Problem Set 2

Tao Wu

October 5, 2023
Professor Suleyman Taspinar
ECON 387: Advanced Econometrics

1. **Define the following terms in your own words**

   (a) The instrumental variable (IV) Z_i is a variable that is correlated with regressor(s) X_i but uncorrelated with the error term u_i. In other words, the covariance between Z_i and X_i should not be 0, and the covariance between Z_i and the error term u_i must be 0.

   (b) An endogenous variable is one that is correlated with the error term u. An exogenous variable is one that is uncorrelated with u.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

   (c) There are threats to internal validity and we need IV to provide a solution to the violation of the zero mean condition assumption. 2 conditions IV must meet before its use. First, the Instrument relevance condition states Z_i must be correlated with regressor X_i.

   (d) Second, Instrument exogeneity condition states Z_i must be uncorrelated with error term u_i

   (e) Omitted variable bias presents a sticky issue in the study of econometrics. Two-stage least squares is a useful technique to mitigate such a problem by first regressing the endogenous variable on the valid IV(s). Then we compute to get its predicted value. In the second stage, we proceed to regress our parameter of interest on the predicted value to get an unbiased estimator, a great way to "purge" endogeneity and other issues.

   (f) In an econometric model, there are k number of endogenous regressors along with r number of W control variables. If there are m number of instrumental variables in the same model, we have descriptive terms when the difference between m and k emerges or does not. If m > k, we say the coefficients are over-identified. if m < k, the coefficients are under-identified. if m = k, the coefficients are exactly identified.

(g) A weak instrument refers to an instrumental variable that is not highly correlated with the endogenous independent variable resulting in inefficient estimation. Second-stage parameter estimators are no longer unbiased and precise. Things like the T-test and confidence interval might be misleading and sampling distribution can not be normal.

(h) First-stage F-statistic is what we can use to test the strength of IVs. When we regress our independent variable on IVs, we can compute the F-statistic for the null hypothesis stating the estimator of all IVs is 0. If the F statistic is greater than 10, we say the IVs are not weak. IVs are weak otherwise.

1. **Consider the cigarette demand study, we covered in this chapter. Suppose we used as an instrument the number of trees per capita in the state. Is this instrument relevant? Is it exogenous? Is it a valid instrument?**

   We should ask if the two conditions are met by the instrument variable considered. The number of trees per capita in the state, to me, does not seem to have instrument relevance without elaborate theories. Our regressor here, namely the average real price per pack of cigarettes including all taxes, does not correlate with such IV. Is it exogenous? No. Because The number of trees per capita is likely to be negatively correlated with the state population variable which is not in the model and by definition, is in the error term. The state population increases, number of trees per capita goes down and vice versa. Conclusion: The number of trees per capita in the state is not a valid instrument.

1. **Consider the IV regression model, where X_i is correlated with u_i and Z_i is an instrument. Suppose that the first three assumptions in the lecture slides (see Slide 17) are satisfied. Which IV assumption is not valid in each case listed below?**

$$Y_i = \beta_1 X_i + \beta_2 W_i + u_i$$

(a) 2 conditions: relevance & exogeneity. if Z_i is independent of (Y_i, X_i, W_i), then the instrument relevance is violated. X_i and Z_i need to be correlated.

(b) if $Z_i = W_i$, Z_i would not be valid instrument. IV has to come from outside the model and cannot replace a control variable that may or may not correlate with the regressor.

(c) if $Z_i = X_i$, the purpose of the model will be changed and no longer suited to answer the original question.

1. **(10 pts) Consider an IV regression model with one regressor, X_i, and two instruments, Z_{1i} and Z_{2i}**

(a) In general, there's one condition that needs to be met before we can consider testing for the exogeneity of instruments, which is the number of the instruments needs to be greater than the number of endogenous regressors. If yes, then we use a test called the overidentifying restrictions test (the J-test). Basically, we would want to find out the predicted TSLS residuals and whether or not the 2 instrument variables have explanatory power on the predicted TSLS residuals. One way we can tell is by computing the F-statistic after regressing the predicted TSLS residuals on the 2 IVs and other control variables. The null hypothesis is that all the coefficients of the Z_i terms are zero. The alternative hypothesis is that there is at least one coefficient that is nonzero and thus explains the predicted TSLS residuals.

