Reduced Form Coding Assignment:
Econ 8250
Due September 4, 2025

The goal of this assignment is to give you experience simulating data with endogeneity issues and get your hands dirty with the various research designs that we discuss in class. My hope is that seeing the result like this will help to make clear why identification is such a central concern in empirical research and you will also have much more clarity about how these different empirical strategies are implemented in practice. For this assignment, you may use STATA, R, Python, or SAS. (I am most comfortable with STATA, so my ability to help you will be highest if you use STATA.) You should return results in a pdf file, made using LaTeX and have formatted tables, rather than just screenshots of an output window.

For each empirical strategy, I want you to simulate out a dataset which has a problem of the type that this empirical strategy fixes. For example, if you are working with the fixed effects model, you would simulate out a "true model" which contains unobserved, time invariant heterogeneity. Then, you will run the regression without an empirical strategy and recover biased estimates. Finally, you will run your model using the empirical strategy and recover the "true" coefficients you set.

While this may be somewhat contrived at times, part of the goal of this assignment is to give you practice writing and describing data, regression results, and making tables pretty. Part of your grade will depend on describing your work carefully. Communication is a critical part of being an economist, if you cannot communicate your work, referees/editors will not make an effort to publish it.

For each 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 Fixed Effects

Follow steps 1-5 above for a situation which fixed effects is meant to solve. You should have unobserved, but time invariant heterogeneity that creates an omitted variable bias problem which produces a sign on a coefficient that is unintuitive. You should have two regression models, one with fixed effects and one without. Aim to have about 1,000 individuals and 50 or so time periods per individual or city.

2 Instrumental Variables

Follow steps 1-5 above for a situation which instrumental variables is meant to solve. Let X denote the endogenous variable. Let Z denote your instrument for X. Let W denote a control variable which is exogenous, uncorrelated with X and the omitted variable. Run the following regressions:

1. Run the OLS regression of X on Y without instruments. You should set this up so it has a large bias in it.

2. Run two-stage least squares regression with Z as an instrument for X using two OLS regressions.

3. Use a canned two-stage least squares command to run the regression.

For each regression above, run one specification with W and one without (6 specifications total). Make sure to comment on the effect that W has on your coefficient estimates. Likewise, discuss the bias of X in OLS and the interpretation of your first stage regression when doing two-stage least squares with OLS.