hetregress — Heteroskedastic linear regression

Description

`hetregress` fits a multiplicative heteroskedastic linear regression by modeling the variance as an exponential function of the specified variables using either maximum likelihood (ML; the default) or Harvey's two-step generalized least-squares (GLS) method.

Quick start

Heteroskedastic regression model of `y` on `x1`, using `x2` to model the variance

```
hetregress y x1, het(x2)
```

Using Harvey's two-step GLS estimator instead of the default ML

```
hetregress y x1, het(x2) twostep
```

With robust standard errors

```
hetregress y x1, het(x2) vce(robust)
```

Perform a Wald test on the variance instead of a likelihood-ratio (LR) test

```
hetregress y x1, het(x2) waldhet
```

Menu

Statistics > Linear models and related > Heteroskedastic linear regression

Syntax

Maximum likelihood estimation

```
hetregress depvar [indepvars] [if] [in] [weight] [, ml_options]
```

Two-step GLS estimation

```
hetregress depvar [indepvars] [if] [in], twostep het(varlist) [ts_options]
```

ml_options	Description
Model	
mle	use maximum likelihood estimator; the default
het(*varlist*)	independent variables to model the variance
noconstant	suppress constant term
constraints(*constraints*)	apply specified linear constraints
SE/Robust	
vce(*vcetype*)	*vcetype* may be oim, robust, cluster *clustvar*, opg, bootstrap, or jackknife
Reporting	
level(#)	set confidence level; default is level(95)
lrmodel	perform the LR model test instead of the default Wald model test
waldhet	perform Wald test on variance instead of LR test
nocnsreport	do not display constraints
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
Maximization	
maximize_options	control the maximization process; seldom used
collinear	keep collinear variables
coeflegend	display legend instead of statistics

ts_options	Description
Model	
* twostep	use two-step GLS estimator; default is maximum likelihood
* het(*varlist*)	independent variables to model the variance
noconstant	suppress constant term
SE	
vce(*vcetype*)	*vcetype* may be conventional, bootstrap, or jackknife
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

*twostep and het() are required.

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

depvar, *indepvars*, and *varlist* may contain time-series operators; see **[U] 11.4.4 Time-series varlists**.

bayes, bayesboot, bootstrap, by, collect, fp, jackknife, rolling, statsby, and svy are allowed; see **[U] 11.1.10 Prefix commands**. For more details, see [BAYES] **bayes: hetregress**.

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

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

vce(), lrmodel, twostep, and weights are not allowed with the svy prefix; see [SVY] **svy**.

aweights, fweights, iweights, and pweights are allowed with maximum likelihood estimation; see **[U] 11.1.6 weight**.

collinear and coeflegend do not appear in the dialog box.

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

Options for maximum likelihood estimation

Model

mle requests that the maximum likelihood estimator be used. This is the default.

het(*varlist*) specifies the independent variables in the variance function. When the het() option is not specified, homoskedasticity is assumed and the waldhet option is not allowed.

noconstant, constraints(*constraints*); see [R] **Estimation options**.

SE/Robust

vce(*vcetype*) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (oim, opg), that are robust to some kinds of misspecification (robust), that allow for intragroup correlation (cluster *clustvar*), and that use bootstrap or jackknife methods (bootstrap, jackknife); see [R] ***vce_option***.

Reporting

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

waldhet specifies that the Wald test of whether lnsigma2 = 0 be performed instead of the LR test.

nocnsreport; 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**.

Maximization

maximize_options: difficult, technique(*algorithm_spec*), iterate(#), [no]log, trace, gradient, showstep, hessian, showtolerance, tolerance(#), ltolerance(#), nrtolerance(#), nonrtolerance, and from(*init_specs*); see [R] **Maximize**. These options are seldom used.

Setting the optimization type to technique(bhhh) resets the default *vcetype* to vce(opg).

The following options are available with hetregress but are not shown in the dialog box:

collinear, coeflegend; see [R] **Estimation options**.

Options for two-step GLS estimation

Model

twostep specifies that the model be fit using Harvey's two-step GLS estimator. This option requires that the independent variables be specified in the het() option to model the variance.

het(*varlist*) specifies the independent variables in the variance function.

noconstant; see [R] **Estimation options**.

