hetoprobit — Heteroskedastic ordered probit regression

Description

`hetoprobit` fits a heteroskedastic ordered probit model for an ordinal dependent variable. `hetoprobit` is a generalization of `oprobit` that allows the variance to be modeled as a function of independent variables and to differ between subjects or groups in the population.

Quick start

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

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

With robust standard errors

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

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

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

Menu

Statistics > Ordinal outcomes > Heteroskedastic ordered probit regression

Syntax

```
hetoprobit depvar [indepvars] [if] [in] [weight] ,
    het(varlist[ , offset(varname_o) ]) [options]
```

options	Description
Model	
* `het(`*varlist*`[...])`	independent variables to model the variance and optional offset variable
`offset(`*varname*`)`	include *varname* in model with coefficient constrained to 1
`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)`
`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
`noheader`	do not display header above coefficient table
`notable`	do not display coefficient table
`collinear`	keep collinear variables
`coeflegend`	display legend instead of statistics

*`het()` is required. The full specification is `het(`*varlist* [, `offset(`*varname*$_o$`)`]`).

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: hetoprobit**.

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

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

`fweights`, `iweights`, and `pweights` are allowed; see **[U] 11.1.6 weight**.

`noheader`, `notable`, `collinear`, and `coeflegend` do not appear in the dialog box.

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

Options

Model

het(*varlist* [, offset(*varname*$_o$)]) specifies the independent variables and, optionally, the offset variable in the variance function. het() is required.

offset(*varname*$_o$) specifies that offset *varname*$_o$ be included in the variance model with the coefficient constrained to be 1.

offset(*varname*), 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(#); see [R] **Estimation options**.

waldhet specifies that a Wald test of whether lnsigma = 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.

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

noheader suppresses the header above the coefficient table.

notable suppresses the display of the coefficient table.

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

Remarks and examples

hetoprobit fits a maximum-likelihood heteroskedastic ordered probit model, which is a generalization of the ordered probit model (see [R] **oprobit**).

In ordinal regression models, the outcome is an ordinal variable—a variable that is categorical and ordered, for instance, "poor", "good", and "excellent". The specific values of the ordinal variable are irrelevant. It matters only that larger values are assumed to correspond to "higher" outcomes. To simplify the discussion in this entry, we assume without loss of generality that the dependent variable takes on the integer values $0, 1, \ldots, H$, for some value $H > 1$.

In ordered probit models, an underlying score is estimated as a linear function of the independent variables and a set of cutpoints. The probability of observing outcome $y_j = h$, where $h = 0, 1, \ldots, H$, corresponds to the probability that the value of the linear function, plus random error, is within the range of the cutpoints associated with the outcome

$$\begin{aligned}\Pr(y_j = h) &= \Pr(\kappa_h < \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + u_j \leq \kappa_{h+1}) \\ &= \Phi\left(\kappa_{h+1} - \mathbf{x}_j\boldsymbol{\beta}\right) - \Phi\left(\kappa_h - \mathbf{x}_j\boldsymbol{\beta}\right)\end{aligned}$$

where $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{kj})$ are the k independent variables that model the mean function; $\boldsymbol{\beta}$ is a column vector of unknown parameters in the mean function; u_j, where $j = 1, \ldots, N$, are normally distributed error terms; κ_h, where $h = 1, \ldots, H$, are the unknown cutpoints that separate the different possible values of h; and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Also, by convention, to complete the intervals for the lowest and highest values of the outcome, $\kappa_0 = -\infty$ and $\kappa_{H+1} = \infty$.

In conventional ordinal probit models, the error term is assumed i.i.d. normal with unit variance for all observations. `hetoprobit` generalizes the ordered probit model by representing the variance of the error term u_j as a multiplicative function of explanatory variables $\mathbf{z}_j = (z_{1j}, z_{2j}, \ldots, z_{mj})$. This approach was introduced by Harvey (1976), though we depart from Harvey slightly by modeling standard deviation rather than variance. More specifically, we model the natural logarithm of the standard deviation as a linear combination of the explanatory variables,

$$\ln \sigma_j = \mathbf{z}_j\boldsymbol{\gamma}$$

where $\boldsymbol{\gamma}$ is a column vector of unknown parameters in the variance function.

