Regression Techniques in Stata

Christopher F Baum

Boston College and DIW Berlin

University of Adelaide, June 2010

Basics of Regression with Stata

A key tool in multivariate statistical inference is *linear regression*, in which we specify the conditional mean of a response variable y as a linear function of k independent variables

$$E[y|x_1, x_2, \ldots, x_k] = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_{i,k} \tag{1}$$

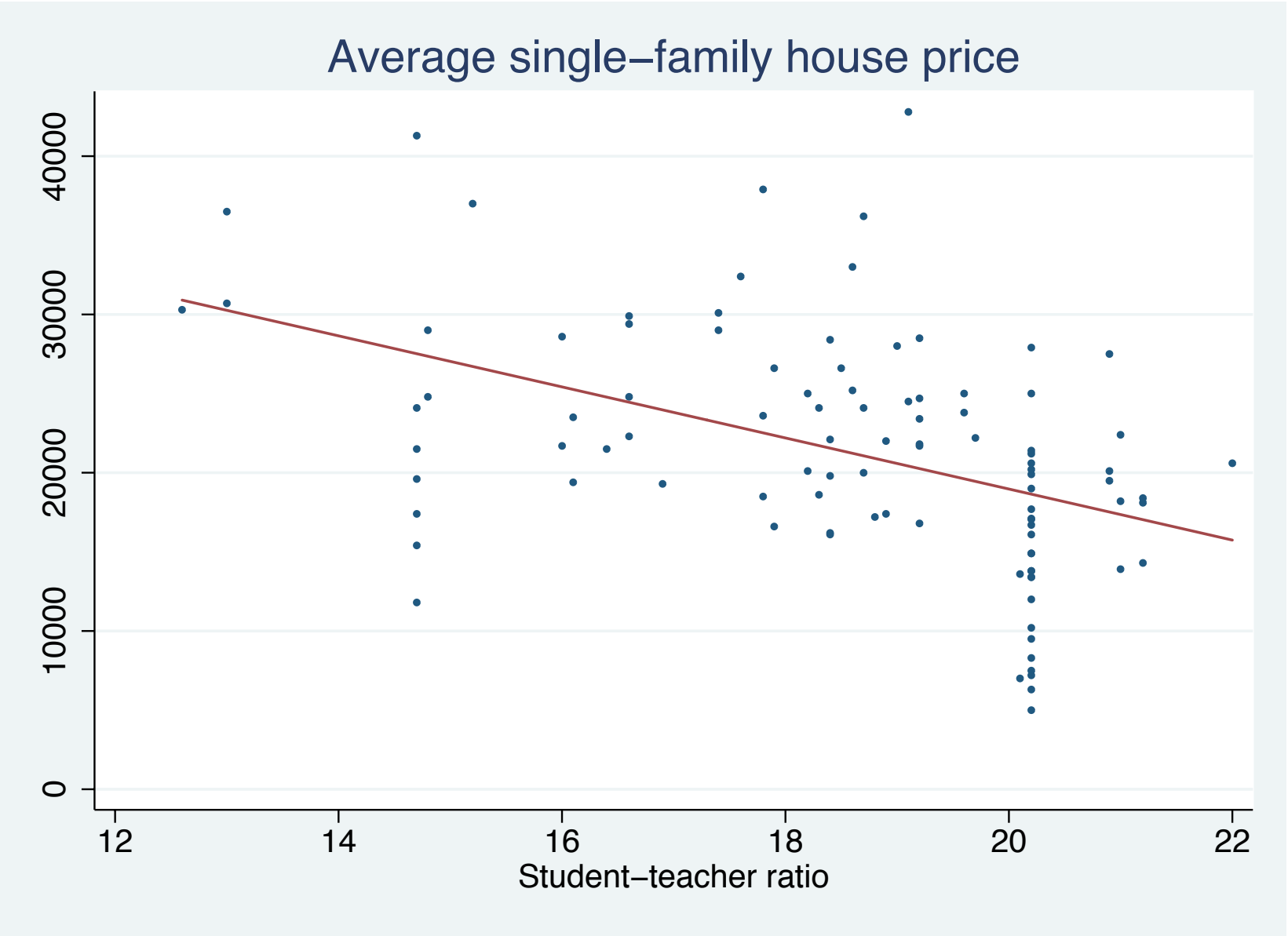
Note that the conditional mean of y is a function of $x_1, x_2, \ldots, x_k$ with fixed parameters $\beta_1, \beta_2, \ldots, \beta_k$. Given values for these βs the linear regression model predicts the average value of y in the population for different values of $x_1, x_2, \ldots, x_k$.

For example, suppose that the mean value of single-family home prices in Boston-area communities, conditional on the communities' student-teacher ratios, is given by

$$E\left[p \mid stratio\right] = \beta_1 + \beta_2\, stratio \tag{2}$$

This relationship reflects the hypothesis that the quality of communities' school systems is capitalized into the price of housing in each community. In this example the population is the set of communities in the Commonwealth of Massachusetts. Each town or city in Massachusetts is generally responsible for its own school system.

Conditional mean of housing price:



We display average single-family housing prices for 100 Boston-area communities, along with the linear fit of housing prices to communities' student-teacher ratios. The conditional mean of *p*, price, for each value of *stratio*, the student-teacher ratio is given by the appropriate point on the line. As theory predicts, the mean house price conditional on the community's student-teacher ratio is inversely related to that ratio. Communities with more crowded schools are considered less desirable. Of course, this relationship between house price and the student-teacher ratio must be considered *ceteris paribus*: all other factors that might affect the price of the house are held constant when we evaluate the effect of a measure of community schools' quality on the house price.

This population regression function specifies that a set of k regressors in X and the stochastic disturbance u are the determinants of the response variable (or regressand) y. The model is usually assumed to contain a constant term, so that x_1 is understood to equal one for each observation. We may write the linear regression model in matrix form as

$$y = X\beta + u \tag{3}$$

where $X = \{x_1, x_2, \ldots, x_k\}$, an $N \times k$ matrix of sample values.

The key assumption in the linear regression model involves the relationship in the population between the regressors X and u. We may rewrite Equation (3) as

$$u = y - X\beta \tag{4}$$

We assume that

$$E(u \mid X) = 0 \tag{5}$$

i.e., that the u process has a *zero conditional mean*. This assumption states that the unobserved factors involved in the regression function are not related in any systematic manner to the observed factors. This approach to the regression model allows us to consider both non-stochastic and stochastic regressors in X without distinction; all that matters is that they satisfy the assumption of Equation (5).

We may use the zero conditional mean assumption (Equation (5)) to define a *method of moments* estimator of the regression function. Method of moments estimators are defined by *moment conditions* that are assumed to hold on the population moments. When we replace the unobservable population moments by their sample counterparts, we derive feasible estimators of the model's parameters. The zero conditional mean assumption gives rise to a set of k moment conditions, one for each x. In the population, each regressor x is assumed to be unrelated to u, or have zero covariance with u.We may then substitute calculated moments from our sample of data into the expression to derive a method of moments estimator for β:

$$\begin{aligned} X'u &= 0 \\ X'(y - X\beta) &= 0 \end{aligned} \tag{6}$$

Substituting calculated moments from our sample into the expression and replacing the unknown coefficients β with estimated values b in Equation (6) yields the *ordinary least squares* (OLS) estimator

$$\begin{aligned} X'y - X'Xb &= 0 \\ b &= (X'X)^{-1}X'y \end{aligned} \tag{7}$$

We may use b to calculate the regression residuals:

$$e = y - Xb \tag{8}$$

Given the solution for the vector b, the additional parameter of the regression problem σ_u^2, the population variance of the stochastic disturbance, may be estimated as a function of the regression residuals e_i:

$$s^2 = \frac{\sum_{i=1}^{N} e_i^2}{N-k} = \frac{e'e}{N-k} \tag{9}$$

where $(N-k)$ are the residual *degrees of freedom* of the regression problem. The positive square root of s^2 is often termed the standard error of regression, or standard error of estimate, or root mean square error. Stata uses the last terminology and displays s as `Root MSE`.

The ordinary least squares (OLS) estimator

The method of moments is not the only approach to deriving the linear regression estimator of Equation (7), which is the well-known formula from which the OLS estimator is derived.

Using the *least squares* or *minimum distance* approach to estimation, we want to solve the sample analogue to this problem as:

$$y = Xb + e \tag{10}$$

where b is the k-element vector of estimates of β and e is the N-vector of least squares residuals. We want to choose the elements of b to achieve the minimum error sum of squares, $e'e$.

The least squares problem may be written as

$$b = \arg\min_b \; e'e = \arg\min_b \; (y - Xb)'(y - Xb) \tag{11}$$

Assuming $N > k$ and linear independence of the columns of X (i.e., X must have full column rank), this problem has the unique solution

$$b = (X'X)^{-1}X'y \tag{12}$$

which are in linear algebraic terms named the least squares *normal equations*: k equations in the k unknowns $b_1, \ldots, b_k$.

The values calculated by least squares in Equation (12) are identical to those computed by the method of moments in Equation (7) since the first-order conditions used to derive the least squares solution above define the moment conditions employed by the method of moments.

To learn more about the sampling distribution of the OLS estimator, we must make some additional assumptions about the distribution of the stochastic disturbance u_i. In classical statistics, the u_i were assumed to be independent draws from the same normal distribution. The modern approach to econometrics drops the normality assumptions and simply assumes that the u_i are independent draws from an identical distribution (*i.i.d.*).

The normality assumption was sufficient to derive the exact finite-sample distribution of the OLS estimator. In contrast, under the *i.i.d.* assumption, one must use large-sample theory to derive the sampling distribution of the OLS estimator. The sampling distribution of the OLS estimator can be shown to be approximately normal using large-sample theory.

Specifically, when the u_i are $i.i.d.$ with finite variance σ_u^2, the OLS estimator b has a large-sample normal distribution with mean β and variance $\sigma_u^2 Q^{-1}$, where Q^{-1} is the variance-covariance matrix of X in the population. We refer this variance-covariance matrix of the estimator as a VCE.

Because it is unknown, we need a consistent estimator of the VCE. While neither σ_u^2 nor Q^{-1} is actually known, we can use consistent estimators of them to construct a consistent estimator of $\sigma_u^2 Q^{-1}$. Given that s^2 consistently estimates σ_u^2 and $(1/N)(X'X)$ consistently estimates Q, $s^2(X'X)^{-1}$ is a VCE of the OLS estimator.

