technische universität
dortmund

Lecture 2c:
Some more notes on instrumental variables

Michael J. Böhm

Empirical Economics

Wintersemester 2025/26

Some things to know about regressions

Simple bivariate regression studies the relationship between two variables Y_i and X_i in a slope parameter β:

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

(Famous) OLS solution: $\hat{\beta} = \frac{Cov(Y_i, X_i)}{Var(X_i)}$ with residual e_i.

Regression produces fitted values (the regression line)

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

and residuals

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i.$$

Some things to know about regressions

Uncorrelated residuals

The first order condition for the slope coefficient $\hat{\beta}$ to the linear regression problem can be written as (e.g. see Angrist/Pischke textbook)

$$\begin{aligned} E\left\{\left(Y_i - \alpha - \hat{\beta}X_i\right)X_i\right\} &= E\left\{e_i X_i\right\} \\ &= Cov(e_i, X_i) \\ &= 0. \end{aligned}$$

The regression residual e_i is uncorrelated with the regressors. This property partitions the variation in Y_i into two orthogonal (uncorrelated) parts:

- Variation related to X_i: the regression line
- Variation unrelated to X_i: the residual

Regression ANOVA

This decomposition leads to

Theorem 1 (Regression ANOVA theorem)

$$
\begin{aligned}
Var(Y_i) &= Var\left(\widehat{Y}_i\right) + Var(e_i) \\
&= Var\left(\alpha + \hat{\beta} X_i\right) + Var(e_i) \\
&= Var\left(\hat{\beta} X_i\right) + Var(e_i) \\
&= \hat{\beta}^2 Var\left(X_i\right) + Var(e_i)
\end{aligned}
$$

Your turn

We know that the variance of the sum of two random variables U and V is

$$Var(U + V) = Var(U) + Var(V) + 2Cov(U, V)$$

What happened to covariance term in the regression ANOVA theorem?

1. A. Regression residuals are uncorrelated with X_i.
2. B. e_i is a constant.
3. C. X_i is a constant.
4. D. β is a constant.

The regression R-square

The regression ANOVA theorem gives us

$$1 = \frac{Var\left(\widehat{Y}_i\right)}{Var(Y_i)} + \frac{Var(e_i)}{Var(Y_i)}.$$

The first term on the right hand side is called the R^2 of the regression

$$\begin{aligned} R^2 &= \frac{Var\left(\widehat{Y}_i\right)}{Var(Y_i)} = 1 - \frac{Var(e_i)}{Var(Y_i)} \\ &= \frac{\hat{\beta}^2 Var\left(X_i\right)}{Var(Y_i)} \\ &= \text{explained part of the variance of } Y_i \end{aligned}$$

The R-square and the correlation coefficient

The R^2 in a one-variable regression is the same as the squared correlation coefficient between Y_i and X_i. The correlation coefficient is

$$\rho = \frac{Cov(Y_i, X_i)}{\sqrt{Var(Y_i)Var(X_i)}}.$$

and easy to show that $\rho^2 = R^2$. Hence:

- R^2 is a good tool for correlation / prediction.
- But not meaningful for causal / policy effects!
- e.g. in next slide we see R^2 from regressing monthly death rates by drowning on ice cream consumption in the U.S.

R-square doesn't tell you whether your regression is meaningful

In Stata:

```
. reg drown icecream

      Source |       SS       df       MS              Number of obs =      12
-------------+------------------------------           F(  1,    10) =   11.67
       Model |   93963.964     1   93963.964           Prob > F      =  0.0066
    Residual |  80487.7027    10  8048.77027           R-squared     =  0.5386
-------------+------------------------------           Adj R-squared =  0.4925
       Total |  174451.667    11  15859.2424           Root MSE      =  89.715

------------------------------------------------------------------------------
      drown |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   icecream |   8.856839   2.592166     3.42   0.007     3.081133    14.63255
      cons |   -502.653    205.482    -2.45   0.034    -960.4955   -44.81047
------------------------------------------------------------------------------
```

Multiple regression solves the OVB problem

When we observe the ommited variable

- Suppose X_{1i} is a variable we are interested in, e.g., the worker's wage rate in the problem set.
- It is correlated with another variable X_{2i} that also affects wages.

If we run $Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i$ we get omitted variable bias:

$$\frac{Cov(Y_i, X_{1i})}{Var(X_{1i})} = \frac{Cov(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i, X_{1i})}{Var(X_{1i})} = \beta_1 + \beta_2 \frac{Cov(X_{2i}, X_{1i})}{Var(X_{1i})}$$

where $\varepsilon_i = \beta_2 X_{2i} + e_i$ and if we assume $Cov(X_{1i}, e_i) = 0$.

Multiple regression solves the OVB problem

The regression anatomy formula

Suppose the other characteristics X_{2i} that are correlated with X_{1i} could be observed (like education).

The regression we want to run

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i.$$

We can always turn a multiple regression into bivariate regressions using the regression anatomy formula:

$$\beta_1 = \frac{Cov(Y_i, \tilde{X}_{1i})}{Var(\tilde{X}_{1i})}$$

where $\tilde{X}_{1i}$ is the residual from a regression of X_{1i} on X_{2i}.

