ECON 8250 - Example of Reduced Form Coding Assignment

1 Controlling for Confounding Variables

1.1 Describe the research question

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.

The research question I am interested in is whether hospitalizations save lives. The endogeneity concern is that people who receive hospital care are sicker than the average person. Therefore, if comparing those with and without hospitalizations, we might find a positive correlation between the probability of dying and hospitalization.

1.2 Describe the data and data generating process

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).

In this case, the I assume that the true model is probability of dying is a function of both being hospitalized and underlying health status. The data is a cross-sectional a sample of individuals. For each individual, we observe (somewhat unrealistically) their probability of dying. I also observe whether each individual has been hospitalized. I cannot observe an individual's underlying health status at the beginning of the year (prior to death). In this setting, I assume that both higher health status and hospitalization means there is a lower probability of dying. However, the endogeneity concern is that worse health status is correlated with hospitalization, so this is an important correlation to account for in my simulation.

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. $\beta = 2$).

P_i is a variable for the probability an individual i dies in a given year. $Health_i$ denotes their prior health status, where a higher value means that an individual is healthier. $Hosp_i$ is an indicator for whether someone has been hospitalized in that year. I assume the true regression model is given by the following linear probability model:

$$P_i = \beta_0 + \beta_1 Health_i + \beta_2 Hosp_i + \nu_i \tag{1}$$

where ν_i is an idiosyncratic health shock that is uncorrelated with prior health status or hospitalization.

The simulated "true" parameters are:

$$\beta_0 = 0.4$$

$$\beta_1 = -0.65$$

$$\beta_2 = -0.1$$

1.3 Describe the naive regression and fix

In words and equations, describe the regressions you are running. Both the regressions that have an endogeneity problem and the ones which you "fix."

I begin by running a regression assuming I do not observe prior health status. Formally, this regression equation is:

$$P_i = \alpha_0 + \alpha_1 Hosp_i + \epsilon_i \tag{2}$$

This will create an endogeneity problem as health status is an omitted variable. To solve this problem, I include health status as a regressor.

$$P_i = \alpha_0 + \alpha_1 Health_i + \alpha_2 Hosp_i + \epsilon_i \tag{3}$$

1.4 Describe the summary stats

Produce a table of summary statistics given 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.

Table 1 presents summary statistics for all the regressors and the outcome variables. The average probability of dying in my sample is five percent, though this (unrealistically) varies between -1.3 and 1.3. Average health status is .5, though this varies between zero and one. 25 percent of the sample was hospitalized.

Table 1: Summary Stats

	mean	sd	min	max
Probability of Dying	0.02	0.36	-1.33	1.31
Prior Health Status	0.50	0.29	0.00	1.00
1(Hospitalized)	0.50	0.50	0.00	1.00
Observations	20000			

1.5 Describe the results

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.

Table 2 presents regression results. The first column shows results without health status. The coefficient of .187 suggests that hospitalization make people 18.7 percentage points *more* likely to die in that year. Under the assumption that hospitalizations improve lives, this seems odd, but this is expected because we aren't controlling for health status. The second column displays results including health status. After controlling for health status, hospitalizations now *reduce* the probability of dying by 9.9 percentage points. Likewise, an increase of .1 in health status reduces health status by .065 percentage points. Finally, notice that the regression results in column 2 are similar to the true coefficients above.

The takeaway from this table is that hospitalizations save lives. However, if one does not control for health status, you might find the opposite result.

Table 2: Regression Results: Hospitalization on Probability of Dying

	(1)	(2)
1(Hospitalized)	0.187***	-0.099***
	(0.005)	(0.007)
Prior Health Status		-0.648***
		(0.010)
Constant	0.002	0.397***
	(0.003)	(0.006)
Observations	20000	20000