logistic — Logistic regression, reporting odds ratios

Syntax

logistic 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, bootstrap, or jackknife

Reporting

level(#) set confidence level; default is level(95)
coef report estimated coefficients
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; seldom used
coefflegend display legend instead of statistics

indepvars may contain factor variables; see [U] 11.4.3 Factor variables.
depvar and indepvars may contain time-series operators; see [U] 11.4.4 Time-series varlists.
bootstrap, by, fp, jackknife, mfp, mi estimate, nestreg, rolling, statsby, stepwise, and svy are allowed; see [U] 11.1.10 Prefix commands.
vce(bootstrap) and vce(jackknife) are not allowed with the mi estimate prefix; see [MI] mi estimate.
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.
coefflegend does not appear in the dialog box.
See [U] 20 Estimation and postestimation commands for more capabilities of estimation commands.
**logistic — Logistic regression, reporting odds ratios**

Menu

Statistics > Binary outcomes > Logistic regression (reporting odds ratios)

**Description**

logistic fits a logistic regression model of \textit{depvar} on \textit{indepvars}, where \textit{depvar} is a 0/1 variable (or, more precisely, a 0/non-0 variable). Without arguments, logistic redisplays the last logistic estimates. logistic displays estimates as odds ratios; to view coefficients, type logit after running logistic. To obtain odds ratios for any covariate pattern relative to another, see [R] lincom.

**Options**

- **Model**
  - noconstant, offset(\textit{varname}), constraints(\textit{constraints}), collinear; see [R] estimation options.

- **SE/Robust**
  - vce(\textit{vcetype}) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (\textit{om}), that are robust to some kinds of misspecification (\textit{robust}), that allow for intragroup correlation (\textit{cluster} \textit{clustvar}), and that use bootstrap or jackknife methods (\textit{bootstrap}, \textit{jackknife}); see [R] vce option.

- **Reporting**
  - level(\#); see [R] estimation options.

  - coef causes logistic to report the estimated coefficients rather than the odds ratios (exponentiated coefficients). coef may be specified when the model is fit or may be used later to redisplay results.

  - nocnsreport; see [R] estimation options.

  - \textit{display_options}: noomitted, vsquish, noemptycells, baselevels, allbaselevels, nolabel, fvwrap(\#), fvwrapon(style), cformat(\%\textit{fmt}), pformat(\%\textit{fmt}), sformat(\%\textit{fmt}), and nolstretch; see [R] estimation options.

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

The following option is available with logistic but is not shown in the dialog box:

- coeflegend; see [R] estimation options.
Remarks and examples

Remarks are presented under the following headings:

- logistic and logit
- Robust estimate of variance
- Video examples

logistic and logit

logistic provides an alternative and preferred way to fit maximum-likelihood logit models, the other choice being logit ([R] logit).

First, let’s dispose of some confusing terminology. We use the words logit and logistic to mean the same thing: maximum likelihood estimation. To some, one or the other of these words connotes transforming the dependent variable and using weighted least squares to fit the model, but that is not how we use either word here. Thus the logit and logistic commands produce the same results.

The logistic command is generally preferred to the logit command because logistic presents the estimates in terms of odds ratios rather than coefficients. To some people, this may seem disadvantageous, but you can type logit without arguments after logistic to see the underlying coefficients. You should be cautious when interpreting the odds ratio of the constant term. Usually, this odds ratio represents the baseline odds of the model when all predictor variables are set to zero. However, you must verify that a zero value for all predictor variables in the model actually makes sense before continuing with this interpretation.

Nevertheless, [R] logit is still worth reading because logistic shares the same features as logit, including omitting variables due to collinearity or one-way causation.

For an introduction to logistic regression, see Lemeshow and Hosmer (2005), Pagano and Gauvreau (2000, 470–487), or Pampel (2000); for a complete but nonmathematical treatment, see Kleinbaum and Klein (2010); and for a thorough discussion, see Hosmer, Lemeshow, and Sturdivant (2013). See Gould (2000) for a discussion of the interpretation of logistic regression. See Dupont (2009) or Hilbe (2009) for a discussion of logistic regression with examples using Stata. For a discussion using Stata with an emphasis on model specification, see Vittinghoff et al. (2012).