Recovering estimation results

The `regress` command shares the features of all estimation (e-class) commands. Saved results from `regress` can be viewed by typing `ereturn list`. All Stata estimation commands save an estimated parameter vector as matrix `e(b)` and the estimated variance-covariance matrix of the parameters as matrix `e(V)`.

One item listed in the `ereturn list` should be noted: `e(sample)` listed as a `function` rather than a `scalar`, `macro` or `matrix`. The `e(sample)` function returns 1 if an observation was included in the estimation sample and 0 otherwise.

The `regress` command honors any *if* and *in* qualifiers and then practices -wise deletion to remove any observations with missing values across the set $\{y, X\}$. Thus, the observations actually used in generating the regression estimates may be fewer than those specified in the `regress` command. A subsequent command such as `summarize` *regressors if* (or *in*) will not necessarily provide the descriptive statistics of the observations on X that entered the regression unless all regressors and the y variable are in the *varlist*.

The set of observations actually used in estimation can easily be determined with the qualifier `if e(sample)`:

```
summarize regressors if e(sample)
```

will yield the appropriate summary statistics from the regression sample. It may be retained for later use by placing it in a new variable:

```
generate byte reg1sample = e(sample)
```

where we use the `byte` data type to save memory since `e(sample)` is an indicator {0,1} variable.

Hypothesis testing in regression

The application of regression methods is often motivated by the need to conduct tests of hypotheses which are implied by a specific theoretical model. In this section we discuss hypothesis tests and interval estimates assuming that the model is properly specified and that the errors are independently and identically distributed (*i.i.d.*). Estimators are random variables, and their sampling distributions depend on that of the error process.

There are three types of tests commonly employed in econometrics: *Wald* tests, *Lagrange multiplier* (LM) tests, and *likelihood ratio* (LR) tests. These tests share the same large-sample distribution, so that reliance on a particular form of test is usually a matter of convenience. Any hypothesis involving the coefficients of a regression equation can be expressed as one or more restrictions on the coefficient vector, reducing the dimensionality of the estimation problem. The Wald test involves estimating the unrestricted equation and evaluating the degree to which the restricted equation would differ in terms of its explanatory power.

The LM (or *score*) test involves estimating the restricted equation and evaluating the curvature of the objective function. These tests are often used to judge whether *i.i.d.* assumptions are satisfied.

The LR test involves comparing the objective function values of the unrestricted and restricted equations. It is often employed in maximum likelihood estimation.

Consider the general form of the Wald test statistic. Given the regression equation

$$y = X\beta + u \tag{13}$$

Any set of linear restrictions on the coefficient vector may be expressed as

$$R\beta = r \tag{14}$$

where R is a $q \times k$ matrix and r is a q-element column vector, with $q < k$. The q restrictions on the coefficient vector β imply that $(k - q)$ parameters are to be estimated in the restricted model. Each row of R imposes one restriction on the coefficient vector; a single restriction may involve multiple coefficients.

For instance, given the regression equation

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \tag{15}$$

We might want to test the hypothesis $H_0 : \beta_2 = 0$. This single restriction on the coefficient vector implies $R\beta = r$, where

$$\begin{aligned} R &= (\ 0\ 1\ 0\ 0\ ) \\ r &= (\ 0\ ) \end{aligned} \tag{16}$$

A test of $H_0 : \beta_2 = \beta_3$ would imply the single restriction

$$\begin{aligned} R &= (\ 0\ 1\ -1\ 0\ ) \\ r &= (\ 0\ ) \end{aligned} \tag{17}$$

Given a hypothesis expressed as $H_0 : R\beta = r$, we may construct the Wald statistic as

$$W = \frac{1}{s^2}(Rb - r)'[R(X'X)^{-1}R']^{-1}(Rb - r) \tag{18}$$

This quadratic form makes use of the vector of estimated coefficients, b, and evaluates the degree to which the restrictions fail to hold: the magnitude of the elements of the vector $(Rb - r)$. The Wald statistic evaluates the sums of squares of that vector, each weighted by a measure of their precision. Its denominator is s^2, the estimated variance of the error process, replacing the unknown parameter σ_u^2.

Stata contains a number of commands for the construction of hypothesis tests and confidence intervals which may be applied following an estimated regression. Some Stata commands report test statistics in the normal and χ^2 forms when the estimation commands are justified by large-sample theory. More commonly, the finite-sample *t* and *F* distributions are reported.

Stata's tests do not deliver verdicts with respect to the specified hypothesis, but rather present the *p-value* (or *prob-value*) of the test. Intuitively, the *p*-value is the probability of observing the estimated coefficient(s) if the null hypothesis is true.

In `regress` output, a number of test statistics and their *p*-values are automatically generated: that of the ANOVA *F* and the *t*-statistics for each coefficient, with the null hypothesis that the coefficients equal zero in the population. If we want to test additional hypotheses after a regression equation, three Stata commands are particularly useful: `test`, `testparm` and `lincom`. The `test` command may be specified as

`test` *coeflist*

where *coeflist* contains the names of one or more variables in the regression model.

A second syntax is

`test` *exp* = *exp*

where *exp* is an algebraic expression in the names of the regressors. The arguments of `test` may be repeated in parentheses in conducting joint tests. Additional syntaxes for `test` are available for multiple-equation models.

The `testparm` command provides similar functionality, but allows wildcards in the coefficient list:

`testparm` *varlist*

where the *varlist* may contain `*` or a hyphenated expression such as `ind1-ind9`. The `lincom` command evaluates linear combinations of coefficients:

`lincom` *exp*

where *exp* is any linear combination of coefficients that is valid in the second syntax of `test`. For `lincom`, the *exp* must *not* contain an equal sign.

If we want to test the hypothesis $H_0 : \beta_j = 0$, the ratio of the estimated coefficient to its estimated standard error is distributed t under the null hypothesis that the population coefficient equals zero. That ratio is displayed by `regress` as the t column of the coefficient table. In the following estimated equation, a test statistic for the significance of a coefficient could be produced by using the commands:

```
. regress lprice lnox ldist rooms stratio
      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  4,   501) =  175.86
       Model |  49.3987735     4  12.3496934             Prob > F      =  0.0000
    Residual |  35.1834974   501  .070226542            R-squared     =  0.5840
-------------+------------------------------           Adj R-squared =  0.5807
       Total |  84.5822709   505  .167489645            Root MSE      =    .265

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnox |    -.95354   .1167418    -8.17   0.000    -1.182904   -.7241762
       ldist |  -.1343401   .0431032    -3.12   0.002    -.2190255   -.0496548
       rooms |   .2545271   .0185303    13.74   0.000     .2181203    .2909338
     stratio |  -.0524512   .0058971    -8.89   0.000    -.0640373   -.0408651
       _cons |   11.08387   .3181115    34.84   0.000     10.45887    11.70886
------------------------------------------------------------------------------

. test rooms
 ( 1)  rooms = 0
       F(  1,   501) =  188.67
            Prob > F =    0.0000
```

In Stata's shorthand this is equivalent to the command `test _b[rooms] = 0` (and much easier to type). If we use the `test` command, we note that the statistic is displayed as *F(1,N-k)* rather than in the t_{N-k} form of the coefficient table.

As many hypotheses to which `test` may be applied involve more than one restriction on the coefficient vector—and thus more than one degree of freedom—Stata routinely displays an *F*-statistic.

If we cannot reject the hypothesis $H_0 : \beta_j = 0$, and wish to restrict the equation accordingly, we remove that variable from the list of regressors.

More generally, we may to test the hypothesis $\beta_j = \beta_j^0 = \theta$, where θ is any constant value. If theory suggests that the coefficient on variable `rooms` should be 0.33, then we may specify that hypothesis in `test`:

```
. qui regress lprice lnox ldist rooms stratio
. test rooms = 0.33
 ( 1)  rooms = .33
       F(  1,   501) =   16.59
            Prob > F =    0.0001
```

The estimates clearly distinguish the estimated coefficient of 0.25 from 0.33.

We might want to compute a point and interval estimate for the sum of several coefficients. We may do that with the `lincom` (linear combination) command, which allows the specification of any linear expression in the coefficients. In the context of our median housing price equation, let us consider an arbitrary restriction: that the coefficients on `rooms, ldist` and `stratio` sum to zero, so that we may write

$$H_0 : \beta_{rooms} + \beta_{ldist} + \beta_{stratio} = 0 \tag{19}$$

It is important to note that although this hypothesis involves *three* estimated coefficients, it only involves *one* restriction on the coefficient vector. In this case, we have unitary coefficients on each term, but that need not be so.

```
. qui regress lprice lnox ldist rooms stratio
. lincom rooms + ldist + stratio
 ( 1)  ldist + rooms + stratio = 0
```

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	.0677357	.0490714	1.38	0.168	-.0286753	.1641468

The sum of the three estimated coefficients is 0.068, with an interval estimate including zero. The *t*-statistic provided by `lincom` provides the same *p*-value as that which `test` would produce.

We may use `test` to consider equality of two of the coefficients, or to test that their ratio equals a particular value:

```
. regress lprice lnox ldist rooms stratio
```

Source	SS	df	MS		
Model	49.3987735	4	12.3496934	Number of obs	= 506
Residual	35.1834974	501	.070226542	F(4, 501)	= 175.86
				Prob > F	= 0.0000
Total	84.5822709	505	.167489645	R-squared	= 0.5840
				Adj R-squared	= 0.5807
				Root MSE	= .265

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnox	-.95354	.1167418	-8.17	0.000	-1.182904	-.7241762
ldist	-.1343401	.0431032	-3.12	0.002	-.2190255	-.0496548
rooms	.2545271	.0185303	13.74	0.000	.2181203	.2909338
stratio	-.0524512	.0058971	-8.89	0.000	-.0640373	-.0408651
_cons	11.08387	.3181115	34.84	0.000	10.45887	11.70886

```
. test ldist = stratio
 ( 1)  ldist - stratio = 0
       F(  1,   501) =    3.63
            Prob > F =    0.0574
. test lnox  = 10 * stratio
 ( 1)  lnox - 10 stratio = 0
       F(  1,   501) =   10.77
            Prob > F =    0.0011
```

Joint hypothesis tests

All of the tests illustrated above are presented as an F-statistic with one numerator degree of freedom since they only involve one restriction on the coefficient vector. In many cases, we wish to test an hypothesis involving multiple restrictions on the coefficient vector. Although the former test could be expressed as a t-test, the latter cannot. Multiple restrictions on the coefficient vector imply a *joint test*, the result of which is not simply a box score of individual tests.

