betareg — Beta regression

Description

betareg estimates the parameters of a beta regression model. This model accommodates dependent variables that are greater than 0 and less than 1, such as rates, proportions, and fractional data.

Quick start

Beta regression of y on x1 and x2

```
    betareg y x1 x2
```

Add categorical variable a using factor-variable syntax

```
    betareg y x1 x2 i.a
```

Add covariates for scale

```
    betareg y x1 x2 i.a, scale(x1 z1)
```

As above, but use probit link for conditional mean and square-root link for conditional scale

```
    betareg y x1 x2 i.a, scale(x1 z1) link(probit) slink(root)
```

Beta regression of y on x1 and x2 with robust standard errors

```
    betareg y x1 x2, vce(robust)
```

Menu

Statistics > Fractional outcomes > Beta regression

Syntax

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

options	Description
Model	
`noconstant`	suppress constant term
`scale(`*varlist*[, `noconstant`]`)`	specify independent variables for scale
`link(`*linkname*`)`	specify link function for the conditional mean; default is `link(logit)`
`slink(`*slinkname*`)`	specify link function for the conditional scale; default is `slink(log)`
`constraints(`*constraints*`)`	apply specified linear constraints
SE/Robust	
`vce(`*vcetype*`)`	*vcetype* may be `oim`, `robust`, `cluster` *clustvar*, `bootstrap`, or `jackknife`
Reporting	
`level(`*#*`)`	set confidence level; default is `level(95)`
`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
`coeflegend`	display legend instead of statistics

linkname	Description
`logit`	logit
`probit`	probit
`cloglog`	complementary log-log
`loglog`	log-log

slinkname	Description
`log`	log
`root`	square root
`identity`	identity

indepvars and *varlist* specified in `scale()` 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.

Options

Model

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

scale(*varlist* [, noconstant]) specifies the independent variables used to model the scale.

noconstant suppresses the constant term in the scale model. A constant term is included by default.

link(*linkname*) specifies the link function used for the conditional mean. *linkname* may be logit, probit, cloglog, or loglog. The default is link(logit).

slink(*slinkname*) specifies the link function used for the conditional scale. *slinkname* may be log, root, or identity. The default is slink(log).

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), 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(#), 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 option is available with betareg but is not shown in the dialog box:

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

Remarks and examples

stata.com

Dependent variables such as rates, proportions, and fractional data are frequently greater than 0 and less than 1. There are a variety of methods to model such variables, including beta regression and fractional logistic regression.

Beta regression is widely used because of its flexibility for modeling variables between 0 and 1 and because its predictions are confined to the same range. However, beta regression models are not appropriate for dependent variables with some observations exactly equal to 0 or 1. See [R] **fracreg** for models when the dependent variable can equal 0 or 1 that also make predictions over the same range. The predictions from linear regression models are not constrained to the 0 to 1 interval; thus they are not widely used for these variables.

These models have applications in a variety of disciplines, such as economics, the social sciences, and health science. For example, Castellani, Pattitoni, and Scorcu (2012) use beta regression to estimate Gini index values for the prices of art by famous and nonfamous artists. In political science, Paolino (2001) explores the advantages of beta regression and reviewed its applicability for a variety of research topics such as the proportion of minority applicants deemed eligible for the Rural Housing Loans program and the proportion of a state's gay and lesbian population that is covered by antidiscrimination laws. In psychology, Smithson, Deady, and Gracik (2007) analyzed the relationship between how jurors judged the probability of a defendant's guilt and the verdict in a trial. Finally, beta regression has been used to model quality-adjusted life years in health-economics outcome studies (Hubben et al. [2008]; Basu and Manca [2012]).

Ferrari and Cribari-Neto (2004) and Smithson and Verkuilen (2006) derived the beta regression estimators implemented in `betareg`. Basu and Manca (2012) also discuss quasimaximum likelihood inference for these estimators. These estimators augment the inherent flexibility of the beta distribution with functional form choices, known as links.

Beta regression is a model of the mean of the dependent variable y conditional on covariates $\mathbf{x}$, which we denote by $\mu_{\mathbf{x}}$. Because y is in $(0,1)$, we must ensure that $\mu_{\mathbf{x}}$ is also in $(0,1)$. We do this by using the link function for the conditional mean, denoted $g(\cdot)$. This is necessary because linear combinations of the covariates are not otherwise restricted to $(0,1)$.