The regression anatomy formula derived

Substitute the expression for the long regression into

$$\begin{aligned}\frac{Cov(Y_i, \tilde{X}_{1i})}{Var(\tilde{X}_{1i})} &= \frac{Cov(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i, \tilde{X}_{1i})}{Var(\tilde{X}_{1i})} \\ &= \frac{\beta_1 Cov(X_{1i}, \tilde{X}_{1i}) + \beta_2 Cov(X_{2i}, \tilde{X}_{1i}) + Cov(e_i, \tilde{X}_{1i})}{Var(\tilde{X}_{1i})}\end{aligned}$$

Residuals are uncorrelated with included regressors so

$$Cov(e_i, \tilde{X}_{1i}) = 0 \text{ and } Cov(X_{2i}, \tilde{X}_{1i}) = 0$$

because $\tilde{X}_{1i}$ is a function of X_{2i} and e_i is the long regression residual, and $\tilde{X}_{1i}$ is the residual from a regression on X_{2i}.

The regression anatomy formula derived

Finally

$$Cov(X_{1i}, \tilde{X}_{1i}) = Var(\tilde{X}_{1i})$$

by the regression ANOVA theorem. Hence

$$\frac{Cov(Y_i, \tilde{X}_{1i})}{Var(\tilde{X}_{1i})} = \beta_1 \frac{Var(\tilde{X}_{1i})}{Var(\tilde{X}_{1i})}.$$

- The process of first regressing X_{1i} on other covariates and then running a bivariate regression is sometimes called partialling out covariates.
- Regression anatomy works for multiple covariates just as well.

IV solves the OVB problem

In lecture 2b we have seen that IV solves the measurement error (ME) problem. Here we show it also solves the omitted variables bias (OVB) problem:

- Suppose D_i is a treatment we are interested in, e.g., the worker's wage rate in the problem set.
- Other characteristics X_i could be observed (like education) or unobserved ε_i (like motivation).

OLS regression of Y_i on D_i will not give the true β, since D_i is correlated with X_i and, most problematically, ε_i.

- We have an instrument Z_i (like winning or not the lottery for the income support program).

IV solves the OVB problem

The regression we want to run

$$Y_i = \alpha + \beta D_i + \gamma X_i + \varepsilon_i.$$

The IV estimator is

$$\lambda = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)} = \frac{\beta Cov(D_i, Z_i) + \gamma Cov(X_i, Z_i) + Cov(\varepsilon_i, Z_i)}{Cov(D_i, Z_i)}.$$

We need

$$\left.\begin{array}{l} Cov(X_i, Z_i) = 0 \\ Cov(\varepsilon_i, Z_i) = 0 \end{array}\right\} \text{ IV randomly assigned}$$

$$Cov(D_i, Z_i) \neq 0 \text{ existence of a first stage}$$

IV solves the OVB problem

With these assumptions

$$\begin{aligned}\lambda &= \frac{\beta Cov(D_i, Z_i) + \gamma \overbrace{Cov(X_i, Z_i)}^{=0} + \overbrace{Cov(\varepsilon_i, Z_i)}^{=0}}{Cov(D_i, Z_i)} \\ &= \frac{\beta Cov(D_i, Z_i)}{Cov(D_i, Z_i)} = \beta.\end{aligned}$$

Where have we used the exclusion restriction?

$$\begin{aligned}Y_i &= \alpha + \beta D_i + \gamma X_i + \theta Z_i + \varepsilon_i \\ \theta &= 0 \Leftrightarrow \text{exclusion restriction}\end{aligned}$$

so Z_i does not itself appear in the regression equation for Y_i.

Your turn

The assumption $Cov\,(X_i, Z_i) = 0$ is equivalent to the IV assumption:

1. Existence of a first stage
2. Exclusion restriction
3. Instrument as good as randomly assigned

(1.–3. are the three assumptions that need to hold for valid IV)

Two stage least squares (2SLS)

Another way of computing λ (yields exactly the same, see again Angrist/Pischke book). Usually used when software like *R* computes your IV estimate.

Start with the first stage $D_i = \alpha_1 + \varphi Z_i + \chi X_i + e_i$ and save the fitted values, $\hat{D}_i$:

$$\hat{D}_i = \alpha_1 + \varphi Z_i + \chi X_i.$$

λ is given by the second stage regression

$$Y_i = \alpha_2 + \lambda_{2sls}\hat{D}_i + \gamma X_i + e_{2i}.$$

This is called the *two stage least squares* (2SLS) method.

2SLS residuals and R^2

- Recall the second stage of 2SLS. This is a regression, and hence $Cov\left(e_{2i}, \hat{D}_i\right) = 0$ by construction (but much less variation in $\hat{D}_i$ than D_i.)
- There is another version of the second stage

$$Y_i = \alpha + \lambda D_i + \gamma X_i + \eta_i$$

  using the orginal D_i as regressor. This is the relationship we are ultimately interested in, and it gives the 2SLS residual η_i. But $Cov(\eta_i, D_i) \neq 0$ because we haven't estimated this equation by OLS (and β_{OLS} and λ_{2SLS} might differ).
- Unlike OLS, 2SLS does not produce a variance decomposition of the variance of Y_i, hence no ANOVA, no (meaningful) R^2.

Basic Readings

- Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton university press, 2009.
- Angrist, Joshua D., and Jörn-Steffen Pischke. Mastering'metrics: The path from cause to effect. Princeton university press, 2014. (*easier undergraduate book*)