

Reduced Form Coding Assignment - ECON 8250

Tate Mason

Fixed Effects Model

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. $\beta = 2$).
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.

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)
unobserved_health <- rnorm(n, mean = 0, sd = 1)
insprem <- 200 + 5 * age + 20 * risk_pref + 10 * unobserved_health
+ rnorm(n, mean = 0, sd = 10)
data <- data.frame(id, age, risk_pref, insprem, unobserved_health)
```

Each agent is a unit, with $n=1000$. The true model is: $InsPrem_i = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot RiskPref_i + \beta_3 \cdot UnobsHealth_i + \epsilon_i$, 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 , $UnobsHealth_i$ is the unobserved health status 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: $InsPrem_i = \alpha_0 + \alpha_1 \cdot Age_i + \alpha_2 \cdot RiskPref_i + u_i$, where u_i is the error term which includes the unobserved health status. The regression that “fixes” the endogeneity problem is: $InsPrem_i = \gamma_0 + \gamma_1 \cdot Age_i + \gamma_2 \cdot RiskPref_i + \gamma_3 \cdot UnobsHealth_i + v_i$, where v_i is the error term.

4. Summary statistics

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

	vars	n	mean	sd	median	trimmed	mad	min	max
id	1	1000	500.50	288.82	500.50	500.50	370.65	1.00	1000.00
age	2	1000	44.03	15.78	44.00	43.98	20.76	18.00	70.00
risk_pref	3	1000	0.01	1.00	0.01	0.01	0.99	-3.08	3.04
insprem	4	1000	420.72	81.85	422.80	420.69	105.79	251.69	608.04
unobserved_health	5	1000	0.04	1.01	0.04	0.04	1.03	-3.34	3.07
	range	skew	kurtosis	se					
id	999.00	0.00	-1.20	9.13					

age	52.00	0.02	-1.25	0.50
risk_pref	6.12	-0.05	-0.03	0.03
insprem	356.36	-0.01	-1.11	2.59
unobserved_health	6.41	-0.02	0.02	0.03

5. Regression results

```
library(lmtest)
```

Loading required package: zoo

Attaching package: 'zoo'

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

as.Date, as.Date.numeric

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

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-33.602	-6.788	0.022	6.983	30.299

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	200.27808	0.95142	210.50	<2e-16 ***
age	5.00220	0.02034	245.89	<2e-16 ***
risk_pref	19.71334	0.32074	61.46	<2e-16 ***

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

Residual standard error: 10.15 on 997 degrees of freedom
 Multiple R-squared: 0.9847, Adjusted R-squared: 0.9846
 F-statistic: 3.2e+04 on 2 and 997 DF, p-value: < 2.2e-16

```
summary(model2)
```

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

Call:

```
lm(formula = insprem ~ age + risk_pref + unobserved_health, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.266e-12	-3.040e-14	-8.600e-15	1.400e-14	9.731e-12

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.000e+02	2.935e-14	6.813e+15	<2e-16 ***
age	5.000e+00	6.276e-16	7.966e+15	<2e-16 ***
risk_pref	2.000e+01	9.899e-15	2.020e+15	<2e-16 ***
unobserved_health	1.000e+01	9.771e-15	1.023e+15	<2e-16 ***

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

Residual standard error: 3.131e-13 on 996 degrees of freedom
 Multiple R-squared: 1, Adjusted R-squared: 1
 F-statistic: 2.276e+31 on 3 and 996 DF, p-value: < 2.2e-16