Problem Set 5 - Solutions

1. **IV Short Questions**

(a) **(Final A Spring 2018)** Consider the simple IV model $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $Cov(X_i, u_i) \neq 0$. Suppose you have an instrument Z_i which satisfies $E(u_i|Z_i) = 0$. Show (i) that Z_i and (ii) that Z_i^2 satisfy the exogeneity condition. *Hint:* Let Y and X be two random variables. Then $E(g(X)Y|X) = g(X)E(Y|X)$ for any function $g(\cdot)$.

Solution

For (i) the exogeneity condition requires that $Cov(Z_i, u_i) = 0$. This condition holds because

$$\begin{aligned} Cov(Z_i, u_i) &= E(Z_i u_i) - E(Z_i)E(u_i) \\ &= E(E(Z_i u_i \mid Z_i)) - E(Z_i)E(E(u_i \mid Z_i)) \\ &= E(Z_i E(u_i \mid Z_i)) - E(Z_i)E(E(u_i \mid Z_i)) \\ &= E(Z_i 0) - E(Z_i)E(0) = 0 \end{aligned}$$

where the second equality follows from the law of iterated expectations and the third equality uses the property of conditional expectations. For (ii) the exogeneity condition requires that $Cov(Z_i^2, u_i) = 0$. This condition holds because

$$\begin{aligned} Cov(Z_i^2, u_i) &= E(Z_i^2 u_i) - E(Z_i^2)E(u_i) \\ &= E(E(Z_i^2 u_i \mid Z_i)) - E(Z_i^2)E(E(u_i \mid Z_i)) \\ &= E(Z_i^2 E(u_i \mid Z_i)) - E(Z_i^2)E(E(u_i \mid Z_i)) \\ &= E(Z_i^2 0) - E(Z_i^2)E(0) = 0 \end{aligned}$$

where the second equality follows from the law of iterated expectations and the third equality uses the property in the hint.

(b) **(Final B Spring 2018)** Consider the following model: $Y_i = \beta_0 + \beta_1 X_i + u_i$ where $X_i \sim N(0, 1)$, $u_i \sim N(0, 1)$, and $Cov(u_i, X_i) = 0.3$. In addition, you have a candidate instrument Z_i. The instrument satisfies relevance, $Cov(X_i, Z_i) = 0.5$, but not exogeneity, $Cov(u_i, Z_i) = 0.3$. Which estimator will be closer to β_1 in large samples, the OLS estimator $\hat{\beta}_1$ or the TSLS estimator $\hat{\beta}_1^{TSLS}$? We assume that the data are iid and that large outliers are unlikely. *Hint:* You can use the following two results without proof: $\hat{\beta}_1 \xrightarrow{p} Cov(Y_i, X_i)/Var(X_i)$ and $\hat{\beta}_1^{TSLS} \xrightarrow{p} Cov(Y_i, Z_i)/Cov(X_i, Z_i)$.

Solution

We derive the probability limit for both estimators:

$$\begin{aligned}
\hat{\beta}_1 \xrightarrow{p} \frac{Cov(Y_i, X_i)}{Var(X_i)} &= \frac{Cov(\beta_0 + \beta_1 X_i + u_i, X_i)}{Var(X_i)} \\
&= \frac{\beta_1 Var(X_i) + Cov(u_i, X_i)}{Var(X_i)} \\
&= \beta_1 + \frac{Cov(u_i, X_i)}{Var(X_i)} \\
&= \beta_1 + 0.3,
\end{aligned}$$

and

$$\begin{aligned}
\hat{\beta}_1^{TSLS} \xrightarrow{p} \frac{Cov(Y_i, Z_i)}{Cov(X_i, Z_i)} &= \frac{Cov(\beta_0 + \beta_1 X_i + u_i, Z_i)}{Cov(X_i, Z_i)} \\
&= \frac{\beta_1 Cov(X_i, Z_i) + Cov(u_i, Z_i)}{Cov(X_i, Z_i)} \\
&= \beta_1 + \frac{Cov(u_i, Z_i)}{Cov(X_i, Z_i)} \\
&= \beta_1 + \frac{0.3}{0.5} \\
&= \beta_1 + 0.6
\end{aligned}$$

Hence, the asymptotic bias (difference between probability limit and β_1) of $\hat{\beta}_1$ is 0.3, whereas the bias for $\hat{\beta}_1^{TSLS}$ is 0.6. This, $\hat{\beta}_1$ is closer to β_1 than $\hat{\beta}_1^{TSLS}$ in large samples.

