Problem Set 3

Francisco Brady

16 Feb 2024

### Service Sector Unionization

In your new job as policy analyst for the union SEIU (Service Employees International Union), you are helping to craft a memo that your boss will use to tout 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 unions.dta contains data from the 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 nine variables:

#### 1. Estimate the relationship between hourly wages and union status using bivariate regression. The results of this regression will be column 1 in your Table 1.

────────────────────────────────────────────────────────────────────────────────────  
 1 2 3 4   
   
────────────────────────────────────────────────────────────────────────────────────  
= 1 if in or cover~n 5.324\*\*\* 4.662\*\*\* 4.140\*\*\* 2.313\*\*\*  
 (0.73) (0.73) (0.76) (0.79)   
Age 0.122\*\*\* 0.123\*\*\* 0.088\*\*\*  
 (0.02) (0.02) (0.02)   
Firm size 0.002\*\*\* 0.002\*\*\*  
 (0.00) (0.00)   
ind\_retail -6.534\*\*\*  
 (0.93)   
ind\_personal -4.477\*\*\*  
 (1.47)   
ind\_health -1.562   
 (1.03)   
ind\_educ -2.327\*\*   
 (0.95)   
\_cons 15.286\*\*\* 10.438\*\*\* 9.229\*\*\* 14.711\*\*\*  
 (0.29) (0.74) (0.79) (1.21)   
────────────────────────────────────────────────────────────────────────────────────  
Adj. R-Squared 0.054 0.091 0.102 0.169   
R-Squared 0.054 0.093 0.104 0.175   
Observations 1000.000 1000.000 1000.000 1000.000   
────────────────────────────────────────────────────────────────────────────────────  
\* p<0.10, \*\* p<0.05, \*\*\* p<0.01

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

The population regression for column 1:

##### b. Interpret the coefficient on union status. [Hint: since union status is a binary variable, it does not make sense to say “A one-unit increase in union status”]

The coefficient on union status is 5.324, this is, on average, the difference of means in hourly wages of those in a union, compared to those not in a union. This indicates that on average, union status is associated with an increase in hourly wages of $5.32.

##### c. Interpret the constant.

The constant is 15.285. This is the average wage associated with non-union workers.

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

To test the null hypothesis we can calculate the test statistic and compare that to the critical value at the 95% level of confidence:

The absolute value of the test statistic exceeds the critical value of 1.96, so we can reject the null hypothesis that the true coefficient on union is equal to 0.

##### e. How much of the variation in wages can be explained by union status?

Using the output, the union status explains around ~5.45% of the variation of wages.

#### 2. After your causal inference training, 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. Given this possibility, explain how the omission of a control for age might bias your estimates of the effect of union status, using the OVB framework.

It is possible age is related to the outcome variable, since in general workers earn more as they get older and more experienced, this would be a positive correlation. Age may also be related to union membership, since union members tend to be older, this is also a positive correlation. Since we only included union membership in the initial model, the coefficient estimate will be too large, this is a positive bias in the simple model.

##### b. Now regress hourly wage on union status and age. The results of this regression will be column 2 in your Table 1. Interpret the coefficient on age. Please discuss the magnitude and the significance.

The coefficient on age is .121, and it is statistically significant. The interpretation is that an additional year of life is associated with an extra 12 cents in hourly wages on average, holding union membership constant.

##### c. Discuss the change in magnitude and statistical significance of the coefficient on union status from column 1 to 2. Was the change consistent with what you predicted above, in part (a)?

The coefficient on union status decreased from 5.324 in the bivariate model to 4.662 in the multivariate model. The coefficient in column 2 is statistically significant at the same level as in the simple model. This change was consistent with the prediction, confirming the positive bias in the simple model. The coefficient on union membership is lower in the model incorporating age, which has a positive and statistically significant coefficient, as expected.

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

The correlation between hourly wages and age is positive. The correlation between age and union status is also positive. We know this because of the positive coefficients in the output, which confirms the positive bias in the bivariate model.

#### 3. You recall from your 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. Now regress hourly wage on union status, age, and empsize, reporting your results in the table, in column 3 of Table 1.