(b) We compare the computed J-statistic (4.45) to a critical value from the chi-squared distribution at a 5% significance level. Since the critical value for a chi-squared distribution with 1 degree of freedom at the 5% significance level is approximately 3.841. The result is that we reject the null hypothesis. This rejection signals that **At least** one of the instruments is endogenous, to begin with.

$$E(u_i|Z_{1i}, Z_{2i}) \neq 0$$

In other words, IVs are not clean and correlated with the error term in the original model. Decisions need to be made to weed out some or all IVs.

1. **In this exercise, you will replicate some of the results in Holian [2020]. In this study, the author explores the effect of urban form (measured by population density) on vehicle ownership. The sample data comes from the American Community Survey (ACS), carried out between 2012 and 2017. The data are contained in the urban.csv file and include the following variables.**

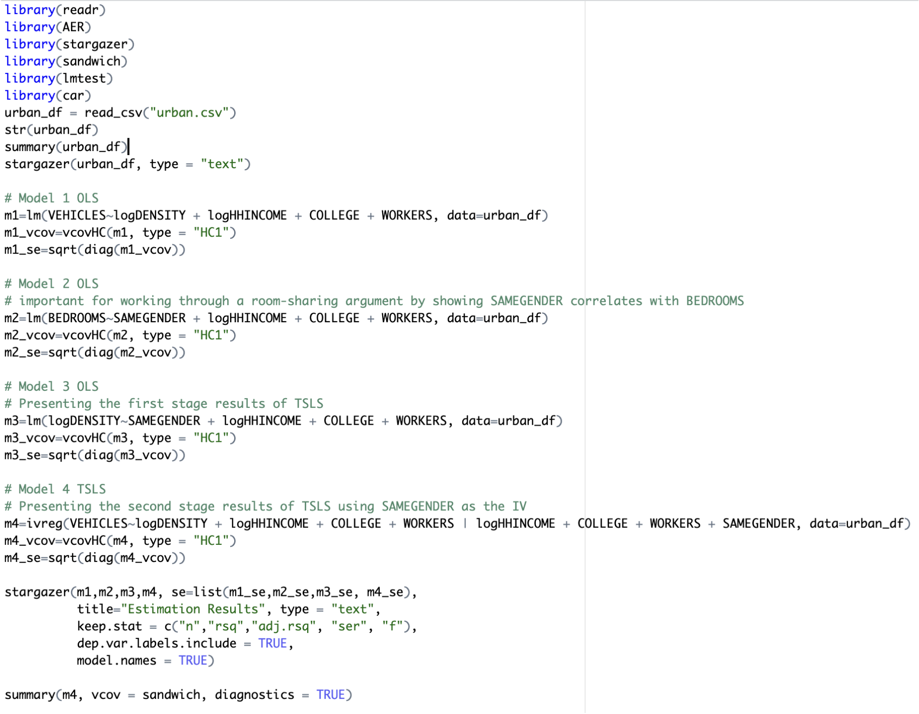
(a) In the introduction, it states, "the main contribution of this paper is to present a compelling instrumental variable (IV) estimate of the effect of residential density on vehicle ownership," which effectively summs up the estimation model used in the paper.

$$Vehicles_i = \beta_0 + \beta_1 logDensity_i + X_i + u_i,$$

, where X_i is a vector of control varibales and u_i is the regression error term. In this model, the regressor, $logDensity_i$ is the population density (people per square mile) in the Public-Use Microdata Area (PUMA) in which household i resides. The author argues that since household location is not randomly assigned— households with unobserved travel preferences self-select where they live, $logDensity_i$ is correlated with the regression error term.

