Problem set 7: Education and wages + public housing and health

YOUR NAME HERE

DATE GOES HERE

Table of Contents

library(tidyverse) # For ggplot, mutate(), filter(), and friends  
library(broom) # For converting models to data frames  
library(estimatr) # For lm\_robust() and iv\_robust()  
library(modelsummary) # For showing side-by-side regression tables

# Task 1: Education, wages, and kids

Let’s look once again at the effect of education on earnings. You’ll use data from the 1976 Current Population Survey run by the US Census. The data is available as wage in the **wooldridge** R package—here I’ve just taken a subset of variables and renamed them. There are three columns:

| Variable name | Description |
| --- | --- |
| wage | Average hourly earnings (in 1976 dollars) |
| education | Years of education |
| n\_kids | Number of dependents living at home |

You’re interested in estimating in:

However, there is an issue with omitted variable bias and endogeneity. Instrumental variables can potentially help address the endogeneity.

## Step 1

Load and look at the dataset

wages <- read\_csv("data/wages.csv")

## Step 2

We need an instrument for education, since part of it is endogenous. Do you think the variable n\_kids (the number of children) would be a valid instrument? Does it meet the three requirements of a valid instrument?

ANSWER HERE

Explain why it passes or fails each of the three requirements for a valid instrument. Test the requirements where possible using scatterplots and regression.

ANSWER HERE and include plots and regression

## Step 3

Assume that the number of children is a valid instrument (regardless of whatever you concluded earlier). Using the number of children (n\_kids) as an instrument for education (education), estimate the effect of education on wages via two-stage least squares (2SLS) instrumental variables (IV).

iv\_reg<-iv\_robust(wage~education|n\_kids, data=wages)  
tidy(iv\_reg)

## term estimate std.error statistic p.value conf.low conf.high df  
## 1 (Intercept) 1.7093781 2.8030095 0.6098367 0.5422343 -3.7971383 7.2158945 524  
## 2 education 0.3331669 0.2221214 1.4999318 0.1342343 -0.1031909 0.7695247 524  
## outcome  
## 1 wage  
## 2 wage

Do this by hand: create a first stage model, extract the predicted education, and use predicted education in the second stage.

Interpret the coefficient that gives the effect of education on wages () and its significance.

mod\_1S<-lm(education~n\_kids,data=wages)  
aug\_reg<-augment\_columns(mod\_1S, wages)|>rename(edu\_hat=.fitted)

second\_stage<-lm(wage~edu\_hat,data = aug\_reg)  
tidy(second\_stage)

## # A tibble: 2 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 1.71 3.40 0.503 0.615  
## 2 edu\_hat 0.333 0.270 1.23 0.218

DO ALL THAT HERE

(Remember that you can also use the iv\_robust() function from the **estimatr** package to run IV/2SLS models in one step with: iv\_robust(y ~ x | z, data = data), where y is the outcome, x is the policy/program, and z is the instrument. Try doing this to check your manual two-stage model.)

## Step 4

Run a naive model predicting the effect of education on wages (i.e. without any instruments). How does this naive model compare with the IV model?

Show the results side-by-side here:

naive<-lm(wage~education,data=wages)  
tidy(naive)

## # A tibble: 2 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -0.902 0.685 -1.32 1.89e- 1  
## 2 education 0.541 0.0533 10.2 3.09e-22

TABLE HERE

## Step 6

Explain which estimates (OLS vs. IV/2SLS) you would trust more (or why you distrust both)

modelsummary(list("Naive model"=naive,"By hand"=second\_stage,"IV regression"=iv\_reg))

|  | Naive model | By hand | IV regression |
| --- | --- | --- | --- |
| (Intercept) | -0.902 | 1.709 | 1.709 |
|  | (0.685) | (3.400) | (2.803) |
| education | 0.541 |  | 0.333 |
|  | (0.053) |  | (0.222) |
| edu\_hat |  | 0.333 |  |
|  |  | (0.270) |  |
| Num.Obs. | 526 | 526 | 526 |
| R2 | 0.164 | 0.003 | 0.140 |
| R2 Adj. | 0.163 | 0.001 | 0.139 |
| AIC | 2778.1 | 2871.1 | 2793.2 |
| BIC | 2790.9 | 2883.9 | 2806.0 |
| Log.Lik. | -1386.050 | -1432.535 |  |
| F | 103.118 | 1.519 |  |
| RMSE | 3.37 | 3.69 | 3.42 |

# Task 2: Public housing and health

[Economic research shows](https://dx.doi.org/10.1002/pam.20288) that there is a potential (albeit weak) connection between health outcomes and residency in public housing. You are interested in finding the effect of public housing assistance on health outcomes. In the absence of experimental data, you must use observational data collected by the Georgia Department of Public Health. You have access to a dataset of 1,000 rows with the following columns:

| Variable name | Description |
| --- | --- |
| HealthStatus | Health status on a scale from 1 = poor to 20 = excellent |
| HealthBehavior | Omitted variable (you can’t actually measure this!) |
| PublicHousing | Number of years spent in public housing |
| Supply | Number of available public housing units in the city per 100 eligible households |
| ParentsHealthStatus | Health status of parents on a scale from 1 = poor to 20 = excellent |
| WaitingTime | Average waiting time before obtaining public housing in the city (in months) |
| Stamp | Dollar amount of food stamps (SNAP) spent each month |
| Age | Age |
| Race | Race; 1 = White, 2 = Black, 3 = Hispanic, 4 = Other |
| Education | Education; 1 = Some high school, 2 = High school, 3 = Bachelor’s, 4 = Master’s |
| MaritalStatus | Marital status; 1 = Single, 2 = Married, 3 = Widow, 4 = Divorced |

(This is simulated data, but it’s based on analysis by [Angela R. Fertig and David A. Reingold](https://dx.doi.org/10.1002/pam.20288))

Your goal is to measure the effect of living in public housing (PublicHousing) on health (HealthStatus). There is omitted variable bias, though, since people who care more about their health might be more likely to self-select into public housing and report a better health status score. The magic variable HealthBehavior measures this omitted variable, and you can use it as reference to make sure you get the models right (this is the same as “ability” in the examples in class), but don’t include it in any of your actual models, since it’s not real.

This data includes four potential instruments:

* Supply: Number of available public housing units in the city per 100 eligible households
* ParentsHealthStatus: Health status of parents on a scale from 1 = poor to 5 = excellent
* WaitingTime: Average waiting time before obtaining public housing in the city (in months)
* Stamp: Dollar amount of food stamps (SNAP) spent each month

You have three tasks:

1. Evaluate the suitability of each of the four potential instruments. Check if they (1) have *relevance* with a scatterplot and model and F-test, (2) meet the *excludability* assumption, and (3) meet the *exogeneity* assumption. Choose one of these as your main instrument and justify why it’s the best. Explain why the other three are not.
2. Estimate a naive model of the effect of public housing on health status (i.e. without any instruments). You can include any control variables you feel appropriate (i.e. that fit in your causal model). If you use variables that are categorical like race, education, or marital status, make sure you wrap them with as.factor() to treat them as categories instead of numbers (e.g. as.factor(education)).
3. Estimate the effect of public housing on health status using 2SLS IV. You can use iv\_robust() to do it all in one step if you want (but you’ll still need to run a first-stage model to find the F statistic). Compare the results with the naive model. Which model do you trust (if any), and why?

housing <- read\_csv("data/public\_housing.csv")

DO ALL THAT HERE.