Assignment 2

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

October 11, 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 at:

1 Part 1: RAND Health Experiment

Research Question: 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.

Answer

First, the *healthRAND.dta* database using the *use* command in STATA. The independent variable chosen for this question is *Income*. By running the command *tabstat health income, by(ins) stat(mean)* we get the following result 1:

Summary statistics: Mean
Group variable: ins

ins	health	income
0	3.828155	32037.75
1	3.785547	31086.03
Total	3.803668	31265.32

Figure 1: Differences in means: Health and Income by Insurance Status

The figure above shows that, between the treated and untreated groups there is no real difference in the health and income means. The results of this table suggest the case that health insurance does not have a significant causal effect on health, or income, in the sample. This would suggest that, on average, having insurance doesn't lead to better health outcomes, or higher income, for the participants in the RAND experiment.

Although, there might be other confounding variables that are influencing both the treatment assignment and the outcome. If these aren't accounted for, a simple mean difference could be misleading. In the case of the RAND experiment, participants were randomized, which helps mitigate this concern, but it's still crucial to check the balance of covariates across treatment groups.

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

Answer

For this question, *educ* - years of education - variable is chosen. To run the t-test, the command *ttest educ, by(ins)* is executed. The results of the test are displayed in the following figure 2:

```
. ttest educ, by(ins)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	552	12.15399	.122436	2.876594	11.91349	12.39448
1	2,387	11.94742	.0613496	2.997355	11.82712	12.06773
Combined	2,939	11.98622	.0548888	2.975661	11.8786	12.09384
diff		.206562	.1405082		-.0689426	.4820665

```
    diff = mean(0) - mean(1)                                  t =   1.4701
H0: diff = 0                     Degrees of freedom =     2937

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9292         Pr(|T| > |t|) = 0.1416          Pr(T > t) = 0.0708
```

Figure 2: t-test: Difference years of education by insurance status

From the figure, we can extract the following information:

(a) Mean educ for uninsured: 12.15399

(b) Mean educ for insured: 11.94742

(c) Difference: .206562

(d) t-statistic: 1.4701

(e) p-value: 0.1416

The interpretation of these results is that, if there was truly no difference in the population, observing a difference as large as .206562 years would happen 14.16% of the time. Given these results, we can conclude there is no statistical evidence to say that there is a difference between the years of education between the treated and untreated populations.

3. Use regression to test whether the mean health score differs between treated and untreated. Report full regression output and comment on the results. Explain whether the analysis answers the research question satisfactorily. Elaborate on whether the assignment to treatment has been randomized or not and how that affects answering the research question.

Answer

First we regress the health score on the insurance status by running the command: *regress health ins*.

```
. regress health ins
```

Source	SS	df	MS	Number of obs	=	3,708
				F(1, 3706)	=	3.78
Model	1.64535031	1	1.64535031	Prob > F	=	0.0520
Residual	1613.42477	3,706	.435354767	R-squared	=	0.0010
				Adj R-squared	=	0.0007
Total	1615.07012	3,707	.435681176	Root MSE	=	.65981

health	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ins	-.042608	.0219171	-1.94	0.052	-.0855789	.0003628
_cons	3.828155	.0166152	230.40	0.000	3.795579	3.860731

Figure 3: Linear regression of Health on Insurance Status

Figure 3 shows the results of the regression. The regression results indicate that, on average, individuals with insurance have a health score that is .042 points *lower* than those without insurance. This difference is statistically significant with a p-value of 0.052 - we could be very strict and reject the significance of the variable, but it is really close to the threshold. The constant term suggests that the average health score for the uninsured group is 3.828115.

Note: One additional approach that was tried involved clustering for family id. The coefficients and confidence intervals in both regression models were identical; the

difference could be seen in the significance of the insurance variable in the clustered model, which was not significant with a p-value of 0.225.

Given that this is based on the RAND Health Insurance Experiment, a randomized controlled trial, the observed difference in health scores can be confidently attributed to the effect of having insurance. The random assignment ensures that other potential confounders are likely balanced across groups. However, the R-squared value is quite low (0.0010), suggesting that the treatment variable only explains a small fraction of the variance in health scores. There might be other factors that play a significant role in determining an individual's health score.

2 Part 2: Angrist and Krueger 91

Research Question: What is the causal effect of schooling on earnings?

QUESTIONS

1. Explain in words why the instrumental variable method solves the confounding problem in this case, and why the provided instrument is appropriate to use.

   Answer

   When assessing the causal effect of schooling on earnings, there's an inherent problem of endogeneity. Factors like ability, motivation, or family background might influence both an individual's decision to pursue more education and their potential earnings. If we simply regress earnings on schooling, the coefficient on schooling could be biased, capturing not just the effect of schooling but also these other unobserved factors.

   The Instrumental Variable method is designed to solve this problem. It requires a variable that's correlated with the potentially endogenous independent variable - schooling - but uncorrelated with the error term in the outcome equation - which includes influences on wages other than schooling.

   Angrist and Krueger proposed using the quarter of birth as an instrument for schooling. The decision is based on schooling laws in the U.S. which dictate that children must remain in school at least until a certain age. Those born later in the calendar year typically start school at an older age and therefor are older when they become eligible to leave school.

   Quarter of birth serves as a good instrument because it follows the principles of relevance, exogeneity and the exclusion restriction. It follows relevance because quarter of birth is correlated with schooling years, as explained above. It follows exogeneity because quarter of birth is random and it's not likely to be correlated with other unobserved variables. Finnaly, it follows the exclusion restriction because the instrument should not be directly related to potential earnings, except for its possible effect on schooling.

2. Test the assumptions of the IV analysis in Stata, report output, and interpret. If a formal test is not appropriate, explain in words whether you think the assumption holds or not, and why.

   Answer

   To answer this question, we will test for relevance and exogeneity assumption. To test for relevance, we can run a regression of schooling on the quarter of birth dummy variable to see if its statistically significant. We do so by running the command: *regress schooling qobi.*