Stata has a variety of commands for performing estimation when the dependent variable is dichotomous or polytomous. See Long and Freese (2014) for a book devoted to fitting these models with Stata. Here is a list of some estimation commands that may be of interest. See help estimation commands for a complete list of all of Stata’s estimation commands.
<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>asclogit</td>
<td>Alternative-specific conditional logit (McFadden's choice) model</td>
</tr>
<tr>
<td>asmprobit</td>
<td>Alternative-specific multinomial probit regression</td>
</tr>
<tr>
<td>asroprobit</td>
<td>Alternative-specific rank-ordered probit regression</td>
</tr>
<tr>
<td>binreg</td>
<td>Generalized linear models for the binomial family</td>
</tr>
<tr>
<td>biprobit</td>
<td>Bivariate probit regression</td>
</tr>
<tr>
<td>blogit</td>
<td>Logit regression for grouped data</td>
</tr>
<tr>
<td>bprobit</td>
<td>Probit regression for grouped data</td>
</tr>
<tr>
<td>clogit</td>
<td>Conditional (fixed-effects) logistic regression</td>
</tr>
<tr>
<td>cloglog</td>
<td>Complementary log-log regression</td>
</tr>
<tr>
<td>exlogistic</td>
<td>Exact logistic regression</td>
</tr>
<tr>
<td>glm</td>
<td>Generalized linear models</td>
</tr>
<tr>
<td>glogit</td>
<td>Weighted least-squares logistic regression for grouped data</td>
</tr>
<tr>
<td>gprobit</td>
<td>Weighted least-squares probit regression for grouped data</td>
</tr>
<tr>
<td>heckoprobit</td>
<td>Ordered probit model with sample selection</td>
</tr>
<tr>
<td>heckprobit</td>
<td>Probit model with sample selection</td>
</tr>
<tr>
<td>hetprobit</td>
<td>Heteroskedastic probit model</td>
</tr>
<tr>
<td>ivprobit</td>
<td>Probit model with endogenous regressors</td>
</tr>
<tr>
<td>logit</td>
<td>Logistic regression, reporting coefficients</td>
</tr>
<tr>
<td>mecloglog</td>
<td>Multilevel mixed-effects complementary log-log regression</td>
</tr>
<tr>
<td>meglm</td>
<td>Multilevel mixed-effects generalized linear model</td>
</tr>
<tr>
<td>melogit</td>
<td>Multilevel mixed-effects logistic regression</td>
</tr>
<tr>
<td>meprobit</td>
<td>Multilevel mixed-effects probit regression</td>
</tr>
<tr>
<td>mlogit</td>
<td>Multinomial (polytomous) logistic regression</td>
</tr>
<tr>
<td>mprobit</td>
<td>Multinomial probit regression</td>
</tr>
<tr>
<td>nlogit</td>
<td>Nested logit regression (RUM-consistent and nonnormalized)</td>
</tr>
<tr>
<td>ologit</td>
<td>Ordered logistic regression</td>
</tr>
<tr>
<td>oprobit</td>
<td>Ordered probit regression</td>
</tr>
<tr>
<td>probit</td>
<td>Probit regression</td>
</tr>
<tr>
<td>rologit</td>
<td>Rank-ordered logistic regression</td>
</tr>
<tr>
<td>scobit</td>
<td>Skewed logistic regression</td>
</tr>
<tr>
<td>slogit</td>
<td>Stereotype logistic regression</td>
</tr>
<tr>
<td>svy: cmd</td>
<td>Survey versions of many of these commands are available;</td>
</tr>
<tr>
<td></td>
<td>see [SVY] svy estimation</td>
</tr>
<tr>
<td>xtcloglog</td>
<td>Random-effects and population-averaged cloglog models</td>
</tr>
<tr>
<td>xtgee</td>
<td>GEE population-averaged generalized linear models</td>
</tr>
<tr>
<td>xtlogit</td>
<td>Fixed-effects, random-effects, and population-averaged logit models</td>
</tr>
<tr>
<td>xtologit</td>
<td>Random-effects ordered logistic models</td>
</tr>
<tr>
<td>xtoprobit</td>
<td>Random-effects ordered probit models</td>
</tr>
<tr>
<td>xtnprobit</td>
<td>Random-effects and population-averaged probit models</td>
</tr>
</tbody>
</table>
Consider the following dataset from a study of risk factors associated with low birthweight described in Hosmer, Lemeshow, and Sturdivant (2013, 24).

```
. use http://www.stata-press.com/data/r13/lbw
(Hosmer & Lemeshow data)
. describe
```

Contains data from http://www.stata-press.com/data/r13/lbw.dta

| obs: 189 | Hosmer & Lemeshow data |
| vars: 11 | 15 Jan 2013 05:01 |
| size: 2,646 |

