**Assignment 3**

Due: Friday, February 16th, 2024 11:59pm

Please upload your assignment to Canvas under the Assignments tab.

This assignment aims to get you comfortable running multivariable regression and interpreting and presenting the output.

For all questions requiring you to use Stata or R, please include your output as part of your homework. In Stata, you can highlight the output, right click, and select “copy as picture” to easily copy and paste Stata output into Word.

**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. *hrwage* - Hourly wage last year (in $). This was estimated by dividing wage and salary income by the approximate number of hours worked last year (weeks worked X 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, but just ignore that for now. Treat it as a continuous variable.
5. And five, mutually exclusive variables indicating industry of employer:
   1. *Ind\_retail-* binary variable indicating working in retail
   2. *Ind\_personal-* binary variable indicating working in personal/service industry
   3. *Ind\_health-* binary variable indicating working in health care industry
   4. *Ind\_educ-* binary variable indicating working in education industry
   5. *Ind\_govt-* binary variable indicating working in government

**Presentation of results**

You should answer the questions below, **but also present the results from all your regressions in a single table (made in Excel, Word, or some other spreadsheet software) 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 (in real words, not abbreviated variable 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 assignment3.txt, replace;

regress y x1 x2, robust;

outreg2 using assignment3.txt, append;

**Questions**

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. 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]
   2. 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"]
   3. Interpret the constant.
   4. Test the null hypotheses that the coefficient on union status in the population regression function is zero.
   5. How much of the variation in wages can be explained by union status?
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.
   1. 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.
   2. 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.
   3. 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)?
   4. 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 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.**
   1. 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.
   2. Do larger firms pay more or less, conditional on union status and age?
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.**
   1. Which industry has the highest wages? How can you tell?
   2. Discuss the changes in magnitude and statistical significance on the coefficient on union status between columns 3 and 4.
   3. 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.
   4. 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?

* 1. 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.

1. 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 |

* 1. What is the consequence of failing to control for sex in your regressions above? (i.e. is there omitted variable bias?)
  2. 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?