With this generalization, the error variance may differ between subjects or between groups in the population, and

$$\Pr(y_j = h) = \Phi\left\{\frac{\kappa_{h+1} - \mathbf{x}_j\boldsymbol{\beta}}{\exp(\mathbf{z}_j\boldsymbol{\gamma})}\right\} - \Phi\left\{\frac{\kappa_h - \mathbf{x}_j\boldsymbol{\beta}}{\exp(\mathbf{z}_j\boldsymbol{\gamma})}\right\}$$

For the model to be identifiable, there can be no constant term in $\mathbf{z}_j\boldsymbol{\gamma}$. Also, as with [R] **oprobit**, there is no constant term in $\mathbf{x}_j\boldsymbol{\beta}$. The role of the constant is subsumed by the cutpoints.

We estimate the coefficients β_1, β_2, …, β_k and γ_1, γ_2, …, γ_m together with the cutpoints κ_1, κ_2, …, κ_H. If the model has no independent variables in $\mathbf{x}_j$, only the cutpoints and the $\boldsymbol{\gamma}$ parameters are estimated.

Modeling of heteroskedastic variance has both constructive and defensive uses. It is known that differences in variance between subjects or between groups in the population can cause biased coefficient estimates and can complicate comparison of distinct groups. Thus, incorporating a model for variance can be necessary for proper inference, even if the variance function itself is not a topic of interest to the researcher. For discussion, see Williams (2010) and the references cited therein. There are also cases where modeling the differences between variances of different subjects or different groups in the population is one of the principal purposes of the study. We will discuss such a scenario in the examples below. See Reardon et al. (2017) and Alvarez and Brehm (1995) for additional examples.

▷ Example 1: Modeling heteroskedasticity of reported health status

In this example, we will use a slightly modified subset of data from the 2015 Eating & Health Module of the American Time Use Survey (ATUS), conducted by the US Bureau of Labor Statistics. Our analysis will not account for the survey design. The ATUS measures the amount of time people spend doing various activities, such as working, caring for children, volunteering, and socializing. Of interest to us is an ordinal response variable, `health`, which contains individuals' self-assessments of their overall health status on a five-point scale: 1 for "poor", 2 for "fair", 3 for "good", 4 for "very good", and 5 for "excellent".

We want to examine the role that age and other factors play in an individual's self-assessment of health. Age is a natural variable to include when modeling mean or typical health status. But we also suspect that the variation in health status is greater in an older population, as compared with a youthful population, which consists mainly of healthy individuals. If our suspicion is true, quantifying the relationship between variation in health status and age may have value, for example, in planning a healthcare strategy that is appropriately tailored both for the older Medicare population and for a younger cohort.

Thus, we will use `hetoprobit` to model heteroskedasticity induced by age. In modeling the variance term, in addition to age, we will include a factor variable, `exercise`, which indicates whether or not an individual exercised during the previous week. For purposes of illustration, imagine that we are not interested in exercise as a topic in its own right, but we are concerned that health variability among those who exercise may differ from the variability among those who do not. Therefore, we include `exercise` in the variance term to help insulate our estimation results against a possible hidden bias.

Our model will include three explanatory variables for the mean function: age, bmi (body mass index), and exercise.

```
. use https://www.stata-press.com/data/r19/eathealth15
(2015 ATUS Eating & Health Module extract)
. hetoprobit health age bmi i.exercise, het(age i.exercise)
  (output omitted)
Fitting ordered probit model:
Iteration 0:  Log likelihood = -2905.7943
Iteration 1:  Log likelihood = -2717.2752
Iteration 2:  Log likelihood = -2716.9679
Iteration 3:  Log likelihood = -2716.9679
Fitting full model:
Iteration 0:  Log likelihood = -2716.9679
Iteration 1:  Log likelihood = -2708.6752
Iteration 2:  Log likelihood = -2708.5492
Iteration 3:  Log likelihood = -2708.5491
Heteroskedastic ordered probit regression              Number of obs =  2,009
                                                       LR chi2(3)    = 366.91
Log likelihood = -2708.5491                            Prob > chi2   = 0.0000
```

health	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
health						
age	-.0083348	.0015969	-5.22	0.000	-.0114646	-.005205
bmi	-.0564072	.0057392	-9.83	0.000	-.0676558	-.0451586
exercise						
Yes	.6493794	.0732137	8.87	0.000	.5058833	.7928755
lnsigma						
age	.0041401	.0011611	3.57	0.000	.0018643	.0064159
exercise						
Yes	-.0773166	.0423038	-1.83	0.068	-.1602305	.0055973
/cut1	-3.903773	.2913163			-4.474742	-3.332803
/cut2	-2.776111	.2262442			-3.219541	-2.33268
/cut3	-1.576396	.174352			-1.918119	-1.234672
/cut4	-.4189882	.1524084			-.7177031	-.1202733

```
LR test of lnsigma=0: chi2(2) = 16.84                  Prob > chi2 = 0.0002
```

