Control Variables

	Dependent variable: Hourly wage			
Regressor	(1)	(2)	(3)	(4)
Union	5.324***	4.662***	4.140***	4.429***
	(0.727)	(0.727)	(0.758)	(0.761)
Age		0.122***	0.123***	0.118***
		(0.018)	(0.018)	(0.018)
Employer Size			0.00171***	0.00157***
			(0.00048)	(0.00047)
Manager				4.522***
				(0.932)
Constant	15.29***	10.44***	9.229***	9.066***
	(0.290)	(0.739)	(0.787)	(0.778)
Observations	1000	1000	1000	1000
R-squared	0.054	0.093	0.104	0.127
Adjusted R-squared	0.054	0.091	0.102	0.124

Notes: Robust standard errors are in parentheses under coefficients. The individual coefficient is statistically significant at the *** 1% level, ** 5% level, or *10% level of significance using a two-sided t-test. Data comes from the 2009 Current Population Survey and includes a random sample of workers in the service industries (retail, health care, education, personal services, and government).

Questions

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

Your table should look something like this. The key things to include are: Descriptive title, all variables (dependent variable and regressors) labeled, regression coefficients, standard errors (or t-stats), significance levels, number of observations, data source and sample, measure of fit (R-squared or adjusted R-squared).

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

$$hrwage_i = \beta_0 + \beta_1 union_i + u_i$$

   b. Interpret the coefficient on union status.

Union workers earn $5.32 more than non-union workers on average. Or union status is associated with $5.32 greater earnings per hour.

c. Interpret the constant.

Non-union workers earn $15.29 per hour on average.

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

$$H_0: \beta_1 = 0 \qquad H_A: \beta_1 \neq 0 \qquad t = \frac{\hat{B}_1}{SE(\hat{B}_1)} = \frac{5.324}{0.727} = 7.32 > 1.96$$

Since the test statistic is much greater than the critical value, we can reject the null hypotheses that union and non-union workers have the same earnings in the population.

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

Union status explains 5% of the variation in hourly wages.

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.

See table.

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

$$hrwage_i = \beta_0 + \beta_1 union_i + \beta_2 Age_i + u_i$$

Note: If you are describing a few different regression specifications, you sometimes may want to use different symbols for coefficients in different specifications since they represent different parameters. I am going to use B's throughout here, and its OK if you do too.

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

Once age is included in the regression as a control variable, the coefficient on union status decreases from $5.32 to $4.66. This implies that the simple bivariate regression suffered from omitted variable bias since age is correlated with both wages and union status. Omitting age from the regression overstates the causal effect of union status on wages – some of the wage advantage experienced by union workers reflects their greater age (and presumably labor market experience).

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

Age is positively correlated with both hourly wage and union status. The first of these can be seen in the positive coefficient on age in the regression. The second follows from the positive correlation between wage and age and the fact that the coefficient on union status decreases when age is included in the regression. The only way that the coefficient on union could decrease when age is added is if age is either (1) positively correlated with both wages and union status (the case here); or (2) negatively correlated with both wages 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.

See table.

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

Once employer size is included in the regression as a control variable, the coefficient on union status decreases further, from $4.66 to $4.14. This implies that the previous regressions suffered from omitted variable bias since employer size is correlated with both wages and union status. Omitting employer size from the regression overstates the causal effect of union status on wages - some of the wage advantage experienced by union workers reflects that they are more likely to work for larger employers.

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

Larger firms tend to pay more, conditional on union status and age, than smaller firms. The coefficient on employer size in regression (3) implies that a 1,000 person increase in employer size is associated with a $1.71 increase in hourly wages, holding age and union status constant. You also could have inferred this from the fact that the coefficient on union status went down when employer size was included and your knowledge that unionization rates are higher at larger firms.

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?

Managers are less likely to be unionized than non-managers in this sample. The fraction unionized is 9.8% for managers versus 18.2% for non-managers. This difference is statistically significant from zero at a 95% level of confidence.

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

Managers have higher earnings than non-managers in this sample. Hourly wages are $20.31 for managers versus $15.80 for non-managers. This difference is statistically significant from zero at a 95% level of confidence.

    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?

Failing to account for management status will cause us to understate the causal effect of unionization on hourly wages, since management status induces a negative omitted variable bias.

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

See table.

e. Interpret the coefficient on union status.

Union status is associated with a $4.30 higher hourly wage, controlling for age, employer size, and management 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?

Failing to control for sex does not affect our estimate of the coefficient on union status. In order for there to be omitted variable bias, sex must be correlated with both (1) unionization; and (2) wages. We can see from the table that men do indeed have higher wages then women, but that the rates of unionization are nearly identical between men and women. Therefore there is no omitted variable bias arising from failing to account for sex.

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

Failing to control for industry (specifically government sector) may cause us to overstate the coefficient on union status in the regressions. In order for there to be omitted variable bias, government sector status must be correlated with both (1) unionization; and (2) wages. We can see from the table that government workers have much higher rates of unionization and also much higher wages then non-government workers. This results in a positive omitted variable bias (e.g. the coefficient on union status will be overstated) if government sector is not accounted for.

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?

(i) Immigration status. Immigrants tend to have lower hourly earnings, possibly due to exploitation by employers. If service-sector unions explicitly target employers with many immigrants (inducing a positive correlation between unionization and immigrant status), then the bivariate regression will understate the true effect of unionization on wages. That is, the bivariate regression will suffer from negative omitted variable bias. Alternatively, if unionization rates are lower for immigrants, then the bivariate regression will suffer from positive omitted variable bias (i.e. be too big). You should rely on your institutional knowledge, expertise, or other data to determine whether unionization is positively or negatively correlated with immigration status is this particular instance.

