Assignment 1

Jonatan Blank Hall & Anwei Luo
Lund University
School of Economics and Management
Economic History: Econometrics II

September 20, 2023

The team has decided to include the GitHub repository where our code is located in order to provide deeper insight on the work done.
The project can be found on this link:

1 Part 1: healthNHIS dataset

In this section we are working with data from the *NHIS Health Survey*. The dataset consists of a treatment variable, an outcome variable, and a number of socio-economic and demographic independent variables.

The research question to work with is: *What is the causal effect of health insurance on health?*

QUESTIONS

1. Load the dataset. Report output of differences in means for the dependent variable and one independent variable of your choice by treatment status. Describe and interpret the reported differences in means by treatment status.

   Solution

   The NHIS dataset contains seven columns, from which we can disregard *family id*, since it provides no insight and only works as a primary key of the base. The base is built by 5000 observations. Three of the columns are dummy variables: female, nonwhite and insurance. The *income* variable is a continuous variable. Both *health* and *education* are discrete variables.

   To answer the question, the independent variable chosen is the *years of education*. Figure 1 shows the output of tabulating the mean of education and health scores by insurance. People in the treatment group,

having health insurance, averagely have 2.84 years of education more than those in the control group. Their health condition is also averagely 0.31 points higher than the average score of the control group. This can be reasonable because people who have more education are more likely to buy health insurance - understanding the risk-benefit dynamics behind being insured. People with a higher degree of education may also pay more attention to keeping fit or following a healthy lifestyle.

```
. tabstat educ health, by(ins) stat(mean) format(%4.2f)

Summary statistics: Mean
Group variable: ins
```

ins	educ	health
0	**11.32**	**3.67**
1	**14.16**	**3.98**
Total	**13.70**	**3.93**

Figure 1: Difference in means: Health and Education by Insurance

2. Choose an independent variable and use a t-test to determine whether an observed difference in means between the treated and untreated is statistically significant. Report and comment on the differences in means and associated p-value.

Solution

The variable *education* is once again chosen to answer the question. Figure 2 displays the output of running a t-test between the means of education years between the treated and control groups.

```
. ttest educ, by(ins)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	**795**	**11.31572**	**.1248655**	**3.520675**	**11.07062**	**11.56083**
1	**4,205**	**14.15577**	**.0409714**	**2.65683**	**14.07544**	**14.23609**
Combined	**5,000**	**13.7042**	**.042388**	**2.997282**	**13.6211**	**13.7873**
diff		**-2.840044**	**.1087454**		**-3.053232**	**-2.626855**

```
    diff = mean(0) - mean(1)                                      t = -26.1165
H0: diff = 0                                     Degrees of freedom =     4998

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

Figure 2: T-test: Education years by Insurance Status

From the t-test, we can observe a significant difference of the average years of education between the treated and untreated group. Averagely, the years of education in treated group is 2.84 higher than the untreated´s group value. In addition, the associated p-value is close to zero. This implies that there may be a selection bias between treatment and control group. Their characteristics in the years of education is not balanced.

Out of curiosity, we decided to run an additional t-test, this time comparing the mean *income* of the treated and control groups. Figure 3 displays the outcome of the test. It is clear - and statistically significant - that there is a huge difference between the mean income of the two groups.

```
. ttest income, by(ins)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	795	44579.16	1316.202	37111.3	41995.52	47162.81
1	4,205	102291.3	844.8298	54783.81	100635	103947.6
Combined	5,000	93115.05	798.5218	56464.02	91549.6	94680.51
diff		-57712.12	2025.595		-61683.17	-53741.06

```
    diff = mean(0) - mean(1)                                  t = -28.4914
H0: diff = 0                                     Degrees of freedom =     4998

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

Figure 3: T-test: Income years by Insurance Status

3. Use regression to test whether the mean health score differs between treated and untreated. Report full regression output and comment on the results.

Solution

```
. reg health ins
```

Source	SS	df	MS	Number of obs	=	5,000
				F(1, 4998)	=	71.92
Model	63.3072192	1	63.3072192	Prob > F	=	0.0000
Residual	4399.24778	4,998	.880201637	R-squared	=	0.0142
				Adj R-squared	=	0.0140
Total	4462.555	4,999	.892689538	Root MSE	=	.93819

health	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ins	.3077124	.0362835	8.48	0.000	.2365808	.3788441
_cons	3.674214	.0332742	110.42	0.000	3.608982	3.739446

Figure 4: Regression: $Health \sim Insurance$

Figure 4 displays the output of fitting a bi-variate regression, explaining the health score by the insurance status. The regression test shows that the mean health score does differ between treated and untreated. In this regression, the mean health score of the treated is 0.3077124 better than the untreated. Health condition, when only considering insurance status in the model, is positively correlated with whether or not to have a health insurance. It is important to notice the value of the R-squared statistic; almost zero. The low value of the statistic indicates that the *insurance* variable explains almost nothing of the variance in health score, making it an incomplete variable to perform predictions.

