**Stat 305A: Linear Models** **Autumn 2019/20**

# Lecture 15: Regression discontinuity and instrumental variables

*Lecturer: Art B. Owen* *November 12*

**Disclaimer**: *These notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students who took stat 305A at Stanford University. They appear in* `http://statweb.stanford.edu/~owen/courses/305a/` *where you may find the most recent version. Stanford University holds the copyright.*

## 15.1 Regression discontinuity

Sometimes one variable $X_1$ is a step function of another variable $X_2$. For instance

$$X_1 = \begin{cases} 1, & X_2 \geqslant \tau \\ 0, & X_2 < \tau. \end{cases}$$

The classical example is that $X_1$ indicates that somebody gets a National Merit Scholar honor and $X_2$ is their score on the PSAT standardized test, and perhaps $\tau = 1700$. We can reason that candidates with $X_2 = 1700$ are not very different from those with $X_2 = 1699$ before the honors are assigned.

If $Y$ would be log income, then a regression of $Y$ on $X_1$ (comparing two groups) might show a difference but that could be due to other factors that predated the honor. If we compare $Y$ between groups with $X_2 = 1699$ and $X_2 = 1700$ then any difference is quite reasonably attributable to the honor, assuming $Y$ is measured after the honor is given.

There are not many people with $X_2 \in \{1699, 1700\}$, so we might expand those windows to $[1699 - \Delta, 1699]$ and $[1700, 1700 + \Delta]$ to get more data, for some $\Delta > 0$. If we make $\Delta$ larger then we start to blur the sharp distinction between the two groups. A common way to address this is via the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

Then there is a linear trend in $X_2$ with a step discontinuity at $X_{i2} = \tau$ and a coefficient $\beta_1$ that, if all goes well, estimates the causal effect of the honor right at the given $\tau$. The effect could certainly be different at $\tau = 1750$ or $\tau = 1650$ than it is at $\tau$. We get what we get.

If the underlying linear model in $X_{i2}$ is not correct, then we may get a biased estimate of the causal impact at the $\tau$ in use. We could use a more flexible model such as a quadratic or cubic. An example in class based on work by Gelman and Zelizer provides a cautionary tale. There a cubic regression plus a discontinuity produced a completely implausible estimate of 2.5 billions years of life lost by some policy related to coal consumption north of the Huai river. Outliers or noise or lack of fit in the regression far from the cutoff point can greatly affect the estimate of the causal effect at the cutoff point.

Another famous regression discontinuity example is from a study of classroom size in Israel. Comparing educational outcomes between small and large classes could be misleading because we might not know why some classes were made small and others large. In Israel there was a policy that no class should have more than 40 elementary students. We can then compare classes of size 39 or 40 to the smaller classes that were formed when $> 40$ students would have otherwise been in a class. Note that we don't compare to all smaller classes, just the ones that were formed for this reason.

© Stanford University 2019

## 15.2 Tie breaker designs

In a tie breaker design there are three groups. If the running variable (what we called $X_2$ above) is large the subject gets the treatment. If it is small, they don't. In between it is random. For instance in Nebraska high scoring high school students got a scholarship, low scoring did not, and there was a middle range where it was random. (!!!!)

## 15.3 Instrumental variables

Suppose that in our basic model

$$Y = Z\beta + \varepsilon$$

it turns out that $\varepsilon$ is correlated with $Z$. That is very disappointing because we assumed they were uncorrelated. However, $\varepsilon$ captures the effect of unmeasured variables and those variables have every right to correlate how they want. Our model is not their constraint. Remember, we just made it up. There is a notion attributed to Meehl (from the 1960s) that everything correlates with everything. Nonzero correlations are to be expected and are actually not that interesting. A zero would be interesting.

Economists like to study the returns to years of education. Suppose that $Y$ is log income and $S$ is years in school. We could fit a regression like

$$Y = \beta_0 + \beta_1 S + \varepsilon.$$

However what we are calling $\varepsilon$ probably includes variables correlated with $S$ and that will bias $\beta_1$ away from what we want. For sake of argument suppose that there is but one missing variable and we call it 'ability'. The same students who stayed longer in school would have actually earned some amount more anyhow without those years in school. Think of the regression

$$Y = \beta_0^* + \beta_1^* S + \beta_2^* A + \varepsilon^*,$$

where $\varepsilon^*$ really is uncorrelated with $S$ and $A$. Then we actually want $\beta_1^*$, not $\beta_1$. We have no chance to reliably measure $A_i$, whatever that is, on any of the people in our data set. For one thing, by the time we see the income data, the school days are long past and ability may change over time. Also what we call ability might be only one of many relevant quantities.

This is where an instrument comes in. Suppose that $W$ is a random variable that satisfies two conditions:

1. $\mathrm{cor}(W, S) \neq 0$, and
2. $\mathrm{cor}(W, A) = \mathrm{cor}(W, \varepsilon^*) = 0$.

Condition 1 is easy to satisfy. Condition 2 is harder because, ordinarily, everything correlates with everything. So $W$ must be special in an interesting way.