The LR test at the bottom of the output is a test of homogeneity of the variance function. The $\chi^2(2)$ statistic of 16.84 is significant, indicating that heteroskedasticity is present. If you prefer the Wald test for heteroskedasticity, you can specify the waldhet option. The coefficients for the variance function are reported in the section of the table labeled lnsigma. Our results indicate that age is a significant contributor to the variance function but that exercise is not significant at a 0.05 level.

The LR test for the model that appears above the coefficient table is a joint test for inclusion of age, bmi, and exercise in the mean function. The null model for this test is the model consisting only of cutpoints and the heteroskedastic term. Coefficients for the mean function are reported in the section of the table labeled health. In this example, age, bmi, and exercise are significant components of the linear predictor of the mean.

The signs of the coefficients in the fitted model are directly interpretable. For example, the negative value for the coefficient of `bmi` implies that higher values of `bmi` predict lower values of health status. However, because of the probit link and the fact that we estimate variance with a log transformation, the numerical relationships between the coefficients of the model and the outcome variables are nonlinear. Postestimation commands recognize and account for these nonlinearities.

◁

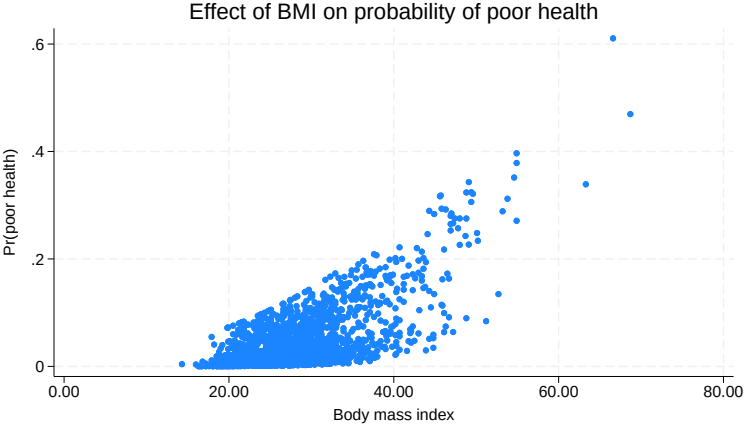
▷ Example 2: Predict the probability of a poor health rating

Ordered probit models allow us to look at the probabilities of different outcomes of interest. Suppose we are interested in predictions of a reported health status of "poor" (`health` = 1) and how it differs across levels of `bmi`. First, we obtain the predicted probability of poor health.

```
. predict pr1, pr outcome(1)
```

We can now visualize how the predicted probability of poor health status differs across the range of `bmi` values in our sample.

```
. twoway scatter pr1 bmi, ytitle("Pr(poor health)")
> title("Effect of BMI on probability of poor health")
```



We see that predicted probabilities of poor health increase as body mass index increases.

◁

▷ Example 3: Predictive margins and average marginal effect

The graph above plots the predicted probability of poor health for each individual in our dataset. We may also want to evaluate how the average predicted probability changes across levels of the covariates in the model. For instance, we can use the `margins` command to obtain the expected probability of having poor health across a range of ages.

