Two-Stage Least Squares

Brenton Kenkel — PSCI 8357
April 7, 2016

Last time, we learned the basics of instrumental variables but left many questions unanswered.

- What if we observe some of the confounding factors and wish to control for them?
- What if we have more than one variable that is correlated with the error term?
- What if we have more than one instrumental variable?

Today we will talk about *two-stage least squares*, a general-purpose instrumental variables estimator that can handle all of these situations. Like last time, these notes draw from Angrist and Pischke (2009, chap. 4) and Angrist and Pischke (2015, chap. 4).

The Estimator

If you guessed that an estimator called "two-stage least squares" would involve running OLS two times, pat yourself on the back—you're right! Assume we want to estimate the coefficients of the linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i,$$

but some of the variables X_{ji} are correlated with the error term. OLS estimation of this equation will be biased and inconsistent, as we have already seen.

Suppose that we have a collection of $q > p$ instruments, $Z_{1i}, \ldots, Z_{qi}$, where each satisfies the following conditions:

1. First stage: Z affects X.
2. Independence: Z is uncorrelated with ϵ.
3. Exclusion restriction: Z only affects Y through its effect on X.

Under these conditions, any exogenous X variable (i.e., any that is uncorrelated with the error term) can be included in Z. Then we just need at least one additional instrument per endogenous variable. We call the instruments that are not themselves covariates the *excluded instruments*, for reasons that will become clear momentarily.

The *two-stage least squares estimator* of β is the following procedure:

1. Regress each X_j on Z and save the predicted values, $\hat{X}_j$. If X_j is included in Z, we will have $\hat{X}_j = X_j$.
2. Estimate β via the OLS estimate of the regression model
$$Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \cdots + \beta_p \hat{X}_{pi} + \epsilon_i.$$

This is obviously easy to implement, and it allows us to incorporate exogenous covariates, multiple endogenous variables, and more instruments than endogenous variables (also called overidentifying restrictions).

Fun fact: letting $\mathbf{X}$ be the $N \times p$ matrix of covariates and $\mathbf{Z}$ be the $N \times q$ matrix of instruments, the instrumental variables estimator can be calculated in a single step via the equation Greene (2003, 78)

$$\hat{\beta}_{\text{2SLS}}(Y, \mathbf{X}, \mathbf{Z}) = [\mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top Y.$$

This is the GLS estimator with $\Omega = [\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top]^{-1}$. In practice, though, you won't directly carry out either two-stage least squares or the GLS formula—you'll feed the covariates and the instruments to the computer and let it do the work for you.

The Intuition

Remember from last time our basic recipe,

$$\text{effect of } T_i \text{ on } Y_i = \frac{\text{effect of } Z_i \text{ on } Y_i}{\text{effect of } Z_i \text{ on } T_i}$$

If the instrument and the treatment are both binary, then the instrumental variables estimator of the average treatment effect is a ratio of differences of means:

$$\hat{\tau}_{\text{IV}} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{T}_{Z=1} - \bar{T}_{Z=0}}.$$

Our goal now is to see that two-stage least squares gives us the same answer.

Consider the first-stage regression,

$$T_i = \gamma_0 + \gamma_1 Z_i + \eta_i.$$

You know from your previous adventures with regression that the OLS estimates will be $\hat{\gamma}_0 = \bar{T}_{Z=0}$ and $\hat{\gamma}_1 = \bar{T}_{Z=1} - \bar{T}_{Z=0}$. The predicted values will therefore be

$$\hat{T}_i = \bar{T}_{Z=0} + (\bar{T}_{Z=1} - \bar{T}_{Z=0}) Z_i.$$

Now suppose we run the second-stage regression,

$$Y_i = \alpha_0 + \alpha_1 \hat{T}_i + \epsilon_i.$$

Our ultimate goal is to show that the estimated coefficient on $\hat{T}_i$ is identical to the IV estimate of the average treatment effect: $\hat{\alpha}_1 = \hat{\tau}_{\text{IV}}$. We can rewrite the second-stage regression as

$$\begin{aligned}
Y_i &= \alpha_0 + \alpha_1 \hat{T}_i + \epsilon_i \\
&= \alpha_0 + \alpha_1 (\bar{T}_{Z=0} + (\bar{T}_{Z=1} - \bar{T}_{Z=0}) Z_i) + \epsilon_i \\
&= \underbrace{\alpha_0 + \alpha_1 \bar{T}_{Z=0}}_{\kappa_0} + \underbrace{\alpha_1 (\bar{T}_{Z=1} - \bar{T}_{Z=0})}_{\kappa_1} Z_i + \epsilon_i \\
&= \kappa_0 + \kappa_1 Z_i + \epsilon_i.
\end{aligned}$$

