

[Description](#)[Remarks and examples](#)[References](#)[Also see](#)

Description

The methods and formulas for the `gsem` command are presented below.

Remarks and examples

stata.com

Remarks are presented under the following headings:

*Introduction**Families of distributions**The Bernoulli family**The beta family**The binomial family**The ordinal family**The multinomial family**The Poisson family**The negative binomial family**The Gaussian family**Reliability**Point mass**Link functions**The logit link**The probit link**The complementary log-log link**The log link**The identity link**Survival distributions**The exponential distribution**The Weibull distribution**The gamma distribution**The loglogistic distribution**The lognormal distribution**Models with continuous latent variables**The likelihood**Gauss–Hermite quadrature**Adaptive quadrature**Laplacian approximation**Survey data**Predictions**Empirical Bayes**Other predictions**Models with categorical latent variables**The likelihood**The EM algorithm**Survey data**Predictions*

Introduction

`gsem` fits generalized linear models with categorical or continuous latent variables via maximum likelihood. Here is a table identifying the family/link combinations that `gsem` allows.

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

Log-likelihood calculations for fitting any model with continuous latent variables require integrating out the latent variables. One widely used modern method is to directly estimate the integral required to calculate the log likelihood by Gauss–Hermite quadrature or some variation thereof. `gsem` implements four different methods for numerically evaluating the integrals.

1. Gauss–Hermite quadrature (GHQ)
2. Mean-variance adaptive quadrature (MVAGH)
3. Mode-curvature adaptive quadrature (MCAGH)
4. Laplacian approximation

The default method is MVAGH. The numerical integration method for MVAGH is based on [Rabe-Hesketh, Skrondal, and Pickles \(2005\)](#), and the other numerical integration methods described in this manual entry are based on [Skrondal and Rabe-Hesketh \(2004, chap. 6.3\)](#).

Families of distributions

`gsem` implements the most commonly used distribution families associated with generalized linear models. `gsem` also implements distributions for ordinal and multinomial outcomes.

In this manual entry, observed endogenous variables are also known as generalized responses or generalized outcomes, but we will simply refer to them as responses or outcomes. The random variable corresponding to a given response will be denoted by Y . An observed value of Y will be denoted by y , and the expected value of Y by μ . For the ordinal and multinomial families, we will refer to a linear prediction, denoted by z , instead of the expected value.

The Bernoulli family

The Bernoulli family is a binary response model. The response Y is assumed to take on the values 0 or 1; however, `gsem` allows any nonzero and nonmissing value to mean 1.

The log of the conditional probability mass function is

$$\log f(y|\mu) = y \log \mu + (1 - y) \log(1 - \mu)$$

where μ is also known as the probability of a success. The default link for the Bernoulli family is the `logit` link.

The beta family

The beta family is a fractional response model. The response Y is assumed to take on the real values between 0 and 1.

The log of the conditional probability density function is

$$\begin{aligned} \log f(y|\mu, s) = & \log\{\Gamma(s)\} - \log\{\Gamma(\mu s)\} - \log\{\Gamma(s - \mu s)\} \\ & + (\mu s - 1) \log y + (s - \mu s - 1) \log(1 - y) \end{aligned}$$

where μ is the mean of Y , and s is the scale parameter. `gsem` fits s in the log scale. The default link for the beta family is the `logit` link.

The binomial family

The binomial family is a count response model and generalizes the Bernoulli family by taking the sum of k independent Bernoulli outcomes. The response Y is assumed to take on the values $0, 1, \dots, k$.

The log of the conditional probability mass function is

$$\begin{aligned} \log f(y|\mu) = & \log\{\Gamma(k + 1)\} - \log\{\Gamma(y + 1)\} - \log\{\Gamma(k - y + 1)\} \\ & + y \log \mu + (1 - y) \log(1 - \mu) \end{aligned}$$

where μ is the expected value for a single Bernoulli outcome. The default link for the binomial family is the `logit` link.

The ordinal family

The ordinal family is a discrete response model. The response Y is assumed to take on one of k unique values. The actual values are irrelevant except that higher values are assumed to correspond to “higher” outcomes. Without loss of generality, we will assume that Y takes on the values $1, \dots, k$. The ordinal family with k outcomes has cutpoints $\kappa_0, \kappa_1, \dots, \kappa_k$, where $\kappa_0 = -\infty$, $\kappa_y < \kappa_{y+1}$, and $\kappa_k = +\infty$.

Given a linear prediction z , the probability that a random response Y takes the value y is

$$\Pr(Y = y|z) = \Pr(Y^* < \kappa_y - z) - \Pr(Y^* < \kappa_{y-1} - z)$$

where Y^* is the underlying stochastic component for Y . The distribution for Y^* is determined by the link function. `gsem` allows `logit`, `probit`, and `cloglog` for the ordinal family. The `logit` link assigns Y^* the extreme value distribution that is synonymous with the logit link for Bernoulli outcomes. The `probit` link assigns Y^* the standard normal distribution that is synonymous with the probit link for Bernoulli outcomes. The `cloglog` link assigns Y^* the distribution that is synonymous with the complementary log-log link for Bernoulli outcomes. The default link for the ordinal family is the `logit` link.

