> **gsem family-and-link options** — Family-and-link options

Description   Syntax   Options   Remarks and examples   Also see

Description

gsem not only allows models of the form $y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i$ but also allows

$$g\{E(y_i)\} = \mathbf{x}_i\boldsymbol{\beta}$$
$$y_i \sim F$$

where you can choose F and $g(\cdot)$ from a menu. F is called the family, and $g(\cdot)$ is called the link. One set of choices is the Gaussian distribution for F and the identity function for $g(\cdot)$. In that case, gsem reproduces linear regression. Other combinations of $g(\cdot)$ and F produce other popular models, including logit (also known as logistic regression), probit, multinomial logit, Poisson regression, and more.

Syntax

gsem *paths* ... , ... *family_and_link_options*

family_and_link_options	Description
<u>f</u>amily(*family*)	distribution family; default is family(gaussian)
<u>l</u>ink(*link*)	link function; default varies per family
cloglog	synonym for family(bernoulli) link(cloglog)
exponential	synonym for family(exponential) link(log)
gamma	synonym for family(gamma) link(log)
logit	synonym for family(bernoulli) link(logit)
loglogistic	synonym for family(loglogistic) link(log)
lognormal	synonym for family(lognormal) link(log)
llogistic	synonym for family(llogistic) link(log)
lnormal	synonym for family(lnormal) link(log)
mlogit	synonym for family(multinomial) link(logit)
nbreg	synonym for family(nbreg mean) link(log)
ocloglog	synonym for family(ordinal) link(cloglog)
ologit	synonym for family(ordinal) link(logit)
oprobit	synonym for family(ordinal) link(probit)
poisson	synonym for family(poisson) link(log)
probit	synonym for family(bernoulli) link(probit)
regress	synonym for family(gaussian) link(identity)
weibull	synonym for family(weibull) link(log)
<u>exp</u>osure(*varname_e*)	include ln(*varname_e*) with coefficient constrained to 1
<u>off</u>set(*varname_o*)	include *varname_o* with coefficient constrained to 1

family	Description
<u>gau</u>ssian [, *options*]	Gaussian (normal); the default
<u>ber</u>noulli	Bernoulli
beta	beta
<u>bin</u>omial [*#* | *varname*]	binomial; default number of binomial trials is 1
<u>o</u>rdinal	ordinal
<u>mu</u>ltinomial	multinomial
<u>poi</u>sson [, *poisson*]	Poisson
<u>nb</u>inomial [mean | <u>cons</u>tant]	negative binomial; default dispersion is mean
<u>e</u>xponential [, *survival*]	exponential
<u>gam</u>ma [, *survival*]	gamma
<u>logl</u>ogistic [, *survival*]	loglogistic
<u>logn</u>ormal [, *survival*]	lognormal
<u>wei</u>bull [, *survival*]	Weibull
pointmass *#*	point-mass density at *#*

link	Description
<u>ident</u>ity	identity
log	log
logit	logit
<u>prob</u>it	probit
<u>clog</u>log	complementary log–log

options	Description
<u>l</u>depvar(*varname*)	lower depvar for interval-response data
<u>u</u>depvar(*varname*)	upper depvar for interval-response data
<u>lc</u>ensored(*varname* | *#*)	lower limit for left-censoring
<u>rc</u>ensored(*varname* | *#*)	upper limit for right-censoring

Only allowed with family(gaussian) with link(identity).

poisson	Description
<u>lt</u>runcated(*varname* | *#*)	lower limit for left-truncation

survival	Description
<u>lt</u>runcated(*varname* | *#*)	lower limit for left-truncation
<u>fa</u>ilure(*varname*)	indicator for failure event
ph	proportional hazards parameterization
aft	accelerated failure-time parameterization

ph is allowed only with families exponential and weibull.

aft is allowed only with families exponential, gamma, loglogistic, lognormal, and weibull.

If you specify both family() and link(), not all combinations make sense. You may choose from the following combinations:

	identity	log	logit	probit	cloglog
Gaussian	D	x			
Bernoulli			D	x	x
beta			D	x	x
binomial			D	x	x
ordinal			D	x	x
multinomial			D		
Poisson		D			
negative binomial		D			
exponential		D			
Weibull		D			
gamma		D			
loglogistic		D			
lognormal		D			
pointmass	D				

D denotes the default.

Options

family(*family*) and link(*linkname*) specify F and $g(\cdot)$. If neither is specified, linear regression is assumed.

Two of the families allow optional arguments:

family(binomial [#|*varname*]) specifies that data are in binomial form, that is, that the response variable records the number of successes from a series of Bernoulli trials. The number of trials is given either as a constant number or as a varname that allows the number of trials to vary over observations, or it is not given at all. In the last case, the number of trials is thus equivalent to specifying family(bernoulli).

family(nbinomial [mean|constant]) specifies a negative binomial model, a Poisson model with overdispersion. Be aware, however, that even Poisson models can have overdispersion if latent variables are included in the model. Let's use the term "conditional overdispersion" to refer to dispersion above and beyond that implied by latent variables, if any.