We know that the OLS estimation of the final equation will give us $\hat{\kappa}_1 = \bar{Y}_{Z=1} - \bar{Y}_{Z=0}$. Therefore, the OLS estimation of the first equation gives us

$$\hat{\alpha}_1 = \frac{\hat{\kappa}_1}{\bar{T}_{Z=1} - \bar{T}_{Z=0}} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{T}_{Z=1} - \bar{T}_{Z=0}} = \hat{\tau}_{\text{IV}},$$

as we wanted.

Standard Errors

Although it is useful to think of the instrumental variables estimator as two-stage least squares, in practice you should not run two separate regression models. One reason why not is that the nominal standard errors for β in

the second-stage regression will be wrong. Instead of running two-stage least squares "by hand", use a command like `ivregress` in Stata or `ivreg()` in the **AER** package in R.

Heteroskedasticity, autocorrelation, and clustering are just as problematic for estimating the standard errors of 2SLS as they are for OLS. Luckily, we can use the same Huber-White corrections as we did for OLS.

Instrument Selection and the Bias-Variance Tradeoff

Most commonly, instrumental variables are a scarce resource. An applied analyst is far more likely to worry about having too few instruments than too many. Suppose, however, you were to find yourself with an abundance of instruments. How should you proceed? There are two related principles to keep in mind.

- There is a bias-variance tradeoff: holding your sample size fixed, an additional instrument usually reduces your standard error but increases the bias. Since the bias vanishes asymptotically, you may think this is not a problem in very large samples, but you would be wrong (Bound, Jaeger, and Baker 1995).

  An informal way to see why this is true is to think about how the 2SLS estimator is constructed. The more instruments we have, the closer the predicted values of the endogenous regressors get to the true values. (Remember that adding a covariate to a regression model always increases the R^2.) With enough instruments, the predicted values are approximately the true values, which means 2SLS is approximately OLS. Since OLS is inconsistent under the assumption of endogeneity, this is not a good thing.

- Weak instruments increase the bias more than they reduce the variance.

Two strong instruments are better than ten weak instruments. It is harder to say whether one weak instrument is better than two weak instruments, or whether OLS might be better than 2SLS if every instrument is weak. It depends on the sample size, the magnitude of the weakness of the instruments, and the plausibility of the independence and exclusion restrictions. When you only have weak instruments available, you should not stake strong claims on a single

specification. The best you can hope for is to have a result that is robust across different permutations of instruments.

For a model with a single endogenous variable, the usual rule of thumb is that the F-statistic of the regression of the endogenous variable on the excluded instruments should be at least 10 (Stock, Wright, and Yogo 2002).

Appendix: Implementation

There are a few different implementations of 2SLS in R. We will use the one from the **AER** package.

```
library("AER")
```

We will reproduce columns 3 and 4 of Table IV in Angrist and Krueger (1991). First, load up the data.

```
AK <- read.csv("AK1991-clean.csv")
head(AK)
```

```
##   year quarter educ wage  age
## 1 1929       3   11 5.02 40.5
## 2 1929       1   12 5.06 41.0
## 3 1928       3   12 5.38 41.5
## 4 1923       4   12 5.18 46.2
## 5 1924       1   16 6.38 46.0
## 6 1923       1   12 5.00 47.0
```

The response is `wage`. The (endogenous) treatment variable is `educ`. The (exogenous) covariates are `age`, `age` squared, and `year` dummies. The excluded instruments are interactions of the `year` dummies with `quarter` dummies.