```
. regress schooling qobi
```

Source	SS	df	MS	Number of obs	=	200,000
				F(1, 199998)	=	30.24
Model	325.409432	1	325.409432	Prob > F	=	0.0000
Residual	2152337.23	199,998	10.7617938	R-squared	=	0.0002
				Adj R-squared	=	0.0001
Total	2152662.64	199,999	10.763367	Root MSE	=	3.2805

schooling	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
qobi	.0937562	.0170501	5.50	0.000	.0603384	.127174
_cons	12.75254	.0084435	1510.34	0.000	12.73599	12.76909

Figure 4: Regression: Education on quarter-of-birth dummy variable

Figure 4 shows the resulting regression. Since the p-value on the *qobi* coefficient is very close to zero, we can say that it is statistically significant that individuals born in the 4th quarter have different schooling years, supporting the relevance assumption.

To test for the exogeneity assumption we must prove that there's no direct correlation between the instrument and the error term in the wage equation. We can do so by regressing the dependent variable on the instrument - reduced form regression. This is achieved by running the command *regress ln_ wage qobi*.

```
. regress ln_wage qobi
```

Source	SS	df	MS	Number of obs	=	200,000
				F(1, 199998)	=	5.10
Model	2.35749145	1	2.35749145	Prob > F	=	0.0239
Residual	92362.8495	199,998	.461818866	R-squared	=	0.0000
				Adj R-squared	=	0.0000
Total	92365.207	199,999	.461828344	Root MSE	=	.67957

ln_wage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
qobi	.0079801	.003532	2.26	0.024	.0010575	.0149028
_cons	5.898721	.0017491	3372.42	0.000	5.895293	5.902149

Figure 5: Reduced form regression

Figure 5 shows and interesting result; quarter of birth is significant; this indicates that being born in the 4th quarter increases ln_wage. This doesn't necessarily violate the exogeneity assumption; it could simply capture the total effect of qobi

on ln_wage through schooling. Additionally, the R-squared statistic is practically zero, meaning that qobi, although significant, explains nothing of the dependent variable.

3. Run a Two-Stage Least Square regression of earnings on schooling. Report output from first and second stages and interpret.

Answer

We use the built-in function in STATA to run a Two-Stage Least Square regression.

