vwls — Variance-weighted least squares

Description

vwls estimates a linear regression using variance-weighted least squares. It differs from ordinary least-squares (OLS) regression in that it does not assume homogeneity of variance, but requires that the conditional variance of depvar be estimated prior to the regression. The estimated variance need not be constant across observations. vwls treats the estimated variance as if it were the true variance when it computes standard errors of the coefficients.

You must supply an estimate of the conditional standard deviation of depvar to vwls by using the sd(varname) option, or you must have grouped data with the groups defined by the indepvars variables. In the latter case, vwls treats all indepvars as categorical variables, computes the mean and standard deviation of depvar separately for each subgroup, and computes the regression of the subgroup means on indepvars.

regress with analytic weights can be used to produce another kind of “variance-weighted least squares”; see Remarks and examples for an explanation of the difference.

Quick start

Variance-weighted least-squares regression of y on x1 and x2, with the estimated conditional std. dev. of y stored in sd

```
vwls y1 x1 x2, sd(sd)
```

Add categorical variable a using factor-variable syntax

```
vwls y1 x1 x2 i.a, sd(sd)
```

As above, but restrict the sample to cases where v is greater than 1

```
vwls y1 x1 x2 i.a if v>1, sd(sd)
```

Variance-weighted least-squares regression for grouped data with subgroups defined by a2 and a3

```
vwls y2 i.a2 i.a3
```

Menu

Statistics > Linear models and related > Other > Variance-weighted least squares
Syntax

\texttt{vwls depvar indepvars [if] [in] [weight] [, options]}

<table>
<thead>
<tr>
<th>options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td></td>
</tr>
<tr>
<td>\texttt{noconstant}</td>
<td>suppress constant term</td>
</tr>
<tr>
<td>\texttt{sd(varname)}</td>
<td>variable containing estimate of conditional standard deviation</td>
</tr>
<tr>
<td>Reporting</td>
<td></td>
</tr>
<tr>
<td>\texttt{level(#)}</td>
<td>set confidence level; default is \texttt{level(95)}</td>
</tr>
<tr>
<td>\texttt{display_options}</td>
<td>control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling</td>
</tr>
<tr>
<td>\texttt{coeflegend}</td>
<td>display legend instead of statistics</td>
</tr>
</tbody>
</table>

\texttt{indepvars} may contain factor variables; see \texttt{[U] 11.4.3 Factor variables.}

bootstrap, by, jackknife, rolling, and \texttt{statsby} are allowed; see \texttt{[U] 11.1.10 Prefix commands.}

Weights are not allowed with the \texttt{bootstrap} prefix; see \texttt{[R] bootstrap.}

\texttt{fweights} are allowed; see \texttt{[U] 11.1.6 weight.}

\texttt{coeflegend} does not appear in the dialog box.

See \texttt{[U] 20 Estimation and postestimation commands} for more capabilities of estimation commands.

Options

\texttt{noconstant}; see \texttt{[R] Estimation options.}

\texttt{sd(varname)} is an estimate of the conditional standard deviation of \texttt{depvar} (that is, it can vary observation by observation). All values of \texttt{varname} must be \texttt{> 0}. If you specify \texttt{sd()}, you cannot use \texttt{fweights}.

If \texttt{sd()} is not given, the data will be grouped by \texttt{indepvars}. Here \texttt{indepvars} are treated as categorical variables, and the means and standard deviations of \texttt{depvar} for each subgroup are calculated and used for the regression. Any subgroup for which the standard deviation is zero is dropped.

