Instrumental Variables Regression: Demand for Cigarettes

Econ 440 - Introduction to Econometrics

Patrick Toche, ptoche@fullerton.edu

05 May 2022

Instrumental Variables Regression

Regression models may suffer from problems like omitted variables, measurement errors and simultaneous causality. If so, the error term is correlated with the regressor of interest and so that the corresponding coefficient is estimated inconsistently.

Adding omitted variables to the regression mitigates the risk of biased estimation of the causal effect of interest.

If there is simultaneous causality, — When causality runs from X to Y and from Y to X, there will be an estimation bias that cannot be corrected for by multiple regression.

A general technique for obtaining a consistent estimator of the coefficient of interest is instrumental variables (IV) regression.

IV regression is used for estimating the elasticity of the demand for cigarettes — a classical example of simultaneous causality.

IV Estimation with a Single Regressor and a Single Instrument

Consider the simple regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad , \quad i = 1, \ldots, n$$

where the error term u_i is correlated with the regressor X_i (X is *endogenous*) such that OLS is inconsistent for the true β_1. In the most simple case, IV regression uses a single instrumental variable Z to obtain a consistent estimator for β_1.

Z must satisfy two conditions to be a valid instrument:

1. **Instrument relevance condition**:
   - The regressor X and its instrument Z *must be* correlated: $\rho_{Z_i, X_i} \neq 0$.
2. **Instrument exogeneity condition**:
   - The instrument Z *must not be* correlated with the error term u: $\rho_{Z_i, u_i} = 0$.

The Two-Stage Least Squares Estimator

- First stage: The variation in the endogenous regressor X is decomposed into a problem-free component that is explained by the instrument Z and a problematic component that is correlated with the error u_i.
- Second stage: Use the problem-free component of the variation in X to estimate β_1.

The first stage regression model is

$$X_i = \pi_0 + \pi_1 Z_i + \nu_i,$$

where $\pi_0 + \pi_1 Z_i$ is the component of X_i that is explained by Z_i while ν_i is the component that cannot be explained by Z_i and exhibits correlation with u_i.

Using the OLS estimates $\widehat{\pi}_0$ and $\widehat{\pi}_1$ we obtain predicted values $\widehat{X}_i, \quad i = 1, \dots, n$. If Z is a valid instrument, $\widehat{X}$ is exogenous. The second stage produces $\widehat{\beta}_0^{TSLS}$ and $\widehat{\beta}_1^{TSLS}$, the TSLS estimates of β_0 and β_1.

For the case of a single instrument one can show that the TSLS estimator of β_1 is:

$$\widehat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Z_i - \overline{Z})}$$

which is the ratio of the sample covariance between Z and Y to the sample covariance between Z and X.

The two-stage least-squares estimator is a consistent estimator for β_1 under the assumption that Z is a valid instrument. The CLT implies that for large n, the distribution of $\widehat{\beta}_1^{TSLS}$ can be approximated by a normal distribution.

Application: The Demand For Cigarettes It is plausible that cigarette consumption can be reduced by taxing cigarettes more heavily. The question is by *how much* taxes must be increased to reach a certain reduction in cigarette consumption. The responsiveness is measured by the price elasticity of demand, which can be estimated. An OLS regression of log quantity on log price cannot be used to identify the elasticity since there is simultaneous causality between demand and supply, but an IV regression can be used to identify the elasticity.

The dataset `CigarettesSW`, from the package `AER`, is a panel data set that contains observations on cigarette consumption and several economic indicators for all 48 continental federal states of the U.S. from 1985 to 1995. Following the book we consider data for the cross section of states in 1995 only.

```
data("CigarettesSW")
summary(CigarettesSW)
```

```
##      state      year         cpi         population           packs
##  AL     : 2   1985:48   Min.   :1.08   Min.   :  478447   Min.   : 49.3
##  AR     : 2   1995:48   1st Qu.:1.08   1st Qu.: 1622606   1st Qu.: 92.5
##  AZ     : 2             Median :1.30   Median : 3697472   Median :110.2
##  CA     : 2             Mean   :1.30   Mean   : 5168866   Mean   :109.2
##  CO     : 2             3rd Qu.:1.52   3rd Qu.: 5901500   3rd Qu.:123.5
##  CT     : 2             Max.   :1.52   Max.   :31493524   Max.   :198.0
##  (Other):84
##      income              tax           price          taxs
##  Min.   :6.89e+06   Min.   :18.0   Min.   : 85   Min.   : 21.3
##  1st Qu.:2.55e+07   1st Qu.:31.0   1st Qu.:103   1st Qu.: 34.8
##  Median :6.17e+07   Median :37.0   Median :138   Median : 41.0
##  Mean   :9.99e+07   Mean   :42.7   Mean   :143   Mean   : 48.3
##  3rd Qu.:1.27e+08   3rd Qu.:50.9   3rd Qu.:176   3rd Qu.: 59.5
##  Max.   :7.71e+08   Max.   :99.0   Max.   :241   Max.   :112.6
##
```