OLS is the same as it ever was.

```
fit_ols <- lm(wage ~ educ + age + I(age^2) + factor(year),
              data = AK)
summary(fit_ols)
```

```
##
## Call:
## lm(formula = wage ~ educ + age + I(age^2) + factor(year), data = AK)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.700 -0.219  0.056  0.307  4.393
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.858292   1.521034    0.56    0.573
## educ              0.080168   0.000355  225.65   <2e-16
## age               0.144552   0.067600    2.14    0.032
## I(age^2)         -0.001542   0.000748   -2.06    0.039
## factor(year)1921 -0.001585   0.009216   -0.17    0.863
## factor(year)1922 -0.011239   0.014724   -0.76    0.445
## factor(year)1923 -0.009737   0.019546   -0.50    0.618
## factor(year)1924 -0.006589   0.023558   -0.28    0.780
## factor(year)1925  0.003161   0.026867    0.12    0.906
## factor(year)1926  0.009874   0.029710    0.33    0.740
## factor(year)1927  0.019409   0.032388    0.60    0.549
## factor(year)1928  0.031107   0.035337    0.88    0.379
## factor(year)1929  0.024737   0.039048    0.63    0.526
##
## Residual standard error: 0.593 on 247186 degrees of freedom
## Multiple R-squared:  0.171,  Adjusted R-squared:  0.171
## F-statistic: 4.25e+03 on 12 and 247186 DF,  p-value: <2e-16
```

It is instructive to run 2SLS by hand, even though we should rely on canned procedures in our published work.

```
X <- model.matrix(~ educ + age + I(age^2) + factor(year),
                  data = AK)
Z <- model.matrix(~ age + I(age^2) + factor(year) * factor(quarter),
                  data = AK)
Y <- AK$wage

ols_first <- lm(X ~ Z)
X_hat <- fitted(ols_first)
ols_second <- lm(Y ~ X_hat)
coef(ols_second)
```

```
##           (Intercept)      X_hat(Intercept)             X_hateduc
##               0.02035                    NA               0.13104
```

```
##              X_hatage         X_hatI(age^2) X_hatfactor(year)1921
##               0.14092              -0.00136               0.00521
## X_hatfactor(year)1922 X_hatfactor(year)1923 X_hatfactor(year)1924
##               0.00953               0.01952               0.03275
## X_hatfactor(year)1925 X_hatfactor(year)1926 X_hatfactor(year)1927
##               0.05602               0.07070               0.09459
## X_hatfactor(year)1928 X_hatfactor(year)1929
##               0.11376               0.11341
```

The canned procedure is the `ivreg()` function. It works like `lm()`, except the model formula is in the form `y ~ x1 + x2 + ... | z1 + z2 + ...`, where the `x` terms are the variables whose coefficients we want to estimate and the `z` terms are the instruments. Any exogenous covariates should be included in both parts of the formula, as in the example below.

```
fit_iv <- ivreg(wage ~ educ + age + I(age^2) + factor(year) |
                    age + I(age^2) + factor(year) * factor(quarter),
                data = AK)
summary(fit_iv)
```

```
##
## Call:
## ivreg(formula = wage ~ educ + age + I(age^2) + factor(year) |
##     age + I(age^2) + factor(year) * factor(quarter), data = AK)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0310 -0.2586  0.0483  0.3365  4.7730
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.020346   1.675494    0.01    0.990
## educ               0.131042   0.033357    3.93  8.6e-05
## age                0.140915   0.070388    2.00    0.045
## I(age^2)          -0.001360   0.000787   -1.73    0.084
## factor(year)1921   0.005211   0.010575    0.49    0.622
## factor(year)1922   0.009533   0.020500    0.47    0.642
## factor(year)1923   0.019516   0.027956    0.70    0.485
## factor(year)1924   0.032753   0.035586    0.92    0.357
## factor(year)1925   0.056019   0.044527    1.26    0.208
## factor(year)1926   0.070699   0.050460    1.40    0.161
```

```
## factor(year)1927  0.094591   0.059713    1.58    0.113
## factor(year)1928  0.113762   0.065490    1.74    0.082
## factor(year)1929  0.113406   0.070928    1.60    0.110
##
## Residual standard error: 0.617 on 247186 degrees of freedom
## Multiple R-Squared: 0.102,   Adjusted R-squared: 0.102
## Wald test: 8.67 on 12 and 247186 DF,  p-value: <2e-16
```