\texttt{level(#)}; see \texttt{[R] Estimation options.}

\texttt{display_options}: \texttt{noci}, \texttt{nopvalues}, \texttt{nomitted}, \texttt{vsquish}, \texttt{noemptycells}, \texttt{baselevels}, \texttt{allbaselevels}, \texttt{nofvlabel}, \texttt{fvwrap(#)}, \texttt{fvwrapon(style)}, \texttt{cformat(\%fmt)}, \texttt{pformat(\%fmt)}, \texttt{sformat(\%fmt)}, and \texttt{nolstretch}; see \texttt{[R] Estimation options.}

The following option is available with \texttt{vwls} but is not shown in the dialog box:

\texttt{coeflegend}; see \texttt{[R] Estimation options.}
Remarks and examples

The \texttt{vwls} command is intended for use with two special—and different—types of data. The first contains data that consist of measurements from physical science experiments in which all error is due solely to measurement errors and the sizes of the measurement errors are known.

You can also use variance-weighted least-squares linear regression for certain problems in categorical data analysis, such as when all the independent variables are categorical and the outcome variable is either continuous or a quantity that can sensibly be averaged. If each of the subgroups defined by the categorical variables contains a reasonable number of subjects, then the variance of the outcome variable can be estimated independently within each subgroup. For the purposes of estimation, \texttt{vwls} treats each subgroup as one observation, with the dependent variable being the subgroup mean of the outcome variable.

The \texttt{vwls} command fits the model

\[ y_i = x_i \beta + \varepsilon_i \]

where the errors $\varepsilon_i$ are independent normal random variables with the distribution $\varepsilon_i \sim N(0, \nu_i)$. The independent variables $x_i$ are assumed to be known without error.

As described above, \texttt{vwls} assumes that you already have estimates $s_i^2$ for the variances $\nu_i$. The error variance is not estimated in the regression. The estimates $s_i^2$ are used to compute the standard errors of the coefficients; see Methods and formulas below.

In contrast, weighted OLS regression assumes that the errors have the distribution $\varepsilon_i \sim N(0, \sigma^2/w_i)$, where the $w_i$ are known weights and $\sigma^2$ is an unknown parameter that is estimated in the regression. This is the difference from variance-weighted least squares: in weighted OLS, the magnitude of the error variance is estimated in the regression using all the data.

Example 1

An artificial, but informative, example illustrates the difference between variance-weighted least squares and weighted OLS.

We measure the quantities $x_i$ and $y_i$ and estimate that the standard deviation of $y_i$ is $s_i$. We enter the data into Stata:

```
. use https://www.stata-press.com/data/r16/vwlsxmpl
. list
```

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.9</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3.2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4.3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4.9</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6.0</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7.2</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>7.9</td>
</tr>
</tbody>
</table>

Because we want observations with smaller variance to carry larger weight in the regression, we compute an OLS regression with analytic weights proportional to the inverse of the squared standard deviations:
If we compute a variance-weighted least-squares regression by using `vwls`, we get the same results for the coefficient estimates but very different standard errors:

```
vwls y x, sd(s)
```

Although the values of $y_i$ were nicely linear with $x_i$, the `vwls` regression used the large estimates for the standard deviations to compute large standard errors for the coefficients. For weighted OLS regression, however, the scale of the analytic weights has no effect on the standard errors of the coefficients—only the relative proportions of the analytic weights affect the regression.

If we are sure of the sizes of our error estimates for $y_i$, using `vwls` is valid. However, if we can estimate only the relative proportions of error among the $y_i$, then `vwls` is not appropriate.

### Example 2

Let’s now consider an example of the use of `vwls` with categorical data. Suppose that we have blood pressure data for $n = 400$ subjects, categorized by gender and race (black or white). Here is a description of the data:
. use https://www.stata-press.com/data/r16/bp
. table gender race, c(mean bp sd bp freq) row col format(%8.1f)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Race</th>
<th>White</th>
<th>Black</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>117.1</td>
<td>118.5</td>
<td>117.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.3</td>
<td>11.6</td>
<td>10.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>100.0</td>
<td>200.0</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>122.1</td>
<td>125.8</td>
<td>124.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.6</td>
<td>15.5</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>100.0</td>
<td>200.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>119.6</td>
<td>122.2</td>
<td>120.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.7</td>
<td>14.1</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200.0</td>
<td>200.0</td>
<td>400.0</td>
</tr>
</tbody>
</table>

