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 activity and risk preference?

2.

```r
set.seed(0219)
n <- 1000
id <- rep(1:n, each = 50)
time <- rep(1:50, times = n)
activity_level <- rnorm(n * 50, mean = 3, sd = 1) # hours of activity per week
risk_pref <- rbinom(n * 50, 1, 0.5) # 0 = low risk preference, 1 = high risk preference
insprem <- 200 + 5 * activity_level + 20 * risk_pref
+ rnorm(n, mean = 0, sd = 10)
data <- data.frame(id = id, activity_level = activity_level, risk_pref = risk_pref, insprem =
```

The unit of observation is an individual with 1000 agents. Each agent has 50 observations over time. In this model, Y is insurance premium, X is activity level, and W is risk preference. The true model is:

$$InsurancePremium_i = \beta_0 + \beta_1 \cdot ActivityLevel_i + \beta_2 \cdot RiskPreference_i + \epsilon_i,$$

where $InsurancePremium_i$ is the insurance premium for individual i, $ActivityLevel_i$ is the activity level of individual i, $RiskPreference_i$ is the risk preference of individual i, and ϵ_i is the error term. The true coefficients are: $\beta_0 = 200$, $\beta_1 = 5$, $\beta_2 = 20$, and $\sigma = 10$. We observe $InsurancePremium_i$, $ActivityLevel_i$, and $RiskPreference_i$.

3. Regressions

The regression with endogeneity problem is:

$$InsurancePremium_i = \alpha_0 + \alpha_1 \cdot ActivityLevel_i + u_i,$$

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

$$InsurancePremium_{it} = \gamma_0 + \gamma_1 \cdot ActivityLevel_{it} + \gamma_2 \cdot RiskPreference_{it} + v_{it},$$

where $InsurancePremium_{it}$ is the insurance premium for individual i at time t, $ActivityLevel_{it}$ is the activity level of individual i at time t, $RiskPreference_{it}$ is the risk preference of individual i at time t, and v_{it} is the error term.

4. Summary statistics

```
library(stargazer)
```

```
Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(data[, -c(1, 2)],
  title = "Summary Statistics",
  type = "latex", digits = 2,
  covariate.labels = c("Insurance Premium", "Risk Preference"),
  summary.stat = c("n", "mean", "sd", "min", "max"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Sep 04, 2025 - 13:00:16

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Insurance Premium	50,000	0.50	0.50	0	1
Risk Preference	50,000	224.98	11.17	196.93	256.46

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
```

```r
library(plm)
library(sandwich)
library(stargazer)

model1 <- lm(insprem ~ activity_level, data = data)
model2 <- lm(insprem ~ activity_level + risk_pref, data = data)

stargazer(model1, model2,
  type = "latex",
  title = "Effect of Activity Level on Insurance Premium",
  dep.var.labels = c("Insurance Premium"),
  covariate.labels = c("Activity Level", "Risk Preference"),
  omit.stat = c("f", "ser"),
  no.space = TRUE)
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Sep 04, 2025 - 13:00:16

Table 2: Effect of Activity Level on Insurance Premium

	Dependent variable:	
	Insurance Premium	
	(1)	(2)
Activity Level	4.978***	5.000***
	(0.045)	(0.000)
Risk Preference		20.000***
		(0.000)
Constant	210.054***	200.000***
	(0.141)	(0.000)
Observations	50,000	50,000
R^2	0.198	1.000
Adjusted R^2	0.198	1.000
Note:	*p<0.1; **p<0.05; ***p<0.01	

Putting perfect collinearity aside, activity level is significant in both regressions, with it being slightly higher in the fixed effects regression. Risk preference is also significant in the fixed

effects regression. The fixed effects regression likely provides more accuarate estimates due to addressing endogeneity concerns, as risk preference is a confounding variable that affects both activity level and insurance premium.

I must have coded the data wrong given the perfect collinearity issue, but I am not sure where the error is. Feedback would be appreciated.

IV Model

1. Research Question

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

2. Dataset

