scobit — Skewed logistic regression

Syntax

`scobit` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

options	Description
Model	
`noconstant`	suppress constant term
`offset(`*varname*`)`	include *varname* in model with coefficient constrained to 1
`asis`	retain perfect predictor variables
`constraints(`*constraints*`)`	apply specified linear constraints
`collinear`	keep collinear variables
SE/Robust	
`vce(`*vcetype*`)`	*vcetype* may be `oim`, `robust`, `cluster` *clustvar*, `opg`, `bootstrap`, or `jackknife`
Reporting	
`level(#)`	set confidence level; default is `level(95)`
`or`	report odds ratios
`nocnsreport`	do not display constraints
display_options	control 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
`coeflegend`	display legend instead of statistics

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

`bootstrap`, `by`, `fp`, `jackknife`, `nestreg`, `rolling`, `statsby`, `stepwise`, and `svy` are allowed; see **[U] 11.1.10 Prefix commands**.

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

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

`fweight`s, `iweight`s, and `pweight`s 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.

Menu

Statistics > Binary outcomes > Skewed logit regression

Description

scobit fits a maximum-likelihood skewed logit model.

See [R] **logistic** for a list of related estimation commands.

Options

Model

noconstant, offset(*varname*), constraints(*constraints*), collinear; see [R] **estimation options**.

asis forces retention of perfect predictor variables and their associated perfectly predicted observations and may produce instabilities in maximization; see [R] **probit**.

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**.

or reports the estimated coefficients transformed to odds ratios, that is, e^b rather than b. Standard errors and confidence intervals are similarly transformed. This option affects how results are displayed, not how they are estimated. or may be specified at estimation or when replaying previously estimated results.

nocnsreport; see [R] **estimation options**.

display_options: 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 option is available with scobit but is not shown in the dialog box:

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

Remarks and examples

stata.com

Remarks are presented under the following headings:

Skewed logistic model
Robust standard errors

Skewed logistic model

`scobit` fits maximum likelihood models with dichotomous dependent variables coded as 0/1 (or, more precisely, coded as 0 and not 0).

▷ Example 1

We have data on the make, weight, and mileage rating of 22 foreign and 52 domestic automobiles. We wish to fit a model explaining whether a car is foreign based on its mileage. Here is an overview of our data:

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)

. keep make mpg weight foreign

. describe

Contains data from http://www.stata-press.com/data/r13/auto.dta
  obs:            74                          1978 Automobile Data
 vars:             4                          13 Apr 2013 17:45
 size:         1,702                          (_dta has notes)
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
make            str18   %-18s                 Make and Model
mpg             int     %8.0g                 Mileage (mpg)
weight          int     %8.0gc                Weight (lbs.)
foreign         byte    %8.0g      origin     Car type
--------------------------------------------------------------------------------
Sorted by:  foreign
     Note:  dataset has changed since last saved

. inspect foreign

foreign:  Car type                           Number of Observations
------------------                     ------------------------------
                                        Total    Integers   Nonintegers
|  #                   Negative            -           -             -
|  #                   Zero               52          52             -
|  #                   Positive           22          22             -
|  #                                   -----       -----         -----
|  #   #               Total              74          74             -
|  #   #               Missing             -
+----------------------                -----
0                    1                    74
   (2 unique values)

      foreign is labeled and all values are documented in the label.
```

The variable `foreign` takes on two unique values, 0 and 1. The value 0 denotes a domestic car, and 1 denotes a foreign car.

The model that we wish to fit is

$$\Pr(\texttt{foreign} = 1) = F(\beta_0 + \beta_1 \texttt{mpg})$$

where $F(z) = 1 - 1/\{1 + \exp(z)\}^\alpha$.

To fit this model, we type

```
. scobit foreign mpg
Fitting logistic model:
Iteration 0:   log likelihood =  -45.03321
Iteration 1:   log likelihood = -39.380959
Iteration 2:   log likelihood = -39.288802
Iteration 3:   log likelihood =  -39.28864
Iteration 4:   log likelihood =  -39.28864
Fitting full model:
Iteration 0:   log likelihood =  -39.28864
Iteration 1:   log likelihood = -39.286393
Iteration 2:   log likelihood = -39.284415
Iteration 3:   log likelihood = -39.284234
Iteration 4:   log likelihood = -39.284197
Iteration 5:   log likelihood = -39.284196
Skewed logistic regression                      Number of obs     =         74
                                                Zero outcomes     =         52
Log likelihood =  -39.2842                      Nonzero outcomes  =         22
```

foreign	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mpg	.1813879	.2407362	0.75	0.451	-.2904463	.6532222
_cons	-4.274883	1.399305	-3.06	0.002	-7.017471	-1.532295
/lnalpha	-.4450405	3.879885	-0.11	0.909	-8.049476	7.159395
alpha	.6407983	2.486224			.0003193	1286.133