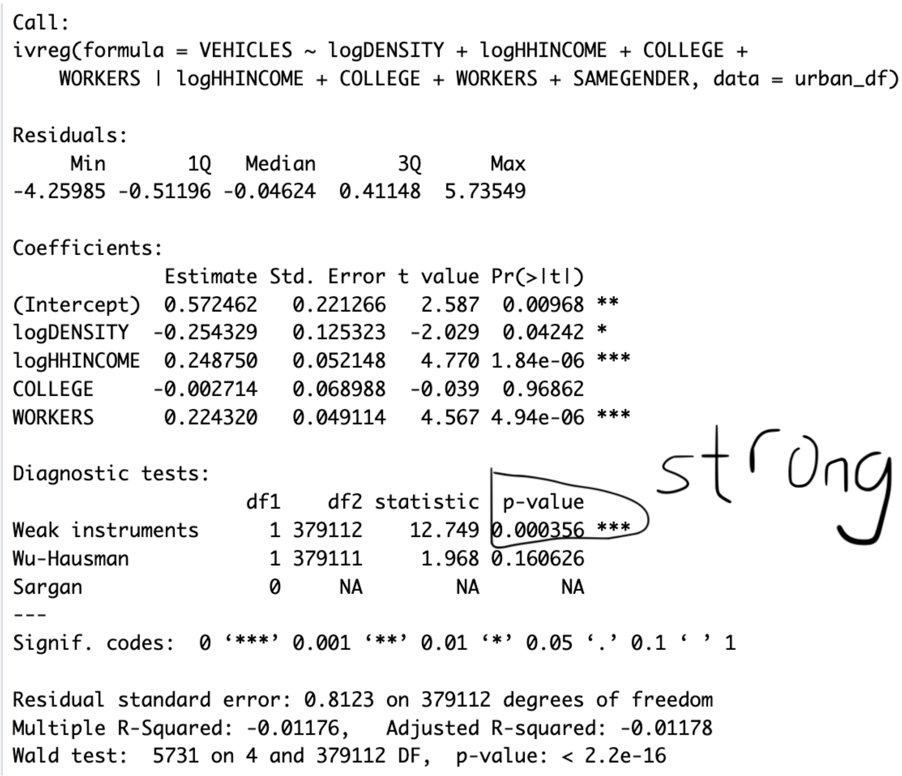
(b) The same-gender indicator is the instrument variable suggested by the author who instruments for $logDensity_i$ using such an indicator for households that have same gender children. The possibility of more easily having children of the same gender share rooms means these households have more options in denser neighborhoods where homes are typically smaller. While this explanation makes sense, let's examine the two conditions that a valid IV must meet. According to the paper, "the same-gender indicator is a relevant instrument as it is a fairly strong predictor of density (F = 12.8), even when using the geographically unrefined measures of population density in the public-use data," this has demonstrated the instrument relevance with F-statistic is greater than 10. Moreover, the author also argues for exogeneity of the instrument variable due to the randomness of a child's gender in a household. Truly random assignment of gender in any given child eliminates any possible correlation with unobserved factors such as travel preferences in the regression error term.