We estimate β_1 in

$$\log(Q_i^{cigarettes}) = \beta_0 + \beta_1 \log(P_i^{cigarettes}) + u_i$$

where $Q_i^{cigarettes}$ is the number of cigarette packs per capita sold and $P_i^{cigarettes}$ is the after-tax average real price per pack of cigarettes in state i.

We use $SalesTax$ as the instrumental variable. $SalesTax$ is measured in dollars per pack. The idea is that $SalesTax$ is a relevant instrument as it is included in the after-tax average price per pack. It is plausible that $SalesTax$ is exogenous since the sales tax does not influence the quantity sold directly, only indirectly through the price.

We deflate the cross data for the year 1995.

The sample correlation between the sales tax and price per pack is approximately 0.614, indicating that $SalesTax$ and $P_i^{cigarettes}$ are correlated: higher sales taxes lead to higher prices. However, a correlation analysis is not sufficient for checking whether the instrument is relevant.

Compute real per capita prices

```
CigarettesSW$rprice <- with(CigarettesSW, price / cpi)
```

Compute the sales tax

```
CigarettesSW$salestax <- with(CigarettesSW, (taxs - tax) / cpi)
```

check the correlation between sales tax and price

```
cor(CigarettesSW$salestax, CigarettesSW$price)
```

```
## [1] 0.61412
```

Generate a subset for the year 1995

```
c1995 <- subset(CigarettesSW, year == "1995")
```

The first stage regression is

$$\log(P_i^{cigarettes}) = \pi_0 + \pi_1 SalesTax_i + \nu_i.$$

We estimate this model in `R` using `lm()`. In the second stage we run a regression of $\log(Q_i^{cigarettes})$ on $\log(\widehat{P_i^{cigarettes}})$ to obtain $\widehat{\beta}_0^{TSLS}$ and $\widehat{\beta}_1^{TSLS}$.

Perform the first stage regression:

```
cig_s1 <- lm(log(rprice) ~ salestax, data = c1995)
coeftest(cig_s1, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.61655    0.02892  159.64  < 2e-16 ***
## salestax     0.03073    0.00484    6.35  8.5e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first stage regression is

$$\widehat{\log(P_i^{cigarettes})} = \underset{(0.03)}{4.62} + \underset{(0.005)}{0.031} SalesTax_i$$

which predicts a positive relation between sales tax price per cigarettes. How much of the observed variation in $\log(P^{cigarettes})$ is explained by the instrument $SalesTax$? The R^2 suggests that about 47% of the variation in after tax prices is explained by the variation of the sales tax across states.

Inspect the R^2 of the first stage regression:

```
summary(cig_s1)$r.squared
```

```
## [1] 0.471
```

Next store $\log(\widehat{P_i^{cigarettes}})$, the fitted values obtained by the first stage regression and run the second stage regression to obtain the TSLS estimate:

```
# store the predicted values
lcigp_pred <- cig_s1$fitted.values
# run the stage 2 regression
cig_s2 <- lm(log(c1995$packs) ~ lcigp_pred)
coeftest(cig_s2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.720      1.703    5.71  7.9e-07 ***
## lcigp_pred    -1.084      0.356   -3.05   0.0038 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The TSLS estimation yields:

$$\widehat{\log(Q_i^{cigarettes})} = \underset{(1.70)}{9.72} + \underset{(0.36)}{1.08}\log(\widehat{P_i^{cigarettes}})$$

The function `ivreg()`, from the package `AER`, can be used to automate the TSLS procedure. It is similar to `lm()`. Instruments can be added to the usual specification of the regression formula using a vertical bar separating the model equation from the instruments. Thus, for the regression at hand the correct formula is `log(packs) ~ log(rprice) | salestax`.

