Instrumental Variables - Solutions

Exercise A - (20 min)

1. Set the seed to `1234` and then generate 5000 iid standard normal draws to serve as your instrument `z`. Next, use `rmvnorm()` to generate 5000 iid draws from a bivariate standard normal distribution with correlation $\rho = 0.5$, independently of `z`.
2. Use the draws you made in the preceding part to generate `x` and `y` according to the IV model with $\pi_0 = 0.5$, $\pi_1 = 0.8$, $\alpha = -0.3$, and $\beta = 1$.
3. Using the formulas from the preceding slides, predict the slope coefficient that you would obtain if you ran a linear regression of `y` on `x`. Run this regression to check.
4. Using the formulas from the preceding slides, predict the slope coefficient that you would obtain if you ran a linear regression of `y` on `z`. Run this regression to check.
5. Use the formulas from the preceding slides to calculate the IV estimate. How does it compare to the true causal effect in your simulation?
6. Try running a regression of `y` on *both* `x` and `z`. What is your estimate of the slope coefficient on `x`? How does it compare to the OLS and IV estimates? What gives?
7. Which will give the most accurate predictions of Y in this example: OLS or IV?

Solution

Parts 1-2

```
set.seed(1234)
n <- 5000
z <- rnorm(n)

library(mvtnorm)
Rho <- matrix(c(1, 0.5,
                0.5, 1), 2, 2, byrow = TRUE)
errors <- rmvnorm(n, sigma = Rho)
colMeans(errors)
```

```
[1] 0.021939478 0.005959135
```

```
var(errors)
```

```
          [,1]      [,2]
[1,] 0.9621788 0.4843148
[2,] 0.4843148 0.9943409
```

```
u <- errors[,1]
v <- errors[,2]
x <- 0.5 + 0.8 * z + v
y <- -0.3 + x + u
```

Solution

Part 3

If we regress `y` on `x`, we will obtain the sample estimate of

$$\beta_{OLS} = \beta + \frac{\text{Cov}(X,U)}{\text{Var}(X)}.$$

Since $\text{Var}(U) = \text{Var}(V) = 1$, $\text{Cov}(X,U) = \text{Cov}(U,V) = \rho$. And since Z and V are uncorrelated,

$$\text{Var}(X) = \pi_1^2\text{Var}(Z) + \text{Var}(V) = \pi_1^2 + 1.$$

Therefore $\beta_{OLS} = \beta + \frac{\rho}{1+\pi_1^2} \approx 1.3$. The sample estimate is quite close to this value:

```
cov(x, y) / var(x)
```

```
[1] 1.305881
```

Solution

Part 4

A regression of `y` on `z` is called the "reduced form." The slope coefficient from this regression is given by

$$\begin{aligned}\gamma_1 \equiv \frac{\text{Cov}(Z,Y)}{\text{Var}(Z)} &= \frac{\text{Cov}(Z,\alpha+\beta X+U)}{\text{Var}(Z)} = \frac{\beta\text{Cov}(Z,X)}{\text{Var}(Z)} \\ &= \frac{\beta\text{Cov}(Z,\pi_0+\pi_1 Z+V)}{\text{Var}(Z)} = \frac{\beta\pi_1\text{Var}(Z)}{\text{Var}(Z)} = \beta\pi_1\end{aligned}$$

since $\text{Cov}(Z,U) = \text{Cov}(Z,V) = 0$. Here $\beta\pi_1 = 1 \times 0.8 = 0.8$. The sample estimate is again quite close to this value:

```
cov(z, y) / var(z)
```

```
[1] 0.8023346
```

Solution

Parts 5-6

OLS is far from the truth; IV is quite close; the regression of `y` on `x` and `z` gives the worst result of all!

```
c(truth = 1,
  OLS = cov(x, y) / var(x),
  IV = cov(z, y) / cov(z, x),
  OLS_x_z = unname(coef(lm(y ~ x + z))[2]))
```

```
   truth      OLS       IV  OLS_x_z
1.000000 1.305881 1.013303 1.487189
```

The result for `OLS_x_z` can be explained as follows. Intuitively, controlling for Z in the regression of Y on X "removes" the variation in X that comes from Z. But since X is endogenous and Z is exogenous, this means we've removed the **good** variation in X and left behind only the **bad** variation, resulting in an estimate that is **worse** than the one we obtained from OLS.