We can work out that

$$\mathrm{cov}(Y, W) = \mathrm{cov}(\beta_0^* + \beta_1^* S + \beta_2^* A + \varepsilon^*, W) = \beta_1^* \mathrm{cov}(S, W).$$

That means that

$$\beta_1^* = \frac{\mathrm{cov}(Y, W)}{\mathrm{cov}(S, W)}$$

and we don't even need $A$ (just our assumptions). The instrumental variables (IV) estimate of $\hat{\beta}_1^*$ is

$$\hat{\beta}_{\mathrm{IV}} = \frac{\widehat{\mathrm{cov}}(Y, W)}{\widehat{\mathrm{cov}}(S, W)} = \frac{SYW}{SSW} = \frac{SYW/SWW}{SSW/SWW} = \frac{\hat{\beta}_{W\to Y}}{\hat{\beta}_{W\to S}}$$

| Response $Y$ | Predictor $X$ | Putative Instrument $W$ |
|---|---|---|
| Earnings | Veteran status | Vietnam era draft lottery number |
| Autism | Early childhood TV watching | # Rainy days (by state) |
| Health | Cardiac treatment | Distance to cardiac care center |
| Birth weight | Maternal smoking | State cigarette tax |
| Test scores | Attendance at charter school | Charter school lottery |

Table 15.1: Some suggested instruments. In class I said that I did not think the one about TV→autism was a good one.

a ratio of the regression slopes we would get regressing $Y$ and $S$ respectively on $W$.

**Warning:** denominators are problematic if they get close to zero. If $\hat{\beta}_{W\to S}$ is not large compared to its standard error from that regression, then $\hat{\beta}_{\mathrm{IV}}$ is poorly determined.

Using the law of large numbers we get

$$\hat{\beta}_{\mathrm{IV}} \to \beta_1^*$$

as long as the true $\beta_{W\to S}$ is not zero.

## 15.4 Instrumental examples

Suppose that $Y$ is a numerical response, that we have on $n$ people. Let $X$ be a treatment 0 or 1 on them, and $W$ is a randomized outcome equal to 0 or 1. Respecting the toss, $X = W$. This $W$ easily satisfies both criteria for an instrument. (Check them.) Then as a ratio of regression coefficients,

$$\hat{\beta}_{\mathrm{IV}} = \frac{\sum_i Y_i(W_i - \bar{W})}{\sum_i W_i(W_i - \bar{W})} \Big/ \frac{\sum_i X_i(W_i - \bar{W})}{\sum_i W_i(W_i - \bar{W})} = \frac{\sum_i Y_i(X_i - \bar{X})}{\sum_i X_i(X_i - \bar{X})}$$

because $W_i = X_i$. Now suppose that there are $n_1$ subjects with $X_i = 1$ and $n_0$ with $X_i = 0$, and to rule out uninteresting complications, $\min(n_0, n_1) > 0$. Then after some algebra

$$\hat{\beta}_{\mathrm{IV}} = \bar{Y}_1 - \bar{Y}_0$$

the difference between the mean responses in the two groups.

I presented an example from Angrist and Pischke (Mostly Harmless Econometrics) relating earnings in the 1980s to schooling in the 1930s using calendar quarter as an instrument. Students born early in the calendar year could reach age 16 (the age when they can quit school) earlier than those born late in the calendar year. Birth month might not really be a good instrument because young vs older students may get different schooling experiences especially in the early grades.

In class we considered some examples in Table 15.1.

## 15.5 Multiple instruments

Now we consider $k \geqslant 1$ predictors and $j \geqslant k$ instruments. Here is how the algebra works out. Suppose that

$$Y = Z\beta + \varepsilon, \quad Z \in \mathbb{R}^{n\times k}, \quad W \in \mathbb{R}^{n\times j}.$$

The variables in $W$ are uncorrelated with $\varepsilon$ but correlated with $Z$ such that $\mathrm{cor}(Z, W)$ has full rank $k$.

We begin by choosing some matrix $R \in \mathbb{R}^{j\times k}$ of rank $k$. (More later). Then

$$R^\mathsf{T} W^\mathsf{T} Y = R^\mathsf{T} W^\mathsf{T} Z\beta + R^\mathsf{T} W^\mathsf{T}\varepsilon.$$

For large $n$ by the LLN and our assumption on $W$,

$$R^\mathsf{T} W^\mathsf{T} Y \doteq R^\mathsf{T} W^\mathsf{T} Z\beta.$$

So

$$\beta \doteq (R^\mathsf{T} W^\mathsf{T} Z)^{-1} R^\mathsf{T} W^\mathsf{T} Y.$$

If $j = k$ then $R$ is invertible and we get

$$\beta \doteq \hat\beta_{\mathrm{IV}} = (W^\mathsf{T} Z)^{-1} W^\mathsf{T} Y,$$

so $R$ doesn't actually matter. Comparing this to the OLS estimator

$$\hat\beta = (Z^\mathsf{T} Z)^{-1} Z^\mathsf{T} Y,$$

we see that the instruments $W$ have taken the place of some $Z$s.

We can make the formula look like the one that was a ratio of regression estimates as follows:

$$\beta \doteq ((W^\mathsf{T} W)^{-1} W^\mathsf{T} Z)^{-1} (W^\mathsf{T} W)^{-1} W^\mathsf{T} Y.$$

Note that $(W^\mathsf{T} W)^{-1} W^\mathsf{T} Y \in \mathbb{R}^k$ while $(W^\mathsf{T} W)^{-1} W^\mathsf{T} Z \in \mathbb{R}^{k\times k}$ (because we are in the $j = k$ case).

If $j > k$ things are more complicated. Suppose we have 10 instruments for 5 predictors. We could use any 5 instruments we like. Or we could use any 5 linear combinations of instruments that we like, each one a column of $R$. Chapter 4 of an econometrics book by McFadden works out what would be the best $R$ in that case, what the resulting variance matrix of $\hat\beta_{\mathrm{IV}}$ would be and how to estimate that best $R$.

The $j = k = 1$ case is the most important one for applications.