In addition, Figure 5 displays the results of performing a t-test between the mean health score of the treated and control group. Even if the difference is expressed in decimals, there is statistical significance for stating that there is a difference in distribution for the two groups' health status.

```
. ttest health, by(ins)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	795	3.674214	.0361438	1.019101	3.603265	3.745162
1	4,205	3.981926	.01422	.9221126	3.954047	4.009805
Combined	5,000	3.933	.0133618	.9448225	3.906805	3.959195
diff		-.3077124	.0362835		-.3788441	-.2365808

```
    diff = mean(0) - mean(1)                                      t =  -8.4808
H0: diff = 0                     Degrees of freedom =     4998

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

Figure 5: T-test: Health Status by Insurance Status

4. Is there a way to satisfactorily answer the research question using the current data? Why, or why not?

   Solution

   We conclude that there is not a satisfactorily answer to the research question, *What is the causal effect of health insurance on health*, using the data provided in the data set. We justify this statement by arguing that the data set doesn't contain enough socioeconomic and demographic independent variables that should be considered when dealing with the omitted variable bias. Some variables that could be useful are: employment status, age, diet and sport activity.

2 Part 2: earnings dataset

In this section we work with a subsample of the *PSID (Panel Study of Income Dynamics)* dataset. It consists of a treatment variable, *college*, an outcome variable, *logwage*, and other socio-economic and demographic independent variables.

The research question to work with is: *What is the causal effect of college education on earnings?*

QUESTIONS

1. Use an appropriate method in Stata to look for evidence of selection into treatment based on gender and ethnicity. Report output and interpret.

   Solution

   Before answering the question, we make a small descriptive analysis of the base. Figure 6 shows the structure of the base; it consists of 7 variables and 2500 observations. We can immediately ignore the variable **id** since it only stores a primary key for the base.

```
. /* we describe the data base */
. describe

Contains data from earnings.dta
 Observations:         2,500                  PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990)
    Variables:             7                  21 Aug 2020 12:05
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
industry        float   %9.0g                 Industry
female          float   %9.0g                 Gender
nonwhite        float   %9.0g                 Ethnicity
logwage         float   %9.0g                 Log of wage
id              float   %9.0g                 Person identification variable
college         float   %9.0g                 College degree
occupation      float   %12.0g     occ2lbl    Occupation
```

Figure 6: Description of PSID base

By making a histogram of the **logwage** variable, we found out the distribution of the wages. As shown in Figure 7, the wages distribution seems to fit a normal distribution very well. This fact is very good since we do not have to work with any skewness or kurtosis in our response variable.

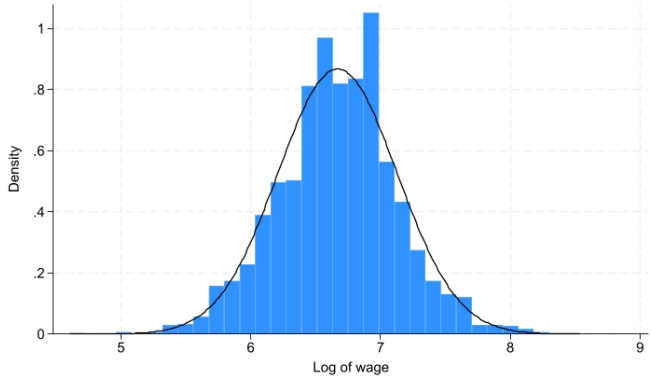

Figure 7: Wages Histogram

Since most of the variables in the base are *dummy* variables, it is possible to count the number of observations within a variable with a *success* value. By doing so, we can observe that all the variables are *balanced* between 0 and 1, except for two: **female** and **nonwhite**. Figure 8 shows this fact. This is an important fact to consider, since there is a big difference in the observed population of men and women, and whites and nonwhites.

-> tabulation of female

Gender	Freq.	Percent	Cum.
0	2,231	89.24	89.24
1	269	10.76	100.00
Total	2,500	100.00	

-> tabulation of nonwhite

Ethnicity	Freq.	Percent	Cum.
0	2,318	92.72	92.72
1	182	7.28	100.00
Total	2,500	100.00	