Let's try to make this intuition a bit more precise. In the population, the coefficient on `x` in the regression of `y` on `x` and `z` is equal to

$$\frac{\text{Cov}(\tilde{X},Y)}{\text{Var}(\tilde{X})}$$

where $\tilde{X}$ is the *residual* when X is regressed on Z and a constant. But we already have a name for this residual: it is $V = X - \pi_0 - \pi_1 Z$. Therefore, since $Y = \alpha + \beta X + U$ and $X = \pi_0 + \pi_1 Z + V$, we have

$$\begin{aligned}\tilde{\beta} \equiv \frac{\text{Cov}(\tilde{X},Y)}{\text{Var}(\tilde{X})} &= \frac{\text{Cov}(V,Y)}{\text{Var}(V)} = \frac{\text{Cov}(V,\alpha+\beta X+U)}{\text{Var}(V)} \\ &= \frac{\text{Cov}(U,V) + \beta\text{Cov}(V,\pi_0+\pi_1 Z+V)}{\text{Var}(V)} \\ &= \frac{\text{Cov}(U,V)}{\text{Var}(V)} + \beta\end{aligned}$$

since $\text{Cov}(Z,V) = 0$ by construction. In our example, $\text{Cov}(U,V)/\text{Var}(V) = 0.5$ and $\beta = 1$ so the result should be 1.5. This is almost *exactly* equal to the estimate we obtained above. Returning to the general case, we can say a bit more here. Recall that

$$\beta_{OLS} = \beta + \frac{\text{Cov}(X,U)}{\text{Var}(X)} = \beta + \frac{\text{Cov}(V,U)}{\text{Var}(X)}.$$

Therefore, the difference between β_{OLS} and $\tilde{\beta}$ comes down to the difference between $\text{Var}(X)$ and $\text{Var}(V)$. But since $X = \pi_0 + \pi_1 Z + V$, we know that $\text{Var}(X)$ must be *larger* than $\text{Var}(V)$. This implies that $|\tilde{\beta} - \beta|$ must be *larger* than $|\beta_{OLS} - \beta|$, just as we found in the numerical example above.

Part 7

In short: for **prediction** use OLS; for **causal inference** use IV. At greater length: there are two issues here. The first is less important, so let's get it out of the way. It can be shown that the OLS estimator has a *lower standard error* than the IV estimator. In other words, IV is a less precise estimator. Of course the real question is: "a less precise estimate of *what*?" If the IV assumptions are satisfied, $\hat{\beta}_I V$ is an estimate of the true causal effect β whereas $\hat{\beta}_{OLS}$ is an estimate of the population linear regression slope. This is the key distinction. To make it clearer, lets put the issue of estimator precision to one side. Suppose I gave you not the estimated OLS and IV coefficients, but the *population* parameters that they estimate. Which should you use to predict Y?

The population linear regression coefficients $(\alpha_{OLS}, \beta_{OLS})$ are the solutions to

$$\min_{a,b} \mathbb{E}[(Y - a - bX)^2].$$

In other words, $(\alpha_{OLS}, \beta_{IV})$ are *by definition* the coefficients that give the best linear prediction of Y based on X, where "best" is defined to mean minimum mean squared error. This tells us everything we need to know. Since IV does *not* solve this optimization problem, it must in general give *worse* predictions of Y.

We can verify this using our simulated data as follows. In the simulation design, the population parameters that IV consistently estimates are $(\alpha = -0.3, \beta = 1)$. Above we showed that $\beta_{OLS} = \beta + \frac{\rho}{1+\pi_1^2}$. We calculate α_{OLS} as follows:

$$\begin{aligned}\alpha_{OLS} &= \mathbb{E}(Y) - \beta_{OLS}\mathbb{E}(X) = [\alpha + \beta\mathbb{E}(X)] - \beta_{OLS}\mathbb{E}(X) \\ &= \alpha + (\beta - \beta_{OLS})\mathbb{E}(X) \\ &= \alpha - (\beta_{OLS} - \beta)[\pi_0 + \pi_1\mathbb{E}(Z)] \\ &= \alpha - \left(\frac{\rho}{1+\pi_1^2}\right)\pi_0\end{aligned}$$

since $\mathbb{E}(Z) = 0$ in the simulation. Using the parameter values from above,

```
alpha <- -0.3
beta <- 1
pi0 <- 0.5
pi1 <- 0.8
rho <- 0.5
beta_OLS <- beta + rho / (1 + pi1^2)
alpha_OLS <- alpha - rho * pi0 / (1 + pi1^2)
c(alpha = alpha, beta = beta, alpha_OLS = alpha_OLS, beta_OLS = beta_OLS)
```

```
    alpha      beta alpha_OLS  beta_OLS
-0.300000  1.000000 -0.452439  1.304878
```