```
Likelihood-ratio test of alpha=1:    chi2(1) =      0.01    Prob > chi2 = 0.9249
Note: likelihood-ratio tests are recommended for inference with scobit models.
```

We find that cars yielding better gas mileage are less likely to be foreign. The likelihood-ratio test at the bottom of the output indicates that the model is not significantly different from a logit model. Therefore, we should use the more parsimonious model.

◁

❑ Technical note

Stata interprets a value of 0 as a negative outcome (failure) and treats all other values (except missing) as positive outcomes (successes). Thus if the dependent variable takes on the values 0 and 1, then 0 is interpreted as failure and 1 as success. If the dependent variable takes on the values 0, 1, and 2, then 0 is still interpreted as failure, but both 1 and 2 are treated as successes.

Formally, when we type `scobit` y x, Stata fits the model

$$\Pr(y_j \neq 0 \mid \mathbf{x}_j) = 1 - 1\Big/\Big\{1 + \exp(\mathbf{x}_j\boldsymbol{\beta})\Big\}^\alpha$$

❑

Robust standard errors

If you specify the vce(robust) option, scobit reports robust standard errors as described in [U] 20.21 Obtaining robust variance estimates. For the model of foreign on mpg, the robust calculation increases the standard error of the coefficient on mpg by around 25%:

```
. scobit foreign mpg, vce(robust) nolog
Skewed logistic regression                      Number of obs     =         74
                                                Zero outcomes     =         52
Log pseudolikelihood =  -39.2842                Nonzero outcomes  =         22
```

foreign	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
mpg	.1813879	.3028487	0.60	0.549	-.4121847	.7749606
_cons	-4.274883	1.335521	-3.20	0.001	-6.892455	-1.657311
/lnalpha	-.4450405	4.71561	-0.09	0.925	-9.687466	8.797385
alpha	.6407983	3.021755			.0000621	6616.919

Without vce(robust), the standard error for the coefficient on mpg was reported to be 0.241, with a resulting confidence interval of $[-0.29, 0.65]$.

Specifying the vce(cluster *clustvar*) option relaxes the independence assumption required by the skewed logit estimator to being just independence between clusters. To demonstrate this, we will switch to a different dataset.

▷ Example 2

We are studying the unionization of women in the United States and have a dataset with 26,200 observations on 4,434 women between 1970 and 1988. For our purposes, we will use the variables age (the women were 14–26 in 1968 and the data thus span the age range of 16–46), grade (years of schooling completed, ranging from 0 to 18), not_smsa (28% of the person-time was spent living outside an SMSA—standard metropolitan statistical area), south (41% of the person-time was in the South), and year. Each of these variables is included in the regression as a covariate along with the interaction between south and year. This interaction, along with the south and year variables, is specified in the scobit command using factor-variables notation, south##c.year. We also have variable union. Overall, 22% of the person-time is marked as time under union membership and 44% of these women have belonged to a union.

We fit the following model, ignoring that women are observed an average of 5.9 times each in these data:

```
. use http://www.stata-press.com/data/r13/union, clear
(NLS Women 14-24 in 1968)
. scobit union age grade not_smsa south##c.year, nrtol(1e-3)
```
(output omitted)
```
Skewed logistic regression                      Number of obs     =      26200
                                                Zero outcomes     =      20389
Log likelihood = -13540.61                      Nonzero outcomes  =       5811
```

union	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0185365	.0043615	4.25	0.000	.0099881	.0270849
grade	.0452803	.0057124	7.93	0.000	.0340842	.0564764
not_smsa	-.1886849	.0317802	-5.94	0.000	-.250973	-.1263968
1.south	-1.422381	.3949298	-3.60	0.000	-2.196429	-.6483327
year	-.0133017	.0049575	-2.68	0.007	-.0230182	-.0035853
south#c.year						
1	.0105663	.0049233	2.15	0.032	.0009168	.0202158
_cons	-10.19247	63.69015	-0.16	0.873	-135.0229	114.6379
/lnalpha	8.972796	63.68825	0.14	0.888	-115.8539	133.7995
alpha	7885.616	502221.1			4.85e-51	1.28e+58

```
Likelihood-ratio test of alpha=1:   chi2(1) =     3.76    Prob > chi2 = 0.0524
Note: likelihood-ratio tests are recommended for inference with scobit models.
```

The reported standard errors in this model are probably meaningless. Women are observed repeatedly, so the observations are not independent. Looking at the coefficients, we find a large southern effect against unionization and a different time trend for the south. The `vce(cluster` *clustvar*`)` option provides a way to fit this model and obtains correct standard errors:

```
. scobit union age grade not_smsa south##c.year, vce(cluster id) nrtol(1e-3)
```
(output omitted)
```
Skewed logistic regression                      Number of obs     =      26200
                                                Zero outcomes     =      20389
Log pseudolikelihood = -13540.61                Nonzero outcomes  =       5811
                              (Std. Err. adjusted for 4434 clusters in idcode)
```

union	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0185365	.0084867	2.18	0.029	.0019029	.0351701
grade	.0452803	.0125764	3.60	0.000	.0206311	.0699296
not_smsa	-.1886849	.0642035	-2.94	0.003	-.3145214	-.0628484
1.south	-1.422381	.5064916	-2.81	0.005	-2.415086	-.4296756
year	-.0133017	.0090621	-1.47	0.142	-.0310632	.0044597
south#c.year						
1	.0105663	.0063172	1.67	0.094	-.0018152	.0229478
_cons	-10.19247	.945772	-10.78	0.000	-12.04615	-8.33879
/lnalpha	8.972796	.7482517	11.99	0.000	7.506249	10.43934
alpha	7885.616	5900.426			1819.377	34178.16

scobit, vce(cluster *clustvar*) is robust to assumptions about within-cluster correlation. That is, it inefficiently sums within cluster for the standard error calculation rather than attempting to exploit what might be assumed about the within-cluster correlation (as do the xtgee population-averaged models; see [XT] **xtgee**).

◁

❑ Technical note

The scobit model can be difficult to fit because of the functional form. Often it requires many iterations, or the optimizer prints out warning and informative messages during the optimization. For example, without the nrtol(1e-3) option, the model using the union dataset will not converge. See [R] **maximize** for details about the optimizer.

❑

❑ Technical note

The main reason for using scobit rather that logit is that the effects of the regressors on the probability of success are not constrained to be the largest when the probability is 0.5. Rather, the independent variables might show their largest impact when the probability of success is 0.3 or 0.6. This added flexibility results because the scobit function, unlike the logit function, can be skewed and is not constrained to be mirror symmetric about the 0.5 probability of success.

As Nagler (1994) pointed out, the point of maximum impact is constrained under the scobit model to fall within the interval $\left(0, 1 - e^{(-1)}\right)$ or approximately (0, 0.63). Achen (2002) notes that if we believe the maximum impact to be outside that range, we can instead estimate the "power logit" model by simply reversing the 0s and 1s of our outcome variable and estimating a scobit model on failure, rather than success. We would need to reverse the signs of the coefficients if we wanted to interpret them in terms of impact on success, or we could leave them as they are and interpret them in terms of impact on failure. The important thing to remember is that the scobit model, unlike the logit model, is not invariant to the choice of which result is assigned to success.

❑

Stored results

scobit 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_aux)	number of auxiliary parameters
e(k_dv)	number of dependent variables
e(ll)	log likelihood
e(ll_c)	log likelihood, comparison model
e(N_f)	number of failures (zero outcomes)
e(N_s)	number of successes (nonzero outcomes)
e(alpha)	alpha
e(N_clust)	number of clusters
e(chi2)	χ^2
e(chi2_c)	χ^2 for comparison test
e(p)	significance
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)	scobit
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(offset)	linear offset variable
e(chi2type)	Wald or LR; type of model χ^2 test
e(chi2_ct)	Wald or LR; type of model 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(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(footnote)	program used to implement the footnote display
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

Methods and formulas

Skewed logit analysis is an alternative to logit that relaxes the assumption that individuals with initial probability of 0.5 are most sensitive to changes in independent variables.

The log-likelihood function for skewed logit is

$$\ln L = \sum_{j \in S} w_j \ln F(\mathbf{x}_j \mathbf{b}) + \sum_{j \notin S} w_j \ln\big\{1 - F(\mathbf{x}_j \mathbf{b})\big\}$$

where S is the set of all observations j such that $y_j \neq 0$, $F(z) = 1 - 1/\big\{1 + \exp(z)\big\}^{\alpha}$, and w_j denotes the optional weights. $\ln L$ 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*.

`scobit` also supports estimation with survey data. For details on VCEs with survey data, see [SVY] **variance estimation**.

References

Achen, C. H. 2002. Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science* 5: 423–450.

Nagler, J. 1994. Scobit: An alternative estimator to logit and probit. *American Journal of Political Science* 38: 230–255.

Also see

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

[R] **cloglog** — Complementary log-log regression

[R] **glm** — Generalized linear models

[R] **logistic** — Logistic regression, reporting odds ratios

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

[U] **20 Estimation and postestimation commands**