(c) **(Final Winter 2019)** Consider the following model $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $X_i = v_i + u_i$ and $Z_i = v_i^2 + \varepsilon_i$. Suppose that v_i, u_i, and ε_i are i.i.d., independent of each other, and normally distributed with mean 0 and variance 1. Is the instrument Z_i relevant? Is it exogenous? *Hint:* Suppose the random variable A is symmetrically distributed with mean zero. Then $E(A^3) = 0$.

Solution

Consider first the exogeneity condition which requires that $Cov(Z_i, u_i) = 0$. Note that

$$Cov(u_i, Z_i) = Cov(u_i, v_i^2 + \varepsilon_i) = 0$$

u_i is independent of both v_i and ε_i. Thus, Z_i satisfies exogeneity.

Consider next the relevance conditions which requires that $Cov(X_i, Z_i) \neq 0$. Note that

$$Cov(X_i, Z_i) = Cov(v_i + u_i, v_i^2 + \varepsilon_i) = Cov(v_i, v_i^2) = E(v_i^3) = 0,$$

where the second equality follows from independence of u_i, v_i, and ε_i and the last equality follows from the hint. Thus, Z_i does not satisfy relevance.

2. **Instrumental variables (Final Winter 2019)**

Your goal is to estimate the demand for coffee in California. You have access to 3 years of daily data on coffee sales (`coffee_sales`) and price of coffee (`coffee_price`). You also decide to include the price of tea (`tea_price`) because it is a substitute for coffee. You want the estimate the following model

$$\texttt{coffee\_sales}_t = \beta_0 + \beta_1 \times \texttt{coffee\_price}_t + \beta_2 \times \texttt{tea\_price}_t + u_t$$

(a) What is the expected sign for β_1 and β_2? Explain.

Solution

β_1 is expected to have negative sign, because of the inverse relationship between price and demand. β_2 is expected to have positive sign, since tea and coffee are substitute goods, and thus an increase in the price of tea, all things equal, will be associated with more demand for coffee.

(b) Consider the following regression output from an OLS regression. Is $\hat{\beta}_1$ a consistent estimator of β_1? Why, or why not?

```
. reg coffee_sales coffee_price tea_price, robust
Linear regression                                   Number of obs     =      1,080
                                                    F(2, 1077)        =      33.19
                                                    Prob > F          =     0.0000
                                                    R-squared         =     0.0576
                                                    Root MSE          =      6.529

------------------------------------------------------------------------------
             |               Robust
coffee_sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
coffee_price |   .1559335   .0398414     3.91   0.000      .077758    .234109
   tea_price |    1.39104   .1874582     7.42   0.000     1.023216   1.758865
       _cons |   2.538022   .2166753    11.71   0.000     2.112869   2.963176
------------------------------------------------------------------------------
```

Solution

The OLS estimator of β_1 in this case is not a consistent estimator, since the dependent variable (coffee_sales) and the regressor (coffee_price) have simultaneous relationship. One may also refer to the case of omitted variable bias such as quality of coffee or consumer's preference.

(c) Suppose that California imports 5% of the coffee beans from Peru, and the remaining 95% from Colombia and Guatemala. You want to use the rainfall in Peru, Colombia and Guatemala as an instrument for the price of coffee. Do you think these instruments could be relevant? Do you think these instruments could be exogenous?

Solution

Exogeneity: Rainfall is a randomly realized weather event, we can assume it is an exogenous IV.

Relevance: Since the rainfall affects supply of the coffee, which in turn changes coffee price, rainfall is a relevant IV.

(d) You first gather data on rainfall on Peru (`peru`) and decide to use it as an instrument for the price of coffee. Below is the output from the first stage regression. Are you worried about weak instruments? Conduct a formal test.

```
. reg coffee_price peru tea_price, robust
Linear regression                                   Number of obs     =      1,080
                                                    F(2, 1077)        =       0.57
                                                    Prob > F          =     0.5648
                                                    R-squared         =     0.0011
                                                    Root MSE          =      5.304
```

```
------------------------------------------------------------------------------
             |               Robust
coffee_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        peru |   .0680046    .153905     0.44   0.659    -.2339831    .3699922
   tea_price |  -.1636786   .1652138    -0.99   0.322     -.487856    .1604988
       _cons |   1.969957   .1620418    12.16   0.000     1.652003     2.28791
------------------------------------------------------------------------------
```

Solution

The first stage F statistic is $0.44^2 = 0.1936$ which is much smaller than the rule-of-thumb value for assessing whether an instrument is weak which is 10. We therefore conclude that the instrument is weak.

