**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
```
Syntax

vwls depvar indepvars [if] [in] [weight] [, options]

options    Description

Model
noconstant  suppress constant term
sd(varname) variable containing estimate of conditional standard deviation

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.
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

noconstant; see [R] Estimation options.

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

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

level(#); see [R] Estimation options.

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

The following option is available with vwls but is not shown in the dialog box:
coeflegend; see [R] Estimation options.
Remarks and examples

The `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, `vwls` treats each subgroup as one observation, with the dependent variable being the subgroup mean of the outcome variable.

The `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, `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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.2</td>
<td>.5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.9</td>
<td>.5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3.2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4.3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4.9</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6.0</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7.2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>7.9</td>
<td>2</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)
```

Variance-weighted least-squares regression
Number of obs = 8
Goodness-of-fit chi2(6) = 0.28
Model chi2(1) = 33.24
Prob > chi2 = 0.9996
Prob > chi2 = 0.0000

| y | Coef.  | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|---|--------|-----------|------|-----|----------------------|
| x | .9824683 | .170409 | 5.77 | 0.000 | .6484728 1.316464 |
| _cons | .1138554 | .51484 | 0.22 | 0.825 | -.8952124 1.122923 |

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:
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    | 116.6486 | 0.9296297 | 125.48| 0.000 | 114.8266 – 118.4707 |

By comparison, an OLS regression gives the following result:

```
regress bp gender race
```

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4485.66639</td>
<td>2</td>
<td>2242.83319</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>58442.7305</td>
<td>397</td>
<td>147.210908</td>
<td>R-squared = 0.0713</td>
</tr>
<tr>
<td>Total</td>
<td>62928.3969</td>
<td>399</td>
<td>157.71528</td>
<td>Root MSE = 12.133</td>
</tr>
</tbody>
</table>

| 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

Methods and formulas

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ 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, $\mathbf{y}$ is the vector of subgroup means, and $s_i$ are the standard deviations for the subgroups.

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

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

and their estimated covariance matrix is

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

A statistic for the goodness of fit of the model is

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

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