The multinomial family

The multinomial family is a discrete response model. The response Y is assumed to take on one of k unique values. The actual values are irrelevant and order does not matter; however, `gsem` requires that the values are nonnegative integers. Without loss of generality, we will assume that Y takes on the values $1, \dots, k$. Each of the k outcomes has its own linear prediction. For the model to be identified, one of the outcomes is chosen to be the base or reference. The linear prediction for the base outcome is constrained to be 0 for all observations. Without loss of generality, we will assume the base outcome is the first outcome. Let z_i be the prediction for outcome i , where $z_1 = 0$ for the base outcome.

Given the k linear predictions $\mathbf{z}' = (z_1, z_2, \dots, z_k)$, the log of the conditional probability mass function is

$$\log f(y|\mathbf{z}) = z_y - \log \left\{ \sum_{i=1}^k \exp(z_i) \right\}$$

The only link allowed for the multinomial family is the `logit` link.

The Poisson family

The Poisson family is a count-data response model. The response Y is assumed to take on nonnegative integer values.

The log of the conditional probability mass function is

$$\log f(y|\mu) = -\mu + y \log \mu - \log \Gamma(y+1)$$

The only link allowed for the Poisson family is the `log` link.

The negative binomial family

The negative binomial family is another count-data response model. It is commonly thought of as a Poisson family with overdispersion. `gsem` allows two parameterizations for the dispersion in this family: mean dispersion and constant dispersion.

The log of the conditional probability mass function is

$$\begin{aligned} \log f(y|\mu, \alpha) &= \log \{ \Gamma(y+m) \} - \log \{ \Gamma(y+1) \} - \log \{ \Gamma(m) \} \\ &\quad + m \log p + y \log(1-p) \end{aligned}$$

where m and p depend on the form of dispersion.

The only link allowed for the negative binomial family is the `log` link.

For mean dispersion, we have

$$m = 1/\alpha$$

$$p = \frac{1}{1 + \alpha\mu}$$

where μ is the expected value of Y and α is the scale parameter. `gsem` fits α in the log scale.

For constant dispersion, we have

$$m = \exp(\log\mu - \log\delta)$$

$$p = \frac{1}{1 + \delta}$$

where μ is the expected value of Y and δ is the scale parameter. `gsem` fits δ in the log scale.

The Gaussian family

The Gaussian family is a continuous response model and is synonymous with the normal distribution.

When the Gaussian family is specified with the `identity` link but no censoring, `gsem` fits this family by using a single multivariate density function and allows the following two special features:

1. `gsem` can fit covariances between the Gaussian error variables.
2. `gsem` can fit paths between Gaussian responses, including nonrecursive systems.

The log of the conditional probability density function is

$$\log f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \left\{ d \log 2\pi + \log |\boldsymbol{\Sigma}| + (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\}$$

where d is the dimension of the observed response vector \mathbf{y} , $\boldsymbol{\mu}$ is the mean of the responses, and $\boldsymbol{\Sigma}$ is the variance matrix of their unexplained errors.

When the Gaussian family is specified with the `log` link or censoring, the two special features described above no longer apply. In addition, the multivariate density function is no longer used. Instead, for each response using the `log` link, the log of the conditional probability density function corresponds to the formula above with $d = 1$. For censored responses, the log likelihood corresponds to the one in the *Methods and formulas* for [R] `intreg`.

Reliability

For a given Gaussian response variable with the `identity` link, the reliability Y may be specified as p or $100 \times p\%$. The variance of Y 's associated error variable is then constrained to $(1 - p)$ times the observed variance of Y .

Point mass

`gsem`, with categorical latent variables, allows an outcome to have a unit probability mass at a single integer value. This feature is useful for fitting models like the zero-inflated Poisson model.

Link functions

Except for the ordinal and multinomial families, the link function defines the transformation between the mean and the linear prediction for a given response. If Y is the random variable corresponding to an observed response variable y , then the link function performs the transformation

$$g(\mu) = z$$

where $\mu = E(Y)$ and z is the linear prediction. In practice, the likelihood evaluator function uses the inverse of the link function to map the linear prediction to the mean.

The logit link

The logit link is

$$g(\mu) = \log\mu - \log(1 - \mu)$$

and its inverse is

$$\mu = g^{-1}(z) = \frac{1}{1 + e^{-z}}$$

The probit link

The probit link is

$$g(\mu) = \Phi^{-1}(\mu)$$

and its inverse is

$$\mu = g^{-1}(z) = \Phi(z)$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution and $\Phi^{-1}(\cdot)$ is its inverse.

The complementary log-log link

The complementary log-log link is

$$g(\mu) = \log\{-\log(1 - \mu)\}$$

and its inverse is

$$\mu = g^{-1}(z) = 1 - \exp\{-\exp(z)\}$$

The log link

The log link is

$$g(\mu) = \log\mu$$

and its inverse is

$$\mu = g^{-1}(z) = e^z$$

The identity link

The identity link is $g(\mu) = \mu$.

Survival distributions

`gsem` also implements the most commonly used parametric distributions associated with survival analysis. This includes support for time-varying covariates and right-censoring. See [\[ST\] streg](#) for more background information on parametric survival models.

In this section, the observed endogenous variables are survival times. The random variable corresponding to a given response will continue to be denoted by Y , with observed value denoted by y and expected value denoted by μ . The response Y is assumed to be a nonnegative real value. The starting time under observation is denoted by y_0 , where $y_0 \geq 0$. The indicator for failure is denoted by d , where $d = 0$ indicates a right-censored survival time, and $d = 1$ indicates failure at observed time y .