Figure 8: Tabulation of Female and Nonwhite variables

We can now answer the question regarding selection into treatment based on gender and ethnicity. One way of answering this is by performing a logit regression where the dependent variable is the treatment, *college*, and the independent variables are gender and ethnicity. Figure 9 shows the output, with coefficients, of the regression. If the coefficient is positive and statistically significant, it suggests that the variable has higher odds of being in the treatment group. If it's neg-

ative and significant, it have lower odds. On the other hand if the coefficient is not significant, it suggests that treatment assignment is likely not influenced by gender or ethnicity. Figure 9 shows that there is no statistical evidence to claim that the treatment is not influenced by gender, but there is strong evidence that it is influenced by race; non-whites have significant lower odds of being in the treatment group.

```
. /* check for selection into treatment based on gender and ethnicity */
. logit college female nonwhite

Iteration 0:  Log likelihood = -1689.1817
Iteration 1:  Log likelihood = -1670.9014
Iteration 2:  Log likelihood = -1670.8566
Iteration 3:  Log likelihood = -1670.8566

Logistic regression                                     Number of obs =  2,500
                                                        LR chi2(2)    =  36.65
                                                        Prob > chi2   = 0.0000
Log likelihood = -1670.8566                             Pseudo R2     = 0.0108

------------------------------------------------------------------------------
     college | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   -.006426   .1377155    -0.05   0.963    -.2763434    .2634913
    nonwhite |  -1.052862   .1919029    -5.49   0.000    -1.428985   -.6767397
       _cons |  -.3108023   .0438204    -7.09   0.000    -.3966887    -.224916
------------------------------------------------------------------------------
```

Figure 9: Logit Regression Female and Nonwhite variables on College

We can double check this fact by tabulating the variables *college* and *nonwhite*. While 1338 whites don't go to college and 980 do, it happens that 145 nonwhites don't go to college and 37 do. In the gender case, it happens that 1314 males don't assist college and 917 do, while 169 women don´t assist college and 100 do. Something worth noticing is that 33% of the females are nonwhite, while only 9% of the men are nonwhite.

2. Run a bivariate regression model of collage degree on earnings. Report output and interpret. Explain whether this analysis answers the research question or not and why.

   Solution

   Running a regression model in Stata is a very straight forward process. By using the command *reg wage college* we obtain the results displayed in Figure 10.

```
. reg logwage college
```

Source	SS	df	MS	Number of obs	=	2,500
				F(1, 2498)	=	270.09
Model	51.6264312	1	51.6264312	Prob > F	=	0.0000
Residual	477.47407	2,498	.191142542	R-squared	=	0.0976
				Adj R-squared	=	0.0972
Total	529.100501	2,499	.21172489	Root MSE	=	.4372

logwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
college	.2925331	.0177999	16.43	0.000	.257629	.3274372
_cons	6.550159	.0113529	576.96	0.000	6.527896	6.572421

Figure 10: Bivariate Regression College on Wages

The coefficient for **college** in the bivariate regression is .2925 and the *P-value* approves it´s significance. This means that on the bi-variate model, college education has a positive effect on earnings. One thing worth-noticing is the value of the *R-squared*: almost zero. Since this indicator represents the proportion of variance in the dependent variable that is explained by the independent variable we can conclude that the college variable has very little predictive power regarding the earnings variable in this specific model.

This bivariate regression gives us an association between having a college degree and earnings. However, a simple bivariate analysis like this does not establish causation. There are two fundamental reasons for this last statement:

(a) **Ommited Variable Bias**: Many confounding variables not included in the model.

(b) **Selection Bias**: Individuals who attend college might be systematically different in unobserved ways from those who don't.

3. Which of the available independent variables would you choose as control variables in a multivariate regression model, and why?

Solution

Two of the independent variables in the base have the potential of being good control variables: **female** and **nonwhite**. The justification for this statement follows:

- There are many economic reports on the gender wage disparity. In many sectors of the economy women have a harder time finding

jobs with good remuneration, mainly because of unjustified discrimination, compromises with family and hostile environments. Additionally, the first question of this section revealed that there is no selection into treatment disparity across gender, making the gender variable a good fit.

- Very similar to the point above, race has been proven to be a discrimination factor in many industries; white people tend to have higher wages. As this could be justified by many social-economical factors, we focus on the fact that it was proven in the first question that nonwhites had lower probabilities of going to college. This last fact, together with the assumption that nonwhites will suffer some sort of discrimination that will be reflected in lower wages, indicates that by not including race in the model, we would overestimate the effect of college education over wage.