SE

vce(*vcetype*) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (conventional) and that use bootstrap or jackknife methods (bootstrap, jackknife); see [R] ***vce_option***.

vce(conventional), the default, uses the two-step variance estimator derived by Heckman.

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 hetregress but is not shown in the dialog box:

coeflegend; see [R] **Estimation options**.

Remarks and examples

Remarks are presented under the following headings:

Introduction

`hetregress` fits a multiplicative heteroskedastic linear regression model using either ML or Harvey's two-step GLS method. Multiplicative heteroskedasticity occurs when the variances of the error terms are assumed to be a multiplicative function of one or more variables. When variables are not specified in the `het()` option, `hetregress` fits a homoskedastic linear regression model.

Heteroskedasticity arises in a regression when the variances of the error terms are not constant across observations. For example, wages may be heteroskedastic when predicted by age group. While there is little variability in wages among workers in their teens and early 20s, wages among workers in their 50s may vary greatly because of a variety of factors. Heteroskedasticity is often found in time-series data and cross-sectional measurements and is a common issue in econometrics, social science, and many other fields. For more detailed information on how to detect the presence of heteroskedasticity, see *Tests for violation of assumptions* in [R] **regress postestimation**.

We can use `hetregress` when the variance is assumed to have a form that is an exponential function of a linear combination of one or more variables. This is known as multiplicative heteroskedasticity and includes most of the useful formulations for variance as special cases. For example, in the special case of groupwise heteroskedasticity, the sample can be divided into groups where each group has a different variance.

A model with multiplicative heteroskedasticity can be written as

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i; \qquad \sigma_i^2 = \exp(\mathbf{z}_i\boldsymbol{\alpha}) \tag{1}$$

where y_i, $i = 1, \ldots, n$, is the dependent variable; $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$ are the k independent variables that model the mean function; and $\mathbf{z}_i = (z_{1i}, z_{2i}, \ldots, z_{mi})$ are the m variables that model the variance function. $\boldsymbol{\beta}$'s are unknown parameters in the mean function, and $\boldsymbol{\alpha}$'s are unknown parameters in the variance function. ϵ_i's are errors that are independent and identically distributed with mean 0 and variance σ_i^2. Groupwise heteroskedasticity is modeled using (1) but where the $\mathbf{z}_i$'s are all indicator (dummy) variables for groups.

Harvey (1976) introduced two methods for dealing with multiplicative heteroskedasticity: ML estimation and two-step GLS estimation. By default, `hetregress` fits the multiplicative heteroskedastic regression model using ML. If the `twostep` option is specified, `hetregress` fits the model using the two-step GLS method. The ML estimates are more efficient than those obtained by the GLS estimator if the mean and variance function are correctly specified and the errors are normally distributed. By contrast, the two-step GLS estimates are more robust if the variance function is incorrect or the errors are nonnormal.

If the form of the variance is completely unknown, we may be better off using the OLS estimator instead of the ML and GLS estimators because it remains unbiased. However, we should then use the robust standard errors to correct for heteroskedasticity. Using robust standard errors for the OLS estimator allows us to make appropriate inferences without specifying any form for the variance. We discuss three modifications of the robust variance calculation in *Robust standard errors* of [R] **regress**.

If the form of the variance is known and does not contain any unknown parameters, we can use the weighted least-squares estimator, also called the generalized least-squares estimator. For example, we can use weighted least squares to correct for heteroskedasticity if the variance is proportional to one of the regressors. See section 9.5.2 of Greene (2018) for details, and also see *Weighted regression* in [R] **regress**.

Greene (2018) and Hill, Griffiths, and Lim (2018) compare the ML estimator and the GLS estimator with the robust OLS estimator. If the form of the heteroskedasticity is specified correctly, the ML and GLS estimators are more efficient than the robust OLS estimator. However, if the form of the heteroskedasticity is misspecified, the robust OLS estimator may be more efficient than the ML and GLS estimators.