##### a. What happens to the coefficient on union status once age and employer size are included in the regression? Explain in terms of omitted variable bias.

The coefficient on union status decreases again from 4.662 to 4.140. This implies that employer size is positively correlated with union status, leading the coefficient estimate in column 2 to be too large. This may be because organizers target larger firms to unionize. In column 3, since the coefficient on firm size is positive, this implies that firm size and wages are positively correlated. Larger firms may be associated with slightly higher wages because they need to compete in larger markets.

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

The coefficient on firm size is 0.002. It is statistically significant, but small, meaning that larger firms are associated on average with slightly higher wages, holding age and union status constant.

#### 4. Next, let’s consider an additional factor related to employment and wages that might bias the union coefficient. Please include indicators for the industry of employment (ind\_retail, ind\_personal, ind\_health, ind\_educ, and ind\_govt; use ind\_govt as the reference category). This will be column 4.

##### a. Which industry has the highest wages? How can you tell?

All of the other industries have lower wages than those associated with government. All of the signs on the coefficients for industry are negative, indicating that on average those wages are lower than the reference category of wages in government, holding union status, age, and firm size constant.

##### b. Discuss the changes in magnitude and statistical significance on the coefficient on union status between columns 3 and 4.

The coefficient on union status decreased again from 4.140 to 2.313. The estimate is statistically significant at the 99.9% level of confidence.

##### c. What are the consequences of not accounting for industry of employment when regressing hourly wages on union status? Explain in terms of omitted variable bias.

on average, different industries could be correlated either positively or negatively with hourly wages, the outcome variable. In addition, various industries have different union concentration, so would be correlated with union status. This would bias the estimates on hourly wages.

##### d. On average, how much would the model predict each of these people to earn:

###### i. A 40-year old unionized worker in a firm of 500 people, working in the health industry?

###### ii. A 55-year old non-unionized government worker in a firm of 1000?

##### e. Test whether the industry variables are jointly significant, controlling for the other variables included in column 3. Report the F-statistic and the p-value and explain in 1-2 sentences the conclusion of the test.

. quietly regress hrwage union age empsize ind\_retail ind\_educ ind\_health ind\_personal, rob  
> ust  
  
. test ind\_retail ind\_educ ind\_health ind\_personal  
  
 ( 1) ind\_retail = 0  
 ( 2) ind\_educ = 0  
 ( 3) ind\_health = 0  
 ( 4) ind\_personal = 0  
  
 F( 4, 992) = 22.52  
 Prob > F = 0.0000

The F-test for joint significance of the industry variables is 22.52, with a p-value of 0.0000. Based on this p-value, we can reject the null hypothesis that all of the coefficients of the industry variables are zero.

#### 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.”

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

|  | Men | Women |
| --- | --- | --- |
| Average Hourly Wage | 17.65 | 15.36 |
| Fraction Unionized | 0.175 | 0.174 |

##### a. What is the consequence of failing to control for sex in your regressions above? (i.e. is there omitted variable bias?)

Omitted variable bias exists when: 1) The variable (sex) is correlated with another explanatory variable, and 2) the variable is correlated with the outcome. In this case, the share of unionization is not significantly different across genders, so we can assume it is uncorrelated. So while gender may be correlated to the outcome variable (hourly wage), it is not correlated with union share, so will not lead to bias in coefficient estimates for union status.

##### b. Describe 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 omitted variable and hourly wages and between the omitted 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?

1. Marital Status: [Research suggests](https://www.stlouisfed.org/on-the-economy/2020/september/taking-closer-look-marital-status-earnings-gap) that marriage has a positive impact on hourly wages. [Other research](https://academic.oup.com/socpro/article-abstract/61/4/625/1631506?redirectedFrom=fulltext) has shown a positive association between union membership and marriage. This implies that the coefficient estimates in a bivariate regression of hourly wages on union status might have a positive bias.
2. Education: Increased years of education is positively associated with increased earnings, and [positively associated](https://docs.iza.org/dp2016.pdf) with union membership. This implies that in a bivariate regression on hourly wages and union status, the union status estimate may be positively biased.