Perform TSLS using `ivreg()`

```
cig_ivreg <- ivreg(log(packs) ~ log(rprice) | salestax, data = c1995)

coeftest(cig_ivreg, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.720      1.528    6.36  8.3e-08 ***
## log(rprice)   -1.084      0.319   -3.40   0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient estimates coincide with those computed earlier.

Two Notes on the Computation of TSLS Standard Errors

1. The standard errors reported for the second-stage regression by `coeftest()` or `summary()`, are *invalid*, because they treat the predictions from the first-stage regression as fixed, not taking into account the uncertainty associated with the first-stage estimation when computing the standard errors from the second-stage regression. Function `ivreg()` performs the necessary adjustments to obtain correct standarderrors.

2. Moreover, as always, it is important to compute heteroskedasticity-robust standard errors with `vcovHC()`.

The TSLS estimate for β_1 suggests that an increase in cigarette prices by one percent reduces cigarette consumption by roughly 1.08 percentage points. But this estimate is almost certainly biased due to omitted variables. A multiple IV regression approach is needed to add appropriate controls to the regression.

The General IV Regression Model

In the general IV model, we distinguish four types of variables: the dependent variable Y, exogenous regressors W, endogenous regressors X and instrumental variables Z.

\begin{align} Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \dots + \beta_{k+r} W_{ri} + u_i, (\#eq:givmodel) \end{align} with $i = 1,$
$dots, n$ is the general instrumental variables regression model where

- Y_i is the dependent variable
- $beta_0,$
  $dots,$
  $beta_{k+1}$ are $1 + k + r$ unknown regression coefficients
- $X_{1i},$
  $dots, X_{ki}$ are k endogenous regressors
- $W_{1i},$
  $dots, W_{ri}$ are r exogenous regressors which are uncorrelated with u_i
- u_i is the error term
- $Z_{1i},$
  $dots, Z_{mi}$ are m instrumental variables

The coefficients are overidentified if $m > k$. If $m < k$, the coefficients are underidentified and when $m = k$ they are exactly identified. For estimation of the IV regression model we require exact identification or overidentification.

Assume that you want to estimate the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + W_{1i} + u_i$$

where X_{1i} and X_{2i} are endogenous regressors that shall be instrumented by Z_{1i}, Z_{2i} and Z_{3i} and W_{1i} is an exogenous regressor. The corresponding data is available in a `data.frame` with column names `y`, `x1`, `x1`, `w1`, `z1`, `z2` and `z3`. It might be tempting to specify the argument `formula` in your call of `ivreg()` as `y ~ x1 + x2 + w1 | z1 + z2 + z3`, but that would not carry out the expected regression. As explained in the documentation of `ivreg()` (see `?ivreg`), it is necessary to list *all* exogenous variables as instruments too, that is joining them by `+`'s on the right of the vertical bar: `y ~ x1 + x2 + w1 | w1 + z1 + z2 + z3` where `w1` is "instrumenting itself".

If there is a large number of exogenous variables it may be convenient to provide an update formula with a `.` (this includes all variables except for the dependent variable) right after the `|` and to exclude all endogenous variables using a `-`. For example, if there is one exogenous regressor `w1` and one endogenous regressor `x1` with instrument `z1`, the appropriate formula would be `y ~ w1 + x1 | w1 + z1` which is equivalent to `y ~ w1 + x1 | . - x1 + z1`.

Similarly to the simple IV regression model, the general IV model can be estimated using the two-stage least squares estimator:

1. **First-stage regression(s)**

   Run an OLS regression for each of the endogenous variables ($X_{1i},$
   $dots, X_{ki}$) on all instrumental variables ($Z_{1i},$

$dots, Z_{mi}$), all exogenous variables (W_{1i},
$dots, W_{ri}$) and an intercept. Compute the fitted values (
$widehatX_{1i}$,
$dots$,
$widehatX_{ki}$).

2. **Second-stage regression**

   Regress the dependent variable on the predicted values of all endogenous regressors, all exogenous variables and an intercept using OLS. This gives
   $widehatbeta_0^{TSLS}$,
   $dots$,
   $widehatbeta_{k+r}^{TSLS}$, the TSLS estimates of the model coefficients.

- *First-stage regression(s)* Run an OLS regression for each of the endogenous variables (X_{1i},
  $dots, X_{ki}$) on all instrumental variables (Z_{1i},
  $dots, Z_{mi}$), all exogenous variables (W_{1i},
  $dots, W_{ri}$) and an intercept. Compute the fitted values (
  $widehatX_{1i}$,
  $dots$,
  $widehatX_{ki}$).
- *Second-stage regression* Regress the dependent variable on the predicted values of all endogenous regressors, all exogenous variables and an intercept using OLS. This gives

  $$\widehat{beta}_0^{TSLS}, \ldots, \widehat{beta}_{k+r}^{TSLS}$$

  the TSLS estimates of the model coefficients.