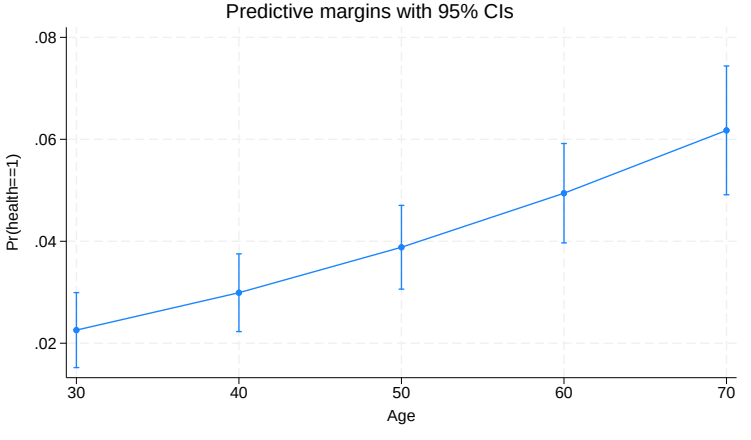
```
. margins, at(age = (30(10)70)) predict(outcome(1))
Predictive margins                                      Number of obs = 2,009
Model VCE: OIM

Expression: Pr(health==1), predict(outcome(1))
1._at: age = 30
2._at: age = 40
3._at: age = 50
4._at: age = 60
5._at: age = 70
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_at						
1	.0225765	.0037522	6.02	0.000	.0152224	.0299306
2	.0299079	.0038887	7.69	0.000	.0222862	.0375297
3	.0388244	.0041936	9.26	0.000	.0306051	.0470438
4	.0494246	.0049748	9.93	0.000	.0396742	.0591751
5	.0617532	.0064443	9.58	0.000	.0491226	.0743839

Based on our model, what would we expect if everyone was 30 years old but had the same distributions of `bmi` and `exercise` that we observed in our data? The first line in this table reports that the average predicted probability of poor health is 0.0226 in this case. The second line shows the average predicted probability of poor health if we set `age = 40`, and so on. We find that for `age = 70`, the average probability of reporting a poor health status has increased to 0.0618. We can visualize this by typing `marginsplot` after `margins`.

```
. marginsplot
Variables that uniquely identify margins: age
```



We have focused on the prediction of poor health. We could instead simultaneously obtain average predicted probabilities of poor, fair, good, very good, and excellent health status and plot them across our requested age range. In that case, we would type

```
. margins, at(age = (30(10)70))
. marginsplot
```

We might also be interested in characterizing the relationship between bmi and the probability of reporting poor health. The coefficients and cutpoints reported in hetoprobit are not easily interpreted. We can, however, use margins to estimate the average marginal effect of bmi on the probability of reporting poor health. Because the average marginal effect depends on the value of bmi, we estimate it across a range of bmi values by typing

```
. margins, dydx(bmi) at(bmi = (20(5)35)) predict(outcome(1))
Average marginal effects                                    Number of obs = 2,009
Model VCE: OIM

Expression: Pr(health==1), predict(outcome(1))
dy/dx wrt:  bmi
1._at: bmi = 20
2._at: bmi = 25
3._at: bmi = 30
4._at: bmi = 35
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
bmi						
_at						
1	.0015875	.0001721	9.23	0.000	.0012503	.0019248
2	.0024512	.0002665	9.20	0.000	.0019288	.0029736
3	.0036447	.0004377	8.33	0.000	.0027868	.0045027
4	.0052124	.0007013	7.43	0.000	.0038379	.006587

The average marginal effect of bmi on the probability of reporting poor health increases as bmi itself increases.

