Title

eivreg — Errors-in-variables regression

Description

`eivreg` fits errors-in-variables regression models when one or more of the independent variables are measured with error. To use `eivreg`, you must have an estimate of each independent variable's reliability or assume it is measured without error.

Quick start

Regression of `y` on `x1`, `x2`, and `x3` adjusted for `x1` measured with 90% reliability

```
eivreg y x1 x2 x3, reliab(x1 .9)
```

As above, but also specify 80% reliability for `x2`

```
eivreg y x1 x2 x3, reliab(x1 .9 x2 .8)
```

Menu

Statistics > Linear models and related > Errors-in-variables regression

Syntax

`eivreg` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

options	Description
Model	
`reliab(`*indepvar # [indepvar # [...]]*`)`	specify measurement reliability for each *indepvar* measured with error
Reporting	
`level(`*#*`)`	set confidence level; default is `level(95)`
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
`coeflegend`	display legend instead of statistics

indepvars may contain factor variables; see **[U] 11.4.3 Factor variables**.

`bootstrap`, `by`, `jackknife`, `rolling`, and `statsby` are allowed; see **[U] 11.1.10 Prefix commands**.

Weights are not allowed with the `bootstrap` prefix; see **[R] bootstrap**.

`aweight`s are not allowed with the `jackknife` prefix; see **[R] jackknife**.

`aweight`s and `fweight`s are allowed; see **[U] 11.1.6 weight**.

`coeflegend` does not appear in the dialog box.

See **[U] 20 Estimation and postestimation commands** for more capabilities of estimation commands.

Options

Model

`reliab(`*indepvar # [indepvar # [...]]*`)` specifies the measurement reliability for each independent variable measured with error. Reliabilities are specified as pairs consisting of an independent variable name (a name that appears in *indepvars*) and the corresponding reliability r, $0 < r \leq 1$. Independent variables for which no reliability is specified are assumed to have reliability 1. If the option is not specified, all variables are assumed to have reliability 1, and the result is thus the same as that produced by `regress` (the ordinary least-squares results).

Reporting

`level(`*#*`)`; see **[R] estimation options**.

display_options: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(`*#*`)`, `fvwrapon(`*style*`)`, `cformat(`*%fmt*`)`, `pformat(`*%fmt*`)`, `sformat(`*%fmt*`)`, and `nolstretch`; see **[R] estimation options**.

The following option is available with `eivreg` but is not shown in the dialog box:

`coeflegend`; see **[R] estimation options**.

Remarks and examples

stata.com

For an introduction to errors-in-variables regression, see Draper and Smith (1998, 89–91) or Kmenta (1997, 352–357). Treiman (2009, 258–261) compares the results of errors-in-variables regression with conventional regression.

Errors-in-variables regression models are useful when one or more of the independent variables are measured with additive noise. Standard regression (as performed by `regress`) would underestimate the effect of the variable, and the other coefficients in the model can be biased to the extent that they are correlated with the poorly measured variable. You can adjust for the biases if you know the reliability:

$$r = 1 - \frac{\text{noise variance}}{\text{total variance}}$$

That is, given the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, for some variable $\mathbf{x}_i$ in $\mathbf{X}$, the $\mathbf{x}_i$ is observed with error, $\mathbf{x}_i = \mathbf{x}_i^* + \mathbf{e}$, and the noise variance is the variance of $\mathbf{e}$. The total variance is the variance of $\mathbf{x}_i$.

▷ Example 1

Say that in our automobile data, the weight of cars was measured with error, and the reliability of our measured weight is 0.85. The result of this would be to underestimate the effect of `weight` in a regression of, say, `price` on `weight` and `foreign`, and it would also bias the estimate of the coefficient on `foreign` (because being of foreign manufacture is correlated with the weight of cars). We would ignore all of this if we fit the model with `regress`:

```
. use http://www.stata-press.com/data/r14/auto
(1978 Automobile Data)

. regress price weight foreign

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(2, 71)        =     35.35
       Model |   316859273         2   158429637   Prob > F        =    0.0000
    Residual |   318206123        71  4481776.38   R-squared       =    0.4989
-------------+----------------------------------   Adj R-squared   =    0.4848
       Total |   635065396        73  8699525.97   Root MSE        =      2117

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   3.320737   .3958784     8.39   0.000     2.531378    4.110096
     foreign |   3637.001    668.583     5.44   0.000     2303.885    4970.118
       _cons |  -4942.844   1345.591    -3.67   0.000    -7625.876   -2259.812
------------------------------------------------------------------------------
```

With `eivreg`, we can account for our measurement error:

```
. eivreg price weight foreign, r(weight .85)

                 assumed                  Errors-in-variables regression
    variable   reliability
------------------------                  Number of obs   =         74
      weight     0.8500                   F(  2,       71) =      50.37
           *     1.0000                   Prob > F         =     0.0000
                                          R-squared        =     0.6483
                                          Root MSE         =    1773.54

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |    4.31985    .431431    10.01   0.000     3.459601    5.180099
     foreign |    4637.32   624.5362     7.43   0.000      3392.03    5882.609
       _cons |  -8257.017   1452.086    -5.69   0.000    -11152.39   -5361.639
------------------------------------------------------------------------------
```

The effect of weight is increased—as we knew it would be—and here the effect of foreign manufacture is also increased. A priori, we knew only that the estimate of `foreign` might be biased; we did not know the direction.