That conditional overdispersion can take one of two forms. In mean overdispersion, the conditional overdispersion is a linear function of the conditional (predicted) mean. Constant overdispersion refers to the conditional overdispersion being, of course, constant.

If you do not specify mean or constant, then mean is assumed.

family(pointmass #) is a special family allowed when lclass() is also specified.

cloglog, exponential, gamma, logit, loglogistic, lognormal, llogistic, lnormal, mlogit, nbreg, ocloglog, ologit, oprobit, poisson, probit, regress, and weibull are shorthands for specifying popular models.

exposure(*varname_e*) and offset(*varname_o*) are most commonly used with families poisson and nbreg, that is, they typically concern count models.

exposure() specifies a variable that reflects the amount of exposure—usually measured in time units—for each observation over which the responses were counted. If one observation was exposed for twice the time of another, and the observations were otherwise identical, one would expect twice as many events to be counted. To assume that, $\ln(varname_e)$ is entered into $\mathbf{x}_i\boldsymbol{\beta}$ with coefficient constrained to be 1.

offset() enters $varname_o$ into $\mathbf{x}_i\boldsymbol{\beta}$ with coefficient constrained to be 1. offset() is just another way of specifying exposure() where the offset variable is the log of amount of exposure.

If neither exposure() nor offset() is specified, observations are assumed to have equal amounts of exposure.

ldepvar(*varname*) and udepvar(*varname*) specify that each observation can be point data, interval data, left-censored data, or right-censored data. The type of data for a given observation is determined by the values in y_i and *varname*. The following specifications are equivalent:

$depvar_1$ <- ... , family(gauss, udepvar($depvar_2$))

$depvar_2$ <- ... , family(gauss, ldepvar($depvar_1$))

Thus only one of ldepvar() or udepvar() is allowed. In either case, $depvar_1$ and $depvar_2$ should have the following form:

Type of data		$depvar_1$	$depvar_2$
point data	$a = [a, a]$	a	a
interval data	$[a, b]$	a	b
left-censored data	$(-\infty, b]$	.	b
right-censored data	$[a, +\infty)$	a	.

lcensored(*varname*|*#*) and rcensored(*varname*|*#*) indicate the lower and upper limits for censoring, respectively. You may specify only one.

lcensored(*arg*) specifies that observations with $y_i \leq arg$ are left-censored and the remaining observations are not.

rcensored(*arg*) specifies that observations with $y_i \geq arg$ are right-censored and the remaining observations are not.

Neither lcensored() nor rcensored() may not be combined with ldepvar() or udepvar().

ltruncated(*varname* |*#*) indicates the lower limits for truncation.

ltruncated(*arg*) specifies that the distribution is truncated on the left at *arg*, meaning that $y_i \leq arg$ is not within the support for the corresponding distribution family. This option rescales the underlying density function to accommodate the truncated support for y_i. This means that values of y_i that are less than or equal to *arg* do not contribute to the likelihood. For survival families, this means that time (time at risk) starts at *arg* instead of at 0.

failure(*varname*) specifies the failure event.

   If failure() is not specified, all observations are assumed to indicate a failure.

   If failure(*varname*) is specified, *varname* is interpreted as an indicator variable; 0 and missing mean censored, and all other values are interpreted as representing failure.

ph specifies the proportional hazards parameterization and is allowed with families exponential and weibull. This is the default parameterization for these families.

aft specifies the accelerated failure-time parameterization and is allowed with families exponential, gamma, loglogistic, lognormal, and weibull. This is an optional parameterization for exponential and weibull but the only parameterization for the others.

Remarks and examples [stata.com]

   In the command language, the family-and-link options may be specified at the end of a command among the global options:

```
. gsem ..., ... family(...) link(...) ...
. gsem ..., ... poisson exposure(time) ...
```

Specified that way, the options apply to all the response variables. Alternatively, they may be specified inside paths to affect single equations:

```
. gsem (y1 <- x1 x2, family(...) poisson(...)) (y2 <- x2 L) ...
. gsem (y1 <- x1 x2, family(...) link(...))        ///
       (y2 <- x2 L,  family(...) link(...)) ...
. gsem (y1 <- x1 x2, poisson exposure(time)) (y2 <- x2 L) ...
. gsem (y1 <- x1 x2, poisson exposure(time))       ///
       (y2 <- x2 L,  family(...) link(...)) ...
```

   On a different topic, it is worth nothing that you can fit exponential-regression models with family(gamma) link(log) if you constrain the log of the scale parameter to be 0 with gsem's constraints() option. For instance, you might type

```
. constraint 1 _b[/y:logs] = 0
. gsem (y <- x1 x2, gamma), constraints(1)
```

The name _b[/y:logs] changes according to the name of the dependent variable. Had y instead been named waitingtime, the parameter would have been named _b[/waitingtime:logs]. Rather than remembering that, remember instead that the best way to discover the names of parameters is to type

```
. gsem (waitingtime <- x1 x2, gamma), noestimate
```

and then look at the output to discover the names. See [SEM] **sem and gsem option constraints( )**.

For examples of generalized response variables, see the following:

Also see