(c) **Replicate the estimation results given in Table 3**



```
library(readr)
library(AER)
library(stargazer)
library(sandwich)
library(lmtest)
library(car)
urban_df = read_csv("urban.csv")
str(urban_df)
summary(urban_df)
stargazer(urban_df, type = "text")

# Model 1 OLS
m1=lm(VEHICLES~logDENSITY + logHHINCOME + COLLEGE + WORKERS, data=urban_df)
m1_vcov=vcovHC(m1, type = "HC1")
m1_se=sqrt(diag(m1_vcov))

# Model 2 OLS
# important for working through a room-sharing argument by showing SAMEGENDER correlates with BEDROOMS
m2=lm(BEDROOMS~SAMEGENDER + logHHINCOME + COLLEGE + WORKERS, data=urban_df)
m2_vcov=vcovHC(m2, type = "HC1")
m2_se=sqrt(diag(m2_vcov))

# Model 3 OLS
# Presenting the first stage results of TSLS
m3=lm(logDENSITY~SAMEGENDER + logHHINCOME + COLLEGE + WORKERS, data=urban_df)
m3_vcov=vcovHC(m3, type = "HC1")
m3_se=sqrt(diag(m3_vcov))

# Model 4 TSLS
# Presenting the second stage results of TSLS using SAMEGENDER as the IV
m4=ivreg(VEHICLES~logDENSITY + logHHINCOME + COLLEGE + WORKERS | logHHINCOME + COLLEGE + WORKERS + SAMEGENDER, data=urban_df)
m4_vcov=vcovHC(m4, type = "HC1")
m4_se=sqrt(diag(m4_vcov))

stargazer(m1,m2,m3,m4, se=list(m1_se,m2_se,m3_se, m4_se),
          title="Estimation Results", type = "text",
          keep.stat = c("n","rsq","adj.rsq", "ser", "f"),
          dep.var.labels.include = TRUE,
          model.names = TRUE)

summary(m4, vcov = sandwich, diagnostics = TRUE)
```

```
Estimation Results
=======================================================================================
                                                   Dependent variable:
                                  -----------------------------------------------------
                                    VEHICLES      BEDROOMS     logDENSITY    VEHICLES
                                       OLS           OLS          OLS      instrumental
                                                                             variable

                                       (1)           (2)          (3)          (4)
---------------------------------------------------------------------------------------
logDENSITY                          -0.091***                                -0.254**
                                     (0.001)                                 (0.125)

SAMEGENDER                                        -0.041***     0.021***
                                                   (0.003)      (0.006)

logHHINCOME                         0.181***      0.361***      0.416***     0.249***
                                     (0.002)       (0.003)      (0.005)      (0.052)

COLLEGE                             -0.093***     0.185***      0.550***      -0.003
                                     (0.003)       (0.003)      (0.007)      (0.069)

WORKERS                             0.288***      -0.011***    -0.391***     0.224***
                                     (0.003)       (0.003)      (0.005)      (0.049)

Constant                            0.286***      0.206***      1.743***     0.572***
                                     (0.023)       (0.029)      (0.053)      (0.221)

---------------------------------------------------------------------------------------
Observations                         379,117       379,117      379,117      379,117
R2                                    0.124         0.135        0.075        -0.012
Adjusted R2                           0.124         0.135        0.075        -0.012
Residual Std. Error (df = 379112)     0.756         0.843        1.815        0.812
F Statistic (df = 4; 379112)      13,363.290*** 14,833.440*** 7,721.925***
=======================================================================================
Note:                                                       *p<0.1; **p<0.05; ***p<0.01
```



```
Call:
ivreg(formula = VEHICLES ~ logDENSITY + logHHINCOME + COLLEGE +
    WORKERS | logHHINCOME + COLLEGE + WORKERS + SAMEGENDER, data = urban_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-4.25985 -0.51196 -0.04624  0.41148  5.73549 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.572462   0.221266   2.587  0.00968 ** 
logDENSITY  -0.254329   0.125323  -2.029  0.04242 *  
logHHINCOME  0.248750   0.052148   4.770 1.84e-06 ***
COLLEGE     -0.002714   0.068988  -0.039  0.96862    
WORKERS      0.224320   0.049114   4.567 4.94e-06 ***

Diagnostic tests:
                    df1    df2 statistic  p-value    
Weak instruments      1 379112    12.749 0.000356 ***
Wu-Hausman            1 379111     1.968 0.160626    
Sargan                0     NA        NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8123 on 379112 degrees of freedom
Multiple R-Squared: -0.01176,   Adjusted R-squared: -0.01178 
Wald test:  5731 on 4 and 379112 DF,  p-value: < 2.2e-16
```

strong