Maximum likelihood estimation

▷ Example 1: Multiplicative heteroskedasticity

Consider the following dataset from a study of household expenditure on food described in Hill, Griffiths, and Lim (2018, chap. 8). We want to investigate the relationship between average household expenditure on food and household income by fitting a model of weekly food expenditure (`food_exp`) on weekly income (`income`) using OLS.

```
. use https://www.stata-press.com/data/r19/foodexp
(Household expenditure on food)

. regress food_exp income
  (output omitted)
```

However, we suspect that the variance for low-income families may be lower than that for high-income families, because low-income families usually have less money to spend on food, while high-income families can choose to spend more or less on food. We plot the least-squares residuals against the value of `income` by using the `rvpplot` command after `regress`.

```
. rvpplot income, yline(0)
```

The graph confirms our suspicions about a relationship between income and the residuals. We believe that the variance is some power function of income. Therefore, we fit a multiplicative heteroskedastic regression model. The model for each observation is

$$\texttt{food\_exp}_i = \beta_0 + \beta_1 \times \texttt{income}_i + \epsilon_i$$

and the variance function can be written as

$$\sigma_i^2 = \sigma^2 \times \texttt{income}_i^{\gamma}$$

where γ is an unknown parameter of the variance function. To ensure that we will get positive values for the variance σ_i^2 for all possible values of the unknown parameter γ, we rewrite this function so that σ_i^2 is an exponential function of a linear combination of $\ln(\texttt{income}_i)$ and a constant term:

$$\sigma_i^2 = \exp\{\alpha_0 + \alpha_1 \times \ln(\texttt{income}_i)\}$$

where $\alpha_0 = \ln(\sigma^2)$ and $\alpha_1 = \gamma$.

To fit this model using hetregress, we first create a variable that contains the logarithm of income (logincome) and use it in the het() option to model the variance function. The constant term in the variance function is always assumed.

```
. generate double logincome = ln(income)
. hetregress food_exp income, het(logincome)
Fitting full model:
Iteration 0:  Log likelihood =   -227.889
Iteration 1:  Log likelihood = -226.61039
Iteration 2:  Log likelihood = -225.72188
Iteration 3:  Log likelihood = -225.71519
Iteration 4:  Log likelihood = -225.71519
Heteroskedastic linear regression               Number of obs     =         40
ML estimation
                                                Wald chi2(1)      =     135.11
Log likelihood = -225.7152                      Prob > chi2       =     0.0000
```

food_exp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
food_exp						
income	10.63444	.9148876	11.62	0.000	8.841295	12.42759
_cons	76.07294	7.369143	10.32	0.000	61.62969	90.5162
lnsigma2						
logincome	2.769762	.4481606	6.18	0.000	1.891383	3.648141
_cons	.4684052	1.310337	0.36	0.721	-2.099809	3.036619

```
LR test of lnsigma2=0: chi2(1) = 19.59                  Prob > chi2 = 0.0000
```

The LR test at the bottom of the output is a test for the parameters of the variance function. The $\chi^2(1)$ statistic of 19.59 is significant, indicating that heteroskedasticity is present. If we had preferred the Wald test for heteroskedasticity instead of the LR test, we would have specified the waldhet option.

In addition to the estimated parameters for the mean function (under food_exp), hetregress reports estimated parameters and test statistics for the variance function. The significant z statistic for logincome also suggests the presence of heteroskedasticity. Relating the output back to our model, $\exp(0.47) \approx 1.60$ is our estimate of σ^2. The coefficient for logincome is 2.77. This is our estimate of γ, and it can be interpreted as the multiplicative factor of the variance associated with income.

We can obtain more formal results by using `nlcom`:

```
. nlcom (sigma2: exp(_b[lnsigma2:_cons]))
      sigma2: exp(_b[lnsigma2:_cons])
```

food_exp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
sigma2	1.597445	2.093191	0.76	0.445	-2.505135	5.700024

Here `sigma2` refers to σ^2 in the variance function given above.