We want to investigate the causes of low birthweight. Here `race` is a categorical variable indicating whether a person is white (`race = 1`), black (`race = 2`), or some other race (`race = 3`). We want indicator (dummy) variables for `race` included in the regression, so we will use factor variables.

```
. logistic low age lwt i.race smoke ptl ht ui
```

Logistic regression

| Number of obs  =  189 |
| LR chi2(8)     = 33.22 |
| Prob > chi2    =  0.0001 |
| Log likelihood =  -100.724 |
| Pseudo R2      =  0.1416 |

| low | Odds Ratio | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|-----|------------|-----------|------|------|----------------------|
| age | .9732636   | .0354759  | -0.74| 0.457| .9061578  1.046339 |
| lwt | .9849634   | .0068217  | -2.19| 0.029| .9716834  0.9984249 |
| race|            |           |      |      |                      |
| black| 3.534767 | 1.860737 | 2.40 | 0.016| 1.259736  9.918406 |
| other| 2.360879 | 1.039949 | 1.96 | 0.050| 1.001356  5.600207 |
| smoke| 2.517698 | 1.00916  | 2.30 | 0.021| 1.147676  5.523162 |
| ptl | 1.719161   | .5952579  | 1.56 | 0.118| .8721455  3.388787 |
| ht  | 6.249602   | 4.322408  | 2.65 | 0.008| 1.611152  24.24199 |
| ui  | 2.1351     | .9808153  | 1.65 | 0.099| .8677528  5.2534 |
| _cons| 1.586014  | 1.910496  | 0.38 | 0.702| 0.1496092 16.8134 |

The odds ratios are for a one-unit change in the variable. If we wanted the odds ratio for `age` to be in terms of 4-year intervals, we would type
. gen age4 = age/4
. logistic low age4 lwt i.race smoke ptl ht ui
(output omitted)

After logistic, we can type logit to see the model in terms of coefficients and standard errors:

. logit
Logistic regression  Number of obs = 189
LR chi2(8) = 33.22
Prob > chi2 = 0.0001
Log likelihood = -100.724  Pseudo R2 = 0.1416

| low    | Coef.  | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|--------|--------|-----------|-------|-------|---------------------|
| age4   | -.1084 | .1458     | -0.74 | 0.457 | -.3941673 - .1773649|
| lwt    | -.0151 | .0069     | -2.19 | 0.029 | -.0287253 - .0015763|
| race   |        |           |       |       |                     |
| black  | 1.2626 | .5264     | 2.40  | 0.016 | .2309024 2.294392   |
| other  | .8620 | .4391     | 1.96  | 0.050 | .0013548 1.722804   |
| smoke  | .9233 | .4008     | 2.30  | 0.021 | .137739 1.708951    |
| ptl    | .5418 | .3462     | 1.56  | 0.118 | -.136799 1.220472   |
| ht     | 1.8325| .6916     | 2.65  | 0.008 | .4769494 3.188086   |
| ui     | .7585 | .4594     | 1.65  | 0.099 | -.1418484 1.658875  |
| _cons  | .4612 | 1.2046    | 0.38  | 0.702 | -.1899729 2.822176  |

If we wanted to see the logistic output again, we would type logistic without arguments.

Example 2

We can specify the confidence interval for the odds ratios with the level() option, and we can do this either at estimation time or when replaying the model. For instance, to see our first model in example 1 with narrower, 90% confidence intervals, we might type

. logistic, level(90)
Logistic regression  Number of obs = 189
LR chi2(8) = 33.22
Prob > chi2 = 0.0001
Log likelihood = -100.724  Pseudo R2 = 0.1416

| low    | Odds Ratio | Std. Err. | z     | P>|z|   | [90% Conf. Interval] |
|--------|------------|-----------|-------|-------|---------------------|
| age4   | .8973      | .1308     | -0.74 | 0.457 | .7059409 1.140448   |
| lwt    | .9849      | .0069     | -2.19 | 0.029 | .9738063 .9962483   |
| race   |            |           |       |       |                     |
| black  | 3.5347     | 1.8607    | 2.40  | 0.016 | 1.487028 8.402379   |
| other  | 2.3681     | 1.0399    | 1.96  | 0.050 | 1.149971 4.876471   |
| smoke  | 2.5177     | 1.0091    | 2.30  | 0.021 | 1.302185 4.867819   |
| ptl    | 1.7191     | 0.9562    | 1.56  | 0.118 | .9726876 3.038505   |
| ht     | 6.2496     | 4.3224    | 2.65  | 0.008 | 2.003487 19.49478   |
| ui     | 2.1351     | .9808     | 1.65  | 0.099 | 1.00291 4.548424    |
| _cons  | 1.5860     | 1.9105    | 0.38  | 0.702 | .2186791 11.50288   |
Robust estimate of variance