Performing a variance-weighted regression using `vwls` gives

. vwls bp gender race

Variance-weighted least-squares regression
Number of obs = 400
Goodness-of-fit chi2(1) = 0.88
Model chi2(2) = 27.11
Prob > chi2 = 0.3486
Prob > chi2 = 0.0000

| bp    | Coef. | Std. Err. | z     | P>|z|    | [95% Conf. Interval] |
|-------|-------|-----------|-------|--------|----------------------|
| gender| 5.876522 | 1.170241 | 5.02  | 0.000  | 3.582892 | 8.170151 |
| race  | 2.372818 | 1.191683 | 1.99  | 0.046  | 0.0371631 | 4.708473 |
| _cons | 116.6496 | .9296297 | 125.48 | 0.000  | 114.8266 | 118.4707 |

By comparison, an OLS regression gives the following result:

. regress bp gender race

Source | SS     | df | MS     | Number of obs = 400
Model   | 4485.66639 | 2  | 2242.83319 | F(2, 397) = 15.24
Residual| 58442.7305 | 397 | 147.210908 | Prob > F = 0.0000
Total   | 62928.3969 | 399 | 157.71528 | R-squared = 0.0713
Adj R-squared = 0.0666

| bp    | Coef. | Std. Err. | t     | P>|t|    | [95% Conf. Interval] |
|-------|-------|-----------|-------|--------|----------------------|
| gender| 6.1775 | 1.213305 | 5.09  | 0.000  | 3.792194 | 8.562806 |
| race  | 2.5875 | 1.213305 | 2.13  | 0.034  | 0.2021938 | 4.972806 |
| _cons | 116.4862 | 1.050753 | 110.86 | 0.000  | 114.4205 | 118.552 |

Note the larger value for the `race` coefficient (and smaller p-value) in the OLS regression. The assumption of homogeneity of variance in OLS means that the mean for black men pulls the regression line higher than in the `vwls` regression, which takes into account the larger variance for black men and reduces its effect on the regression.
Stored results

vwls stores the following in e():

Scalars
- e(N): number of observations
- e(df_m): model degrees of freedom
- e(chi2): model $\chi^2$
- e(df_gf): goodness-of-fit degrees of freedom
- e(chi2_gf): goodness-of-fit $\chi^2$
- e(rank): rank of e(V)

Macros
- e(cmd): vwls
- e(cmdline): command as typed
- e(depvar): name of dependent variable
- e(properties): b V
- e(predict): program used to implement predict
- 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 $y = (y_1, y_2, \ldots, y_n)'$ be the vector of observations of the dependent variable, where $n$ is the number of observations. When sd() is specified, let $s_1, s_2, \ldots, s_n$ be the standard deviations supplied by sd(). For categorical data, when sd() is not given, the means and standard deviations of $y$ for each subgroup are computed, and $n$ becomes the number of subgroups, $y$ is the vector of subgroup means, and $s_i$ are the standard deviations for the subgroups.

Let $V = \text{diag}(s_1^2, s_2^2, \ldots, s_n^2)$ denote the estimate of the variance of $y$. Then the estimated regression coefficients are

$$b = (X'V^{-1}X)^{-1}X'V^{-1}y$$

and their estimated covariance matrix is

$$\hat{\text{Cov}}(b) = (X'V^{-1}X)^{-1}$$

A statistic for the goodness of fit of the model is

$$Q = (y - Xb)'V^{-1}(y - Xb)$$

where $Q$ has a $\chi^2$ distribution with $n - k$ degrees of freedom ($k$ is the number of independent variables plus the constant, if any).
References


Also see

[R] `vwls postestimation` — Postestimation tools for `vwls`

[R] `regress` — Linear regression

[U] 11.1.6 `weight`

[U] 20 Estimation and postestimation commands