And, indeed, the OLS and IV coefficients agree with the estimated values:

```
coef(AER::ivreg(y ~ x | z))
```

```
(Intercept)           x
 -0.2847239   1.0133034
```

```
coef(lm(y ~ x))
```

```
(Intercept)           x
 -0.4312689   1.3058811
```

Now we can use the population IV and OLS coefficients, respectively, to approximate the predictive mean-squared error:

```
c(IV = mean((y - alpha - beta * x)^2),
  OLS = mean((y - alpha_OLS - beta_OLS * x)^2))
```

```
       IV       OLS
0.9624677 0.8118119
```

So we see that, indeed, OLS gives more accurate estimates. Here's yet another way to think about this. For learning the causal effect β, it's a *problem* that X is correlated with U. For *prediction*, on the other hand, it's *good* that X is correlated with U. The OLS slope "picks" up some of the effect of U because of its correlation with X and this improves our predictions of Y.

Exercise B - (20 min)

1. Set the seed to `1234` and then generate 10000 draws of (Z_1, Z_2, W) from a trivariate standard normal distribution in which each pair of RVs has correlation 0.3. Then generate (U, V) independently of (Z_1, Z_2, W) as in Exercise A above.
2. Use the draws you made in the preceding part to generate `x` and `y` according to the IV model from above with coefficients $(\pi_0, \pi_1, \pi_2, \pi_3) = (0.5, 0.2, -0.15, 0.25)$ for the first-stage and $(\beta_0, \beta_1, \beta_2) = (-0.3, 1, -0.7)$ for the causal model.
3. Run TSLS "by hand" by carrying out two regressions with `lm()`. Compare your estimated coefficients and standard errors to those from `AER::ivreg()`.
4. Run TSLS "by hand" but this time omit `w` from your first-stage regression, including it only in your second-stage regression. What happens? Why?
5. What happens if you drop Z_1 from your TSLS regression in `ivreg()`? Explain.
6. What happens if you omit `w` from *both* your first-stage and causal model formulas in `ivreg()`? Are there any situations in which this would work? Explain.

Solution

Parts 1-2

```
set.seed(1234)
n <- 10000

R_zw <- matrix(c(1, 0.3, 0.3,
                 0.3, 1, 0.3,
                 0.3, 0.3, 1), 3, 3, byrow = TRUE)
zw <- rmvnorm(n, sigma = R_zw)
z1 <- zw[,1]
z2 <- zw[,2]
w <- zw[,3]

R_uv <- matrix(c(1, 0.5,
                0.5, 1), 2, 2, byrow = TRUE)
errors <- rmvnorm(n, sigma = R_uv)
u <- errors[,1]
v <- errors[,2]

x <- 0.5 + 0.2 * z1 - 0.15 * z2 + 0.25 * w + v
y <- -0.3 + x - 0.7 * w + u
```

Solution

Part 3

```
# TSLS "by hand"
first_stage <- lm(x ~ z1 + z2 + w)
xhat <- fitted.values(first_stage)
second_stage <- lm(y ~ xhat + w)

# TSLS using AER::ivreg
tsls <- AER::ivreg(y ~ x + w | z1 + z2 + w)

library(broom)
library(tidyverse)

tidy(second_stage) |>
  knitr::kable(digits = 2, caption = 'Second Stage')
```

Second Stage

term	estimate	std.error	statistic	p.value
(Intercept)	-0.32	0.05	-7.11	0
xhat	1.06	0.08	12.67	0
w	-0.71	0.03	-24.54	0

```
tidy(tsls) |>
  knitr::kable(digits = 2, caption = 'TSLS Results')
```