The log likelihood for these models is given by

$$l(y, y_0, d) = d \log f(y) + (1 - d) \log S(y) - \log S(y_0)$$

where $S(\cdot)$ is the survivor function for Y , and $f(\cdot)$ is the corresponding density function. The relationship between the survivor function and the density is

$$f(y) = -\frac{\partial S(y)}{\partial y}$$

In the following, we provide formulas for the survivor functions with respect to a linear prediction, denoted by z .

The `log` link is the only link allowed with the survival distributions. Except for the gamma distribution, the use of the `log` link is merely a nod to the fact that the expected value of the outcome involves exponentiating the linear prediction. A formula for the expected value of Y , as a function of the linear prediction, is provided for each survival distribution.

The exponential distribution

The survivor function for the exponential distribution is

$$S(y) = \exp(-\lambda y)$$

where λ is a function of z and is determined by one of the two parameterizations implemented in `gsem`. The proportional hazards parameterization is

$$\lambda = \exp(z)$$

and the accelerated failure-time parameterization is

$$\lambda = \exp(-z)$$

The expected value of Y is

$$\mu = \frac{1}{\lambda}$$

The Weibull distribution

The survivor function for the Weibull distribution is

$$S(y) = \exp(-\lambda y^s)$$

where s is a scale parameter, and λ is a function of z that is determined by one of the two parameterizations implemented in `gsem`. `gsem` fits s in the log scale. The proportional hazards parameterization is

$$\lambda = \exp(z)$$

with expected value

$$\mu = \Gamma(1 + 1/s) \exp(-z/s)$$

and the accelerated failure-time parameterization is

$$\lambda = \exp(-sz)$$

with expected value

$$\mu = \Gamma(1 + 1/s) \exp(z)$$

The gamma distribution

The survivor function for the gamma distribution is

$$S(y) = \int_y^\infty \frac{1}{\Gamma(s-2) \{e^z s^2\}^{s-2}} t^{s-2-1} \exp\left(-\frac{t}{e^z s^2}\right) \partial t$$

where s is a scale parameter. `gsem` fits s in the log scale.

The expected value for this distribution is

$$\mu = \exp(z)$$

The loglogistic distribution

The survivor function for the loglogistic distribution is

$$S(y) = \left\{ 1 + \exp\left(\frac{\log y - z}{s}\right) \right\}^{-1}$$

where s is a scale parameter. `gsem` fits s in the log scale.

The expected value for this distribution is

$$\mu = \frac{\pi s \exp(z)}{\sin(\pi s)}$$

The lognormal distribution

The survivor function for the lognormal distribution is

$$S(y) = 1 - \Phi\left(\frac{\log y - z}{s}\right)$$

where s is a scale parameter. `gsem` fits s in the log scale.

The expected value for this distribution is

$$\mu = \exp\left(z + \frac{1}{2}s^2\right)$$

Models with continuous latent variables

The likelihood

`gsem` fits generalized linear models with continuous latent variables via maximum likelihood. The likelihood for the specified model is derived under the assumption that each response variable is independent and identically distributed across the estimation sample. The response variables are also assumed to be independent of each other. These assumptions are conditional on the latent variables and the observed exogenous variables.

The likelihood is computed by integrating out the latent variables. Let θ be the vector of model parameters, \mathbf{y} be the vector of observed response variables, \mathbf{x} be the vector of observed exogenous variables, and \mathbf{u} be the $r \times 1$ vector of latent variables. Then, the marginal likelihood looks something like

$$\mathcal{L}(\theta) = \int_{\mathbb{R}^r} f(\mathbf{y}|\mathbf{x}, \mathbf{u}, \theta) \phi(\mathbf{u}|\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u) \partial \mathbf{u}$$

where \mathbb{R} denotes the set of values on the real line, \mathbb{R}^r is the analog in r -dimensional space, θ is a vector of the unique model parameters, $f(\cdot)$ is the conditional probability density function for the observed response variables, $\phi(\cdot)$ is the multivariate normal density for \mathbf{u} , $\boldsymbol{\mu}_u$ is the expected value of \mathbf{u} , and $\boldsymbol{\Sigma}_u$ is the covariance matrix for \mathbf{u} . All auxiliary parameters are fit directly without any further parameterization, so we simply acknowledge that the auxiliary parameters are among the elements of θ .

The \mathbf{y} variables are assumed to be independent, conditionally on \mathbf{x} and \mathbf{u} , so $f(\cdot)$ is the product of the individual conditional densities. One exception to this is when \mathbf{y} contains two or more Gaussian response variables with the identity link, in which case the Gaussian responses are actually modeled using a multivariate normal density to allow for correlated errors and nonrecursive systems among Gaussian responses. This one exception does not change how the integral is numerically evaluated, so we make no effort to represent this distinction in the formulas.

For a single-level model with n response variables, the conditional joint density function for a given observation is

$$f(\mathbf{y}|\mathbf{x}, \mathbf{u}, \theta) = \prod_{i=1}^n f_i(y_i|\mathbf{x}, \mathbf{u}, \theta)$$

For a two-level model, the likelihood is computed at the cluster level, so the conditional density is also a product of the observation-level density contributions within a given cluster,

$$f(\mathbf{y}|\mathbf{x}, \mathbf{u}, \theta) = \prod_{i=1}^n \prod_{j=1}^t f_i(y_{ij}|\mathbf{x}_j, \mathbf{u}, \theta)$$

where t is the number of individuals within the cluster. This extends to more levels by expanding the products down to the observations nested within the hierarchical groups. Because the single-level model is a special case of a two-level model where all the groups have a single observation, we will now use the two-level notation and subscripts.

