Cheat Sheet for Instrumental Variables

PBAF 529 - Dr. Marieka Klawitter

by Jarrad Fjelstad and Erik Rose

Why use an instrumental variable?

BLUE assumptions mean that the error term in an OLS regression averages to zero and is not correlated with any regressor variable. OLS prediction becomes inconsistent when the value of a regressor variable correlates with the error term. This can indicate selection bias, omitted variable bias, or model misspecification. An instrumental variable (IV) is a technique for eliminating error correlation.

Exogenous regressors are variables with values independent of the error term in the structural model.

Endogenous regressors are variables with values that correlate with the error values in the structural model. Endogenous variables create systematic inconsistencies in the estimation of *ß*s for the explanatory variables, because unobserved characteristics distort the predicted effect size.

An instrumental variable tries to control for the effects of unobserved characterists. We predict new values for the endogenous variable using the exogenous variables in our structural model and adding one or more instrumental variables. This is called the first-stage regression. IVs help us triangulate what the average values for the endogenous variable would be, based upon similarities between observations in the sample, without the influence of unobserved characteristics.

Two key assumptions:

- The IV must correlate with the endogenous variable.
- Values of the IVs and exogenous variables are unrelated to error values in the structural model.

The new predicted values for the endogenous variable will then be free of selection bias or unobserved characteristics, assuming a correctly specified model. If the IVs correlate with the regressor, but not the with error term of the dependent variable, then the predicted outcomes are not as likely to carry the biases of unobserved characteristics present in the OLS model.

First-stage models with fewer IVs than endogenous variables are under-identified, and we cannot determine the values of the parameters of the model based on what we observe.

When the # of IVs = the # of endogenous variables, the model is **just-identified**.

When the # of IVs > the # of endogenous variables, the model is **over-identified**.

In cases of multiple endogenous variables, you can also build a structural model for each endogenous term using standard rank and order conditions.

The Math:

If **y** is the dependent variable, **x** is the independent variable, and **u** is the error term, OLS assumes the error correlates with **y** but not **x**: $\begin{tikzcd} x \arrow[r] & y \\ u \arrow[ur] & \end{tikzcd}$ An endogenous **x** correlates with **u** directly: $\begin{tikzcd} x \arrow[r] & y \\ u \arrow[u] \arrow[ur] & \end{tikzcd}$ Then $\hat{ß}$ becomes inconsistent for ß**x** because of **u**'s indirect effect on **y** through **x**. The instrumental variable **z** predicts new values for **x** =E(x_i) where: $\begin{tikzcd} z \arrow[r] & x \arrow[r] & y \\ & u \arrow[u] \arrow[ur] & \end{tikzcd}$ Then $\hat{ß}_{IV} = \sum_i z_i y_i / \sum_i z_i x_i$ is consistent for ß, so long as **z** is correlated with **x** but not with **u**.

The Structural Equation: $y_{1i} = y'_{2i}ß_1 + x'_{1i}ß_2 + u_i$

if y_{1i} is the dependent, y'_{2i} are endogenous regressors, x'_{1i} are exogenous regressors, and i=1....N.

The First-Stage Equation: $y_{2ji} = x'_{1i}\pi_{1j} + x'_{2i}\pi_{1j} + v_{ji}$

where y_{2ji} estimates y'_{2i}, x'_{1i} are the exogenous regressors from the structural equation, x'_{2i} are the instruments for y'_{2i}, and v_{ji} is the error term for y_{2ji}. j=1....m where there are m endogenous variables.

Assume $E\{z'_i(y_i-x'_iß)\}=0$. In just-identified models, if $\sum_{i=1}^{N} z'_i(y_i-z'_iß)=0$ then $\hat{ß}_{IV}=(Z'X)^{-1}Z'y$

where X is the matrix of the stack of x'_i vectors, y is the stack of y_i vectors, and Z is the stack of z'_i vectors, so Z′(y-Xß)=0.

Types of IV Regression:

IV - Just-Identified
The first-stage regression measures: the ratio of the covariance of the instrumental variables and the dependant variable and the covariance of the instrumental variable and the regressor variables.

2SLS - Two Stage Least Squares
This is the default method for regressing over-identified models.

GMM - Generalized Method of Moments, and Optimal GMM (OGMM)
2SLS is a special case of GMM. GMM uses different weighted matrices between the 'numerator' and 'denominator' sides of the covariance ratio. The most common alternative to 2SLS is Optimal GMM.

LIML - Limited Information Maximum Likelihood

This estimator is more efficient and consistent than 2SLS for smaller sample sizes. As the number of instruments relative to the sample size grows larger, LIML becomes a stronger choice of estimator.

JIVE - Jackknife IV

This estimator is also better at estimating effect sizes for variables compared to 2SLS when the sample size is smaller.

3SLS - Three Stage Least Squares

This estimator is identical to 2SLS, except that you must specify a structural equation for each first-stage regression using standard rules of rank and order. 3SLS becomes inconsistent if errors are heteroskedastic.

Using IVs in Stata:

The primary command in stata for running a regression using instrumental variables is `ivregress`. You can also use `treatreg`, `jive`, `condivreg`, `ivreg2`, and `reg3`.