```
library(truncnorm)
set.seed(0219)
n <- 1000
weather <- rbinom(n, 1, 0.5)
socializing <- rtruncnorm(n, a = 0, b = 30, mean = 10, sd = 2)
motivation <- rtruncnorm(n, a = 0, b = 10, mean = 5, sd = 1)
exercise_freq <- 5 + 3 * weather + 0.5 * socializing + 0.2 * motivation + rnorm(n, mean = 0,
mental_health_true <- 50 + 2 * exercise_freq + 1.5 * socializing + 3 * motivation + rnorm(n,
mental_health_obs <- 50 + 2 * exercise_freq + 1.5 * socializing + rnorm(n, mean = 0, sd = 5)
data_iv <- data.frame(weather = weather,
  exercise_freq = exercise_freq,
  socializing = socializing,
  motivation = motivation,
  mental_health_true = mental_health_true,
  mental_health_obs = mental_health_obs)
```

The unit of observation is an individual with n = 1000. In this model, Y is mental health score, X is exercise frequency, Z is weather (instrument), and W is socializing frequency. The true model is:

$$MentalHealth_i = \beta_0 + \beta_1 \cdot ExerciseFreq_i + \beta_2 \cdot Socializing_i + \beta_3 \cdot Motivation_i + \epsilon_i,$$

where $MentalHealth_i$ is the mental health score for individual i, $ExerciseFreq_i$ is the exercise frequency of individual i, $Socializing_i$ is the socializing frequency of individual i, $Motivation_i$

is the motivation level of individual i, and ϵ_i is the error term. The true coefficients are: $\beta_0 = 50$, $\beta_1 = 2$, $\beta_2 = 1.5$, $\beta_3 = 3$, and $\sigma = 5$. We observe $MentalHealth_i$, $ExerciseFreq_i$, and $Socializing_i$, but we do not observe $Motivation_i$, so the model is misspecified as

$$MentalHealth_i = \alpha_0 + \alpha_1 \cdot ExerciseFreq_i + \alpha_2 \cdot Socializing_i + u_i,$$

where u_i is the error term which includes the motivation parameter.

3. Regressions

The regression with endogeneity problem is:

$$MentalHealth_i = \alpha_0 + \alpha_1 \cdot ExerciseFreq_i + \alpha_2 \cdot Socializing_i + u_i,$$

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

$$MentalHealth_i = \gamma_0 + \gamma_1 \cdot \widehat{ExerciseFreq}_i + \gamma_2 \cdot Socializing_i + v_i,$$

where $\widehat{ExerciseFreq}_i$ is the predicted exercise frequency from the first stage regression using weather as an instrument, and v_i is the error term.

4. Summary statistics

```
library(stargazer)
stargazer(data_iv,
  title = "Summary Statistics",
  type = "latex", digits = 2,
  covariate.labels = c("Weather", "Exercise Frequency", "Socializing Frequency", "Motivation
  summary.stat = c("n", "mean", "sd", "min", "max"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Sep 04, 2025 - 13:00:16

Table 3: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Weather	1,000	0.48	0.50	0	1
Exercise Frequency	1,000	12.39	2.73	2.83	20.05
Socializing Frequency	1,000	9.96	2.04	3.84	15.80
Motivation	1,000	5.01	1.04	1.48	9.10
Mental Health (True)	1,000	104.73	9.76	78.14	135.67
Mental Health (Observed)	1,000	89.78	8.76	54.27	115.83

5. Regression results

Exercise frequency is not signifcant in the IV regression, but is in the OLS. Socializing frequency is significant in both regressions. The IV regression likely provides more accuarate estimates due to addressing endogeneity concerns, as weather is a valid instrument that affects exercise frequency but is not directly related to mental health. The results suggest that socialization is a more important factor for mental health than exercise frequency when accounting for endogeneity.

```
library(AER)
```

```
Loading required package: car
```

```
Loading required package: carData
```

```
Loading required package: survival
```

```
library(stargazer)

model1_ols <- lm(mental_health_obs ~ exercise_freq, data = data_iv)
model2_ols <- lm(mental_health_obs ~ exercise_freq + socializing, data = data_iv)

model1_2sls <- lm(exercise_freq ~ weather, data = data_iv)
data_iv$x_now <- model1_2sls$fitted.values
model2_2sls <- lm(exercise_freq ~ weather + socializing, data = data_iv)
data_iv$x_w2 <- model2_2sls$fitted.values

model1_step2 <- lm(mental_health_obs ~ x_now, data = data_iv)
model2_step2 <- lm(mental_health_obs ~ x_w2 + socializing, data = data_iv)
```

```r
model_iv <- ivreg(mental_health_obs ~ exercise_freq + socializing | weather + socializing, da

stargazer(model1_ols, model2_ols, model1_step2, model2_step2, model_iv,
  type = "latex",
  title = "Effect of Exercise Frequency on Mental Health",
  dep.var.labels = c("Mental Health (Observed)"),
  column.labels = c("OLS (No W)", "OLS (With W)", "2SLS (No W)", "2SLS (With W)", "IV Regress
  covariate.labels = c("Exercise Frequency", "Socializing Frequency", "Socializing Frequency
  omit.stat = c("f", "ser"),
  no.space = TRUE,
  flip = TRUE,
  float.env = "sidewaystable")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Sep 04, 2025 - 13:00:17 % Requires LaTeX packages: rotating

Table 4: Effect of Exercise Frequency on Mental Health

	Dependent variable:				
	Mental Health (Observed)				
	OLS				*instrumental variable*
	OLS (No W)	OLS (With W)	2SLS (No W)	2SLS (With W)	IV Regression
	(1)	(2)	(3)	(4)	(5)
Exercise Frequency	2.462***	2.047***			1.995***
	(0.065)	(0.061)			(0.101)
Socializing Frequency				1.995***	
				(0.133)	
Socializing Frequency (2SLS)		1.479***		1.505***	1.505***
		(0.082)		(0.120)	(0.091)
Exercise Frequency (2SLS)			1.945***		
			(0.174)		
Constant	59.271***	49.680***	65.678***	50.064***	50.064***
	(0.828)	(0.892)	(2.168)	(1.413)	(1.073)
Observations	1,000	1,000	1,000	1,000	1,000
R^2	0.588	0.690	0.112	0.462	0.690
Adjusted R^2	0.587	0.689	0.111	0.461	0.689
Note:					*p<0.1; **p<0.05; ***p<0.01