Except for the ordinal and multinomial families, we use the link function to map the conditional mean

$$\mu_{ij} = E(y_{ij}|\mathbf{x}_j, \mathbf{u})$$

to the linear prediction

$$z_{ij} = \mathbf{x}'_j \boldsymbol{\beta}_i + \mathbf{x}'_j \boldsymbol{\Lambda}_i \mathbf{u}$$

where $\boldsymbol{\beta}_i$ is the vector of the fixed-effect coefficients and $\boldsymbol{\Lambda}_i$ is the matrix of the latent loadings for y_{ij} . For notational convenience, we will overload the definitions of $f(\cdot)$ and $f_i(\cdot)$ so that they are functions of the responses and model parameters through the linear predictions $\mathbf{z}' = (z_1, \dots, z_n)$. Thus $f(\mathbf{y}|\mathbf{x}, \mathbf{u}, \boldsymbol{\theta})$ is equivalently specified as $f(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta})$, and $f_i(y_{ij}|\mathbf{x}_j, \mathbf{u}, \boldsymbol{\theta})$ is equivalently specified as $f_i(y_{ij}, z_{ij}, \boldsymbol{\theta})$. In this new notation, the likelihood for a given cluster is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \int_{\mathbb{R}^r} f(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) \phi(\mathbf{u}|\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u) \partial \mathbf{u} \\ &= \frac{1}{(2\pi)^{r/2} \sqrt{|\boldsymbol{\Sigma}_u|}} \int_{\mathbb{R}^r} \exp \left\{ \log f(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) - \frac{1}{2} (\mathbf{u} - \boldsymbol{\mu}_u)' \boldsymbol{\Sigma}_u^{-1} (\mathbf{u} - \boldsymbol{\mu}_u) \right\} \partial \mathbf{u} \end{aligned} \quad (2)$$

`gsem` allows nonrecursive systems between Gaussian response variables with the identity link, but non-Gaussian responses and Gaussian responses with the log link are not allowed to participate in any nonrecursive systems. This means that if a given response y is specified with a family other than Gaussian or a link other than identity, then y cannot have a path that ultimately leads back to itself. Any response may participate in a recursive system because the participating responses may be treated as exogenous variables when predicting other responses in a recursive system.

The latent vector \mathbf{u} consists of stacked collections of the latent variables from each level. Within each level, the latent endogenous variables $\boldsymbol{\eta}$ are stacked over the latent exogenous variables $\boldsymbol{\xi}$. Within a given level, the latent exogenous variables and latent endogenous errors are assumed independent and multivariate normal,

$$\begin{aligned} \boldsymbol{\xi} &\sim N(\boldsymbol{\kappa}, \boldsymbol{\Phi}) \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, \boldsymbol{\Psi}) \end{aligned}$$

so according to the linear relationship

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}$$

we have that the latent variables are jointly multivariate normal. This linear relationship implies that `gsem` allows latent variables to predict each other but only within level. It also means that `gsem` allows paths from observed variables to latent variables; however, the observed variable must be constant within group if the path is to a group-level latent variable.

For our two-level model, we have

$$\mathbf{u} \sim N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u)$$

where

$$\boldsymbol{\mu}_u = \begin{pmatrix} \boldsymbol{\mu}_\eta \\ \boldsymbol{\kappa} \end{pmatrix} \quad \boldsymbol{\Sigma}_u = \begin{pmatrix} \boldsymbol{\Sigma}_{\eta\eta} & \boldsymbol{\Sigma}_{\eta\xi} \\ \boldsymbol{\Sigma}_{\xi\eta} & \boldsymbol{\Phi} \end{pmatrix}$$

$$\boldsymbol{\mu}_\eta = (\mathbf{I} - \mathbf{B})^{-1} (\boldsymbol{\Gamma}\boldsymbol{\kappa} + \mathbf{A}\mathbf{x})$$

$$\boldsymbol{\Sigma}_{\eta\eta} = (\mathbf{I} - \mathbf{B})^{-1} (\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}) \{(\mathbf{I} - \mathbf{B})^{-1}\}'$$

$$\boldsymbol{\Sigma}_{\eta\xi} = (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Gamma}\boldsymbol{\Phi}$$

The vector θ is therefore the set of unique model parameters taken from the following:

β_i is the vector of fixed-effect coefficients for y_{ij} .

Λ_i is the matrix of latent loadings for y_{ij} .

\mathbf{B} is the matrix of latent endogenous coefficients.

Γ is the matrix of latent exogenous coefficients.

\mathbf{A} is the matrix of latent fixed-effect coefficients.

κ is the vector of latent exogenous means.

Φ is the matrix of latent exogenous variances and covariances.

Ψ is the matrix of latent endogenous error variances and covariances.

Auxiliary parameters are the result from some of the distribution families.

Each level of a multilevel model will have its own set of the following parameters: \mathbf{B} , Γ , \mathbf{A} , κ , Φ , and Ψ . For multilevel models, Σ_u is a block-diagonal matrix with a block for each level.