◁

❑ Technical note

Swept under the rug in our example is how we would determine the reliability, r. We can easily see that a variable is measured with error, but we may not know the reliability because the ingredients for calculating r depend on the unobserved noise.

For our example, we made up a value for r, and in fact we do not believe that weight is measured with error at all, so the reported `eivreg` results have no validity. The `regress` results were the statistically correct results here.

But let's say that we do suspect that weight is measured with error and that we do not know r. We could then experiment with various values of r to describe the sensitivity of our estimates to possible error levels. We may not know r, but r does have a simple interpretation, and we could probably produce a sensible range for r by thinking about how the data were collected.

If the reliability, r, is less than the R^2 from a regression of the poorly measured variable on all the other variables, including the dependent variable, the information might as well not have been collected; no adjustment to the final results is possible. For our automobile data, running a regression of `weight` on `foreign` and `price` would result in an R^2 of 0.6743. Thus the reliability must be at least 0.6743 here. If we specify a reliability that is too small, `eivreg` will inform us and refuse to fit the model:

```
. eivreg price weight foreign, r(weight .6742)
reliability r() too small
r(399);
```

Returning to our problem of how to estimate r, too small or not, if the measurements are summaries of scaled items, the reliability may be estimated using the `alpha` command; see [MV] **alpha**. If the score is computed from factor analysis and the data are scored using `predict`'s default options (see [MV] **factor postestimation**), the square of the standard deviation of the score is an estimate of the reliability.

❑

❑ Technical note

Consider a model with more than one variable measured with error. For instance, say that our model is that `price` is a function of `weight`, `foreign`, and `mpg` and that both `weight` and `mpg` are measured with error.

```
. eivreg price weight foreign mpg, r(weight .85 mpg .9)
                  assumed                  Errors-in-variables regression
   variable     reliability
------------------------------             Number of obs     =         74
     weight       0.8500                   F(  3,        70) =     429.14
        mpg       0.9000                   Prob > F          =     0.0000
          *       1.0000                   R-squared         =     0.9728
                                           Root MSE          =     496.41

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   12.88302   .6820532    18.89   0.000     11.52271    14.24333
     foreign |   8268.951   352.8719    23.43   0.000      7565.17    8972.732
         mpg |   999.2043   73.60037    13.58   0.000      852.413    1145.996
       _cons |  -56473.19   3710.015   -15.22   0.000    -63872.58    -49073.8
------------------------------------------------------------------------------
```

❑

Stored results

eivreg stores the following in e():

Scalars	
e(N)	number of observations
e(df_m)	model degrees of freedom
e(df_r)	residual degrees of freedom
e(r2)	*R*-squared
e(F)	*F* statistic
e(rmse)	root mean squared error
e(rank)	rank of e(V)
Macros	
e(cmd)	eivreg
e(cmdline)	command as typed
e(depvar)	name of dependent variable
e(rellist)	*indepvars* and associated reliabilities
e(wtype)	weight type
e(wexp)	weight expression
e(properties)	b V
e(predict)	program used to implement predict
e(marginsok)	predictions allowed by margins
e(asbalanced)	factor variables fvset as asbalanced
e(asobserved)	factor variables fvset as asobserved
Matrices	
e(b)	coefficient vector
e(Cns)	constraints matrix
e(V)	variance–covariance matrix of the estimators
Functions	
e(sample)	marks estimation sample

Methods and formulas

Let the model to be fit be

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}$$

$$\mathbf{X} = \mathbf{X}^* + \mathbf{U}$$

where $\mathbf{X}^*$ are the true values and $\mathbf{X}$ are the observed values. Let $\mathbf{W}$ be the user-specified weights. If no weights are specified, $\mathbf{W} = \mathbf{I}$. If weights are specified, let $\mathbf{v}$ be the specified weights. If fweight frequency weights are specified, then $\mathbf{W} = \text{diag}(\mathbf{v})$. If aweight analytic weights are specified, then $\mathbf{W} = \text{diag}\{\mathbf{v}/(\mathbf{1}'\mathbf{v})(\mathbf{1}'\mathbf{1})\}$, meaning that the weights are normalized to sum to the number of observations.

The estimates $\mathbf{b}$ of $\boldsymbol{\beta}$ are obtained as $\mathbf{A}^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$, where $\mathbf{A} = \mathbf{X}'\mathbf{W}\mathbf{X} - \mathbf{S}$. $\mathbf{S}$ is a diagonal matrix with elements $N(1-r_i)s_i^2$. N is the number of observations, r_i is the user-specified reliability coefficient for the ith explanatory variable or 1 if not specified, and s_i^2 is the (appropriately weighted) variance of the variable.

The variance–covariance matrix of the estimators is obtained as $s^2\mathbf{A}^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{A}^{-1}$, where the root mean squared error $s^2 = (\mathbf{y}'\mathbf{W}\mathbf{y} - \mathbf{b}\mathbf{A}\mathbf{b}')/(N-p)$, where p is the number of estimated parameters.

References

Draper, N., and H. Smith. 1998. *Applied Regression Analysis*. 3rd ed. New York: Wiley.

Kmenta, J. 1997. *Elements of Econometrics*. 2nd ed. Ann Arbor: University of Michigan Press.

Treiman, D. J. 2009. *Quantitative Data Analysis: Doing Social Research to Test Ideas*. San Francisco: Jossey-Bass.

Also see

[R] **eivreg postestimation** — Postestimation tools for eivreg

[R] **regress** — Linear regression

[SEM] **example 24** — Reliability

[U] **20 Estimation and postestimation commands**