Two-step GLS estimation

▷ Example 2: Groupwise heteroskedasticity

Here we will use a dataset of 725 faculty members' salaries described in DeMaris (2004) to determine whether there is evidence of a difference in salaries between male faculty and female faculty. In addition to sex (`female`), other variables that might affect the salaries are prior experience (`priorexp`), years in rank (`yrrank`), years at the university (`yrbg`), and marketability of discipline (`salfac`). We will treat `female` as a factor variable and all other variables as continuous variables.

We could fit this model with `regress` by including main effects and the interaction terms between `female` and all other variables (by using factor-variable notation).

```
. use https://www.stata-press.com/data/r19/salary, clear
(DeMaris (2004) - Faculty salaries)
. regress salary i.female##(c.priorexp c.yrrank c.yrbg c.salfac)
```

Source	SS	df	MS	Number of obs	=	725
				F(9, 715)	=	135.23
Model	8.9287e+10	9	9.9207e+09	Prob > F	=	0.0000
Residual	5.2453e+10	715	73360978.8	R-squared	=	0.6299
				Adj R-squared	=	0.6253
Total	1.4174e+11	724	195773163	Root MSE	=	8565.1

salary	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
1.female	5735.113	4987.433	1.15	0.251	-4056.65	15526.88
priorexp	1042.845	82.62092	12.62	0.000	880.6366	1205.054
yrrank	-47.80904	99.85936	-0.48	0.632	-243.8617	148.2436
yrbg	1009.12	75.5161	13.36	0.000	860.8607	1157.38
salfac	33601.3	2531.61	13.27	0.000	28631.02	38571.58
female# c.priorexp 1	-662.7972	159.6171	-4.15	0.000	-976.1715	-349.4228
female# c.yrrank 1	-447.8621	218.8695	-2.05	0.041	-877.5659	-18.15833
female#c.yrbg 1	115.4784	157.9372	0.73	0.465	-194.5976	425.5545
female# c.salfac 1	-7618.157	5142.204	-1.48	0.139	-17713.78	2477.467
_cons	2137.718	2634.266	0.81	0.417	-3034.103	7309.539

However, we believe that the variances differ between female faculty and male faculty. In this case, we will use `estat hettest` to perform the Breusch and Pagan (1979) test for heteroskedasticity. See *Tests for violation of assumptions* in [R] **regress postestimation** for more detailed information.

```
. estat hettest i.female

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: i.female

H0: Constant variance

    chi2(1) =  11.80
Prob > chi2 = 0.0006
```

The results above suggest the presence of heteroskedasticity with respect to sex. This is a case of groupwise heteroskedasticity and can be modeled using `hetregress` by treating the sex variable (`female`) as a factor variable (`i.female`) in the `het()` option.

We add the twostep option to obtain two-step GLS estimates instead of ML estimates.

```
. hetregress salary i.female##(c.priorexp c.yrrank c.yrbg c.salfac),
> het(i.female) twostep
Heteroskedastic linear regression               Number of obs     =        725
Two-step GLS estimation

                                                Wald chi2(9)      =    1270.02
                                                Prob > chi2       =     0.0000
```

salary	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
salary						
1.female	5735.113	4459.648	1.29	0.198	-3005.635	14475.86
priorexp	1042.845	89.0886	11.71	0.000	868.2348	1217.456
yrrank	-47.80904	107.6765	-0.44	0.657	-258.8511	163.233
yrbg	1009.12	81.4276	12.39	0.000	849.5253	1168.716
salfac	33601.3	2729.788	12.31	0.000	28251.01	38951.59
female# c.priorexp						
1	-662.7972	142.2286	-4.66	0.000	-941.5601	-384.0342
female# c.yrrank						
1	-447.8621	191.2937	-2.34	0.019	-822.7908	-72.93342
female#c.yrbg						
1	115.4784	138.9659	0.83	0.406	-156.8898	387.8467
female# c.salfac						
1	-7618.157	4544.735	-1.68	0.094	-16525.67	1289.359
_cons	2137.718	2840.48	0.75	0.452	-3429.52	7704.956
lnsigma2						
1.female	-.5676939	.1808783	-3.14	0.002	-.9222088	-.2131789
_cons	17.90879	.0982708	182.24	0.000	17.71618	18.1014

```
Wald test of lnsigma2=0: chi2(1) = 9.85                  Prob > chi2 = 0.0017
```

The Wald test for heteroskedasticity is reported at the bottom of the coefficient table instead of the LR test because there is no likelihood computed for the two-step GLS estimation.

