Reduced Form Coding Assignment - ECON 8250

Tate Mason

1. Start by coming up with a research question where you might use this research design. This does not need to be too creative, I just want an example. However, I would prefer if it wasn't the one I used in class and was vaguely related to health. Intuitively and in words, describe what the endogeneity concern might be with a question like this.
2. Tell me basics about the dataset you simulate. What is your unit of observation? Then, in words, describe what variables you are assuming constitute the "true model" and which variables you are assuming you can observe and cannot observe. Describe any important correlations between variables. Also, describe other variables, like policies (for diff-in-diff), thresholds (for RD), or instruments (for IV). Give me an equation for your "true model" and introduce all the letters you are using. I want an equation, written like they would be written in a paper, not STATA code. Then separately, tell me what your "true" coefficients are (i.e. = 2).
3. In words and equations, describe the regressions you are running. Both the regressions that have an endogeneity problem and the ones which you "fix."
4. Produce a table of summary statistics with the mean, standard deviation, number of observations, min and max of each variable you use. This is both regressors and outcome variables. You do not need to show me summary statistics for fixed effects.
5. Produce regression results in nice table layout, with intuitive variable labels (i.e. not stata variable names), and not too many variables (i.e. don't display fixed effects). Describe the regression results for each of your regressions in words.

Fixed Effects Model

1. Research Question

How does insurance premium rise with age and risk preference?

2.

```
set.seed(0219)
n <- 1000
id <- 1:n
age <- sample(18:70, n, replace = TRUE)
risk_pref <- rnorm(n, mean = 0, sd = 1)
insprem <- 200 + 5 * age + 20 * risk_pref
+ rnorm(n, mean = 0, sd = 10)
data <- data.frame(id, age, risk_pref, insprem)
```

Each agent is a unit, with n=1000. The true model is:

$$InsPrem_i = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot RiskPref_i + \epsilon_i,$$

where $InsPrem_i$ is the insurance premium for agent i, Age_i is the age of agent i, $RiskPref_i$ is the risk preference of agent i, and ϵ_i is the error term. The true coefficients are: $\beta_0 = 200$, $\beta_1 = 5$, $\beta_2 = 20$, $\beta_3 = 10$.

3. Regressions

The regression with endogeneity problem is:

$$InsPrem_i = \alpha_0 + \alpha_1 \cdot Age_i + u_i,$$

where u_i is the error term which includes the risk preference parameter. The regression that "fixes" the endogeneity problem is:

$$InsPrem_i = \gamma_0 + \gamma_1 \cdot Age_i + \gamma_2 \cdot RiskPref_i + v_i,$$

where v_i is the error term.

4. Summary statistics

```
library(psych)
describe(data)
```

```
          vars    n   mean     sd median trimmed    mad   min     max  range
id           1 1000 500.50 288.82 500.50  500.50 370.65  1.00 1000.00 999.00
age          2 1000  44.03  15.78  44.00   43.98  20.76 18.00   70.00  52.00
```

```
risk_pref    3 1000   0.01   1.00   0.01    0.01   0.99  -3.08    3.04   6.12
insprem      4 1000 420.34  81.25 423.39  420.37 105.08 247.82  591.71 343.89
           skew kurtosis   se
id         0.00    -1.20 9.13
age        0.02    -1.25 0.50
risk_pref -0.05    -0.03 0.03
insprem   -0.01    -1.12 2.57
```

5. Regression results

```r
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
library(sandwich)
model1 <- lm(insprem ~ age, data = data)
model2 <- lm(insprem ~ age + risk_pref, data = data)
summary(model1)
```

```

Call:
lm(formula = insprem ~ age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-61.888 -13.420  -0.143  13.437  60.780

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 200.67257    1.87784   106.9   <2e-16 ***
age           4.98957    0.04015   124.3   <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.03 on 998 degrees of freedom
Multiple R-squared:  0.9393,	Adjusted R-squared:  0.9392 
F-statistic: 1.544e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
summary(model2)
```

```
Warning in summary.lm(model2): essentially perfect fit: summary may be
unreliable


Call:
lm(formula = insprem ~ age + risk_pref, data = data)

Residuals:
       Min         1Q     Median         3Q        Max 
-1.154e-12 -1.651e-14 -1.220e-15  1.350e-14  2.388e-12 

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)    
(Intercept) 2.000e+02  8.146e-15 2.455e+16   <2e-16 ***
age         5.000e+00  1.742e-16 2.871e+16   <2e-16 ***
risk_pref   2.000e+01  2.746e-15 7.283e+15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.688e-14 on 997 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 4.369e+32 on 2 and 997 DF,  p-value: < 2.2e-16
```

IV Model

1. Research Question

How does exercise frequency affect mental health, using weather as an instrument?

