IV Estimation: Strangetown

Eyayaw Beze

July 11, 2020

Let y_i be a measure of good health, and D_i an indicator for smoking

$$D_i = \begin{cases} 1, & if\ i\ smokes \\ 0, & otherwise \end{cases}$$

$$Z_i = \begin{cases} 1, & if\ i \text{ received a pack of cigarettes} \\ 0, & otherwise \end{cases}$$

The true model reads:

$$y_i = \alpha + \beta_1^{\text{true}} D_i + \beta_2 w_i + \epsilon_i \tag{1}$$

with w_i denoting individual income. But since you do not observe income you can only estimate the following model:

$$y_i = \alpha + \beta_1 D_i + u_i \tag{2}$$

a) Over- or under-estimate β_1^{true} if we don't observe wage in (1)? Show the inconsistency of the coefficient if we estimate (2) instead?

```
# data generating process
dgp <- function(n, y.bar, d, z) {
  sigma <- 1 / sqrt(n)
  y <- y.bar + sigma * scale(rnorm(n)) # with fixed mean and sd;
                                        # mean(y) == y.bar, sd(x) == sigma
  return(data.frame(y, D = rep(d, n), Z = rep(z, n)))
}

# to vectorize over n, y.bar, D, and Z
map_df <- function(..., f, binder = rbind) {
  return(as.data.frame(do.call(binder, Map(f, ...))))
}
```

```
n <- 70 ## sample size
# Di = Zi = 0
n00 <- 30
y00.bar <- 1.0

# Di = 0; Zi = 1
n01 <- 10
y01.bar <- 0.8
# Di = 1; Zi = 0
n10 <- 20
y10.bar <- 1.5
```

```r
# Di = Zi = 1
n11 <- 10
y11.bar <- 1.2

n.vec <- c(n00, n01, n10, n11)
y.bar <- c(y00.bar, y01.bar, y10.bar, y11.bar)
D <- c(0, 0, 1, 1)
Z <- c(0, 1, 0, 1)

set.seed(123) # for reproducibility
toy_data <- map_df(n.vec, y.bar, D, Z, f = dgp)

head(toy_data) # view 6 rows of the data
```

```
#>         y D Z
#> 1 0.90446 0 0
#> 2 0.96593 0 0
#> 3 1.29885 0 0
#> 4 1.02189 0 0
#> 5 1.03283 0 0
#> 6 1.32795 0 0
```

```r
subsets <- alist(
  D == 0 & Z == 0,
  D == 0 & Z == 1,
  D == 1 & Z == 0,
  D == 1 & Z == 1
)

# check whether the mean of the generated data matches the sample means
sapply(subsets, function(x) mean(toy_data[eval(x, toy_data), "y"]))
```

```
#> [1] 1.0 0.8 1.5 1.2
```

```r
y.bar
```

```
#> [1] 1.0 0.8 1.5 1.2
```

c) Calculate β_1^{OLS} that you obtain by estimating equation (2) by OLS. Interpret the coefficient.

$$
\begin{aligned}
\beta_1^{OLS} &= \frac{Cov(y_i, D_i)}{Var(D_i)} = \frac{E\left[y_i D_i\right] - E\left[y_i\right] E\left[D_i\right]}{E[D_i^2] - (E[D_i])^2} = \frac{E\left[y_i \mid D_i = 1\right] - E\left[y_i\right] E[D_i]}{E[D_i] - (E[D_i])^2} \\
&= \frac{\frac{(n_{10}\bar{y}_{10} + n_{11}\bar{y}_{11})}{n} - \frac{(n_{00}\bar{y}_{00} + n_{01}\bar{y}_{01} + n_{10}\bar{y}_{10} + n_{11}\bar{y}_{11})}{n} \frac{(n_{10}+n_{11})}{n}}{\left(\frac{n_{10}+n_{11}}{n}\right) - \left(\frac{n_{10}+n_{11}}{n}\right)^2} \\
&= \frac{(1/70)\left(20 \cdot 1.5 + 10 \cdot 1.2)\right) - \left[(1/70)\left(30 \cdot 1.0 + 10 \cdot 0.8 + 20 \cdot 1.5 + 10 \cdot 1.2)\right]\left[(1/70)(20+10)\right]}{\left(\frac{20+10}{70}\right) - \left(\frac{20+10}{70}\right)^2}
\end{aligned}
\tag{3}
$$

Note: $E[D_i] = p$ and $Var(D_i) = p(1-p)$ where p is the probability that D_i takes on 1—**since D_i is a Bernoulli random variable**. In our case, $Pr(D_i = 1) = p = \frac{n_{10}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}} = (20+10)/70 = 3/7$.

IV Estimation: LATE/Wald

By (3),

```
b_ols <- ((1 / n) * (n10 * y10.bar + n11 * y11.bar) -
  ((1 / n) * (n00 * y00.bar + n01 * y01.bar + n10 * y10.bar + n11 * y11.bar) * (n10 + n11) / n
  )) /
  ((n10 + n11) / n - ((n10 + n11) / n)^2)
```

$\hat{\beta}_1^{OLS} = 0.45$.

Or using the `covariance` and `variance` formula:

```
beta_ols <- with(toy_data, cov(y, D) / var(D))
alpha <- with(toy_data, mean(y) - beta_ols * mean(D))
```

$\hat{\beta}_1^{OLS} = 0.45$ and $\hat{\alpha} = 0.95$.