```
. ivregress 2sls ln_wage (schooling = qobi), first

First-stage regressions
```

Number of obs	200,000
F(1, 199998)	30.24
Prob > F	0.0000
R-squared	0.0002
Adj R-squared	0.0001
Root MSE	3.2805

schooling	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
qobi	.0937562	.0170501	5.50	0.000	.0603384	.127174
_cons	12.75254	.0084435	1510.34	0.000	12.73599	12.76909

Instrumental variables 2SLS regression

Number of obs	200,000
Wald chi2(1)	5.75
Prob > chi2	0.0165
R-squared	0.1126
Root MSE	.64016

ln_wage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
schooling	.0851158	.0354875	2.40	0.016	.0155616	.1546699
_cons	4.813279	.4533734	10.62	0.000	3.924684	5.701875

Endogenous: **schooling**
Exogenous: **qobi**

Figure 6: TSLSR using quarter of birth as instrument

Figure 10 shows the result of running the TSLR. The results are divided in two tables. The first one, *First-stage regressions*, shows the effect of quarter of year birth on schooling. We can see that quarter of birth is very significant and has a coefficient of .0937562. The second table, *Instrumental variables 2SLS regression*, displays the effect of *schooling* on *log wage* after accounting for endogeneity using the instrument. The coefficient is statistically significant suggesting that schooling has a causal effect on earnings. The coefficient is .085, indicating there is an 8.5% increase in wage for an additional year of schooling.

3 Part 3: Minimum Legal Drinking Age

Research Question: What is the causal effect of alcohol consumption on young adult mortality?

QUESTIONS

1. In this RD setting, define the following:
   (a) Causal variable of interest
   (b) Running variable
   (c) At least one potential confounding factor

   Answer

   (a) Causal variable of interest: a dummy variable indicating whether an individual is above MLDA or not, it is a function of age.
   (b) Running variable:age of the individual (which is agecell in the dataset)
   (c) deaths from external causes might also change with the running variable.

2. Describe the confounding problem in this setting, and how this specific design can potentially create a causal estimate to answer the research question.

   Answer

   Various factors could influence both alcohol consumption and mortality. For example economic status, health behavior and social environment. In this RD setup, confounding factors that typically plague observational studies of alcohol consumption and mortality are minimized by focusing on the age-based discontinuity at 21. While individuals just below and just above 21 are likely similar in many respects, their alcohol consumption patterns might differ notably due to the legal restriction.

   Some confounding problems that might happen in this experiment are:

   (a) Non-compliance with the MLDA law
   (b) Other age-related changes at 21

   The age of 21 is not only significant for legal alcohol consumption. Other privileges or responsibilities that kick in at this age might also affect mortality. Additionally, not everyone follows the MLDA. Many below 21 consume alcohol, and not all those who turn 21 start drinking. This non-compliance can confuse the actual causal effect of alcohol on mortality.

3. Present two scatterplots, showing the running variable on the x-axis, and:
   (a) Motor vehicle accident mortality on the y-axis
   (b) Homicide mortality on the y-axis

Comment briefly on the graphs and what the reason behind differences between the two graphs may be.

Answer

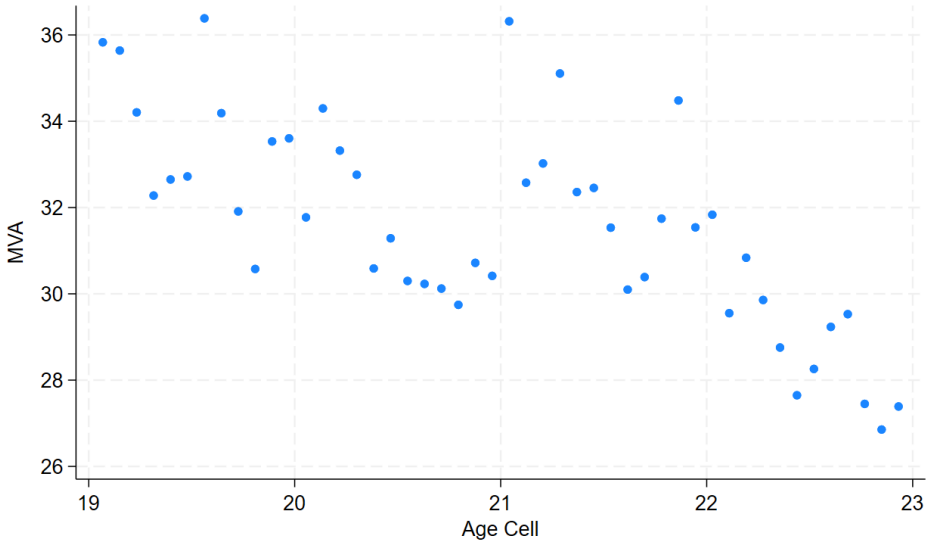

Figure 7: scatterplot with mva

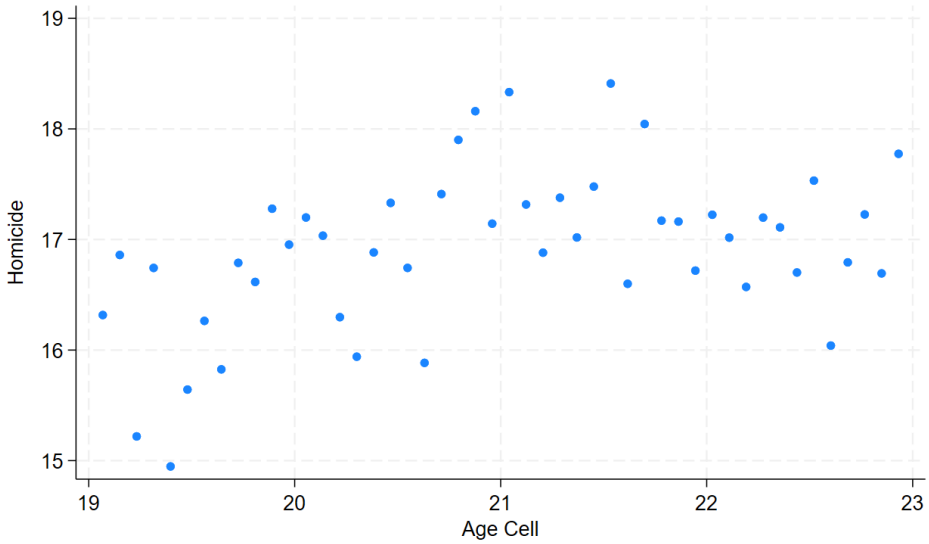

Figure 8: scatterplot with homicide

The scatterplot with motor vehicle accident mortality on the y axis shows different patterns of plots located at the right of 21 and the left of 2, while the scatterplot with homicide mortality on the y axis doesn't show such kind of division. It looks like there is a jump of the motor vehicle accident mortality at the cutoff point of age 21. The reason behind this may be the reach of the minimum legal drinking age

makes young people drunk and resulting in increased motor vehicle accident which contributes to the mortality.

4. Generate a treatment dummy that switches in accordance with the rule that we are exploiting. Run two RD regressions with the running variable and the RD dummy as independent variables, and:

   (a) Motor vehicle accident mortality as outcome

   (b) Homicide mortality as outcome

   Report the regression results and comment on the magnitude and significance of the RD dummy coefficient in the two models.

Answer

