Optional Reading: IV and 2SLS

(Jing Li, Miami University)

1. The takeaway of FW theorem is: do not eat a whole fish, eat fish body only. That means, only part of x can be used to infer its causal effect on y. When we run a multiple regression, the fish body is the residual of step-one regression, which captures the part of x that has nothing to do with w

2. But what if we have no data for w? Then controlling for w with a multiple regression is not doable. For instance, w can be school district. We believe school district matters for house price, and school district and baths are related (houses in good district tend to be nicer and have more bathrooms). So ignoring school district will result in bias. However, house data does not have that district variable

3. Situations like school district are very common—we want to control for confounders that are unobserved (no data). One solution is using instrumental variable (IV) and two-stage-least-square (2SLS) estimator

4. Recall the formula implied by FW theorem

$$\hat{\beta}_1^{multiple\ regression} = \frac{\sum \hat{r}_i y_i}{\sum \hat{r}_i^2} \tag{1}$$

   see `https://www.fsb.miamioh.edu/lij14/311r_or_fw.pdf` for proof. The key insight is that the step-one-regression-residual $\hat{r}$ is only part of x_1—it represents the fish body, and lies outside the w or x_2 circle, see Venn diagram on page 18 of this note `https://www.fsb.miamioh.edu/lij14/311r_slide_ch4n.pdf`. But when x_2 has no data, we cannot get $\hat{r}$ via step-one regression. IV can help!

5. Let z denote the IV. 2SLS estimator aims to use z to extract a portion of fish body. The 2SLS formula is

$$\hat{\beta}_1^{2SLS} = \frac{\sum \hat{x}_i y_i}{\sum \hat{x}_i^2} \tag{2}$$

   where $\hat{x}$ is the fitted value of regressing x onto z

$$\hat{x} = \hat{\alpha}_0 + \hat{\alpha}_1 z \qquad (first\ stage\ regression) \tag{3}$$

   Formula (2) and formula (1) look similar. That is not coincidence, since 2SLS makes

use of fish body (the good part of fish) as well

6. So what is IV? It should satisfy three conditions:

   (a) It is correlated with x, $cov(z, x) \neq 0$, (relevance condition)

   (b) It is uncorrelated with w, $cov(z, w) = 0$, (exogeneity condition)

   (c) It has no direct effect on y, (exclusion condition)

7. From (2) it is obvious that we can obtain 2SLS results through a two-step procedure, which is different from the two-step procedure of FW theorem

   - Step I: run the first stage regression (3), and save fitted value $\hat{x}$
   - Step II: regressing y onto that fitted value $\hat{x}$ (second stage)

   To see why that works, consider the structural model

$$y_i = \beta_1 x_i + u_i \tag{4}$$

   where the error term u contains the unobserved confounder w. Multiply each term by $\hat{x}_i$ and do summation. We will get

$$\sum \hat{x}_i y_i = \beta_1 \sum \hat{x}_i x_i + \sum \hat{x}_i u_i \tag{5}$$

   The last term $\sum \hat{x}_i u_i$ converges to zero because of exogeneity condition $cov(w, z) = 0$. Then we can solve for β_1

$$\beta_1 = \frac{\sum \hat{x}_i y_i}{\sum \hat{x}_i x_i} = \frac{\sum \hat{x}_i y_i}{\sum \hat{x}_i^2} \tag{6}$$

   The last fraction justifies using $\hat{x}$ to predict y, i.e., running the second stage regression.

8. Next we use simulation as an illustration

```
> set.seed(12345)
> n = 1000
> w = rnorm(n)
> z = rnorm(n)
> x = w + z + rnorm(n)  # z+rnorm(n) is fish body or exogenous part
> y = 2*x - 5*w + rnorm(n)
```

```
> simple.reg = lm(y~x)
> coef(simple.reg)
(Intercept)           x
-0.2316563    0.2592358

> multiple.reg = lm(y~x+w)
> coef(multiple.reg)
(Intercept)           x           w
0.01673683   1.98788464 -5.01317305
```

(a) The blue line states that x depends on both w and z. So $cov(w, x) \neq 0$, and $cov(z, x) \neq 0$. That latter implies that relevance condition is met.

(b) The orange line states that y depends on x and w. However, z is absent in the orange line. Thus, z has no direct effect on y, and the exclusion condition is met.

(c) Also notice that z and w are generated independently, so exogeneity condition is met

(d) The estimated slope in simple regression 0.2592358 (in red) is far away from the true causal effect 2 (see orange line), so is biased. That bias is attributed to the confounder w.

(e) The estimated slope in multiple regression 1.98788464 (in green) is close to the true causal effect 2, so is unbiased. OVB disappears because the multiple regression controls for w.

9. Now, pretend the data of w is unavailable. As a result, it becomes infeasible to fit the multiple regression. In that case we need to run the first stage regression (3), save the fitted value, and uses that fitted value to predict y

```
> first.stage = lm(x~z)
> xhat = fitted(first.stage)
> second.stage = lm(y~xhat)
> coef(second.stage)
(Intercept)        xhat
-0.2165239    1.8168767
```

The slope coefficient estimated by 2SLS is 1.8168767, close to the true causal effect 2. 2SLS works! Good news!