Gauss–Hermite quadrature

The integral in (2) is generally not tractable, so we must use numerical methods. In the univariate case, the integral of a function multiplied by the kernel of the standard normal distribution can be approximated using Gauss–Hermite quadrature (GHQ). For q -point GHQ, let the abscissa and weight pairs be denoted by (a_k^*, w_k^*) , $k = 1, \dots, q$. The GHQ approximation is then

$$\int_{-\infty}^{\infty} f(x) \exp(-x^2) dx \approx \sum_{k=1}^q w_k^* f(a_k^*)$$

Using the standard normal distribution yields the approximation

$$\int_{-\infty}^{\infty} f(x) \phi(x) dx \approx \sum_{k=1}^q w_k f(a_k)$$

where $a_k = \sqrt{2}a_k^*$ and $w_k = (w_k^*)/\sqrt{\pi}$.

We can use a change-of-variables technique to transform the multivariate integral (2) into a set of nested univariate integrals. Each univariate integral can then be evaluated using GHQ. Let \mathbf{v} be a random vector whose elements are independently standard normal, and let \mathbf{L} be the Cholesky decomposition of Σ_u ; that is, $\Sigma_u = \mathbf{L}\mathbf{L}'$. In the distribution, we have that $\mathbf{u} = \boldsymbol{\mu}_u + \mathbf{L}\mathbf{v}$, and the linear predictions vector as a function of \mathbf{v} is

$$z_{ij} = \mathbf{x}'_j \beta_i + \mathbf{x}'_j \Lambda_i (\boldsymbol{\mu}_u + \mathbf{L}\mathbf{v})$$

so the likelihood for a given cluster is

$$\mathcal{L}(\theta) = (2\pi)^{-r/2} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \exp \left\{ \log f(\mathbf{y}, \mathbf{z}, \theta) - \frac{1}{2} \sum_{k=1}^r v_k^2 \right\} dv_1 \dots dv_r \quad (3)$$

where r is the number of latent variables.

Consider an r -dimensional quadrature grid containing q quadrature points in each dimension. Let the vector of abscissas $\mathbf{a}_k = (a_{k_1}, \dots, a_{k_r})'$ be a point in this grid, and let $\mathbf{w}_k = (w_{k_1}, \dots, w_{k_r})'$ be the vector of corresponding weights. The GHQ approximation to the likelihood for a given cluster is

$$\mathcal{L}^{\text{GHQ}}(\boldsymbol{\theta}) = \sum_{k_1=1}^q \dots \sum_{k_r=1}^q \left[\exp \left\{ \sum_{i=1}^n \log f_i(y_{ij}, z_{ijk}, \boldsymbol{\theta}) \right\} \prod_{s=1}^r w_{k_s} \right]$$

where

$$z_{ijk} = \mathbf{x}'_j \boldsymbol{\beta} + \mathbf{x}'_j \boldsymbol{\Lambda}_i (\boldsymbol{\mu}_u + \mathbf{L} \boldsymbol{\alpha}_k)$$

Adaptive quadrature

This section sets the stage for mean–variance adaptive Gauss–Hermite quadrature (MVAGH) and mode-curvature adaptive Gauss–Hermite quadrature (MCAGH).

Let's reconsider the likelihood in (3). If we fix the observed variables and the model parameters, we see that the posterior density for \mathbf{v} is proportional to

$$\phi(\mathbf{v}) f(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta})$$

It is reasonable to assume that this posterior density can be approximated by a multivariate normal density with mean vector $\boldsymbol{\mu}_v$ and variance matrix $\boldsymbol{\tau}_v$. Instead of using the prior density of \mathbf{v} as the weighting distribution in the integral, we can use our approximation for the posterior density,

$$\mathcal{L}(\boldsymbol{\theta}) = \int_{\mathbb{R}^r} \frac{f(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) \phi(\mathbf{v})}{\phi(\mathbf{v}, \boldsymbol{\mu}_v, \boldsymbol{\tau}_v)} \phi(\mathbf{v}, \boldsymbol{\mu}_v, \boldsymbol{\tau}_v) d\mathbf{v}$$

The likelihood is then approximated with

$$\mathcal{L}^*(\boldsymbol{\theta}) = \sum_{k_1=1}^q \dots \sum_{k_r=1}^q \left[\exp \left\{ \sum_{i=1}^n \log f_i(y_{ij}, z_{ijk}^*, \boldsymbol{\theta}) \right\} \prod_{s=1}^r \omega_{k_s} \right]$$

where

$$z_{ijk}^* = \mathbf{x}'_j \boldsymbol{\beta} + \mathbf{x}'_j \boldsymbol{\Lambda}_i (\boldsymbol{\mu}_u + \mathbf{L} \boldsymbol{\alpha}_k)$$

and $\boldsymbol{\alpha}_k$ and the ω_{k_s} are the adaptive versions of the abscissas and weights after an orthogonalizing transformation, which eliminates posterior covariances between the latent variables. $\boldsymbol{\alpha}_k$ and the ω_{k_s} are functions of \mathbf{a}_k and \mathbf{w}_k and the adaptive parameters $\boldsymbol{\mu}_v$ and $\boldsymbol{\tau}_v$.

For MVAGH, $\boldsymbol{\mu}_v$ is the posterior mean and $\boldsymbol{\tau}_v$ is the posterior variance of \mathbf{v} . They are computed iteratively by updating the posterior moments by using the MVAGH approximation, starting with a 0 mean vector and identity variance matrix.