The syntax for ***ivregress*** *is:*

ivregress *estimator depvar* [exogenous regressors] (endogenous regressors = istrumental vars) [*if*] [*in*] [*weight*] [, *options*]

where *estimator* is either **2sls**, **gmm** (optimal GMM), or **liml**, and *depvar* is the dependant variable. *options* include **first**, which yields output from the first-stage regression, and **vce**(*type*) where *type* includes **robust** (for robust standard errors), **unadjusted** (for nonrobust standard errors), **cluster** *clustvar* (where *clustvar* identifies the cluster variable), **bootstrap**, and **hac** *kernel*.

If using estimator **gmm** on an over-identified model, option **wmatrix**(*wmtype*) specifies the weighting matrix where *wmtype* is: **robust, cluster** *clustvar*, also **hac** *kernel*, **hac** *kernel #*, or **hac** *kernel opt* (for time-series data with heteroskedasticity- and autocorrelation-consistent errors), where **vce**() is set to **wmatrix**() if not specified.

Use **ivregress liml** when the sample size is smaller. Then it will be more accurate than 2SLS, especially for weak instruments.

ivreg2 is a user-written command (type “findit ivreg2”), where the syntax and function is like **ivregress** except you specify the estimator (**2sls**, **gmm**, or **liml**) in the *options* after the comma. It provides additional statistical tests compared to **ivregress** and stores the results in e().

The syntax for ***treatreg*** *is*:

treatreg *depvar* [*indepvars*] [*if*] [*in*] [*weight*], **treat**(endogenous variable = instrumental vars + exogenous vars) [**twostep**]

This estimator is for the special case when endogenous regressor is binary. Unless you specify **twostep**, **treatreg** uses maximum likelihood to determine fits. **treatreg** offers increased precision but greater chance of misspecification error, and becomes inconsistent if errors are heteroskedastic.

Jackknife IV (JIVE):

The syntax for **jive** is the same for **ivregress**, except you specify the estimator as an option, either **ujive1** (the default), **ujive2**, **jive1**, or **jive2**. **jive** is a user-written command (type “findit jive”).

jive eliminates the correlation beween first-stage fitted values and the structural equation's error term by subtracting one observation from the sample, recalculating the first-stage, and using an evaluation equation to determine the effect if each observation were removed in turn and an average effect calculated from there.

JIVE can be more consistent in smaller sample sizes or when using many weak instruments, but C & T suggests LIML is more reliable.

condivreg is another user-written command (type "findit condivreg"). When the sample size is smaller, **condivreg** estimates more precise confidence levels. The syntax is the same for **ivregress**, but without the estimator (2sls, gmm, or liml). **condivreg** uses the likeood-ratio test statistic. Use *option* **lm** to employ the Lagrange Multiplier, and **ar** to use the Anderson-Rubin test statistic. **level**(#) sets the confidence level and **test**(#) sets the null hypthesis to p=# instead of p=0. The model assumes i.i.d. errors - heteroskedastic errors can create problems.

3SLS is only for over-identified models. Specify a structural model for each endogenous regressor using rank and order conditions. Use **reg3** with each structural equation following in parenthesis. This estimator uses the cross-correlation of errors to produce a more precise estimation than 2SLS, provided errors are i.i.d. If errors are heteroskedastic, 3SLS becomes inconsistent.

Diagnostics:

IV estimators are biased in finite samples, even with consistent instruments:

1. When the number of instruments is very large relative to the sample size and the first stage regression fits very well, the IV estimator may approach the OLS estimator and be similarly biased.

2. When when the correlation between the structural-error u and some components of the first-stage regression errors is high, then asymptotic theory may be a poor guide to the finite-sample distribution.

3. With weak instruments, asymptotic theory may be a poor fit to the finite-sample distribution of the IV estimator, even if the sample has 1,000s of observations.

Endogeneity Diagnostics:

After running an IV regression, use **estat endogenous** to test the hypothesis that the endogenous variables are actually exogenous using the DWH test. The variable is endogenous if p < .05. You can also use the **hausman** command to perform a Hausman test comparing your IV regression to your OLS regression, but this test is inconsistent when errors are heteroskedastic, and C & T recommends you perform a Hausman bootstrap (see C&T 13.4.6).

Weak Instrument Diagnostics:

Use **correlate** to test gross correlation between endogenous variables and instruments. (For over-identified models, what matters is partial correlation after controlling for other instruments.) Low correlation leads to efficiency loss compared to OLS (high SEs and low t-scores), and very low correlation can indicate weak instruments.

For **ivregress**, insert the option **first** at the end to produce the first-stage regression results in the output. Instruments may be weak when the r^2 fit is low, and if the t-scores of your instruments are low.

After an IV regression, use **estat firststage** [, *options*] to produce Shea's partial r^2 (the minimum eigenvalue of a matrix analog of the F statistic of each first-stage regression). If the r^2 is low, this can indicate weak instruments. This test also produces an F statistic. If you are using robust standard errors, use the *option* **forcenonrobust** to calculate critical values, where the % shown is the size distortion you are willing to tolerate in the Wald statistic, and each critical value is the minimum F value required. If your F statistic is smaller than your desired critical value, then your instrument is too weak. Use the *option* **all** to include results for each endogenous regressor as well.

After running **ivregress gmm** on an overidentified model, use **estat overid** to test your instruments for validity. IVs are invalid if $p < .05$.

Finally, test the sensitivity of your instruments by running different types of IV regressions (2sls, gmm, liml, jive), saving the results, also saving the **estat firststage** results, and then compare the values in a table. Large fluctuations in the results can indicate weak instruments or model misspecification.