Compared with the OLS results obtained using regress, the estimated coefficients for the mean function are not affected by heteroskedasticity, but their standard errors are. Also, the estimated variance in salaries for female faculty is about $\exp(-0.6) \approx 0.5$ times the estimated variance in salaries for male faculty.

The results above suggest that priorexp, yrbg, and salfac have significant effects on the salary of male faculty. We see also that the effects of priorexp and yrrank on salaries are significantly different between males and females. For example, each additional year of experience for male faculty increases their salary by $1,042.85, and the estimated difference in effect is $662.80 less for female faculty than for male faculty.

To obtain an estimate for female faculty of the effect of experience on salary, we can use lincom.

```
. lincom priorexp + 1.female#c.priorexp
 ( 1)  [salary]priorexp + [salary]1.female#c.priorexp = 0
```

salary	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
(1)	380.0481	110.8702	3.43	0.001	162.7465	597.3497

We see that each additional year of experience increases salary by only $380.05 and that this effect is significant.

We can estimate the effect of each of the other variables on the salaries of female faculty if we wish.

```
. lincom yrrank + 1.female#c.yrrank
 ( 1)  [salary]yrrank + [salary]1.female#c.yrrank = 0
```

salary	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
(1)	-495.6711	158.1108	-3.13	0.002	-805.5627	-185.7796

The effect of yrrank is associated with a decrease in salary. The effect is significant for female faculty but not for male faculty.

◁

Stored results

hetregress (ML) stores the following in e():

Scalars

e(N)	number of observations
e(k)	number of parameters
e(k_eq)	number of equations in e(b)
e(k_eq_model)	number of equations in overall model test
e(k_dv)	number of dependent variables
e(df_m)	model degrees of freedom
e(ll)	log likelihood, full model
e(ll_0)	log likelihood, constant-only model
e(ll_c)	log likelihood, comparison model
e(N_clust)	number of clusters
e(chi2)	χ^2 for mean model test
e(chi2_c)	χ^2 for heteroskedasticity test
e(p_c)	p-value for heteroskedasticity test
e(df_m_c)	degrees of freedom for heteroskedasticity test
e(p)	p-value for the mean model test
e(rank)	rank of e(V)
e(rank0)	rank of e(V) for constant-only model
e(ic)	number of iterations
e(rc)	return code
e(converged)	1 if converged, 0 otherwise

Macros

e(cmd)	hetregress
e(cmdline)	command as typed
e(depvar)	name of dependent variable
e(wtype)	weight type
e(wexp)	weight expression

e(title)	title in estimation output
e(title2)	secondary title in estimation output
e(clustvar)	name of cluster variable
e(chi2type)	Wald or LR; type of model χ^2 test
e(chi2_ct)	Wald or LR; type of heteroskedastic χ^2 test corresponding to e(chi2_c)
e(vce)	*vcetype* specified in vce()
e(vcetype)	title used to label Std. err.
e(opt)	type of optimization
e(which)	max or min; whether optimizer is to perform maximization or minimization
e(method)	ml
e(ml_method)	type of ml method
e(user)	name of likelihood-evaluator program
e(technique)	maximization technique
e(properties)	b V
e(predict)	program used to implement predict
e(marginsok)	predictions allowed by margins
e(marginsnotok)	predictions disallowed 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(ilog)	iteration log (up to 20 iterations)
e(gradient)	gradient vector
e(V)	variance–covariance matrix of the estimators
e(V_modelbased)	model-based variance

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.

hetregress (two-step GLS) stores the following in e():

Scalars

e(N)	number of observations
e(k)	number of parameters
e(df_m)	model degrees of freedom
e(chi2)	χ^2 for mean model test
e(chi2_c)	χ^2 for heteroskedasticity test
e(p_c)	p-value for heteroskedasticity test
e(df_m_c)	degrees of freedom for heteroskedasticity test
e(p)	p-value for the mean model test
e(rank)	rank of e(V)