A joint test is usually constructed in Stata by listing each hypothesis to be tested in parentheses on the `test` command. As presented above, the first syntax of the `test` command, `test` *coeflist*, perfoms the joint test that two or more coefficients are jointly zero, such as $H_0 : \beta_2 = 0$ and $\beta_3 = 0$.

It is important to understand that this joint hypothesis is not at all the same as $H_0' : \beta_2 + \beta_3 = 0$. The latter hypothesis will be satisfied by a locus of $\{\beta_2, \beta_3\}$ values: all pairs that sum to zero. The former hypothesis will only be satisfied at the point where *each coefficient* equals zero. The joint hypothesis may be tested for our median house price equation:

```
. regress lprice lnox ldist rooms stratio
      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  4,   501) =  175.86
       Model |  49.3987735     4  12.3496934             Prob > F      =  0.0000
    Residual |  35.1834974   501  .070226542            R-squared     =  0.5840
-------------+------------------------------           Adj R-squared =  0.5807
       Total |  84.5822709   505  .167489645            Root MSE      =    .265

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnox |    -.95354   .1167418    -8.17   0.000    -1.182904   -.7241762
       ldist |  -.1343401   .0431032    -3.12   0.002    -.2190255   -.0496548
       rooms |   .2545271   .0185303    13.74   0.000     .2181203    .2909338
     stratio |  -.0524512   .0058971    -8.89   0.000    -.0640373   -.0408651
       _cons |   11.08387   .3181115    34.84   0.000     10.45887    11.70886
------------------------------------------------------------------------------

. test lnox ldist
 ( 1)  lnox = 0
 ( 2)  ldist = 0
       F(  2,   501) =   58.95
            Prob > F =    0.0000
```

The data overwhelmingly reject the joint hypothesis that the model excluding `lnox` and `ldist` is correctly specified relative to the full model.

Tests of nonlinear hypotheses

What if the hypothesis tests to be conducted cannot be written in the linear form

$$H_0 : R\beta = r \tag{20}$$

for example, if theory predicts a certain value for the product of two coefficients in the model, or for an expression such as $(\beta_2/\beta_3 + \beta_4)$? Two Stata commands are analogues to those we have used above: `testnl` and `nlcom`.

The former allows specification of nonlinear hypotheses on the β values, but unlike `test`, the syntax `_b[`*varname*`]` must be used to refer to each coefficient value. If a joint test is to be conducted, the equations defining each nonlinear restriction must be written in parentheses, as illustrated below.

The `nlcom` command permits us to compute nonlinear combinations of the estimated coefficients in point and interval form, similar to `lincom`. Both commands employ the *delta method*, an approximation to the distribution of a nonlinear combination of random variables appropriate for large samples which constructs Wald-type tests. Unlike tests of linear hypotheses, nonlinear Wald-type tests based on the delta method are sensitive to the scale of the y and X data.

```
. regress lprice lnox ldist rooms stratio
      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  4,   501) =  175.86
       Model |  49.3987735     4  12.3496934            Prob > F      =  0.0000
    Residual |  35.1834974   501  .070226542           R-squared     =  0.5840
-------------+------------------------------           Adj R-squared =  0.5807
       Total |  84.5822709   505  .167489645           Root MSE      =    .265

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnox |    -.95354   .1167418    -8.17   0.000    -1.182904   -.7241762
       ldist |  -.1343401   .0431032    -3.12   0.002    -.2190255   -.0496548
       rooms |   .2545271   .0185303    13.74   0.000     .2181203    .2909338
     stratio |  -.0524512   .0058971    -8.89   0.000    -.0640373   -.0408651
       _cons |   11.08387   .3181115    34.84   0.000     10.45887    11.70886
------------------------------------------------------------------------------

. testnl _b[lnox] * _b[stratio] = 0.06
  (1)  _b[lnox] * _b[stratio] = 0.06
               F(1, 501) =        1.44
                Prob > F =        0.2306
```

In this example, we consider a restriction on the product of the coefficients of `lnox` and `stratio`. The product of these coefficients cannot be distinguished from 0.06.

We may also test a joint nonlinear hypothesis:

```
. regress lprice lnox ldist rooms stratio
      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  4,   501) =  175.86
       Model |  49.3987735     4  12.3496934            Prob > F      =  0.0000
    Residual |  35.1834974   501  .070226542           R-squared     =  0.5840
-------------+------------------------------           Adj R-squared =  0.5807
       Total |  84.5822709   505  .167489645           Root MSE      =    .265

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnox |    -.95354   .1167418    -8.17   0.000    -1.182904   -.7241762
       ldist |  -.1343401   .0431032    -3.12   0.002    -.2190255   -.0496548
       rooms |   .2545271   .0185303    13.74   0.000     .2181203    .2909338
     stratio |  -.0524512   .0058971    -8.89   0.000    -.0640373   -.0408651
       _cons |   11.08387   .3181115    34.84   0.000     10.45887    11.70886
------------------------------------------------------------------------------

. testnl ( _b[lnox] * _b[stratio] = 0.06 ) ///
>        ( _b[rooms] / _b[ldist] = 3 * _b[lnox] )
  (1)  _b[lnox] * _b[stratio] = 0.06
  (2)  _b[rooms] / _b[ldist] = 3 * _b[lnox]
           F(2, 501) =        5.13
            Prob > F =        0.0062

.
```

The joint hypothesis may be rejected at the 99% level.

Computing residuals and predicted values

After estimating a linear regression model with `regress` we may compute the regression residuals or the predicted values.

Computation of the residuals for each observation allows us to assess how well the model has done in explaining the value of the response variable for that observation. Is the in-sample prediction $\hat{y}_i$ much larger or smaller than the actual value y_i?

Computation of predicted values allows us to generate in-sample predictions: the values of the response variable generated by the estimated model. We may also want to generate out-of-sample predictions: that is, apply the estimated regression function to observations that were not used to generate the estimates. This may involve hypothetical values of the regressors or actual values. In the latter case, we may want to apply the estimated regression function to a separate sample (e.g., to Springfield-area communities rather than Boston-area communities) to evaluate its applicability beyond the regression sample.

If a regression model is well specified, it should generate reasonable predictions for any sample from the population. If out-of-sample predictions are poor, the model's specification may be too specific to the original sample.

Neither the residuals nor predicted values are calculated by Stata's `regress` command, but either may be computed immediately thereafter with the `predict` command. This command is given as

```
predict [type] newvar [if] [in] [, choice]
```

where *choice* specifies the quantity to be computed for each observation.

For linear regression, `predict`'s default action is the computation of predicted values. These are known as the *point predictions*. If the residuals are required, the command

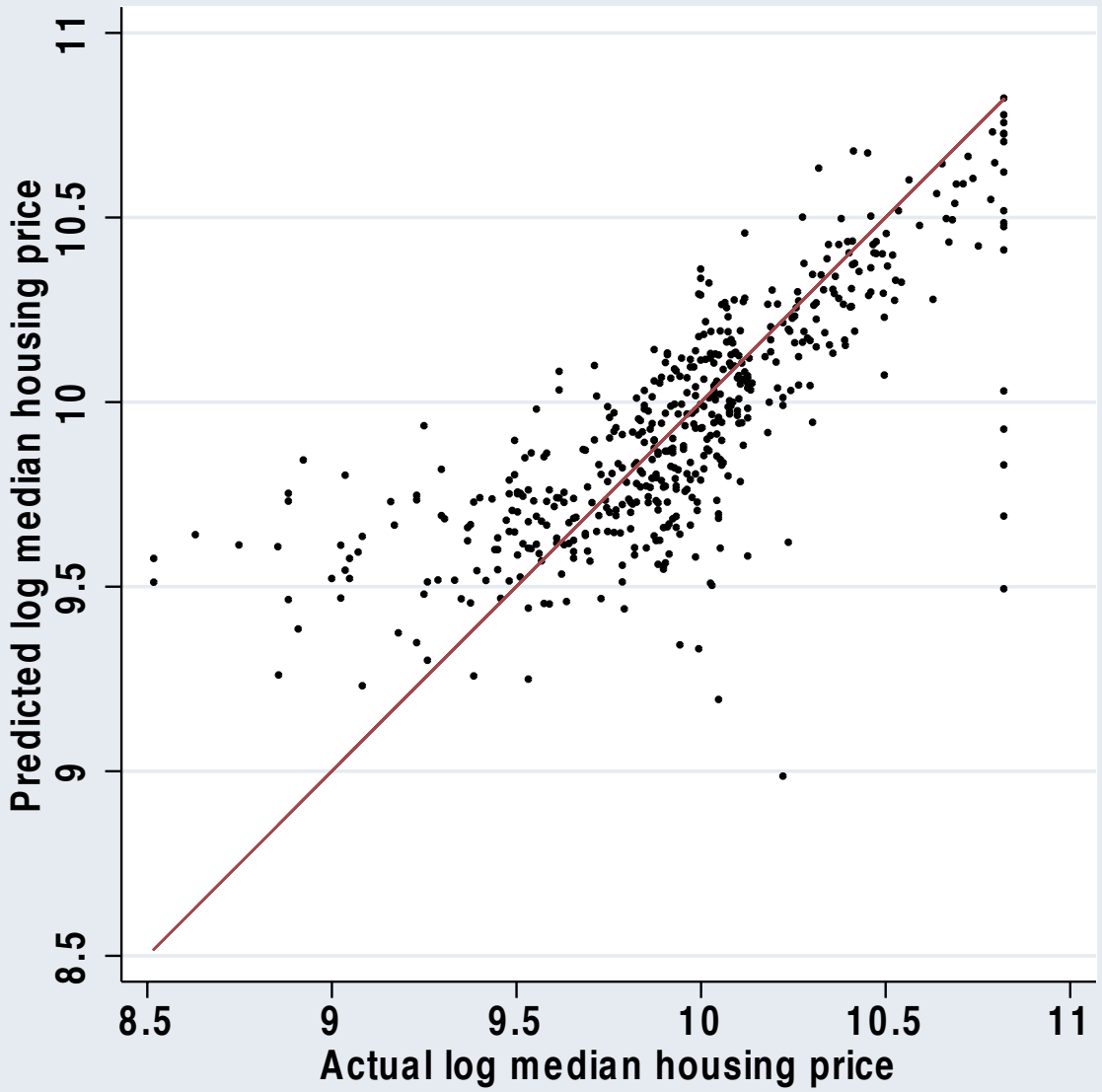
```
predict double lpriceeps, residual
```

should be used.