Algebraically,

$$g(\mu_{\mathbf{x}}) = \mathbf{x}\boldsymbol{\beta}$$

or, equivalently,

$$\mu_{\mathbf{x}} = g^{-1}(\mathbf{x}\boldsymbol{\beta})$$

where $g^{-1}(\cdot)$ is the inverse function of $g(\cdot)$. For example, the default logit link implies that

$$\ln\{\mu_{\mathbf{x}}/(1-\mu_{\mathbf{x}})\} = \mathbf{x}\boldsymbol{\beta}$$

and that

$$\mu_{\mathbf{x}} = \exp(\mathbf{x}\boldsymbol{\beta})/\{1+\exp(\mathbf{x}\boldsymbol{\beta})\}$$

Using a link function to keep the conditional-mean model inside an interval is common in the statistical literature; see [R] **glm** for additional applications of link functions.

The conditional variance of the beta distribution is

$$\mathrm{Var}(y|\mathbf{x}) = \{\mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})\}/(1+\psi)$$

The parameter ψ is known as the scale factor because it rescales the conditional variance. We use the scale link to ensure that $\psi > 0$.

▷ Example 1: Beta regression model of a rate

Suppose we wish to know whether offering a summer instruction program increases a school's pass rate for a mandatory state exam administered to students. The school-wide pass rate must be between 0 and 1. It is unlikely that any school will have either no students pass or all students pass, so we consider estimating the effect of the summer instruction program using beta regression.

The dataset `sprogram` contains fictional data on the pass rate of 1,000 schools (`prate`). We begin by reading in the data and verifying that `prate` contains no 0s or 1s.

```
. use http://www.stata-press.com/data/r14/sprogram
(Fictional summer program data)

. summarize prate

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       prate |      1,000    .8150803    .1233684   .2986041   .9973584
```

`prate` ranges from 0.299 to 0.997, so we proceed with our choice of a beta regression model.

We model `prate` as a function of a binary indicator for whether the school offered voluntary, half-day instruction to students during the past two summers (`summer`). The summer program should raise the scores of disadvantaged children who otherwise would not have access to programs that maintain their skills through the summer, thereby increasing the total proportion of students who pass the exam the next year.

We include the fraction of students receiving free or reduced-price meals (`freemeals`) and the sum of parents' monetary donations to the school two years earlier (`pdonations`) as additional covariates that measure affluence of the students' parents. We estimate the parameters of this model using the default logit link for the conditional mean and log link for the conditional scale.

```
. betareg prate i.summer freemeals pdonations
initial:       log likelihood =  781.55846
rescale:       log likelihood =  781.55846
rescale eq:    log likelihood =  781.55846
(setting technique to bhhh)
Iteration 0:   log likelihood =  781.55846
Iteration 1:   log likelihood =  891.57913
Iteration 2:   log likelihood =  892.99578
Iteration 3:   log likelihood =  893.02725
Iteration 4:   log likelihood =   893.0279
Iteration 5:   log likelihood =  893.02792

Beta regression                                 Number of obs     =      1,000
                                                LR chi2(3)        =     164.61
                                                Prob > chi2       =     0.0000

Link function    :  g(u) = log(u/(1-u))         [Logit]
Slink function   :  g(u) = log(u)               [Log]

Log likelihood =  893.02792

------------------------------------------------------------------------------
       prate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prate        |
      summer |
        yes  |   .5560171   .0480307    11.58   0.000     .4618787    .6501555
   freemeals |  -.4564181   .0834885    -5.47   0.000    -.6200525   -.2927836
  pdonations |   .0449706   .0097781     4.60   0.000      .025806    .0641353
       _cons |   1.175013   .0642797    18.28   0.000     1.049027    1.300999
-------------+----------------------------------------------------------------
scale        |
       _cons |   2.375433   .0443005    53.62   0.000     2.288606    2.462261
------------------------------------------------------------------------------
```

The output table reports the estimated coefficients of the covariates and an estimated scale parameter. The coefficient of the factor variable for `summer==1`, shown as `yes` under `summer`, is significant and positive. Thus we conclude that the summer program was effective at increasing a school's pass rate.

However, we cannot determine the magnitude of the effect from these results. In general, when you use `betareg`, the best way to obtain interpretable effect sizes for the covariates is by using `margins`. See example 1 in [R] **betareg postestimation** for more information.

◁

❑ Technical note

The results displayed in example 1 can be written concisely in algebraic form. Because we used the default logit link, the estimated conditional mean is

$$\widehat{\mu}_{\mathbf{x}} = \exp(\mathbf{x}\widehat{\boldsymbol{\beta}})/\{1 + \exp(\mathbf{x}\widehat{\boldsymbol{\beta}})\}$$

where the estimated $\mathbf{x}\boldsymbol{\beta}$ is

$$\mathbf{x}\widehat{\boldsymbol{\beta}} = -0.456 \times \texttt{freemeals} + 0.045 \times \texttt{pdonations} + 0.556 \times (\texttt{summer==1}) + 1.18$$

The estimated ψ with the default log link for the scale is $\widehat{\psi} = \exp(2.38) = 10.80$, which is simply substituted into the formula given above to express the conditional variance.