Macros

e(cmd)	hetregress
e(cmdline)	command as typed
e(depvar)	name of dependent variable
e(title)	title in estimation output
e(title2)	secondary title in estimation output
e(chi2type)	Wald; type of model χ^2 test
e(chi2_ct)	Wald; type of heteroskedastic χ^2 test corresponding to e(chi2_c)

e(vce)	*vcetype* specified in vce()
e(method)	twostep
e(properties)	b V
e(predict)	program used to implement predict
e(marginsok)	predictions allowed by margins
e(marginsnotok)	predictions disallowed 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

Methods and formulas are presented under the following headings:

Maximum likelihood estimation

By default, hetregress fits a multiplicative heteroskedastic regression using ML estimation. The log-likelihood function is

$$\ln L = \sum_{i=1}^{n} \frac{w_i}{2} \left\{ \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{\exp(\mathbf{z}_i\boldsymbol{\alpha})} - \ln(2\pi) - \mathbf{z}_i\boldsymbol{\alpha} \right\}$$

where y_i, $i = 1, \ldots, n$, is the dependent variable; $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$ are the k independent variables that model the mean function; $\mathbf{z}_i = (z_{1i}, z_{2i}, \ldots, z_{mi})$ are the m variables that model the variance function; and w_i are the weights. $\boldsymbol{\beta}$ is a column vector of unknown parameters in the mean function, and $\boldsymbol{\alpha}$ is a column vector of unknown parameters in the variance function. The GLS estimates $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ and $\widehat{\boldsymbol{\alpha}}_{\text{GLS}}$ (described below) are used as the initial values in ML estimation. The $\ln L$ function is maximized as described in [R] **Maximize**.

This command supports the Huber/White/sandwich estimator of the variance and its clustered version using vce(robust) and vce(cluster *clustvar*), respectively. See [P] **_robust**, particularly *Maximum likelihood estimators* and *Methods and formulas*.

hetregress also supports estimation with survey data. For details on VCEs with survey data, see [SVY] **Variance estimation**.

Two-step GLS estimation

`hetregress` uses two-step GLS estimation when the `twostep` option is specified. Harvey (1976) describes the procedure in detail, but here are the main steps.

1. Use OLS to estimate regression coefficients $\boldsymbol{\beta}$ and compute residuals $\mathbf{e}_i$, $i = 1, \ldots, n$.
2. Use OLS to regress the log-squared residuals, $\ln(\mathbf{e}_i^2)$, on $\mathbf{z}$ and estimate $\boldsymbol{\alpha}$.
3. Perform correction for the OLS estimates of $\boldsymbol{\alpha}$ to obtain $\widehat{\boldsymbol{\alpha}}_c$ and their covariance matrix based on Harvey (1976).
4. Compute $\widehat{\sigma_i}^2 = \exp(\mathbf{z}_i\widehat{\boldsymbol{\alpha}}_c)$, $i = 1, \ldots, n$.
5. Refit the original regression model using $\widehat{\sigma_i}^2$'s as weights to obtain the GLS estimates $\widehat{\boldsymbol{\alpha}}_{\text{GLS}}$ and $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$.

References

Breusch, T. S., and A. R. Pagan. 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47: 1287–1294. https://doi.org/10.2307/1911963.

DeMaris, A. 2004. *Regression with Social Data: Modeling Continuous and Limited Response Variables*. Hoboken, NJ: Wiley. https://doi.org/10.1002/0471677566.

Greene, W. H. 2018. *Econometric Analysis*. 8th ed. New York: Pearson.

Harvey, A. C. 1976. Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44: 461–465. https://doi.org/10.2307/1913974.

Hill, R. C., W. E. Griffiths, and G. C. Lim. 2018. *Principles of Econometrics*. 5th ed. Hoboken, NJ: Wiley.

Also see

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

[R] **regress** — Linear regression

[BAYES] **bayes: hetregress** — Bayesian heteroskedastic linear regression

[SVY] **svy estimation** — Estimation commands for survey data

[U] 20 Estimation and postestimation commands