2. Dataset

```
set.seed(0219)
n <- 1000
id <- 1:n
exercise_freq <- rnorm(n, mean = 3, sd = 1)
weather <- rnorm(n, mean = 0, sd = 1)
mental_health <- 50 + 2 * exercise_freq + 5 * weather + rnorm(n, mean = 0, sd = 5)
data_iv <- data.frame(id, exercise_freq, weather, mental_health)
```

Each agent is a unit, with n=1000. The true model is:

$$MentalHealth_i = \beta_0 + \beta_1 \cdot ExerciseFreq_i + \beta_2 \cdot Weather_i + \epsilon_i,$$

where $MentalHealth_i$ is the mental health score for agent i, $ExerciseFreq_i$ is the exercise frequency of agent i, $Weather_i$ is the weather condition for agent i, and ϵ_i is the error term. The true coefficients are: $\beta_0 = 50$, $\beta_1 = 2$, $\beta_2 = 5$.

3. Regressions

The regression with endogeneity problem is:

$$MentalHealth_i = \alpha_0 + \alpha_1 \cdot ExerciseFreq_i + u_i,$$

where u_i is the error term which includes the weather parameter. The regression that "fixes" the endogeneity problem using IV is: First stage:

$$ExerciseFreq_i = \pi_0 + \pi_1 \cdot Weather_i + w_i,$$

Second stage:

$$MentalHealth_i = \gamma_0 + \gamma_1 \cdot Exerc\hat{i}seFreq_i + v_i,$$

where $Exerc\hat{i}seFreq_i$ is the predicted exercise frequency from the first stage, and v_i is the error term.

4. Summary statistics

```r
describe(data_iv)
```

```
              vars    n   mean     sd median trimmed    mad   min     max
id               1 1000 500.50 288.82 500.50  500.50 370.65  1.00 1000.00
exercise_freq    2 1000   2.92   1.00   2.92    2.91   0.98  0.18    6.17
weather          3 1000   0.04   1.03   0.04    0.04   1.04 -3.52    4.10
mental_health    4 1000  55.93   7.78  56.12   55.94   7.82 30.29   80.26
               range  skew kurtosis   se
id            999.00  0.00    -1.20 9.13
exercise_freq   5.99  0.14    -0.06 0.03
weather         7.62  0.07     0.14 0.03
mental_health  49.97 -0.02     0.02 0.25
```

5. Regression results

```r
library(AER)
```

```
Loading required package: car

Loading required package: carData


Attaching package: 'car'

The following object is masked from 'package:psych':

    logit

Loading required package: survival
```

```r
model1_iv <- lm(mental_health ~ exercise_freq, data = data_iv)
model2_iv <- ivreg(mental_health ~ exercise_freq | weather, data = data_iv)
summary(model1_iv)
```

```
Call:
lm(formula = mental_health ~ exercise_freq, data = data_iv)

Residuals:
     Min       1Q   Median       3Q      Max 
-24.6977  -4.9843  -0.1016   4.9495  23.9425 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    48.4043     0.7170   67.51   <2e-16 ***
exercise_freq   2.5732     0.2321   11.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.342 on 998 degrees of freedom
Multiple R-squared:  0.1097,	Adjusted R-squared:  0.1088 
F-statistic: 122.9 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
summary(model2_iv)
```

```
Call:
ivreg(formula = mental_health ~ exercise_freq | weather, data = data_iv)

Residuals:
     Min       1Q   Median       3Q      Max 
-254.751  -49.604    1.171   54.552  222.524 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)    -183.80     111.53  -1.648   0.0997 .
exercise_freq    82.01      38.15   2.150   0.0318 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79.89 on 998 degrees of freedom
Multiple R-Squared: -104.4,	Adjusted R-squared: -104.5 
Wald test: 4.622 on 1 and 998 DF,  p-value: 0.03181 
```

As can be seen from the regression results, the first model without the instrument shows a biased estimate of the effect of exercise frequency on mental health due to omitted variable bias. The second model using weather as an instrument provides a more accurate estimate of the causal effect of exercise frequency on mental health.