```
. reg mva over21 centered_age
```

Source	SS	df	MS	Number of obs	=	48
				F(2, 45)	=	53.14
Model	187.819794	2	93.909897	Prob > F	=	0.0000
Residual	79.5215648	45	1.76714588	R-squared	=	0.7025
				Adj R-squared	=	0.6893
Total	267.341359	47	5.68811402	Root MSE	=	1.3293

mva	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
over21	4.534033	.7679953	5.90	0.000	2.987211	6.080855
centered_age	-3.148829	.3372437	-9.34	0.000	-3.828073	-2.469585
_cons	29.35597	.4292665	68.39	0.000	28.49138	30.22055

Figure 9: Model 1

```
.
. reg homicide over21 centered_age
```

Source	SS	df	MS	Number of obs	=	48
				F(2, 45)	=	4.65
Model	4.29084615	2	2.14542307	Prob > F	=	0.0146
Residual	20.7542314	45	.461205142	R-squared	=	0.1713
				Adj R-squared	=	0.1345
Total	25.0450775	47	.53287399	Root MSE	=	.67912

homicide	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
over21	.1043579	.3923462	0.27	0.791	-.6858679	.8945837
centered_age	.2218874	.1722879	1.29	0.204	-.1251181	.568893
_cons	16.85989	.2192996	76.88	0.000	16.4182	17.30158

Figure 10: Model 2

In the first model, the coefficient of treatment variable is significant with the magnitude of 4.54033, which is not low. This implies that there is a causal relationship

between alcohol consumption on young adult mortality through affecting motor vehicle accident mortality. In the second model, the coefficient is neither high nor significant. This implies that little evidence shows that homicide mortality can be influenced by the consumption of alcohol.