For MCAGH, $\boldsymbol{\mu}_v$ is the posterior mode for \mathbf{v} , and $\boldsymbol{\tau}_v$ is the curvature at the mode. They are computed by optimizing the integrand in (3) with respect to \mathbf{v} .

Laplacian approximation

Let's reconsider the likelihood in (2) and denote the argument in the exponential function by

$$\begin{aligned} h(\mathbf{u}) &= \log f(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) - \frac{1}{2}(\mathbf{u} - \boldsymbol{\mu}_u)' \boldsymbol{\Sigma}_u^{-1}(\mathbf{u} - \boldsymbol{\mu}_u) \\ &= \sum_{i=1}^n \sum_{j=1}^t \log f_i(y_{ij}, z_{ij}, \boldsymbol{\theta}) - \frac{1}{2}(\mathbf{u} - \boldsymbol{\mu}_u)' \boldsymbol{\Sigma}_u^{-1}(\mathbf{u} - \boldsymbol{\mu}_u) \end{aligned}$$

where

$$z_{ij} = \mathbf{x}'_j \boldsymbol{\beta}_i + \mathbf{x}'_j \boldsymbol{\Lambda}_i \mathbf{u}$$

The Laplacian approximation is based on a second-order Taylor expansion of $h(\mathbf{u})$ about the value of \mathbf{u} that maximizes it. The first and second partials with respect to \mathbf{u} are

$$\begin{aligned} h'(\mathbf{u}) &= \frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} = \sum_{i=1}^n \sum_{j=1}^t \frac{\partial \log f_i(y_{ij}, z_{ij}, \boldsymbol{\theta})}{\partial z_{ij}} \boldsymbol{\Lambda}'_i \mathbf{x}_j - \boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu}_u) \\ \mathbf{H}(\mathbf{u}) &= \frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'} = \sum_{i=1}^n \sum_{j=1}^t \mathbf{x}'_j \boldsymbol{\Lambda}_i \frac{\partial^2 \log f_i(y_{ij}, z_{ij}, \boldsymbol{\theta})}{\partial z_{ij} \partial z_{ij}} \boldsymbol{\Lambda}'_i \mathbf{x}_j - \boldsymbol{\Sigma}^{-1} \end{aligned}$$

The maximizer of $h(\mathbf{u})$ is $\hat{\mathbf{u}}$ such that $h'(\hat{\mathbf{u}}) = \mathbf{0}$. The integral in (2) is proportional to the posterior density of \mathbf{u} given the data, so $\hat{\mathbf{u}}$ is also the posterior mode.

The second-order Taylor approximation then takes the form

$$h(\mathbf{u}) \approx h(\hat{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \hat{\mathbf{u}})' \mathbf{H}(\hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}}) \quad (4)$$

because the first-order derivative term is 0. The integral is approximated by

$$\int_{\mathfrak{R}^r} \exp\{h(\mathbf{u})\} d\mathbf{u} \approx \exp\{h(\hat{\mathbf{u}})\} (2\pi)^{r/2} |\mathbf{H}(\hat{\mathbf{u}})|^{-1/2}$$

because the second term in (4) is the kernel of a multivariate normal density once it is exponentiated. The Laplacian approximation for the log likelihood is

$$\log \mathcal{L}^{\text{Lap}}(\boldsymbol{\theta}) = -\frac{1}{2} \log |\boldsymbol{\Sigma}_u| - \frac{1}{2} \log |\mathbf{H}(\hat{\mathbf{u}})| + h(\hat{\mathbf{u}})$$

Survey data

In the presence of sampling weights, following [Rabe-Hesketh and Skrondal \(2006\)](#), the weighted log pseudolikelihood for a two-level model is given as

$$\log \mathcal{L}(\boldsymbol{\theta}) = \sum_{j=1}^M w_j \log \int_{-\infty}^{\infty} \exp \left\{ \sum_{i=1}^{n_j} w_{i|j} \log f(y_{ij} | \eta_{ij}) \right\} \phi(\mathbf{v}_{j1}) d\mathbf{v}_{j1}$$

where w_j is the inverse of the probability of selection for the j th cluster; $w_{i|j}$ is the inverse of the conditional probability of selection of individual i , given the selection of cluster j ; $f(\cdot)$ is as defined previously; and $\phi(\cdot)$ is the standard multivariate normal density.

Weighted estimation is achieved through the direct application of w_j and $w_{i|j}$ into the likelihood calculations as detailed above to reflect replicated clusters for w_j and replicated observations within clusters for $w_{i|j}$. Because this estimation is based on replicated clusters and observations, frequency weights are handled similarly.

Weights are not allowed with crossed models or the Laplacian approximation.

Predictions

We begin by considering the prediction of the latent variables \mathbf{u} for a given cluster in a two-level model. Prediction of latent variables in multilevel generalized linear models involves assigning values to the latent variables, and there are many methods for doing so; see [Skrondal and Rabe-Hesketh \(2009\)](#) and [Skrondal and Rabe-Hesketh \(2004, chap. 7\)](#) for a comprehensive review. Stata offers two methods of predicting latent variables: empirical Bayes means (also known as posterior means) and empirical Bayes modes (also known as posterior modes). Below we provide more details about the two methods.