(ii) Region. It is likely that unionization rates vary throughout the country. For instance, workers in the Northeast are probably more likely to be unionized than the South. It is also likely that wages are higher in the Northeast than in the South, regardless of unionization. If so, failing to account for region (for instance, by not including a dummy variable for living in the Northeast) will cause the coefficient on union status in the bivariate regression to suffer from positive omitted variable bias (i.e. be too big).

Do file

```
#delimit ;
clear all;
set mem 300m;
capture log close;
log using assignment3solns.log, text replace;

* ================================================================
* Public Policy 639 Assignment 3 - Solutions
* ================================================================;

use unions.dta, replace;

regress hrwage union, robust;
outreg2 using assignment4.txt, addstat(Adjusted R-squared, `e(r2_a)') replace;

regress hrwage union age, robust;
outreg2 using assignment4.txt, addstat(Adjusted R-squared, `e(r2_a)') append;

regress hrwage union age empsize, robust;
outreg2 using assignment4.txt, addstat(Adjusted R-squared, `e(r2_a)') append;

ttest hrwage, by(manager);
ttest union, by(manager);

regress hrwage union age empsize manager, robust;
outreg2 using assignment4.txt, addstat(Adjusted R-squared, `e(r2_a)') append;

log close;
```

Log file

```
. * ================================================================
> * Public Policy 639 Assignment 3 - Solutions
> * ================================================================;
. use unions.dta, replace;
```

```
. regress hrwage union, robust;

Linear regression                                       Number of obs =    1000
                                                        F(  1,   998) =   53.65
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0545
                                                        Root MSE      =  8.4188

------------------------------------------------------------------------------
             |               Robust
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   5.324295   .7268916     7.32   0.000     3.897883    6.750706
       _cons |    15.2858   .2901237    52.69   0.000     14.71648    15.85512
------------------------------------------------------------------------------

. outreg2 using assignment4.txt, addstat(Adjusted R-squared, `e(r2_a)') replace;
dir : seeout

. regress hrwage union age, robust;

Linear regression                                       Number of obs =    1000
                                                        F(  2,   997) =   48.95
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0930
                                                        Root MSE      =  8.2494

------------------------------------------------------------------------------
             |               Robust
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   4.662216   .7270729     6.41   0.000     3.235447    6.088985
         age |   .1218292   .0183556     6.64   0.000      .085809    .1578493
       _cons |   10.43756   .7389661    14.12   0.000     8.987454    11.88767
------------------------------------------------------------------------------

. outreg2 using assignment4.txt, addstat(Adjusted R-squared, `e(r2_a)') append;
dir : seeout

. regress hrwage union age empsize, robust;

Linear regression                                       Number of obs =    1000
                                                        F(  3,   996) =   38.98
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.1045
                                                        Root MSE      =  8.2015

------------------------------------------------------------------------------
             |               Robust
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   4.140148    .758009     5.46   0.000      2.65267    5.627626
         age |   .1231519   .0182505     6.75   0.000     .0873381    .1589657
     empsize |   .0017124   .0004772     3.59   0.000      .000776    .0026488
       _cons |   9.229249   .7867947    11.73   0.000     7.685283    10.77321
------------------------------------------------------------------------------

. outreg2 using assignment4.txt, addstat(Adjusted R-squared, `e(r2_a)') append;
dir : seeout

. ttest hrwage, by(manager);

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
```

```
---------+--------------------------------------------------------------------
       0 |     908    15.79662      .28316     8.53247     15.2409    16.35235
       1 |      92    20.31407    .9191141    8.815833    18.48836    22.13977
---------+--------------------------------------------------------------------
combined |    1000    16.21223    .2736502    8.653578    15.67523    16.74922
---------+--------------------------------------------------------------------
    diff |           -4.517444    .9364202               -6.355022   -2.679865
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -4.8242
Ho: diff = 0                                     degrees of freedom =      998

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

. ttest union, by(manager);

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     908    .1817181     .012804    .3858245    .1565891     .206847
       1 |      92    .0978261    .0311424    .2987072    .0359656    .1596866
---------+--------------------------------------------------------------------
combined |    1000        .174    .0119945    .3792992    .1504627    .1975373
---------+--------------------------------------------------------------------
    diff |             .083892    .0414355                .0025813    .1652026
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.0246
Ho: diff = 0                                     degrees of freedom =      998

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9784         Pr(|T| > |t|) = 0.0432          Pr(T > t) = 0.0216

. regress hrwage union age empsize manager, robust;

Linear regression                               Number of obs =    1000
                                                F(  4,   995) =   36.49
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1271
                                                Root MSE      =  8.1014

------------------------------------------------------------------------------
             |               Robust
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   4.429469   .7613091     5.82   0.000     2.935514    5.923425
         age |   .1182091   .0179572     6.58   0.000     .0829709    .1534474
     empsize |   .0015729    .000472     3.33   0.001     .0006467    .0024991
     manager |   4.522259   .9322119     4.85   0.000     2.692932    6.351586
       _cons |   9.065699    .778256    11.65   0.000     7.538488    10.59291
------------------------------------------------------------------------------


. outreg2 using assignment3.txt, addstat(Adjusted R-squared, `e(r2_a)') append;
dir : seeout

. log close;
-------------------------------------------------------------------------------
```