If you specify `vce(robust)`, Stata reports the robust estimate of variance described in [U] 20.21 Obtaining robust variance estimates. Here is the model previously fit with the robust estimate of variance:

```
. logistic low age lwt i.race smoke ptl ht ui, vce(robust)
```

Logistic regression

| low | Odds Ratio | Std. Err. | z    | P>|z| | 95% Conf. Interval |
|-----|------------|-----------|------|-----|-------------------|
| age | 0.9732636  | 0.0329376 | -0.80| 0.423 | 0.9108015 - 1.040009 |
| lwt | 0.9849634  | 0.0070209 | -2.13| 0.034 | 0.9712984 - 0.9988206 |
| race|            |           |      |      |                   |
| black| 3.534767 | 1.793616 | 2.49 | 0.013 | 1.307504 - 9.556051 |
| other| 2.368079 | 1.026563 | 1.99 | 0.047 | 1.012512 - 5.538501 |
| smoke| 2.517698 | 0.9736417 | 2.39 | 0.017 | 1.179852 - 5.372537 |
| ptl | 1.719161  | 0.7072902 | 1.32 | 0.188 | 0.7675715 - 3.850476 |
| ht | 6.249602 | 4.102026 | 2.79 | 0.005 | 1.726445 - 22.6231 |
| ui | 2.1351 | 1.042775 | 1.55 | 0.120 | 0.8197749 - 5.560858 |
| _cons | 1.586014 | 1.939482 | 0.38 | 0.706 | 0.144345 - 17.42658 |

Also you can specify `vce(cluster clustvar)` and then, within cluster, relax the assumption of independence. To illustrate this, we have made some fictional additions to the low-birthweight data.

Say that these data are not a random sample of mothers but instead are a random sample of mothers from a random sample of hospitals. In fact, that may be true—we do not know the history of these data.

Hospitals specialize, and it would not be too incorrect to say that some hospitals specialize in more difficult cases. We are going to show two extremes. In one, all hospitals are alike, but we are going to estimate under the possibility that they might differ. In the other, hospitals are strikingly different. In both cases, we assume that patients are drawn from 20 hospitals.

In both examples, we will fit the same model, and we will type the same command to fit it. Below are the same data we have been using but with a new variable, `hospid`, that identifies from which of the 20 hospitals each patient was drawn (and which we have made up):
logistic — Logistic regression, reporting odds ratios

. use http://www.stata-press.com/data/r13/hospid1, clear
. logistic low age lwt i.race smoke ptl ht ui, vce(cluster hospid)

Logistic regression                 Number of obs =    189
Wald chi2(8) =    49.67
Prob > chi2 =    0.0000
Log pseudolikelihood =  -100.724
Pseudo R2 =    0.1416

(Std. Err. adjusted for 20 clusters in hospid)

Odds Ratio Std. Err. z   P>|z|   [95% Conf. Interval]
--- ------- -------- ------ -------- -----------------------
age .9732636 .0397476 -0.66 0.507 .898396 1.05437
lwt .9849634 .0057101 -2.61 0.009 .9738352 .9962187
race
  black 3.534767 2.013285 2.22 0.027 1.157563 10.79386
  other 2.368079 1.8451325 2.42 0.016 1.176562 4.766257
smoke
  smoke 2.517698 .8284259 2.81 0.005 1.321062 4.79826
  ptl 1.719161 .6676221 1.40 0.163 .8030814 3.680219
  ht 6.249602 4.066275 2.82 0.005 1.74591 22.37086
  ui 2.1351 1.093144 1.48 0.138 .7827337 5.824014
_cons 1.586014 1.661913 0.44 0.660 .2034094 12.36639

The standard errors are similar to the standard errors we have previously obtained, whether we used the robust or conventional estimators. In this example, we invented the hospital IDs randomly.