See *Methods and formulas* for the functional forms of the other links implemented in `betareg` for the conditional mean and scale.

❑

▷ Example 2: Modeling conditional variance

Some processes require that we model the scale parameter as a function of covariates. For example, we might believe that the proportion of students with free or reduced meals influences the variance of the estimated mean.

We augment example 1 by modeling the scale parameter as a function of `freemeals`.

```
. betareg prate i.summer freemeals pdonations, scale(freemeals)
  (output omitted)
Beta regression                                 Number of obs     =      1,000
                                                LR chi2(4)        =     169.38
                                                Prob > chi2       =     0.0000

Link function  :  g(u) = log(u/(1-u))           [Logit]
Slink function :  g(u) = log(u)                 [Log]

Log likelihood =  895.41544

------------------------------------------------------------------------------
       prate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prate        |
      summer |
         yes |   .5571133   .0480378    11.60   0.000     .4629609    .6512658
   freemeals |  -.5291892   .0896511    -5.90   0.000    -.7049021   -.3534762
  pdonations |   .0454228   .0097809     4.64   0.000     .0262527    .0645929
       _cons |   1.209179   .0649585    18.61   0.000     1.081863    1.336495
-------------+----------------------------------------------------------------
scale        |
   freemeals |  -.3598137   .1644214    -2.19   0.029    -.6820737   -.0375536
       _cons |   2.547047   .0882327    28.87   0.000     2.374114     2.71998
------------------------------------------------------------------------------
```

Again, we conclude that having a summer program increases the pass rate. The effect of an increase in the proportion of students receiving free meals on the conditional variance is ambiguous because it is in both equations. We can use `margins` to estimate the effect of the program on the conditional mean or the conditional variance.

◁

The estimators in betareg are consistent and efficient when the model is correctly specified. Smithson and Verkuilen (2006) discuss model selection for beta regression and note that selecting the model that minimizes the Bayesian information criterion (BIC) will select the correct model in large samples. Selecting the model that minimizes the BIC is a general approach to model selection; see Cameron and Trivedi (2005) for more details.

▷ Example 3: Model selection

We fit the models quietly and use estimates store to store the results under the names model1, model2, model3, and model4.

```
. quietly betareg prate i.summer freemeals pdonations, scale(freemeals)

. estimates store model1

. quietly betareg prate i.summer freemeals pdonations, scale(freemeals)
> link(cloglog)

. estimates store model2

. quietly betareg prate i.summer freemeals pdonations, scale(freemeals)
> slink(root)

. estimates store model3

. quietly betareg prate i.summer freemeals pdonations, scale(freemeals)
> link(cloglog) slink(root)

. estimates store model4
```

Next, we use estimates table to display the coefficients, standard errors, and the BIC for each model.

```
. estimates table model1 model2 model3 model4, stats(bic) se

--------------------------------------------------------------------
    Variable |   model1       model2       model3       model4
-------------+------------------------------------------------------
prate        |
      summer |
         yes |  .55711332    .27762093    .55719742    .27765283
             |  .04803785    .02460121    .04803698    .02459953
             |
   freemeals | -.52918917   -.25685191    -.5300549   -.25729221
             |  .08965112    .04308385    .08978883    .04314336
  pdonations |  .04542281    .02162612    .04542462    .02163225
             |  .00978086    .00444813    .00978099    .00444817
       _cons |  1.2091789    .37961457    1.2094991    .37977162
             |  .06495845    .03212448    .06496946    .03212509
-------------+------------------------------------------------------
scale        |
   freemeals | -.35981368   -.36808486   -.59912234   -.61295521
             |  .16442142    .16448631    .27259725    .27273851
       _cons |  2.5470469    2.5516626    3.5692049    3.5771444
             |   .0882327    .08821157    .15204032    .15218042
-------------+------------------------------------------------------
Statistics   |
         bic | -1749.3843   -1749.9903     -1749.444   -1750.0511
--------------------------------------------------------------------
                                                        legend: b/se
```

We select model4, the model with the complementary log-log link for the conditional mean and the square-root link for the conditional variance, because it has the lowest BIC.

◁

Stored results

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
`e(ll_0)`	log likelihood, constant-only model
`e(N_clust)`	number of clusters
`e(chi2)`	χ^2
`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)`	`betareg`
`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(link)`	link function in the conditional mean equation
`e(slink)`	link function in the conditional scale equation
`e(clustvar)`	name of cluster variable
`e(chi2type)`	`Wald` or `LR`; type of model χ^2 test
`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(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

Beta regression models were proposed by Ferrari and Cribari-Neto (2004) and extended by Smithson and Verkuilen (2006) to allow the scale parameter to depend on covariates.