Or using `lm`—a linear model estimation workhorse in `R`:

```
lm(y ~ D, toy_data)$coefficients
```

```
#> (Intercept)           D
#>        0.95        0.45
```

(d) Calculate β_1^{IV} and discuss your result w.r.t. to the previous findings. Using The Wald Estimator:

$$\beta_1^{IV} = \frac{\mathbb{E}\left(Y_i \mid z_i = 1\right) - \mathbb{E}\left(Y_i \mid z_i = 0\right)}{\mathbb{E}\left(D_i \mid z_i = 1\right) - \mathbb{E}\left(D_i \mid z_i = 0\right)}$$

```
# Wald Estimator
with(
  toy_data,
  (mean(y[Z == 1]) - mean(y[Z == 0])) / (mean(D[Z == 1]) - mean(D[Z == 0]))
)
```

```
#> [1] -2
```

Using matrix notation:

$$\widehat{\boldsymbol{\beta}}_{IV} = \left[\mathbf{Z}'\mathbf{X}\right]^{-1}\left[\mathbf{Z}'y\right]$$

```
# the [1, 1, 1,...]' is instrumented with itself
with(toy_data, {
  constant <- rep(1, length(D))
  Z.mat <- cbind(constant, Z) # Z.mat = [[1, 1,..., 1]', Z]
  D.mat <- cbind(constant, D) # [[1, 1,..., 1]', D]
  solve(t(Z.mat) %*% D.mat) %*% t(Z.mat) %*% y
})
```

```
#>          [,1]
#> constant    2
#> D          -2
```

The Wald estimand of $\widehat{\beta}_{IV}$ can be interpreted as the effect of smoking on those whose treatment status can be changed by the instrument. The effect of smoking on health of those who smoked because they were given packs of cigarettes, but would not otherwise have smoked. This obviously excludes voluntary smokers and those who did not, but it includes smokers for whom receiving the cigarettes was important.

```r
# Or using AER package
library(AER)
library(stargazer)
model_iv <- ivreg(y ~ D | Z, data = toy_data)
stargazer(model_iv, header = FALSE)
```

Table 1:

	Dependent variable:
	y
D	-2.000
	(3.327)
Constant	2.000
	(1.434)
Observations	70
R^2	-12.293
Adjusted R^2	-12.489
Residual Std. Error	1.257 (df = 68)
Note:	*p<0.1; **p<0.05; ***p<0.01

- **Compliers**. The subpopulation with $d_{1i} = 1$ and $d_{0i} = 0$:
- **Always-takers**. The subpopulation with $d_{1i} = d_{0i} = 1$:
- **Never-takers**. The subpopulation with $d_{1i} = d_{0i} = 0$:

Using the exclusion restriction, we can define potential outcomes indexed solely against treatment status using the single-index ($\mathrm{Y}_{1i}, \mathrm{Y}_{0i}$) notation. In particular,

$$\begin{aligned} \mathrm{Y}_{1i} &\equiv \mathrm{Y}_i(1,1) = \mathrm{Y}_i(1,0) \\ \mathrm{Y}_{0i} &\equiv \mathrm{Y}_i(0,1) = \mathrm{Y}_i(0,0) \end{aligned}$$

The observed outcome, Y_i, can therefore be written in terms of potential outcomes as:

$$\begin{aligned} \mathrm{Y}_i &= \mathrm{Y}_i\left(0, \mathrm{Z}_i\right) + \left[\mathrm{Y}_i\left(1, \mathrm{Z}_i\right) - \mathrm{Y}_i\left(0, \mathrm{Z}_i\right)\right] \mathrm{D}_i \\ &= \mathrm{Y}_{0i} + \left(\mathrm{Y}_{1i} - \mathrm{Y}_{0i}\right) \mathrm{D}_i \end{aligned}$$

A random-coefficients notation for this is

$$\mathrm{Y}_i = \alpha_0 + \rho_i \mathrm{D}_i + \eta_i$$

with $\alpha_0 \equiv E\left[\mathrm{Y}_{0i}\right]$ and $\rho_i \equiv \mathrm{Y}_{1i} - \mathrm{Y}_{0i}$

```r
always.takers <- with(toy_data, sum(D == 1)) # n(d,z): n11 + n10 -> 30
never.takers <- with(toy_data, sum(D == 0)) # n00 + n01  -> 40
compliers <- with(toy_data,
```

```r
                   sum(D == 0 & Z == 0) +
                     sum(D == 1 & Z == 1)) # n11 + n00 -> 40
defiers <- with(toy_data,
                 sum(D == 0 & Z == 1) +
                   sum(D == 1 & Z == 0)) # n01 + n10 -> 30

# population average treatment effect
with(toy_data, mean(y[D == 1]) - mean(y[D == 0]))
```

```
#> [1] 0.45
```

```r
# The treatment effect on the treated
(ATT <- with(subset(toy_data, D == 1),
             mean(y[Z == 1]) - mean(y[Z == 0])))
```

```
#> [1] -0.3
```

```r
# TE on compliers
with(subset(toy_data, (D == 1 & Z == 1) | (D == 0 & Z == 0)),
     mean(y[Z == 1]) - mean(y[Z == 0]))
```

```
#> [1] 0.2
```

```r
# TE on always-takers
with(toy_data, mean(y[(D == 1 & Z == 1)]) - mean(y[(D == 1 & Z == 0)]))
```

```
#> [1] -0.3
```

```r
# proportions of subpopulation
props <- with(toy_data,
              c(
    sum(Z[D == 1] == 0) / length(Z[D == 1]), # D_0i = 1 | D_i = 1
    sum(Z[D == 1] == 1) / length(Z[D == 1]), # D_1i = 1 | D_i = 1
    sum(Z[D == 0] == 0) / length(Z[D == 0]), # D_0i = 0 | D_i = 0
    sum(Z[D == 0] == 1) / length(Z[D == 0])  # D_1i = 0 | D_i = 0
  ))
props
```

```
#> [1] 0.66667 0.33333 0.75000 0.25000
```