# **eivreg** — Errors-in-variables regression


Description | Quick start | Menu | Syntax | Options
Remarks and examples | Stored results | Methods and formulas | References | Also see


## 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)
```

Same 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, collect, jackknife, rolling, and statsby are allowed; see [U] **11.1.10 Prefix commands**.

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

aweights are not allowed with the jackknife prefix; see [R] **jackknife**.

aweights and fweights 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

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. Also see Lockwood and McCaffrey (2020) for how to use sem (see [SEM] **sem**) to fit errors-in-variables 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 https://www.stata-press.com/data/r19/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 | Coefficient  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, reliab(weight .85)
Errors-in-variables regression
                 Assumed
   Variable   reliability
----------------------------
                                        Number of obs     =         74
     weight        0.8500               F(  2,        71) =      18.46
          *        1.0000               Prob > F          =     0.0000
                                        R-squared         =     0.6483
                                        Root MSE          =    1773.54

-------------------------------------------------------------------------------
       price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+-----------------------------------------------------------------
      weight |    4.31985   .7134251     6.06   0.000      2.89732    5.742379
     foreign |    4637.32   849.0221     5.46   0.000     2944.418    6330.222
       _cons |  -8257.017   2390.337    -3.45   0.001    -13023.21   -3490.821
-------------------------------------------------------------------------------
```

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, reliab(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.

❑

### ▷ Example 2

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, reliab(weight .85 mpg .95)
Errors-in-variables regression

                  Assumed
    Variable   reliability
                                            Number of obs     =         74
      weight        0.8500                  F(   3,        70) =       9.58
         mpg        0.9500                  Prob > F          =     0.0000
           *        1.0000                  R-squared         =     0.8522
                                            Root MSE          =    1158.04

       price   Coefficient  Std. err.      t    P>|t|     [95% conf. interval]

      weight     9.69903   3.768985     2.57   0.012     2.182027    17.21603
     foreign    6918.624   2259.531     3.06   0.003     2412.132    11425.12
         mpg    627.6764   431.0284     1.46   0.150    -231.9826    1487.335
       _cons   -38545.27   20960.72    -1.84   0.070    -80350.11    3259.564
```

◁

## 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)` | $R^2$ |
| `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(V)` | variance–covariance matrix of the estimators |

Functions

| | |
|---|---|
| `e(sample)` | marks estimation sample |

In addition to the above, the following is stored in `r()`:

Matrices
`r(table)` matrix containing the coefficients with their standard errors, test statistics, $p$-values, and confidence intervals

Note that results stored in `r()` are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

## Methods and formulas

Let the model to be fit be

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

where $\mathbf{X}^*$ are the true values and $\mathbf{X}$ are the observed values. $\boldsymbol{\epsilon}$ and $\mathbf{U}$ are assumed to be independent and have zero means and finite fourth moments. Var($\mathbf{U}$) is assumed to be diagonal.

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 $(1 - r_j)s_j^2$. Here $r_j$ is the user-specified reliability coefficient for the $j$th explanatory variable (or 1 if not specified), and $s_j^2$ is the (appropriately weighted) sample variance of the variable.

The root mean squared error is $(\mathbf{y}'\mathbf{W}\mathbf{y} - \mathbf{b}\mathbf{A}\mathbf{b}')/(n - p)$, where $n$ is the number of observations and $p$ is the number of estimated parameters. The variance–covariance matrix of the estimators is obtained based on the formulas provided in Stefanski and Boos (2002), Buonaccorsi (2010), and Fuller (1987). For each $i = 1, 2, \ldots, n$, let residual $e_i = y_i - \mathbf{x}_i\mathbf{b}$, where $\mathbf{x}_i$ is the $i$th row of $\mathbf{X}$. Consider matrix $\mathbf{H}$, where the $i$th row of $\mathbf{H}$, $\mathbf{h}_i$, is

$$\mathbf{h}_i' = \begin{pmatrix} e_i x_{i1} + (x_{i1} - \overline{x}_1)^2(1 - r_1)b_1 \\ e_i x_{i2} + (x_{i2} - \overline{x}_2)^2(1 - r_2)b_2 \\ \vdots \\ e_i x_{ip} + (x_{ip} - \overline{x}_p)^2(1 - r_p)b_p \end{pmatrix}$$

where $\overline{x}_j$ is the weighted mean of the $j$th variable.

If analytic weights, `aweights`, are specified, the variance–covariance matrix is $\mathbf{A}^{-1}\ \mathbf{H}'\mathbf{W}\mathbf{W}\mathbf{H}\mathbf{A}^{-1}$; otherwise, it is $\mathbf{A}^{-1}\mathbf{H}'\mathbf{W}\mathbf{H}\mathbf{A}^{-1}$.

## References

Buonaccorsi, J. P. 2010. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781420066586.

Draper, N., and H. Smith. 1998. *Applied Regression Analysis*. 3rd ed. New York: Wiley. https://doi.org/10.1002/9781118625590.

Erickson, T., R. Parham, and T. M. Whited. 2017. Fitting the errors-in-variables model using high-order cumulants and moments. *Stata Journal* 17: 116–129.

Fuller, W. A. 1987. *Measurement Error Models*. New York: Wiley. https://doi.org/10.1002/9780470316665.

Kmenta, J. 1997. *Elements of Econometrics*. 2nd ed. Ann Arbor: University of Michigan Press. https://doi.org/10.3998/mpub.15701.

Lee, Y. J., and D. Wilhelm. 2020. Testing for the presence of measurement error in Stata. *Stata Journal* 20: 382–404.

Lockwood, J. R., and D. F. McCaffrey. 2020. Recommendations about estimating errors-in-variables regression in Stata. *Stata Journal* 20: 116–130.

Stefanski, L. A., and D. D. Boos. 2002. The calculus of M-estimation. *American Statistician* 56: 29–38. https://doi.org/10.1198/000313002753631330.

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**



For suggested citations, see the FAQ on citing Stata documentation.