The result of adjusting a regression on wage considering college, gender and race is displayed in Figure 11. We can see that we were right on the assumption that not considering race would overestimate college influence on wages. We also see that the R-squared value increased. Finally, we show that both female and nonwhite characteristics have a negative effect on wage.

```
. reg logwage college female nonwhite
```

Source	SS	df	MS	Number of obs	=	2,500
				F(3, 2496)	=	214.70
Model	108.529427	3	36.1764758	Prob > F	=	0.0000
Residual	420.571074	2,496	.168498026	R-squared	=	0.2051
				Adj R-squared	=	0.2042
Total	529.100501	2,499	.21172489	Root MSE	=	.41049

logwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
college	.2767829	.0168261	16.45	0.000	.2437884	.3097773
female	-.4487711	.027073	-16.58	0.000	-.501859	-.3956832
nonwhite	-.1422965	.0325	-4.38	0.000	-.2060262	-.0785668
_cons	6.615213	.0113679	581.92	0.000	6.592921	6.637504

Figure 11: Multivariate Regression on Wages considering college, gender and race

4. Run your chosen model and store the estimates. Then, add one or more variables that you consider a "bad control", run and store the model. Report the models side by side and compare. Explain what

changed and why.

Solution

Lets consider three models:

(a) Model 1: Bivariate model $Wage \sim College$

(b) Model 2: Good-Variables model $Wage \sim College + Female + NonWhite$

(c) Model 3: Bad-Variables model $Wage \sim College + Female + NonWhite + Occupation$

```
. esttab m1 m2 m3
```

	(1) logwage	(2) logwage	(3) logwage
college	0.293***	0.277***	0.153***
	(16.43)	(16.45)	(7.45)
female		-0.449***	-0.482***
		(-16.58)	(-18.02)
nonwhite		-0.142***	-0.133***
		(-4.38)	(-4.18)
occupation			-0.205***
			(-10.17)
_cons	6.550***	6.615***	6.775***
	(576.96)	(581.92)	(351.16)
N	2500	2500	2500

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 12: Comparison of the three models

We fit the three models and display the results side by side. Figure 12 shows this. It´s possible to observe that in all three models the variables considered are significant. Going from model 1 to model 2, we observe that the *college* variable decreases, but only a small bit. We argue that the racial variable has a negative effect on both wages and college, overestimating the college variable when absent in model 1. Then, going from model 2 to model 3 there is a big change in the college coefficient. Occupation has a negative coefficient since blue-collar jobs, represented by 1, usually pay less than white-collar jobs,

represented by 0. Therefore occupation has a negative effect on wage. There is also a negative effect between college and occupation, since people that go to college tend to have a white-collar job. Therefore, when omitting the occupation variable, the collage coefficient was being overestimated.

We argue that the third model contains a bad control variable, *occupation*, since occupation itself is very dependent on college. Wage, the dependent variable, is directly dependent on occupation, and one could argue that it would be possible to predict occupation given wage. Therefore it is a bad control variable and should not be included in the predictive model since it will *absorb* some of the information explained by the college variable.

5. Considering the second model. It is reasonable to think that an individual's ability would affect the level of income. If omitting the variable *ability* in the regression, will that bias the estimate of college? If yes, in what direction?

   Solution

   If we were to consider that the variable *ability* is available in the dataset, in order for it to bias the estimate of college we would have to make some assumptions. First of all, we would have to assume that the ability of a person has an effect on the income of that person. We could further assume that it has a positive effect on the income. Additionally, we have to assume that ability has an effect on the decision of the individual in chasing a college education. We could assume that it does, and that person with ability are more prompt to study college.

   If we assume the points mentioned in the paragraph above, then we would say that not including the *ability* variable will bias the estimate of college. Since ability has a positive effect on both college and wage outcomes, then omitting the ability variable would result in overestimating the effect of college on wage.

6. Explain shortly, verbally and with equations, why random assignment to treatment solves the selection problem.

   Solution

   The selection problem arises when individuals in a study are not randomly assigned to treatment and control groups. This can lead to biased estimates of the treatment effect due to confounding variables, which might be related to both the treatment and the outcome.

Random assignment ensures that both observed and unobserved characteristics are, on average, distributed between the treatment and control groups.

Let Y_{i1} be the potential outcome for individual i if treated, and Y_{i0} be the potential outcome for individual i if not treated. The observed outcome Y_i for each individual can be represented as:

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0} \tag{1}$$

where D_i is an indicator for receiving the treatment (1 if treated, 0 if not).

Without random assignment, $E[Y_{i1}|D_i = 1] - E[Y_{01}|D_i = 0]$ might not be equal to $E[Y_{i1} - Y_{i0}]$, the *average treatment effect*. The reason behind this is that the groups $D_i = 0$ and $D_i = 1$ might differ in ways other than just the treatment. With random assignment, we assure that:

- $E[Y_{i1}|D_i = 1] \simeq E[Y_{i1}|D_i = 0]$
- $E[Y_{i0}|D_i = 1] \simeq E[Y_{i0}|D_i = 0]$

This makes the treatment and control groups comparable, so the difference in observed outcomes between them can be attributed to the treatment effect:

$$E[Y_{i1}|D_i = 1] - E[Y_{01}|D_i = 0] = E[Y_{i1} - Y_{i0}] \tag{2}$$