Beta regression is only appropriate for a dependent variable that is strictly greater than 0 and strictly less than 1 because the beta distribution only has support on the interval $(0, 1)$. The density of a beta-distributed dependent variable y conditional on covariates $\mathbf{x}$ can be written as

$$f(y;\mu_{\mathbf{x}},\psi_{\mathbf{x}}) = \frac{\Gamma(\psi_{\mathbf{x}})}{\Gamma(\mu_{\mathbf{x}}\psi_{\mathbf{x}})\Gamma\{(1-\mu_{\mathbf{x}})\psi_{\mathbf{x}}\}} y^{\mu_{\mathbf{x}}\psi_{\mathbf{x}}-1}(1-y)^{(1-\mu_{\mathbf{x}})\psi_{\mathbf{x}}-1}$$

where $\mu_{\mathbf{x}} = \mathbf{E}(y|\mathbf{x})$, $\mu_{\mathbf{x}}$ is linked to the covariates by the link function $g(\mu_{\mathbf{x}}) = \mathbf{x}\boldsymbol{\beta}$, $\psi_{\mathbf{x}}$ scales the conditional variance according to

$$\text{Var}(y|\mathbf{x}) = \mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})/(1+\psi_{\mathbf{x}})$$

and $\psi_{\mathbf{x}}$ is linked to the covariates by link function $h(\psi_{\mathbf{x}}) = \mathbf{x}\boldsymbol{\gamma}$.

This parameterization yields a log-likelihood function of

$$\sum_{i=1}^{N} \omega_i \Big(\ln\{\Gamma(\psi_{\mathbf{x},i})\} - \ln\{\Gamma(\mu_{\mathbf{x},i}\psi_{\mathbf{x},i})\} - \ln[\Gamma\{(1-\mu_{\mathbf{x},i})\psi_{\mathbf{x},i}\}] \\ + (\mu_{\mathbf{x},i}\psi_{\mathbf{x},i}-1)\ln(y_i) + \{(1-\mu_{\mathbf{x},i})\psi_{\mathbf{x},i}-1\}\ln(1-y_i)\Big)$$

The definitions of the link functions are

Name	Function
`logit`	$g(\mu_{\mathbf{x}}) = \ln\{\mu_{\mathbf{x}}/(1-\mu_{\mathbf{x}})\}$
`probit`	$g(\mu_{\mathbf{x}}) = \Phi^{-1}(\mu_{\mathbf{x}})$
`cloglog`	$g(\mu_{\mathbf{x}}) = \ln\{-\ln(1-\mu_{\mathbf{x}})\}$
`loglog`	$g(\mu_{\mathbf{x}}) = -\ln\{-\ln(\mu_{\mathbf{x}})\}$

The definitions of the scale-link functions are

Name	Function
`log`	$h(\psi_{\mathbf{x}}) = \ln(\psi_{\mathbf{x}})$
`root`	$h(\psi_{\mathbf{x}}) = \sqrt{\psi_{\mathbf{x}}}$
`identity`	$h(\psi_{\mathbf{x}}) = \psi_{\mathbf{x}}$

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*.

Acknowledgments

We thank Maarten L. Buis of the University of Konstanz, Germany, Nicholas J. Cox of Durham University, UK, and Stephen P. Jenkins of the London School of Economics and Political Science, UK, who wrote `betafit`, an early implementation of beta regression in Stata.

References

Basu, A., and A. Manca. 2012. Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making* 32: 56–69.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Castellani, M., P. Pattitoni, and A. E. Scorcu. 2012. Visual artist price heterogeneity. *Economics and Business Letters* 1(3): 16–22.

Ferrari, S. L. P., and F. Cribari-Neto. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31: 799–815.

Hubben, G. A. A., D. Bishai, P. Pechlivanoglou, A. M. Cattelan, R. Grisetti, C. Facchin, F. A. Compostella, J. M. Bos, M. J. Postma, and A. Tramarin. 2008. The societal burden of HIV/AIDS in Northern Italy: An analysis of costs and quality of life. *AIDS Care: Psychological and Socio-medical Aspects of AIDS/HIV* 20: 449–455.

Paolino, P. 2001. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis* 9: 325–346.

Smithson, M., S. Deady, and L. Gracik. 2007. Guilty, not guilty, or . . . ? Multiple options in jury verdict choices. *Journal of Behavioral Decision Making* 20: 481–498.

Smithson, M., and J. Verkuilen. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11: 54–71.

Also see

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

[R] **fracreg** — Fractional response regression

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

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

[U] 20 Estimation and postestimation commands