TSLS Results

term	estimate	std.error	statistic	p.value
(Intercept)	-0.32	0.03	-12.66	0
x	1.06	0.05	22.56	0
w	-0.71	0.02	-43.70	0

Solution

Part 4

```
first_stage <- lm(x ~ z1 + z2)
xhat <- fitted.values(first_stage)
second_stage <- lm(y ~ xhat + w)
coef(second_stage)
```

```
(Intercept)        xhat           w 
 -0.2171103   0.8459799  -0.4675632 
```

This doesn't work: when we included `w` in the first-stage, our point estimates were very close to the truth, but now they're noticeably incorrect. To understand why, consider the population linear regression of X on Z_1, Z_2, and W, namely

$$X = \pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + \pi_3 W + V = \tilde{X} + V.$$

By construction, V is uncorrelated with (Z_1, Z_2, W). So when we substitute $X = \tilde{X} + V$ into our causal model, following the TSLS logic, we obtain

$$\begin{aligned} Y &= \beta_0 + \beta_1(\tilde{X} + V) + \beta_2 W + U \\ &= \beta_0 + \beta_1 \tilde{X} + \beta_2 W + (U + \beta_1 V) \\ &= \beta_0 + \beta_1 \tilde{X} + \beta_2 W + \epsilon. \end{aligned}$$

In this regression, both $\tilde{X}$ and W are uncorrelated with ϵ. This follows because $\tilde{X}$ is just a linear function of (Z_1, Z_2, W) while ϵ is a linear function of U and V. By assumption (Z_1, Z_2, W) are uncorrelated with U and by construction they are uncorrelated with V.

Now consider an *alternative* first-stage regression, one that excludes W:

$$X = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \eta = X^* + \eta.$$

By construction (Z_1, Z_2) are uncorrelated with η but the same *is not necessarily true* of W, since it was not included in the regression. Accordingly, when we substitute into our causal model, we obtain

$$\begin{aligned} Y &= \beta_0 + \beta_1(X^* + \eta) + \beta_2 W + U \\ &= \beta_0 + \beta_1 X^* + \beta_2 W + (U + \beta_1 \eta) \\ &= \beta_0 + \beta_1 \tilde{X} + \beta_2 W + \nu. \end{aligned}$$

Since ν includes η and η will in general be correlated with W, the regressor W is *endogenous* in this regression. Since W is endogenous, all the coefficients are messed up!

Solution

Part 5

```
AER::ivreg(y ~ x + w | z2 + w) |>
  tidy() |>
  knitr::kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.30	0.05	-5.65	0
x	1.01	0.10	9.76	0
w	-0.70	0.03	-22.94	0

This works fine: both Z_1 and Z_2 are valid and relevant instruments, but we only have one endogenous regressor so we can include either of them or both. The point estimates are similar but slightly different. The standard errors have increased, because our first-stage now picks up less of the exogenous variation in X. (We "leave behind" the endogenous variation that is due to Z_1.)

Solution

Part 6

```
AER::ivreg(y ~ x | z1 + z2) |>
  tidy() |>
  knitr::kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.01	0.03	-0.22	0.83
x	0.43	0.06	7.35	0.00

This doesn't work. Excluding W gives a causal model with a "new" error term:

$$Y = \beta_0 + \beta_1 X + \tilde{U}, \quad \tilde{U} \equiv U + \beta_2 W.$$

Now, (Z_1, Z_2) are still relevant instruments because they're correlated with X. The question is whether they remain *exogenous* instruments in the causal model with error term $\tilde{U}$. By assumption $\text{Cov}(Z_1, U) = \text{Cov}(Z_2, U) = 0$ so the relevant question is whether the instruments are correlated with W. If not, they are exogenous in the causal model with error term $\tilde{U}$. In our simulation, however, Z_1 and Z_2 are both correlated with W. This is why our TSLS estimates from above are so obviously wrong.

Exercise C - (10 min)

In this exercise you will need to work with the columns `critics`, `order`, and `ranking` from the `qe` tibble.

1. Add a variable called `first` to `qe` that takes on the value `TRUE` if `order` equals one. You will need this variable in the following parts.
2. Do musicians who perform *first* receive different average rankings from the jury than other musicians? Discuss briefly.
3. Is it possible to estimate the causal effect of performing *first* on subsequent ratings by critics using this dataset? If so how and what is the effect?
4. Estimate the causal effect of ranking in the competition on future success as measured by critics' rating two ways: via OLS and via IV using `first` to instrument for ranking. Discuss your findings.