If you want heteroskedasticity-consistent standard errors, you can use the `vcovHC()` function. Unfortunately, the dataset here is so big that it crashes the function, at least on my computer. So let's take a subsample of the data, run 2SLS on it, and then calculate the HC1 estimator of the standard errors.

```
set.seed(14)
AK_sample <- AK[sample(1:nrow(AK), 10000), ]
fit_iv_sample <- update(fit_iv,
                        data = AK_sample)
summary(fit_iv_sample)
```

```
##
## Call:
## ivreg(formula = wage ~ educ + age + I(age^2) + factor(year) |
##     age + I(age^2) + factor(year) * factor(quarter), data = AK_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3300 -0.2210  0.0511  0.3028  2.8224
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.24938    7.51985   -0.17    0.868
## educ              0.06966    0.03687    1.89    0.059
## age               0.24466    0.33181    0.74    0.461
## I(age^2)         -0.00265    0.00367   -0.72    0.471
## factor(year)1921 -0.02980    0.04549   -0.66    0.512
## factor(year)1922 -0.06312    0.07279   -0.87    0.386
## factor(year)1923 -0.06202    0.09690   -0.64    0.522
## factor(year)1924 -0.04359    0.11917   -0.37    0.715
## factor(year)1925 -0.00222    0.13836   -0.02    0.987
## factor(year)1926 -0.03198    0.15054   -0.21    0.832
## factor(year)1927  0.00534    0.16801    0.03    0.975
```

```
## factor(year)1928  0.00969    0.18185    0.05    0.957
## factor(year)1929  0.01441    0.20238    0.07    0.943
##
## Residual standard error: 0.585 on 9987 degrees of freedom
## Multiple R-Squared: 0.179,   Adjusted R-squared: 0.178
## Wald test: 1.11 on 12 and 9987 DF,  p-value: 0.344
```

```
summary(fit_iv_sample,
        vcov = vcovHC(fit_iv_sample, type = "HC1"))
```

```
##
## Call:
## ivreg(formula = wage ~ educ + age + I(age^2) + factor(year) |
##     age + I(age^2) + factor(year) * factor(quarter), data = AK_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3300 -0.2210  0.0511  0.3028  2.8224
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.24938    7.30492   -0.17     0.86
## educ              0.06966    0.03556    1.96     0.05
## age               0.24466    0.32713    0.75     0.45
## I(age^2)         -0.00265    0.00366   -0.72     0.47
## factor(year)1921 -0.02980    0.04646   -0.64     0.52
## factor(year)1922 -0.06312    0.07691   -0.82     0.41
## factor(year)1923 -0.06202    0.10327   -0.60     0.55
## factor(year)1924 -0.04359    0.12571   -0.35     0.73
## factor(year)1925 -0.00222    0.14477   -0.02     0.99
## factor(year)1926 -0.03198    0.15756   -0.20     0.84
## factor(year)1927  0.00534    0.17263    0.03     0.98
## factor(year)1928  0.00969    0.18451    0.05     0.96
## factor(year)1929  0.01441    0.19851    0.07     0.94
##
## Residual standard error: 0.585 on 9987 degrees of freedom
## Multiple R-Squared: 0.179,   Adjusted R-squared: 0.178
## Wald test: 1.18 on 12 and 9987 DF,  p-value: 0.291
```

What if you have panel data? You can get instrumental variables estimates with fixed effects and/or clustered standard errors through `plm()`, by using

the same kind of two-part formula that `ivreg()` takes.

References

Angrist, Joshua D, and Alan B Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106 (4): 979–1014.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. An Empiricist's Companion. Princeton University Press.

———. 2015. *Mastering 'Metrics*. The Path from Cause to Effect. Princeton University Press.

Bound, John, David A Jaeger, and Regina M Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90 (430): 443.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Prentice Hall.

Stock, James H, Jonathan H Wright, and Motohiro Yogo. 2002. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business & Economic Statistics* 20 (4): 518–29.