The regression estimates are only available to `predict` until another estimation command (e.g., `regress`) is issued. If these series are needed, they should be computed at the earliest opportunity. The use of `double` as the optional *type* in these commands ensures that the series will be generated with full numerical precision, and is strongly recommended.

We often would like to evaluate the quality of the regression fit in graphical terms. With a single regressor, a plot of actual and predicted values of y_i versus x_i will suffice. In multiple regression, the natural analogue is a plot of actual y_i versus the predicted $\hat{y}_i$ values.

Actual vs. predicted housing prices:



The aspect ratio has been constrained to unity so that points on the 45° line represent perfect predictions. Note that the model systematically overpredicts the (log) price of relatively low-priced houses, which may give cause for concern about the applicability of the model to lower-income communities. There are also a number of very high median prices that appear to be seriously underpredicted by the model.

Regression with non-i.i.d. errors

If the regression errors are independently and identically distributed (*i.i.d.*), OLS produces consistent estimates; their sampling distribution in large samples is normal with a mean at the true coefficient values and their *VCE* is consistently estimated by the standard formula.

If the zero conditional mean assumption holds but the errors are not *i.i.d.*, OLS produces consistent estimates whose sampling distribution in large samples is still normal with a mean at the true coefficient values, but whose variance cannot be consistently estimated by the standard formula.

We have two options when the errors are not *i.i.d*. First, we can use the consistent OLS point estimates with a different estimator of the *VCE* that accounts for non-*i.i.d*. errors. Alternatively, if we can specify how the errors deviate from *i.i.d*. in our regression model, we can use a different estimator that produces consistent and more efficient point estimates.

The tradeoff between these two methods is that of *robustness* versus *efficiency*. In a *robust* approach we place fewer restrictions on the estimator: the idea being that the consistent point estimates are good enough, although we must correct our estimator of their *VCE* to account for non-*i.i.d*. errors. In the *efficient* approach we incorporate an explicit specification of the non-*i.i.d*. distribution into the model. If this specification is appropriate, the additional restrictions which it implies will produce a more efficient estimator than that of the robust approach.

Robust standard errors

We will only discuss the robust approach. If the errors are conditionally heteroskedastic and we want to apply the robust approach, we use the Huber–White–sandwich estimator of the variance of the linear regression estimator, available in most Stata estimation commands as the `robust` option.

If the assumption of homoskedasticity is valid, the non-robust standard errors are more efficient than the robust standard errors. If we are working with a sample of modest size and the assumption of homoskedasticity is tenable, we should rely on non-robust standard errors. But since robust standard errors are very easily calculated in Stata, it is simple to estimate both sets of standard errors for a particular equation and consider whether inference based on the non-robust standard errors is fragile. In large data sets, it has become increasingly common practice to report robust (or Huber–White–sandwich) standard errors.

To illustrate the use of the robust estimator of the *VCE*, we use a dataset (`fertil2`) that contains data on 4,361 women from a developing country. The average woman in the sample is 30 years old, first bore a child at 19 and has had 3.2 children, with just under three children in the household. We expect that the number of children ever born is increasing in the mother's age and decreasing in their age at first birth, since the latter measure indicates when they began childbearing. The use of contraceptives is expected to decrease the number of children ever born.

We want to model the number of children ever born (`ceb`) to each woman based on their `age`, their age at first birth (`agefbrth`) and an indicator of whether they regularly use a method of contraception (`usemeth`).

For later use we employ `estimates store` to preserve the results of this (non-displayed) regression. We then estimate the same model using robust standard errors (the `robust` option on `regress`), saving those results with `estimates store`. The `estimates table` command is then used to display the two sets of coefficient estimates, standard errors and *t*-statistics of the regression model of `ceb`.

```
. qui regress ceb age agefbrth usemeth, robust
. estimates store Robust
. estimates table nonRobust Robust, b(%9.4f) se(%5.3f) t(%5.2f) ///
>   title(Estimates of CEB with OLS and Robust standard errors)
Estimates of CEB with OLS and Robust standard errors
```

Variable	nonRobust	Robust
age	0.2237	0.2237
	0.003	0.005
	64.89	47.99
agefbrth	-0.2607	-0.2607
	0.009	0.010
	-29.64	-27.26
usemeth	0.1874	0.1874
	0.055	0.061
	3.38	3.09
_cons	1.3581	1.3581
	0.174	0.168
	7.82	8.11

legend: b/se/t

Our priors are borne out by the estimates, although the effect of contraceptive use appears to be marginally significant. The robust estimates of the standard errors are quite similar to the non-robust estimates, suggesting that heteroskedasticity may not be a problem in this sample. Naturally, we might want to conduct a formal test of homoskedasticity.

The Newey–West estimator of the VCE

In an extension to Huber–White–sandwich robust standard errors, we may employ the *Newey–West* estimator that is appropriate in the presence of arbitrary heteroskedasticity and autocorrelation, thus known as the *HAC* estimator. Its use requires us to specify an additional parameter: the maximum order of any significant autocorrelation in the disturbance process, or the *maximum lag L*. One rule of thumb that has been used is to choose $L = \sqrt[4]{N}$. This estimator is available as the Stata command `newey`, which may be used as an alternative to `regress` for estimation of a regression with *HAC* standard errors.

Like the `robust` option, application of the *HAC* estimator does not modify the point estimates; it only affects the *VCE*. Test statistics based on the *HAC VCE* are robust to arbitrary heteroskedasticity and autocorrelation as well.

Testing for heteroskedasticity

After estimating a regression model we may base a test for heteroskedasticity on the regression residuals. If the assumption of homoskedasticity conditional on the regressors holds, it can be expressed as:

$$H_0 : Var\left(u|X_2, X_3, ..., X_k\right) = \sigma_u^2 \tag{21}$$

A test of this null hypothesis can evaluate whether the variance of the error process appears to be independent of the explanatory variables. We cannot observe the variances of each element of the disturbance process from samples of size one, but we can rely on the squared residual, e_i^2, to be a consistent estimator of σ_i^2. The logic behind any such test is that although the squared residuals will differ in magnitude across the sample, they should not be systematically related to *anything*, and a regression of squared residuals on any candidate Z_i should have no meaningful explanatory power.

One of the most common tests for heteroskedasticity is derived from this line of reasoning: the *Breusch–Pagan* test. The BP test, a Lagrange Multiplier (LM) test, involves regressing the squares of the regression residuals on a set of variables in an auxiliary regression

$$e_i^2 = d_1 + d_2 Z_{i2} + d_3 Z_{i3} + \ldots d_\ell Z_{i\ell} + v_i \tag{22}$$

The Breusch–Pagan (Cook–Weisberg) test may be executed with `estat hettest` after `regress`. If no regressor list (of Zs) is provided, `hettest` employs the fitted values from the previous regression (the $\hat{y}_i$ values). As mentioned above, the variables specified in the set of Zs could be chosen as measures which did not appear in the original regressor list.

We consider the potential scale-related heteroskedasticity in a model of median housing prices where the scale can be thought of as the average size of houses in each community, roughly measured by its number of rooms.

After estimating the model, we calculate three test statistics: that computed by `estat hettest` without arguments, which is the Breusch–Pagan test based on fitted values; `estat hettest` with a variable list, which uses those variables in the auxiliary regression; and White's general test statistic from `whitetst`.

```
 . qui regress lprice rooms crime ldist
. hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lprice
         chi2(1)      =   140.84
         Prob > chi2  =   0.0000
. hettest rooms crime ldist
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: rooms crime ldist
         chi2(3)      =   252.60
         Prob > chi2  =   0.0000
. whitetst
White´s general test statistic :  144.0052  Chi-sq( 9)  P-value =  1.5e-26
```

Each of these tests indicates that there is a significant degree of heteroskedasticity related to scale in this model.

Testing for serial correlation in the error distribution

How might we test for the presence of serially correlated errors? Just as in the case of pure heteroskedasticity, we base tests of serial correlation on the regression residuals. In the simplest case, autocorrelated errors follow the so-called $AR(1)$ model: an *autoregressive process* of order one, also known as a first-order Markov process:

$$u_t = \rho u_{t-1} + v_t, \ |\rho| < 1 \tag{23}$$

where the v_t are uncorrelated random variables with mean zero and constant variance.

If we suspect that there might be autocorrelation in the disturbance process of our regression model, we could use the estimated residuals to diagnose it. The empirical counterpart to u_t in Equation (23) will be the e_t series produced by `predict`. We estimate the auxiliary regression of e_t on e_{t-1} without a constant term, as the residuals have mean zero.

The resulting slope estimate is a consistent estimator of the first-order autocorrelation coefficient ρ of the u process from Equation (23). Under the null hypothesis, $\rho = 0$, so that a rejection of this null hypothesis by this Lagrange Multiplier (LM) test indicates that the disturbance process exhibits $AR(1)$ behavior.

A generalization of this procedure which supports testing for higher-order autoregressive disturbances is the Lagrange Multiplier (*LM*) test of Breusch and Leslie Godfrey. In this test, the regression residuals are regressed on the original X matrix augmented with p lagged residual series. The null hypothesis is that the errors are serially independent up to order p.

We illustrate the diagnosis of autocorrelation using a time series dataset `ukrates` of monthly short-term and long-term interest rates on UK government securities (Treasury bills and gilts), 1952m3–1995m12.

The model expresses the monthly change in the short rate `rs`, the Bank of England's monetary policy instrument as a function of the prior month's change in the long-term rate `r20`. The regressor and regressand are created on the fly by Stata's time series operators `D.` and `L.` The model represents a monetary policy reaction function.

```
. regress D.rs LD.r20
```

Source	SS	df	MS		Number of obs =	524
					F(1, 522) =	52.88
Model	13.8769739	1	13.8769739		Prob > F =	0.0000
Residual	136.988471	522	.262430021		R-squared =	0.0920
					Adj R-squared =	0.0902
Total	150.865445	523	.288461654		Root MSE =	.51228

D.rs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
r20						
LD.	.4882883	.0671484	7.27	0.000	.356374	.6202027
_cons	.0040183	.022384	0.18	0.858	-.0399555	.0479921

```
. predict double eps, residual
(2 missing values generated)
```

The Breusch–Godfrey test performed here considers the null of serial independence up to sixth order in the disturbance process, and that null is soundly rejected. We also present an unconditional test—the Ljung–Box *Q* test, available as command `wntestq`.

```
. estat bgodfrey, lags(6)
Breusch-Godfrey LM test for autocorrelation
```

lags(*p*)	chi2	df	Prob > chi2
6	17.237	6	0.0084

H0: no serial correlation

```
. wntestq eps
Portmanteau test for white noise
---------------------------------------
 Portmanteau (Q) statistic =    82.3882
 Prob > chi2(40)           =     0.0001
```

Both tests decisively reject the null of no serial correlation.