Solution

Setup

```
data_url <- 'https://ditraglia.com/data/Ginsburgh-van-Ours-2003.csv'
```

```r
qe <- read_csv(data_url)
```

Parts 1-2

Yes: musicians who perform first tend to have rankings that are around 3.4 *higher*. In other words, the judges rate them *worse*. This difference is large and highly statistically significant. Because performance order is random, this may suggest some kind of systematic bias in judging. An alternative explanation is that performing earlier in the competition *causes* participants to perform worse. For our IV exercise, it doesn't actually matter whether order has an effect on rankings because of judge bias or actual difference in performance. What matters is that, performance order is independent of talent *by definition*. It also seems quite plausible that performance order does not have a direct effect of its own on future career success. In other words, holding constant how highly a person is ranked in the competition, we would not expect people who perform earlier to for some reason have more success later in their career. I would say that this is an *unusually plausible* instrumental variable!

```r
qe <- qe |>
  mutate(first = order == 1)

tidy_me <- function(results, mytitle) {
  # Helper function for making little tables in this solution
  results |>
    tidy() |>
    select(term, estimate, std.error) |>
    knitr::kable(digits = 2, caption = mytitle)
}

lm(ranking ~ first, qe) |>
  tidy_me('First stage')
```

First stage

term	estimate	std.error
(Intercept)	6.21	0.30
firstTRUE	3.42	1.05

Solution

Part 3

Yes: since performance order is random assigned, whether a musician performs *first* is unrelated to any other observed or unobserved characteristics. We can estimate this effect via the reduced form regression of `ranking` on `first`. We estimate that performers who appear first are subsequently rated about 8 points lower by critics.

```r
lm(critics ~ first, qe) |>
  tidy_me('Reduced Form')
```

Reduced Form

term	estimate	std.error
(Intercept)	14.31	1.08
firstTRUE	-7.94	3.74

This effect is statistically significant, but it's hard to interpret the magnitude. To get a better sense of the meaning of "8 points lower" we can center and standardize `critics` using the base R function `scale`:

```r
lm(scale(critics) ~ first, qe) |>
  tidy_me('Reduced Form - Standardized Outcome')
```

Reduced Form - Standardized Outcome

term	estimate	std.error
(Intercept)	0.05	0.09
firstTRUE	-0.66	0.31

We see that musicians who perform first are rated about two-thirds of a standard deviation *lower* than musicians who do not perform first. This is a moderately large effect size.

Solution

Part 4

I will compute these results two different ways: first using the "raw" outcome `critics` for comparability with the paper, and second using the centered and scaled version of the same to aid interpretation. The unscaled results are as follows:

```r
ols <- lm(critics ~ ranking, qe)
iv <- AER::ivreg(critics ~ ranking | first, data = qe)

tidy_me(ols, 'OLS results')
```

OLS results

term	estimate	std.error
(Intercept)	23.23	2.03
ranking	-1.47	0.28

```r
tidy_me(iv, 'IV results')
```

IV results

term	estimate	std.error
(Intercept)	28.73	6.86
ranking	-2.32	1.04

These estimates match the values in Table 3 of the paper after flipping the signs: a *lower* ranking means a musician has performed *better* relative to the other in the judges eyes. The scaled results are as follows:

```r
ols_scaled <- lm(scale(critics) ~ ranking, qe)
iv_scaled <- AER::ivreg(scale(critics) ~ ranking | first, data = qe)

tidy_me(ols_scaled, 'OLS results - standardized outcome')
```

OLS results - standardized outcome

term	estimate	std.error
(Intercept)	0.80	0.17
ranking	-0.12	0.02

```r
tidy_me(iv_scaled, 'IV results - standardized outcome')
```

IV results - standardized outcome

term	estimate	std.error
(Intercept)	1.25	0.57
ranking	-0.19	0.09

The OLS and IV results are qualitatively similar but the IV estimate is larger. Being ranked one place *higher*, i.e. having a lower numerical ranking, causes approximately a 0.2 standard deviation increase in critics' scores.