10. In practice, it is better to use ivreg function in the AER package

```
> library(AER)
> iv.model = ivreg(y~x|z)
> summary(iv.model)$coef
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.2165239  0.1559854 -1.388104 1.654153e-01
x            1.8168767  0.1499046 12.120223 1.220129e-31
```

11. The downside of IV estimation is that, typically, the t value from IV estimation is less than that from the multiple regression, see results below, and compare two orange t values $12.120223 < 91.3376628$.

```
> summary(multiple.reg)$coef
               Estimate Std. Error      t value  Pr(>|t|)
(Intercept)  0.01673683 0.03019406    0.5543088 0.5794918
x            1.98788464 0.02176413   91.3376628 0.0000000
w           -5.01317305 0.03756947 -133.4374194 0.0000000
```

This is because the fish body used by multiple regression is $z + rnorm(n)$, whereas the fish body used by IV is only z. Put differently, IV only uses portion of fish body shown in the Venn diagram

12. We can introduce more terminology. $z + rnorm(n)$ is called the exogenous part of x, while w is endogenous part. The exogenous part has nothing to do with w, so is useful to disentangle the causal effect of x on y from the causal effect of w on y. The process of locating the exogenous part is called identification. A researcher should explain where the exogenous variation come from. IV is just one method to obtain the exogenous variation.

13. There is an alternative way to use IV. We can show

$$cov(y, z) = cov(\beta_1 x + u, z) = \beta_1 cov(x, z) \tag{7}$$

since $cov(u, z) = cov(w, z) = 0$. It follows that

$$\hat{\beta}_1^{IV} = \frac{cov(y, z)}{cov(x, z)} = \frac{cov(y, z)/var(z)}{cov(x, z)/var(z)} = \frac{reduced\ form\ slope}{first\ stage\ slope} \tag{8}$$

```
> first.stage = coef(lm(x~z))
> reduced.form = coef(lm(y~z))
> reduced.form[2]/first.stage[2]
1.816877
```

The estimate 1.816877 is the same as 2SLS.

14. From (8), it is clear that we will run into trouble if $cov(x, z) \approx 0$. That situation is called weak IV. Weak IV is especially problematic when the exogeneity condition is violated $cov(u, z) \neq 0$. We can show the IV estimate is biased in that case, and the bias can be inflated substantially by $cov(x, z) \approx 0$.

$$\hat{\beta}_1^{IV} = \beta_1 + \frac{cov(u, z)}{cov(x, z)} \tag{9}$$

15. Next we redo the simulation to demonstrate the issue of weak IV and failed exogeneity. The changes are in red

```
> set.seed(12345)
> n = 1000
> z = rnorm(n)
> w = z + rnorm(n)
> x = w + 0.01*z + rnorm(n)
> y = 2*x - 5*w + rnorm(n)
> iv.model = ivreg(y~x|z)
> summary(iv.model)$coef
              Estimate Std. Error     t value     Pr(>|t|)
(Intercept) -0.1093493  0.1543973  -0.7082333 4.789657e-01
x           -3.0326412  0.1490719 -20.3434780 3.077905e-77

> coef(lm(y~x))
(Intercept)           x
-0.09516661 -1.49987102
```

Now $cov(w, z) \neq 0$, so exogeneity is violated. Moreover, $cov(x, z) \approx 0.01 \approx 0$, so IV is weak. The IV estimate -3.0326412 is more biased than the simple regression estimate -1.49987102 since the former is further away from the true value 2.

16. In some sense, the consequence of weak IV is similar to multicollinearity—the fish body is tiny, and estimation becomes imprecise

17. More valid IVs can help

```
> set.seed(12345)
> n = 1000
> w = rnorm(n)
> z1 = rnorm(n)
> z2 = rnorm(n)
> x = w + z1 + 3*z2 + rnorm(n)
> y = 2*x - 5*w + rnorm(n)

> summary(ivreg(y~x|z1))$coef
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.2467107  0.1562325 -1.579125 1.146242e-01
x            1.7689401  0.1513297 11.689313 1.123921e-29

> summary(ivreg(y~x|z1+z2))$coef
              Estimate Std. Error   t value      Pr(>|t|)
(Intercept) -0.2373454 0.16094140 -1.474732  1.406001e-01
x            1.9793483 0.05344686 37.033951 1.354828e-189
```

    The results from using two IVs (in orange) are better than using just one IV (in blue). Intuitively, that is because a bigger portion of fish body is used for estimation.

18. To sum up, valid IVs satisfying all three conditions should provide good estimate. However, if any of those condition fails, IV may produce result even worse than simple regression! Use IV with caution.

19. In a famous paper `https://davidcard.berkeley.edu/papers/geo_var_schooling.pdf`, the author wants to estimate the causal effect of going to college on future earning. The unobserved confounders are people's motivation, capability, family background, etc. The IV considered by the author is the distance (proximity) between a person's home and college. Can you argue that the three requirements (exogeneity, relevance, exclusion) are fulfilled by this IV?

20. In another famous research, people use weather as IV to investigate the causal effect of watching TV on developing autism. See https://www.wsj.com/articles/SB117131554110006323

21. The third example is that draft lottery serves as IV to show how serving in army affects earning. See https://economics.mit.edu/sites/default/files/publications/Angrist%201990%20-%20Lifetime%20Earnings%20and%20the%20Vietname%20.pdf