Empirical Bayes

Let $\hat{\boldsymbol{\theta}}$ denote the estimated model parameters. Empirical Bayes (EB) predictors of the latent variables are the means or modes of the empirical posterior distribution with the parameter estimates $\boldsymbol{\theta}$ replaced with their estimates $\hat{\boldsymbol{\theta}}$. The method is called “empirical” because $\hat{\boldsymbol{\theta}}$ is treated as known. EB combines the prior information about the latent variables with the likelihood to obtain the conditional posterior distribution of latent variables. Using Bayes’s theorem, the empirical conditional posterior distribution of the latent variables for a given cluster is

$$\begin{aligned} \omega(\mathbf{u} | \mathbf{y}, \mathbf{x}; \hat{\boldsymbol{\theta}}) &= \frac{f(\mathbf{y} | \mathbf{u}, \mathbf{x}; \hat{\boldsymbol{\theta}}) \phi(\mathbf{u}; \hat{\boldsymbol{\mu}}_{\mathbf{u}}, \hat{\boldsymbol{\Sigma}}_{\mathbf{u}})}{\int f(\mathbf{y} | \mathbf{u}, \mathbf{x}; \hat{\boldsymbol{\theta}}) \phi(\mathbf{u}; \hat{\boldsymbol{\mu}}_{\mathbf{u}}, \hat{\boldsymbol{\Sigma}}_{\mathbf{u}}) d\mathbf{u}} \\ &= \frac{f(\mathbf{y} | \mathbf{u}, \mathbf{x}; \hat{\boldsymbol{\theta}}) \phi(\mathbf{u}; \hat{\boldsymbol{\mu}}_{\mathbf{u}}, \hat{\boldsymbol{\Sigma}}_{\mathbf{u}})}{\mathcal{L}(\hat{\boldsymbol{\theta}})} \end{aligned}$$

The denominator is just the likelihood contribution of the given cluster.

EB mean predictions of latent variables, $\tilde{\mathbf{u}}$, also known as posterior means, are calculated as

$$\tilde{\mathbf{u}} = \int_{\mathbb{R}^r} \mathbf{u} \omega(\mathbf{u} | \mathbf{y}, \mathbf{x}; \hat{\boldsymbol{\theta}}) d\mathbf{u}$$

where we use the notation $\tilde{\mathbf{u}}$ rather than $\hat{\mathbf{u}}$ to distinguish predicted values from estimates. This multivariate integral is approximated by MVAGH. If you have multiple latent variables within a level or latent variables across levels, the calculation involves orthogonalizing transformations with the Cholesky transformation because the latent variables are no longer independent under the posterior distribution.

When all the response variables are normal, the posterior density is multivariate normal, and EB means are also best linear unbiased predictors (BLUPs); see [Skrondal and Rabe-Hesketh \(2004, 227\)](#). In generalized mixed-effects models, the posterior density tends to multivariate normal as cluster size increases.

EB modal predictions can be approximated by solving for $\tilde{\mathbf{u}}$ such that

$$\frac{\partial}{\partial \mathbf{u}} \log \omega(\mathbf{u} | \mathbf{y}, \mathbf{x}; \hat{\boldsymbol{\theta}}) \Big|_{\mathbf{u}=\tilde{\mathbf{u}}} = \mathbf{0}$$

Because the denominator in $\omega(\cdot)$ does not depend on \mathbf{u} , we can omit it from the calculation to obtain the EB mode. The calculation of EB modes does not require numerical integration; thus they are often used in place of EB means. As the posterior density gets closer to being multivariate normal, EB modes get closer and closer to EB means.

Just like there are many methods of assigning values to the random effects, there are many methods of calculating standard errors of the predicted random effects; see [Skrondal and Rabe-Hesketh \(2009\)](#) for a comprehensive review.

Stata uses the posterior standard deviation as the standard error of the posterior means predictor of random effects. For a given level, the EB posterior covariance matrix of the random effects is given by

$$\text{Cov}(\tilde{\mathbf{u}} | \mathbf{y}, \mathbf{x}; \hat{\boldsymbol{\theta}}) = \int_{\mathfrak{R}^r} (\mathbf{u} - \tilde{\mathbf{u}})(\mathbf{u} - \tilde{\mathbf{u}})' \omega(\mathbf{u} | \mathbf{y}, \mathbf{x}; \hat{\boldsymbol{\theta}}) d\mathbf{u}$$

The posterior covariance matrix and the integrals are approximated by MVAGH.

Conditional standard errors for the estimated posterior modes are derived from standard theory of maximum likelihood, which dictates that the asymptotic variance matrix of $\tilde{\mathbf{u}}$ is the negative inverse of the Hessian matrix.

Other predictions

In what follows, we show formulas with the posterior means estimates of random effects $\tilde{\mathbf{u}}$, which are used by default or if the `conditional(ebmeans)` option is specified. If the `conditional(ebmodes)` option is specified, $\tilde{\mathbf{u}}$ are simply replaced with the posterior modes $\hat{\mathbf{u}}$ in these formulas.

For the i th response in the j th observation within a given cluster in a two-level model, the linear predictor is computed as

$$\hat{z}_{ij} = \mathbf{x}'_j \hat{\boldsymbol{\beta}} + \mathbf{x}'_j \hat{\boldsymbol{\Lambda}}_i \tilde{\mathbf{u}}$$

The linear predictor includes the offset or exposure variable if one was specified during estimation, unless the `nooffset` option is specified. If the `fixedonly` option is specified, the linear predictor is computed as

$$\hat{z}_{ij} = \mathbf{x}'_j \hat{\boldsymbol{\beta}}$$

The predicted mean, conditional on the predicted latent variables, is

$$\hat{\mu}_{ij} = g^{-1}(\hat{z}_{ij})$$

where $g^{-1}(\cdot)$ is the inverse link function defined in *Link functions* above. For the ordinal and multinomial families, the predicted mean is actually a probability, and `gsem` can produce a probability for each outcome value as described in *The ordinal family* and *The multinomial family* above. For the survival distributions, the predicted mean is computed using the formulas provided in *Survival distributions* above.