◁

▷ Example 4: Interpreting the variance function

From the output of our hetoprobit command, we determined that variance of health status is affected by age. Let's consider to what extent. In this example, we assess the effect of age on the variance by using the margins command. We use the predict(sigma) option to obtain the average predicted standard deviation of the errors. We will look at ages 15 and 85, which are the youngest and oldest ages, respectively, in our dataset.

```
. margins, predict(sigma) at(age = (15,85)) noatlegend
Predictive margins                                          Number of obs = 2,009
Model VCE: OIM

Expression: Heteroskedastic standard deviation, predict(sigma)
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_at						
1	1.014732	.0320905	31.62	0.000	.9518354	1.077628
2	1.355853	.1398144	9.70	0.000	1.081822	1.629884

Variation increases with age. The expected standard deviation of the error term changes from 1.015 at age 15 to 1.356 at age 85.

◁

Stored results

hetoprobit stores the following in e():

Scalars

e(N)	number of observations
e(k_cat)	number of categories
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_aux)	number of auxiliary parameters
e(k_dv)	number of dependent variables
e(df_m)	model degrees of freedom ($x\beta$ term)
e(ll)	log likelihood
e(ll_0)	log likelihood, cutpoint-only (heteroskedastic) model
e(ll_c)	log likelihood, comparison (homoskedastic) model
e(N_clust)	number of clusters
e(chi2)	χ^2
e(chi2_c)	χ^2 for heteroskedasticity test
e(p)	p-value for model test
e(p_c)	p-value for heteroskedasticity test
e(df_m_c)	degrees of freedom for heteroskedasticity test
e(rank)	rank of e(V)
e(ic)	number of iterations
e(rc)	return code
e(converged)	1 if converged, 0 otherwise

Macros

e(cmd)	hetoprobit
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(clustvar)	name of cluster variable
e(offset1)	offset for ordered probit equation
e(offset2)	offset for variance equation
e(chi2type)	LR; type of model χ^2 test
e(chi2_ct)	LR or Wald; type of heteroskedasticity 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(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(marginsdefault)	default predict() specification for 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(cat)	category values
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.

Methods and formulas

hetoprobit fits a cumulative probit model with heteroskedastic variance using maximum likelihood estimation. Namely, the model is that, for a subject with explanatory variables $\mathbf{x}$ and $\mathbf{z}$,

$$\Pr(Y \le h) = \Phi\left\{\frac{\kappa_{h+1} - \mathbf{x}\boldsymbol{\beta}}{\exp(\mathbf{z}\boldsymbol{\gamma})}\right\}$$

where Y is an ordinal outcome taking on values $h = 0, 1, \ldots, H$, and $\Phi(\cdot)$ is the cdf of the standard normal distribution. The value κ_{h+1} is a cutpoint that separates the region corresponding to $Y = h$ from regions for higher-valued categories. The effects $\boldsymbol{\beta}$ and the effects $\boldsymbol{\gamma}$ are the same for each cumulative probability.

The log-likelihood function is

$$\ln L = \sum_{j=1}^{N} w_j \sum_{h=0}^{H} I_h(y_j) \ln\left[\Phi\left\{\frac{\kappa_{h+1} - \mathbf{x}_j\boldsymbol{\beta}}{\exp(\mathbf{z}_j\boldsymbol{\gamma})}\right\} - \Phi\left\{\frac{\kappa_h - \mathbf{x}_j\boldsymbol{\beta}}{\exp(\mathbf{z}_j\boldsymbol{\gamma})}\right\}\right]$$

where

$$I_h(y_j) = \begin{cases} 1 & \text{if } y_j = h \\ 0 & \text{otherwise} \end{cases}$$

and y_j, where $j = 1, \ldots, N$, is an observed value of Y; w_j are optional weights; $\kappa_0 = -\infty$ and $\kappa_{H+1} = \infty$; and all other terminology is defined in *Remarks and examples* above.

The log-likelihood 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*.

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

References

Aitchison, J., and S. D. Silvey. 1957. The generalization of probit analysis to the case of multiple responses. *Biometrika* 44: 131–140. https://doi.org/10.2307/2333245.

Allison, P. D. 1999. Comparing logit and probit coefficients across groups. *Sociological Methods and Research* 28: 186–208. https://doi.org/10.1177/0049124199028002003.

Alvarez, R. M., and J. Brehm. 1995. American ambivalence towards abortion policy: Development of a heteroskedastic probit model of competing values. *American Journal of Political Science* 39: 1055–1082. https://doi.org/10.2307/2111669.

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

Long, J. S., and J. Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 3rd ed. College Station, TX: Stata Press.

McCullagh, P. 1980. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, B ser., 42: 109–142. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x.

Reardon, S. F., B. R. Shear, K. E. Castellano, and A. D. Ho. 2017. Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics* 42: 3–45. https://doi.org/10.3102/1076998616666279.

Williams, R. 2009. Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods and Research* 37: 531–559. https://doi.org/10.1177/0049124109335735.

———. 2010. Fitting heterogeneous choice models with oglm. *Stata Journal* 10: 540–567.

Yatchew, A., and Z. Griliches. 1985. Specification error in probit models. *Review of Economics and Statistics* 67: 134–139. https://doi.org/10.2307/1928444.

Also see

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

[R] **hetprobit** — Heteroskedastic probit model

[R] **oprobit** — Ordered probit regression

[BAYES] **bayes: hetoprobit** — Bayesian heteroskedastic ordered probit regression

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

[U] 20 Estimation and postestimation commands