(e) You run a TSLS regression using rainfall in Peru as an instrument for the price of coffee. Below are the results. Why do you think the 95% confidence interval on the coefficient for `coffee_price` is so wide?

```
. ivregress 2sls coffee_sales (coffee_price=peru) tea_price, robust
Instrumental variables (2SLS) regression          Number of obs   =      1,080
                                                  Wald chi2(2)    =      25.34
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
                                                  Root MSE        =     9.2408

------------------------------------------------------------------------------
             |               Robust
coffee_sales |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
coffee_price |  -1.080299   3.928601    -0.27   0.783    -8.780214    6.619617
   tea_price |   1.191753   .6819136     1.75   0.081    -.1447727     2.52828
       _cons |   4.968845   7.729655     0.64   0.520      -10.181    20.11869
------------------------------------------------------------------------------
Instrumented:  coffee_price
Instruments:   tea_price peru
```

Solution

Note that

$$Var(\hat{\beta}_{1,TSLS}) = \frac{Var((Z_i E(Z_i))u_i)}{n(Cov(Z_i, X_i))^2}$$

The equation above shows that if the correlation between the IV (Z_i) and the endogenous regressor (X_i) is too low (i.e. the IV is weak), we the variance of the TSLS estimator of β_1 can explode. In that sense, the weak IV issue raised in the previous part of this problem can be a reason of the large standard error presented in the table and the wide confidence band of $\hat{\beta}_{1,TSLS}$.

(f) Given the results in the previous regressions, you gather more data on rainfall, this time from Colombia (`colombia`) and Guatemala (`guatemala`). Consider the following output:

```
. ivregress 2sls coffee_sales (coffee_price=peru colombia guatemala) tea_price, robust

Instrumental variables (2SLS) regression          Number of obs   =      1,080
                                                  Wald chi2(2)    =     160.22
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
```

```
                                                  Root MSE        =     8.6837

------------------------------------------------------------------------------
             |               Robust
coffee_sales |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
coffee_price |   -.926853   .0780546   -11.87   0.000    -1.079837   -.7738687
   tea_price |    1.21649   .2529215     4.81   0.000     .7207727    1.712207
       _cons |   4.667122   .3086696    15.12   0.000     4.062141    5.272103
------------------------------------------------------------------------------
Instrumented:  coffee_price
Instruments:   tea_price peru colombia guatemala

. predict uhat, resid

. reg uhat peru colombia guatemala tea_price

      Source |       SS           df       MS      Number of obs   =     1,080
-------------+----------------------------------   F(4, 1075)      =      0.01
       Model |  1.88111645         4  .470279112   Prob > F        =    0.9999
    Residual |  81438.1308     1,075  75.7564008   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =   -0.0037
       Total |   81440.012     1,079  75.4773049   Root MSE        =    8.7038

------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        peru |  -.0105625   .2612422    -0.04   0.968    -.5231649    .5020398
    colombia |  -.0060195   .0896098    -0.07   0.946    -.1818494    .1698103
   guatemala |   .0253868   .1845148     0.14   0.891    -.3366632    .3874367
   tea_price |   .0016505   .2621378     0.01   0.995    -.5127092    .5160103
       _cons |   -.027946   .3289844    -0.08   0.932    -.6734704    .6175784
------------------------------------------------------------------------------

. test peru colombia guatemala

 ( 1)  peru = 0
 ( 2)  colombia = 0
 ( 3)  guatemala = 0

       F(  3,  1075) =    0.01
            Prob > F =    0.9990
```

Based on the information given, conduct a formal test for instrument exogeneity at the 10%-level. What do you conclude?

Solution

Here, since we have three IVs, $m = 3$. And the value of the F statistics to test all of those three IV has no statistical relationship with the regression error is 0.01. Therefore, the value of the J-test statistics is $J = 3 \times (0.01) = 0.03$. And the critical value (10%) is $\chi^2_{2,0.90} = 4.61$. Since we have $J < \chi^2_{2,0.90}$, we cannot reject the null hypothesis and thus we cannot find any statistical evidence against the exogeneity of the three IVs.