Regression with indicator variables

Economic and financial data come in three flavors: quantitative (or cardinal), ordinal (or ordered) and qualitative. Regression analysis handles quantitative data where both regressor and regressand may take on any real value. We also may work with *ordinal* or ordered data. They are distinguished from cardinal measurements in that an ordinal measure can only express inequality of two items, and not the magnitude of their difference.

We frequently encounter data that are purely *qualitative*, lacking any obvious ordering. If these data are coded as string variables, such as `M` and `F` for survey respondents' genders, we are not likely to mistake them for quantitative values. But in other cases, where a quality may be coded numerically, there is the potential to misuse this qualitative factor as quantitative.

One-way ANOVA

In order to test the hypothesis that a qualitative factor has an effect on a response variable, we must convert the qualitative factor into a set of *indicator variables*, or dummy variables. We then conduct a *joint test* on their coefficients. If the hypothesis to be tested includes a single qualitative factor, the estimation problem may be described as a one-way analysis of variance, or *one-way ANOVA*. ANOVA models may be expressed as linear regressions on an appropriate set of indicator variables.

This notion of the equivalence of one-way ANOVA and linear regression on a set of indicator variables that correspond to a single qualitative factor generalizes to multiple qualitative factors.

If there are two qualitative factors (e.g., race and sex) that are hypothesized to affect income, a researcher would regress income on two appropriate sets of indicator variables, each representing one of the qualitative factors. This is then an example of *two-way ANOVA*.

Using factor variables

One of the biggest innovations in Stata version 11 is the introduction of *factor variables*. Just as Stata's time series operators allow you to refer to lagged variables (`L.` or differenced variables (`D.`), the `i.` operator allows you to specify factor variables for any non-negative integer-valued variable in your dataset.

In the `auto.dta` dataset, where `rep78` takes on values 1...5, you could `list rep78 i.rep78`, or `summarize i.rep78`, or `regress mpg i.rep78`. Each one of those commands produces the appropriate indicator variables 'on-the-fly': not as permanent variables in your dataset, but available for the command.

For the `list` command, the variables will be named `1b.rep78`, `2.rep78` `...5.rep78`. The `b.` is the base level indicator, by default assigned to the smallest value. You can specify other base levels, such as the largest value, the most frequent value, or a particular value.

For the `summarize` command, only levels 2...5 will be shown; the base level is excluded from the list. Likewise, in a regression on `i.rep78`, the base level is the variable excluded from the regressor list to prevent perfect collinearity. The conditional mean of the excluded variable appears in the constant term.

Interaction effects

If this was the only feature of factor variables (being instantiated when called for) they would not be very useful. The real advantage of these variables is the ability to define `interaction effects` for both integer-valued and continuous variables. For instance, consider the indicator `foreign` in the `auto` dataset. We may use a new operator, `#`, to define an interaction:

```
regress mpg i.rep78 i.foreign i.rep78#i.foreign
```

All combinations of the two categorical variables will be defined, and included in the regression as appropriate (omitting base levels and cells with no observations).

In fact, we can specify this model more simply: rather than

```
regress mpg i.rep78 i.foreign i.rep78#i.foreign
```

we can use the *factorial interaction* operator, `##`:

```
regress mpg i.rep78##i.foreign
```

which will provide exactly the same regression, producing all first-level and second-level interactions. Interactions are not limited to pairs of variables; up to eight factor variables may be included.

Furthermore, factor variables may be interacted with continuous variables to produce analysis of covariance models. The continuous variables are signalled by the new `c.` operator:

```
regress mpg i.foreign i.foreign#c.displacement
```

which essentially estimates two regression lines: one for domestic cars, one for foreign cars. Again, the factorial operator could be used to estimate the same model:

```
regress mpg i.foreign##c.displacement
```

As we will see in discussing marginal effects, it is very advantageous to use this syntax to describe interactions, both among categorical variables and between categorical variables and continuous variables. Indeed, it is likewise useful to use the same syntax to describe squared (and cubed...) terms:

```
regress mpg i.foreign c.displacement c.displacement#c.displacement
```

In this model, we allow for an intercept shift for `foreign`, but constrain the slopes to be equal across foreign and domestic cars. However, by using this syntax, we may ask Stata to calculate the marginal effect $\partial mpg/\partial displacement$, taking account of the squared term as well, as Stata understands the mathematics of the specification in this explicit form.

Computing marginal effects

With the introduction of factor variables in Stata 11, a powerful new command has been added: `margins`, which supersedes earlier versions' `mfx` and `adjust` commands. Those commands remain available, but the new command has many advantages. Like those commands, `margins` is used after an estimation command.

In the simplest case, `margins` applied after a simple one-way ANOVA estimated with `regress i.rep78`, with `margins i.rep78`, merely displays the conditional means for each category of `rep78`.

```
. regress mpg i.rep78
```

Source	SS	df	MS
Model	549.415777	4	137.353944
Residual	1790.78712	64	27.9810488
Total	2340.2029	68	34.4147485

Number of obs	= 69
F(4, 64)	= 4.91
Prob > F	= 0.0016
R-squared	= 0.2348
Adj R-squared	= 0.1869
Root MSE	= 5.2897

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rep78						
2	-1.875	4.181884	-0.45	0.655	-10.22927	6.479274
3	-1.566667	3.863059	-0.41	0.686	-9.284014	6.150681
4	.6666667	3.942718	0.17	0.866	-7.209818	8.543152
5	6.363636	4.066234	1.56	0.123	-1.759599	14.48687
_cons	21	3.740391	5.61	0.000	13.52771	28.47229

```
. margins i.rep78
Adjusted predictions                              Number of obs   =         69
Model VCE    : OLS
Expression   : Linear prediction, predict()
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
rep78						
1	21	3.740391	5.61	0.000	13.66897	28.33103
2	19.125	1.870195	10.23	0.000	15.45948	22.79052
3	19.43333	.9657648	20.12	0.000	17.54047	21.3262
4	21.66667	1.246797	17.38	0.000	19.22299	24.11034
5	27.36364	1.594908	17.16	0.000	24.23767	30.4896

We now estimate a model including both displacement and its square:

```
. regress mpg i.foreign c.displacement c.displacement#c.displacement
```

Source	SS	df	MS		
Model	1416.01205	3	472.004018	Number of obs =	74
Residual	1027.44741	70	14.6778201	F(3, 70) =	32.16
				Prob > F =	0.0000
				R-squared =	0.5795
				Adj R-squared =	0.5615
Total	2443.45946	73	33.4720474	Root MSE =	3.8312

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.foreign	-2.88953	1.361911	-2.12	0.037	-5.605776	-.1732833
displacement	-.1482539	.0286111	-5.18	0.000	-.2053169	-.0911908
c. displacement# c. displacement	.0002116	.0000583	3.63	0.001	.0000953	.0003279
_cons	41.40935	3.307231	12.52	0.000	34.81328	48.00541

`margins` can then properly evaluate the regression function for domestic and foreign cars at selected levels of `displacement`:

```
. margins i.foreign, at(displacement=(100 300))
Adjusted predictions                              Number of obs   =         74
Model VCE    : OLS
Expression   : Linear prediction, predict()
1._at        : displacement    =         100
2._at        : displacement    =         300
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at#foreign						
1 0	28.69991	1.216418	23.59	0.000	26.31578	31.08405
1 1	25.81038	.8317634	31.03	0.000	24.18016	27.44061
2 0	15.97674	.7014015	22.78	0.000	14.60201	17.35146
2 1	13.08721	1.624284	8.06	0.000	9.903668	16.27074

In earlier versions of Stata, calculation of marginal effects in this model required some programming due to the nonlinear term `displacement`. Using `margins, dydx`, that is now simple. Furthermore, and most importantly, the default behavior of `margins` is to calculate average marginal effects (AMEs) rather than marginal effects at the average (MAE) or at some other point in the space of the regressors. In Stata 10, the user-written command `margeff` (Tamas Bartus, on the SSC Archive) was required to compute AMEs.

Current practice favors the use of AMEs: the computation of each observation's marginal effect with respect to an explanatory factor, averaged over the estimation sample, to the computation of MAEs (which reflect an average individual: e.g. a family with 2.3 children).

We illustrate by computing average marginal effects (AMEs) for the prior regression:

```
. margins, dydx(foreign displacement)
Average marginal effects                        Number of obs   =         74
Model VCE    : OLS
Expression   : Linear prediction, predict()
dy/dx w.r.t. : 1.foreign displacement
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.foreign	-2.88953	1.361911	-2.12	0.034	-5.558827	-.2202327
displacement	-.0647596	.007902	-8.20	0.000	-.0802473	-.049272

Note: dy/dx for factor levels is the discrete change from the base level.

Alternatively, we may compute elasticities or semi-elasticities:

```
. margins, eyex(displacement) at(displacement=(100(100)400))
Average marginal effects                        Number of obs   =         74
Model VCE    : OLS
Expression   : Linear prediction, predict()
ey/ex w.r.t. : displacement
1._at        : displacement    =         100
2._at        : displacement    =         200
3._at        : displacement    =         300
4._at        : displacement    =         400
```

	ey/ex	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
displacement _at						
1	-.3813974	.0537804	-7.09	0.000	-.486805	-.2759898
2	-.6603459	.0952119	-6.94	0.000	-.8469578	-.473734
3	-.4261477	.193751	-2.20	0.028	-.8058926	-.0464028
4	.5613844	.4817784	1.17	0.244	-.3828839	1.505653

Consider a model where we specify a factorial interaction between categorical and continuous covariates:

```
regress mpg i.foreign i.rep78##c.displacement
```

In this specification, each level of `rep78` has its own intercept and slope, whereas `foreign` only shifts the intercept term.

We may compute elasticities or semi-elasticities with the `over` option of `margins` for all combinations of `foreign` and `rep78`:

```
. margins, eyex(displacement) over(foreign rep78)
Average marginal effects                        Number of obs   =         69
Model VCE    : OLS
Expression   : Linear prediction, predict()
ey/ex w.r.t. : displacement
over         : foreign rep78
```

	ey/ex	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
displacement						
foreign# rep78						
0 1	-.7171875	.5342	-1.34	0.179	-1.7642	.3298253
0 2	-.5953046	.219885	-2.71	0.007	-1.026271	-.1643379
0 3	-.4620597	.0999242	-4.62	0.000	-.6579077	-.2662118
0 4	-.6327362	.1647866	-3.84	0.000	-.955712	-.3097604
0 5	-.8726071	.0983042	-8.88	0.000	-1.06528	-.6799345
1 3	-.128192	.0228214	-5.62	0.000	-.1729213	-.0834628
1 4	-.1851193	.0380458	-4.87	0.000	-.2596876	-.110551
1 5	-1.689962	.3125979	-5.41	0.000	-2.302642	-1.077281

The `margins` command has many other capabilities which we will not discuss here. Perusal of the reference manual article on `margins` would be useful to explore its additional features.

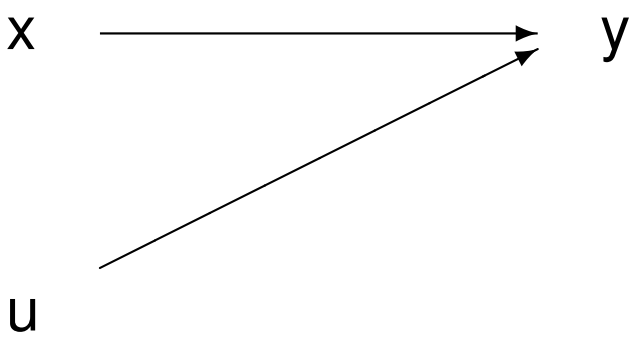
Regression with instrumental Variables

What are instrumental variables (IV) methods? Most widely known as a solution to *endogenous regressors*: explanatory variables correlated with the regression error term, IV methods provide a way to nonetheless obtain consistent parameter estimates.

Although IV estimators address issues of endogeneity, the violation of the zero conditional mean assumption caused by endogenous regressors can also arise for two other common causes: measurement error in regressors (errors-in-variables) and omitted-variable bias. The latter may arise in situations where a variable known to be relevant for the data generating process is not measurable, and no good proxies can be found.

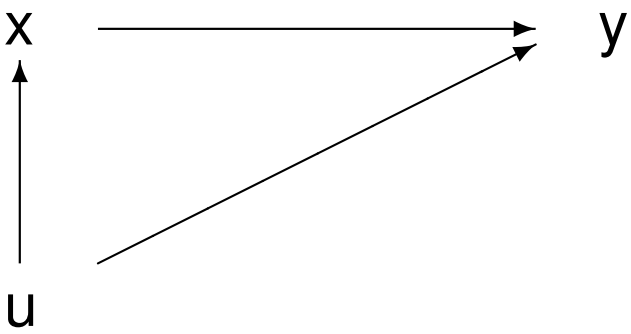
First let us consider a path diagram illustrating the problem addressed by IV methods. We can use ordinary least squares (OLS) regression to consistently estimate a model of the following sort.

Standard regression: $y = xb + u$
no association between x and u; OLS consistent



However, OLS regression breaks down in the following circumstance:

Endogeneity: $y = xb + u$
correlation between x and u; OLS inconsistent



The correlation between x and u (or the failure of the zero conditional mean assumption $E[u|x] = 0$) can be caused by any of several factors.

Endogeneity

We have stated the problem as that of *endogeneity*: the notion that two or more variables are jointly determined in the behavioral model. This arises naturally in the context of a *simultaneous equations model* such as a supply-demand system in economics, in which price and quantity are jointly determined in the market for that good or service.

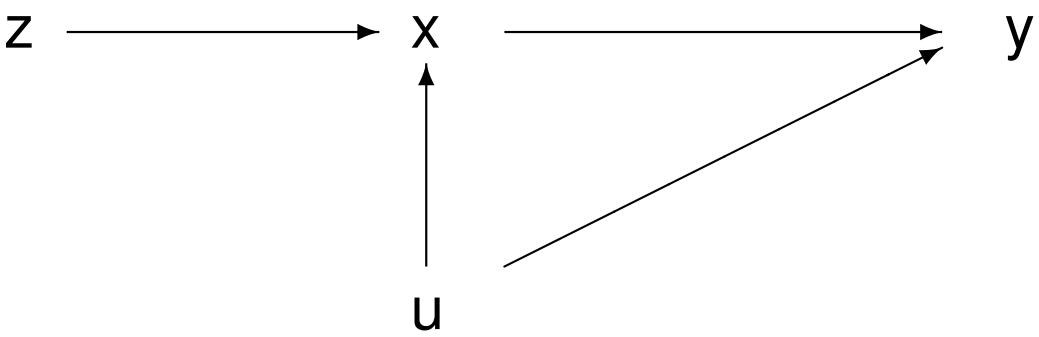
A shock or disturbance to either supply or demand will affect both the equilibrium price and quantity in the market, so that by construction both variables are correlated with any shock to the system. OLS methods will yield inconsistent estimates of any regression including both price and quantity, however specified.

As a different example, consider a cross-sectional regression of public health outcomes (say, the proportion of the population in various cities suffering from a particular childhood disease) on public health expenditures *per capita* in each of those cities. We would hope to find that spending is effective in reducing incidence of the disease, but we also must consider the *reverse causality* in this relationship, where the level of expenditure is likely to be partially determined by the historical incidence of the disease in each jurisdiction.

In this context, OLS estimates of the relationship will be biased even if additional controls are added to the specification. Although we may have no interest in modeling public health expenditures, we must be able to specify such an equation in order to *identify* the relationship of interest, as we discuss henceforth.

The solution provided by IV methods may be viewed as:

Instrumental variables regression: $y = xb + u$
z uncorrelated with u, correlated with x



The additional variable *z* is termed an *instrument* for *x*. In general, we may have many variables in *x*, and more than one *x* correlated with *u*. In that case, we shall need at least that many variables in *z*.

Choice of instruments

To deal with the problem of *endogeneity* in a supply-demand system, a candidate *z* will affect (e.g.) the quantity supplied of the good, but not directly impact the demand for the good. An example for an agricultural commodity might be temperature or rainfall: clearly exogenous to the market, but likely to be important in the production process.

For the public health example, we might use *per capita* income in each city as an instrument or *z* variable. It is likely to influence public health expenditure, as cities with a larger tax base might be expected to spend more on all services, and will not be directly affected by the unobserved factors in the primary relationship.

But why should we not always use IV?

It may be difficult to find variables that can serve as valid instruments. Many variables that have an effect on included endogenous variables also have a direct effect on the dependent variable.

IV estimators are innately *biased*, and their finite-sample properties are often problematic. Thus, most of the justification for the use of IV is asymptotic. Performance in small samples may be poor.

The precision of IV estimates is lower than that of OLS estimates (least squares is just that). In the presence of *weak instruments* (excluded instruments only weakly correlated with included endogenous regressors) the loss of precision will be severe, and IV estimates may be no improvement over OLS. This suggests we need a method to determine whether a particular regressor must be treated as endogenous.

IV estimation as a GMM problem

Before discussing further the motivation for various weak instrument diagnostics, we define the setting for IV estimation as a Generalized Method of Moments (GMM) optimization problem. Economists consider GMM to be the invention of Lars Hansen in his 1982 *Econometrica* paper, but as Alistair Hall points out in his 2005 book, the method has its antecedents in Karl Pearson's *Method of Moments* [MM] (1895) and Neyman and Egon Pearson's *minimum Chi-squared estimator* [MCE] (1928). Their MCE approach overcomes the difficulty with MM estimators when there are more moment conditions than parameters to be estimated. This was recognized by Ferguson (*Ann. Math. Stat.* 1958) for the case of *i.i.d.* errors, but his work had no impact on the econometric literature.

We consider the model

$$y = X\beta + u, \quad u \sim (0, \Omega)$$

with X $(N \times k)$ and define a matrix Z $(N \times \ell)$ where $\ell \geq k$. This is the Generalized Method of Moments IV (IV-GMM) estimator. The ℓ instruments give rise to a set of ℓ moments:

$$g_i(\beta) = Z_i' u_i = Z_i'(y_i - x_i\beta), \; i = 1, N$$

where each g_i is an ℓ-vector. The method of moments approach considers each of the ℓ moment equations as a sample moment, which we may estimate by averaging over N:

$$\bar{g}(\beta) = \frac{1}{N} \sum_{i=1}^{N} z_i(y_i - x_i\beta) = \frac{1}{N} Z'u$$

The GMM approach chooses an estimate that solves $\bar{g}(\hat{\beta}_{GMM}) = 0$.

If $\ell = k$, the equation to be estimated is said to be *exactly identified* by the *order condition* for identification: that is, there are as many excluded instruments as included right-hand endogenous variables. The method of moments problem is then k equations in k unknowns, and a unique solution exists, equivalent to the standard IV estimator:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

In the case of *overidentification* ($\ell > k$) we may define a set of k instruments:

$$\hat{X} = Z'(Z'Z)^{-1}Z'X = P_Z X$$

which gives rise to the *two-stage least squares* (2SLS) estimator

$$\hat{\beta}_{2SLS} = (\hat{X}'X)^{-1}\hat{X}'y = (X'P_Z X)^{-1}X'P_Z y$$

which despite its name is computed by this single matrix equation.

The IV-GMM approach

In the 2SLS method with overidentification, the ℓ available instruments are "boiled down" to the k needed by defining the P_Z matrix. In the IV-GMM approach, that reduction is not necessary. All ℓ instruments are used in the estimator. Furthermore, a *weighting matrix* is employed so that we may choose $\hat{\beta}_{GMM}$ so that the elements of $\bar{g}(\hat{\beta}_{GMM})$ are as close to zero as possible. With $\ell > k$, not all ℓ moment conditions can be exactly satisfied, so a criterion function that weights them appropriately is used to improve the efficiency of the estimator.

The GMM estimator minimizes the criterion

$$J(\hat{\beta}_{GMM}) = N\, \bar{g}(\hat{\beta}_{GMM})' W\, \bar{g}(\hat{\beta}_{GMM})$$

where W is a $\ell \times \ell$ symmetric weighting matrix.

Solving the set of FOCs, we derive the IV-GMM estimator of an overidentified equation:

$$\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y$$

which will be identical for all W matrices which differ by a factor of proportionality. The *optimal* weighting matrix, as shown by Hansen (1982), chooses $W = S^{-1}$ where S is the covariance matrix of the moment conditions to produce the most *efficient* estimator:

$$S = E[Z'uu'Z] = lim_{N\to\infty}\ N^{-1}[Z'\Omega Z]$$

With a consistent estimator of S derived from 2SLS residuals, we define the feasible IV-GMM estimator as

$$\hat{\beta}_{FEGMM} = (X'Z\ \hat{S}^{-1}Z'X)^{-1}X'Z\ \hat{S}^{-1}Z'y$$

where *FEGMM* refers to the *feasible efficient* GMM estimator.

The derivation makes no mention of the form of Ω, the variance-covariance matrix (*vce*) of the error process *u*. If the errors satisfy all classical assumptions are *i.i.d.*, $S = \sigma_u^2 I_N$ and the optimal weighting matrix is proportional to the identity matrix. The IV-GMM estimator is merely the standard IV (or 2SLS) estimator.

If there is heteroskedasticity of unknown form, we usually compute *robust* standard errors in any Stata estimation command to derive a consistent estimate of the *vce*. In this context,

$$\hat{S} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 Z_i' Z_i$$

where $\hat{u}$ is the vector of residuals from any consistent estimator of β (e.g., the 2SLS residuals). For an overidentified equation, the IV-GMM estimates computed from this estimate of *S* will be more efficient than 2SLS estimates.

We must distinguish the concept of IV/2SLS estimation with robust standard errors from the concept of estimating the same equation with IV-GMM, allowing for arbitrary heteroskedasticity. Compare an overidentified regression model estimated (a) with IV and classical standard errors and (b) with robust standard errors. Model (b) will produce the same point estimates, but different standard errors in the presence of heteroskedastic errors.

However, if we reestimate that overidentified model using the GMM two-step estimator, we will get different point estimates because we are solving a different optimization problem: one in the ℓ-space of the instruments (and moment conditions) rather than the k-space of the regressors, and $\ell > k$. We will also get different standard errors, and in general smaller standard errors as the IV-GMM estimator is more efficient. This does not imply, however, that summary measures of fit will improve.

If errors are considered to exhibit arbitrary intra-cluster correlation in a dataset with M clusters, we may derive a *cluster-robust* IV-GMM estimator using

$$\hat{S} = \sum_{j=1}^{M} \hat{u}_j' \hat{u}_j$$

where

$$\hat{u}_j = (y_j - x_j \hat{\beta}) X' Z (Z' Z)^{-1} z_j$$

The IV-GMM estimates employing this estimate of S will be both robust to arbitrary heteroskedasticity and intra-cluster correlation, equivalent to estimates generated by Stata's `cluster(`*varname*`)` option. For an overidentified equation, IV-GMM cluster-robust estimates will be more efficient than 2SLS estimates.

The IV-GMM approach may also be used to generate *HAC standard errors*: those robust to arbitrary heteroskedasticity and autocorrelation. Although the best-known *HAC* approach in econometrics is that of Newey and West, using the Bartlett kernel (per Stata's `newey`), that is only one choice of a *HAC* estimator that may be applied to an IV-GMM problem. Baum–Schaffer–Stillman's `ivreg2` (from the SSC Archive) and Stata 10's `ivregress` provide several choices for kernels. For some kernels, the kernel *bandwidth* (roughly, number of lags employed) may be chosen automatically in either command.

The ivreg2 command

The estimators we have discussed are available from Baum, Schaffer and Stillman's *ivreg2* package (`ssc describe ivreg2`). The `ivreg2` command has the same basic syntax as Stata's older `ivreg` command:

```
ivreg2 depvar [varlist1] (varlist2=instlist) ///
           [if] [in]  [, options]
```

The ℓ variables in `varlist1` and `instlist` comprise Z, the matrix of instruments. The k variables in `varlist1` and `varlist2` comprise X. Both matrices by default include a units vector.

By default `ivreg2` estimates the IV estimator, or 2SLS estimator if $\ell > k$. If the `gmm2s` option is specified in conjunction with `robust`, `cluster()` or `bw()`, it estimates the IV-GMM estimator.

With the `robust` option, the *vce* is heteroskedasticity-robust.

With the `cluster(`*varname*`)` option, the *vce* is cluster-robust.

With the `robust` and `bw( )` options, the *vce* is *HAC* with the default Bartlett kernel, or "Newey–West". Other `kernel( )` choices lead to alternative *HAC* estimators. In `ivreg2`, both `robust` and `bw( )` options must be specified for *HAC*. Estimates produced with `bw( )` alone are robust to arbitrary autocorrelation but assume homoskedasticity.

Example of IV and IV-GMM estimation

We illustrate with a wage equation estimated from the Griliches dataset (`griliches76`) of very young men's wages. Their log(wage) is explained by completed years of schooling, experience, job tenure and IQ score.

The IQ variable is considered endogenous, and instrumented with three factors: their mother's level of education (`med`), their score on a standardized test (`kww`) and their `age`. The estimation in `ivreg2` is performed with

```
ivreg2 lw s expr tenure (iq = med kww age)
```

where the parenthesized expression defines the *included endogenous* and *excluded exogenous* variables. You could also use official Stata's `ivregress 2sls`.

```
. esttab, label stat(rmse) mtitles(IV IVrob IVGMMrob) nonum
```

	IV	IVrob	IVGMMrob
iq score	-0.00509	-0.00509	-0.00676
	(-1.06)	(-1.01)	(-1.34)
completed years of~g	0.122***	0.122***	0.128***
	(7.68)	(7.51)	(7.88)
experience, years	0.0357***	0.0357***	0.0368***
	(5.15)	(5.10)	(5.26)
tenure, years	0.0405***	0.0405***	0.0443***
	(4.78)	(4.51)	(4.96)
Constant	4.441***	4.441***	4.523***
	(14.22)	(13.21)	(13.46)
rmse	0.366	0.366	0.372

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

These three columns compare standard IV (2SLS) estimates, IV with robust standard errors, and IV-GMM with robust standard errors, respectively. Notice that the coefficients' point estimates change when IV-GMM is employed, and that their *t*-statistics are larger than those of robust IV. Note also that the IQ score is not significant in any of these models.

Tests of overidentifying restrictions

If and only if an equation is *overidentified*, we may test whether the excluded instruments are appropriately independent of the error process. That test should always be performed when it is possible to do so, as it allows us to evaluate the validity of the instruments.

A test of *overidentifying restrictions* regresses the residuals from an IV or 2SLS regression on all instruments in Z. Under the null hypothesis that all instruments are uncorrelated with u, the test has a large-sample $\chi^2(r)$ distribution where r is the number of overidentifying restrictions.

Under the assumption of *i.i.d*. errors, this is known as a *Sargan test*, and is routinely produced by `ivreg2` for IV and 2SLS estimates. It can also be calculated after `ivreg` estimation with the `overid` command, which is part of the `ivreg2` suite. After `ivregress`, the command `estat overid` provides the test.

If we have used IV-GMM estimation in `ivreg2`, the test of overidentifying restrictions becomes *J*: the GMM criterion function. Although *J* will be identically zero for any exactly-identified equation, it will be positive for an overidentified equation. If it is "too large", doubt is cast on the satisfaction of the moment conditions underlying GMM.

The test in this context is known as the *Hansen test* or *J test*, and is calculated by `ivreg2` when the `gmm2s` option is employed.

The Sargan–Hansen test of overidentifying restrictions should be performed routinely in any overidentified model estimated with instrumental variables techniques. Instrumental variables techniques are powerful, but if a strong rejection of the null hypothesis of the Sargan–Hansen test is encountered, you should strongly doubt the validity of the estimates.

For instance, let's rerun the last IV-GMM model we estimated and focus on the test of overidentifying restrictions provided by the Hansen *J* statistic. The model is overidentified by two degrees of freedom, as there is one endogenous regressor and three excluded instruments. We see that the *J* statistic strongly rejects its null, casting doubts on the quality of these estimates.

Let's reestimate the model excluding `age` from the instrument list and see what happens. We will see that the sign and significance of the key endogenous regressor changes as we respecify the instrument list, and the p-value of the *J* statistic becomes large when `age` is excluded.

Example: Test of overidentifying restrictions

```
. esttab, label stat(j jdf jp) mtitles(age no_age) nonum
```

	age	no_age
iq score	-0.00676	0.0181**
	(-1.34)	(2.97)
completed years of~g	0.128***	0.0514**
	(7.88)	(2.63)
experience, years	0.0368***	0.0440***
	(5.26)	(5.58)
tenure, years	0.0443***	0.0303***
	(4.96)	(3.48)
Constant	4.523***	2.989***
	(13.46)	(7.58)
j	49.84	0.282
jdf	2	1
jp	1.50e-11	0.595

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

```
. sjlog close
```

We may be quite confident of some instruments' independence from *u* but concerned about others. In that case a *GMM distance* or *C* test may be used. The `orthog( )` option of `ivreg2` tests whether a *subset* of the model's overidentifying restrictions appear to be satisfied.

This is carried out by calculating two Sargan–Hansen statistics: one for the full model and a second for the model in which the listed variables are (a) considered endogenous, if included regressors, or (b) dropped, if excluded regressors. In case (a), the model must still satisfy the order condition for identification. The difference of the two Sargan–Hansen statistics, often termed the *GMM distance* or *C statistic*, will be distributed χ^2 under the null hypothesis that the specified orthogonality conditions are satisfied, with d.f. equal to the number of those conditions.

A variant on this strategy is implemented by the `endog( )` option of `ivreg2`, in which one or more variables considered endogenous can be tested for exogeneity. The *C* test in this case will consider whether the null hypothesis of their exogeneity is supported by the data.

If all endogenous regressors are included in the `endog( )` option, the test is essentially a test of whether IV methods are required to estimate the equation. If OLS estimates of the equation are consistent, they should be preferred. In this context, the test is equivalent to a *Hausman test* comparing IV and OLS estimates, as implemented by Stata's `hausman` command with the `sigmaless` option. Using `ivreg2`, you need not estimate and store both models to generate the test's verdict.

For instance, with the model above, we might question whether IV techniques are needed. We can conduct the *C* test via:

```
ivreg2 lw s expr tenure (iq=med kww), gmm2s robust endog(iq)
```

where the `endog(iq)` option tests the null hypothesis that `iq` is properly exogenous in this model. The test statistic has a p-value of 0.0108, suggesting that the data overwhelmingly reject the use of OLS in favor of IV. At the same time, the *J* statistic (with a p-value of 0.60) indicates that the overidentifying restrictions are not rejected.

The weak instruments problem

Instrumental variables methods rely on two assumptions: the excluded instruments are distributed independently of the error process, and they are sufficiently correlated with the included endogenous regressors. Tests of overidentifying restrictions address the *first* assumption, although we should note that a rejection of their null may be indicative that the exclusion restrictions for these instruments may be inappropriate. That is, some of the instruments have been improperly excluded from the regression model's specification.

The specification of an instrumental variables model asserts that the excluded instruments affect the dependent variable only *indirectly*, through their correlations with the included endogenous variables. If an excluded instrument exerts both direct and indirect influences on the dependent variable, the exclusion restriction should be rejected. This can be readily tested by including the variable as a regressor.

In our earlier example we saw that including `age` in the excluded instruments list caused a rejection of the *J* test. We had assumed that `age` could be treated as excluded from the model. Is that assumption warranted?

If `age` is entered as a regressor, it has a t-statistic over 8. Thus, its rejection as an excluded instrument may well reflect the misspecification of the equation, omitting age.

To test the *second* assumption—that the excluded instruments are sufficiently correlated with the included endogenous regressors—we should consider the goodness-of-fit of the "first stage" regressions relating each endogenous regressor to the entire set of instruments.

It is important to understand that the theory of single-equation ("limited information") IV estimation requires that all columns of X are conceptually regressed on all columns of Z in the calculation of the estimates. We cannot meaningfully speak of "this variable is an instrument for that regressor" or somehow restrict which instruments enter which first-stage regressions. Stata's `ivregress` or `ivreg2` will not let you do that because such restrictions only make sense in the context of estimating an entire system of equations by full-information methods (for instance, with `reg3`).

The `first` and `ffirst` options of `ivreg2` present several useful diagnostics that assess the first-stage regressions. If there is a single endogenous regressor, these issues are simplified, as the instruments either explain a reasonable fraction of that regressor's variability or not. With multiple endogenous regressors, diagnostics are more complicated, as each instrument is being called upon to play a role in each first-stage regression.

With sufficiently weak instruments, the asymptotic identification status of the equation is called into question. An equation identified by the order and rank conditions in a finite sample may still be *effectively unidentified*.

As Staiger and Stock (*Econometrica*, 1997) show, the weak instruments problem can arise even when the first-stage *t*- and *F*-tests are significant at conventional levels in a large sample. In the worst case, the bias of the IV estimator is the same as that of OLS, IV becomes inconsistent, and instrumenting only aggravates the problem.

Beyond the informal "rule-of-thumb" diagnostics such as $F > 10$, `ivreg2` computes several statistics that can be used to critically evaluate the strength of instruments. We can write the first-stage regressions as

$$X = Z\Pi + v$$

With X_1 as the endogenous regressors, Z_1 the excluded instruments and Z_2 as the included instruments, this can be partitioned as

$$X_1 = [Z_1 Z_2]\ [\Pi'_{11} \Pi'_{12}]' + v_1$$

The rank condition for identification states that the $L \times K_1$ matrix Π_{11} must be of full column rank.

We do not observe the true Π_{11}, so we must replace it with an estimate. Anderson's (John Wiley, 1984) approach to testing the rank of this matrix (or that of the full Π matrix) considers the *canonical correlations* of the *X* and *Z* matrices. If the equation is to be identified, all *K* of the canonical correlations will be significantly different from zero.

The squared canonical correlations can be expressed as eigenvalues of a matrix. Anderson's *CC* test considers the null hypothesis that the minimum canonical correlation is zero. Under the null, the test statistic is distributed χ^2 with $(L - K + 1)$ d.f., so it may be calculated even for an exactly-identified equation. Failure to reject the null suggests the equation is unidentified. `ivreg2` reports this Lagrange Multiplier (LM) statistic.

The Cragg–Donald statistic is a closely related test of the rank of a matrix. While the Anderson *CC* test is a LR test, the C–D test is a Wald statistic, with the same asymptotic distribution. The C–D statistic plays an important role in Stock and Yogo's work (see below). Both the Anderson and C–D tests are reported by `ivreg2` with the `first` option.

Recent research by Kleibergen and Paap (KP) (*J. Econometrics*, 2006) has developed a robust version of a test for the rank of a matrix: e.g. testing for *underidentification*. The statistic has been implemented by Kleibergen and Schaffer as command `ranktest`. If non-*i.i.d*. errors are assumed, the `ivreg2` output contains the K–P `rk` statistic in place of the Anderson canonical correlation statistic as a test of underidentification.

Stock and Yogo (Camb. U. Press festschrift, 2005) propose testing for weak instruments by using the *F*-statistic form of the C–D statistic. Their null hypothesis is that the estimator is weakly identified in the sense that it is subject to bias that the investigator finds unacceptably large.

Their test comes in two flavors: maximal relative bias (relative to the bias of OLS) and maximal size. The former test has the null that instruments are weak, where weak instruments are those that can lead to an asymptotic relative bias greater than some level *b*. This test uses the finite sample distribution of the IV estimator, and can only be calculated where the appropriate moments exist (when the equation is suitably overidentified: the m^{th} moment exists iff $m < (L - K + 1)$). The test is routinely reported in `ivreg2` and `ivregress` output when it can be calculated, with the relevant critical values calculated by Stock and Yogo.

The second test proposed by Stock and Yogo is based on the performance of the Wald test statistic for the endogenous regressors. Under weak identification, the test rejects too often. The test statistic is based on the rejection rate r tolerable to the researcher if the true rejection rate is 5%. Their tabulated values consider various values for r. To be able to reject the null that the size of the test is unacceptably large (versus 5%), the Cragg–Donald F statistic must exceed the tabulated critical value.

The Stock–Yogo test statistics, like others discussed above, assume *i.i.d.* errors. The Cragg–Donald F can be robustified in the absence of *i.i.d.* errors by using the Kleibergen–Paap `rk` statistic, which `ivreg2` reports in that circumstance.

When you may (and may not!) use IV

A common inquiry on Statalist: what should I do if I have an endogenous regressor that is a dummy variable? Should I, for instance, fit a probit model to generate the "hat values", estimate the model with OLS including those "hat values" instead of the 0/1 values, and puzzle over what to do about the standard errors?

An aside: you really do not want to do two-stage least squares "by hand", for one of the things that you must then deal with is getting the correct *VCE* estimate. The *VCE* and *RMSE* computed by the second-stage regression are not correct, as they are generated from the "hat values", not the original regressors. But back to our question.

Should I fit a probit model to generate the "hat values", estimate the model with OLS including those "hat values" instead of the 0/1 values, and puzzle over what to do about the standard errors?

No, you should just estimate the model with `ivreg2` or `ivregress`, treating the dummy endogenous regressor like any other endogenous regressor. This yields consistent point and interval estimates of its coefficient. There are other estimators (notably in the field of selection models or treatment regression) that explicitly deal with this problem, but they impose additional conditions on the problem. If you can use those methods, fine. Otherwise, just run IV. This solution is also appropriate for count data.

Another solution to the problem of an endogenous dummy (or count variable), as discussed by Cameron and Trivedi, is due to Basmann (*Econometrica*, 1957). Obtain fitted values for the endogenous regressor with appropriate nonlinear regression (logit or probit for a dummy, Poisson regression for a count variable) using all the instruments (included and excluded). Then do regular linear IV using the fitted value as an instrument, but the original dummy (or count variable) as the regressor. This is also a consistent estimator, although it has a different asymptotic distribution than does that of straight IV.

Equation nonlinear in endogenous variables

A second FAQ: what if my equation includes a nonlinear function of an endogenous regressor? For instance, from Wooldridge, *Econometric Analysis of Cross Section and Panel Data* (2002), p. 231, we might write the supply and demand equations for a good as

$$\begin{aligned} \log q^s &= \gamma_{12} \log(p) + \gamma_{13}[\log(p)]^2 + \delta_{11} z_1 + u_1 \\ \log q^d &= \gamma_{22} \log(p) + \delta_{22} z_2 + u_2 \end{aligned}$$

where we have suppressed intercepts for convenience. The exogenous factor z_1 shifts supply but not demand. The exogenous factor z_2 shifts demand but not supply. There are thus two exogenous variables available for identification.

This system is still *linear in parameters*, and we can ignore the log transformations on p, q. But it is, in Wooldridge's terms, *nonlinear in endogenous variables*, and identification must be treated differently.

If we used these equations to obtain $\log(p) = y_2$ as a function of exogenous variables and errors (the reduced form equation), the result would not be linear. $E[y_2|z]$ would not be linear unless $\gamma_{13} = 0$, assuming away the problem, and $E[y_2^2|z]$ will not be linear in any case. We might imagine that y_2^2 could just be treated as an additional endogenous variable, but then we need at least one more instrument. Where do we find it?

Given the nonlinearity, other functions of z_1 and z_2 will appear in a linear projection with y_2^2 as the dependent variable. Under linearity, the reduced form for y_2 involves z_1, z_2 and combinations of the errors. Square that reduced form, and $E[y_2^2|z]$ is a function of z_1^2, z_2^2 and $z_1 z_2$ (and the expectation of the squared composite error). Given that this relation has been derived under assumptions of linearity and homoskedasticity, we should also include the levels of z_1, z_2 in the projection (first stage regression).

The supply equation may then be estimated with instrumental variables using z_1, z_2, z_1^2, z_2^2 and $z_1 z_2$ as instruments. You could also use higher powers of the exogenous variables.

The mistake that may be made in this context involves what Hausman dubbed the *forbidden regression*: trying to mimic 2SLS by substituting fitted values for some of the endogenous variables inside the nonlinear functions. Nether the conditional expectation of the linear projection nor the linear projection operator passes through nonlinear functions, and such attempts "...rarely produce consistent estimators in nonlinear systems." (Wooldridge, p. 235)

In our example above, imagine regressing y_2 on exogenous variables, saving the predicted values, and squaring them. The "second stage" regression would then regress $\log(q)$ on $\hat{y}, \hat{y}^2, z_1$.

This two-step procedure does not yield the same results as estimating the equation by 2SLS, and it generally cannot produce consistent estimates of the structural parameters. The linear projection of the square is not the square of the linear projection, and the "by hand" approach assumes they are identical.

Further reading

There are many important considerations relating to the use of IV techniques, including LIML (limited-information maximum likelihood estimation) and GMM-CUE (continuously updated GMM estimates). For more details, please see

- Enhanced routines for instrumental variables/GMM estimation and testing. Baum CF, Schaffer ME, Stillman S, *Stata Journal* 7:4, 2007. Boston College Economics working paper no. 667, available from `http://ideas.repec.org`.
- *An Introduction to Modern Econometrics Using Stata*, Baum CF, Stata Press, 2006 (particularly Chapter 8).
- Instrumental variables and GMM: Estimation and testing. Baum CF, Schaffer ME, Stillman S, *Stata Journal* 3:1–31, 2003. Freely available from `http://stata-journal.com`.