Here are the results of the estimation with the same data but with a different set of hospital IDs:

. use http://www.stata-press.com/data/r13/hospid2
. logistic low age lwt i.race smoke ptl ht ui, vce(cluster hospid)

Logistic regression                 Number of obs =    189
Wald chi2(8) =    7.19
Prob > chi2 =    0.5167
Log pseudolikelihood =  -100.724
Pseudo R2 =    0.1416

(Std. Err. adjusted for 20 clusters in hospid)

Odds Ratio Std. Err. z   P>|z|   [95% Conf. Interval]
--- ------- -------- ------ -------- -----------------------
age .9732636 .0293064 -0.90 0.368 .9174862 1.032432
lwt .9849634 .0106123 -1.41 0.160 .9643817 1.005984
race
  black 3.534767 3.120385 1.43 0.153 .6265521 19.9418
  other 2.368079 1.297738 1.57 0.116 .8089594 6.932114
smoke
  smoke 2.517698 1.570287 1.48 0.139 .7414969 8.548655
  ptl 1.719161 .6799153 1.40 0.163 .7919045 3.732161
  ht 6.249602 7.165454 1.60 0.110 .660558 59.12808
  ui 2.1351 1.093144 1.48 0.138 .7827337 5.824014
_cons 1.586014 1.946253 0.38 0.707 .1431423 17.573

Note the strikingly larger standard errors. What happened? In these data, women most likely to have low-birthweight babies are sent to certain hospitals, and the decision on likeliness is based not just on age, smoking history, etc., but on other things that doctors can see but that are not recorded in our data. Thus merely because a woman is at one of the centers identifies her to be more likely to have a low-birthweight baby.
Video examples

Logistic regression, part 1: Binary predictors
Logistic regression, part 2: Continuous predictors
Logistic regression, part 3: Factor variables

Stored results

logistic stores the following in e():

Scalars

- \( e(N) \) number of observations
- \( e(N_{\text{cds}}) \) number of completely determined successes
- \( e(N_{\text{cdf}}) \) number of completely determined failures
- \( e(k) \) number of parameters
- \( e(k_{\text{eq}}) \) number of equations in \( e(b) \)
- \( e(k_{\text{eq}_\text{model}}) \) number of equations in overall model test
- \( e(k_{\text{dv}}) \) number of dependent variables
- \( e(df_m) \) model degrees of freedom
- \( e(r^2_p) \) pseudo-\( R^2 \)-squared
- \( e(ll) \) log likelihood
- \( e(ll_0) \) log likelihood, constant-only model
- \( e(N_{\text{clust}}) \) number of clusters
- \( e(\chi^2) \) \( \chi^2 \) test
- \( e(p) \) significance of model test
- \( e(rank) \) rank of \( e(V) \)
- \( e(ic) \) number of iterations
- \( e(rc) \) return code
- \( e(\text{converged}) \) 1 if converged, 0 otherwise

Macros

- \( e(cmd) \) logistic
- \( e(cmdline) \) 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 \( \chi^2 \) test
- \( e(vcetype) \) \( vce() \) 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
- \( e(\text{estat\_cmd}) \) program used to implement \( \text{estat} \)
- \( e(predict) \) program used to implement \( \text{predict} \)
- \( e(marginsnotok) \) predictions disallowed by \( \text{margins} \)
- \( e(asbalanced) \) factor variables \( \text{fvset} \) as \( \text{asbalanced} \)
- \( e(asobserved) \) factor variables \( \text{fvset} \) as \( \text{asobserved} \)
Methods and formulas

Define $x_j$ as the (row) vector of independent variables, augmented by 1, and $b$ as the corresponding estimated parameter (column) vector. The logistic regression model is fit by logit; see [R] logit for details of estimation.

The odds ratio corresponding to the $i$th coefficient is $\psi_i = \exp(b_i)$. The standard error of the odds ratio is $s_i^{\psi} = \psi_i s_i$, where $s_i$ is the standard error of $b_i$ estimated by logit.

Define $I_j = x_j b$ as the predicted index of the $j$th observation. The predicted probability of a positive outcome is

$$p_j = \frac{\exp(I_j)}{1 + \exp(I_j)}$$

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.

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

References


Also see

[R] logistic postestimation — Postestimation tools for logistic
[R] brier — Brier score decomposition
[R] cloglog — Complementary log-log regression
[R] exlogistic — Exact logistic regression
[R] logit — Logistic regression, reporting coefficients
[R] roc — Receiver operating characteristic (ROC) analysis
[M] estimation — Estimation commands for use with mi estimate
[SVY] svy estimation — Estimation commands for survey data
[XT] xtlogit — Fixed-effects, random-effects, and population-averaged logit models
[U] 20 Estimation and postestimation commands