Models with categorical latent variables

The likelihood

`gsem` fits generalized linear models with categorical latent variables via maximum likelihood. The likelihood for the specified model is derived under the assumption that each response variable is independent and identically distributed across the estimation sample. The response variables are also assumed to be independent of each other. These assumptions are conditional on the categorical latent variables and the observed exogenous variables.

The likelihood is computed by combining the conditional likelihoods from each latent class (level combinations of the categorical latent variables) weighted by the associated latent class probabilities. Let θ be the vector of model parameters. For a given observation, let \mathbf{y} be the vector of observed response variables and \mathbf{x} be the vector of observed exogenous variables. Without loss of generality, we will assume a single categorical latent variable C with k levels $1, \dots, k$. The marginal likelihood for a given observation looks something like

$$\mathcal{L}_C(\theta) = \sum_{i=1}^k \pi_i f_i(\mathbf{y}|\mathbf{x}, c_i = 1, \theta)$$

where π_i is the probability for the i th latent class, $f_i(\cdot)$ is the conditional probability density function for the observed response variables in the i th latent class, and $\mathbf{c}' = (c_1, \dots, c_k)$ is the vector of latent class indicators. When $c_i = 1$, all other elements of \mathbf{c} are 0. All auxiliary parameters are fit directly without any further parameterization, so we simply acknowledge that the auxiliary parameters are among the elements of θ .

The \mathbf{y} variables are assumed to be independent, conditionally on \mathbf{x} and C , so $f_i(\cdot)$ is the product of the individual conditional densities. One exception to this is when \mathbf{y} contains two or more Gaussian response variables with the identity link, in which case the Gaussian responses are actually modeled using a multivariate normal density to allow for correlated errors and nonrecursive systems among Gaussian responses. This one exception does not meaningfully change the following discussion, so we make no effort to represent this distinction in the formulas.

For the i th latent class with n response variables, the conditional joint density function for a given observation is

$$f_i(\mathbf{y}|\mathbf{x}, \theta) = \prod_{j=1}^n f_{ij}(y_{ij}|\mathbf{x}, \theta)$$

Except for the [ordinal](#) and [multinomial](#) families, we use the link function to map the conditional mean

$$\mu_{ij} = E(y_{ij} | \mathbf{x}, c_i = 1)$$

to the linear prediction

$$z_{ij} = \mathbf{x}'\boldsymbol{\beta}_{ij}$$

where $\boldsymbol{\beta}_{ij}$ is the vector of the coefficients for y_{ij} . For notational convenience, we will overload the definitions of $f_i(\cdot)$ and $f_{ij}(\cdot)$ so that they are functions of the responses and model parameters through the linear predictions $\mathbf{z}'_i = (z_{i1}, \dots, z_{in})$. Thus $f_i(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$ is equivalently specified as $f_i(\mathbf{y}, \mathbf{z}_i, \boldsymbol{\theta})$, and $f_{ij}(y_{ij} | \mathbf{x}, \boldsymbol{\theta})$ is equivalently specified as $f_{ij}(y_{ij}, z_{ij}, \boldsymbol{\theta})$. In this new notation, the likelihood for a given observation is

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^k \pi_i \prod_{j=1}^n f_{ij}(y_{ij}, z_{ij}, \boldsymbol{\theta}) \tag{1}$$

`gsem` uses the multinomial distribution to model the probabilities for the latent classes. For the i th latent class, the probability is given by

$$\pi_i = \Pr(c_i = 1 | \mathbf{x}) = \frac{\exp(z_i)}{\sum_a \exp(z_a)}$$

where the linear prediction for the i th latent class is

$$z_i = \mathbf{x}'\boldsymbol{\gamma}_i$$

and $\boldsymbol{\gamma}_i$ is the associated vector of coefficients. `gsem` assumes the first latent class is the base level; $\boldsymbol{\gamma}_1$ is a vector of zeros so that $z_1 = 0$ and $\exp(z_1) = 1$.

The vector $\boldsymbol{\theta}$ is therefore the set of unique model parameters taken from the following:

$\boldsymbol{\gamma}_i$ is the vector of coefficients for the i th latent class.

$\boldsymbol{\beta}_{ij}$ is the vector of coefficients for y_{ij} .

Auxiliary parameters that result from some of the distribution families.

Each latent class will have its own set of these parameters.

The EM algorithm

`gsem` uses the EM algorithm to refine starting values before maximizing the likelihood in (1).

The EM algorithm uses the complete-data likelihood, a likelihood where it is as if we have observed values for the latent class indicator variables \mathbf{c} . In the complete-data case, the likelihood for a given observation is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^k \{\pi_i f_i(\mathbf{y}, \mathbf{z}_i, \boldsymbol{\theta})\}^{c_i}$$

so the complete-data log likelihood is

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^k c_i \{ \log \pi_i + \log f_i(\mathbf{y}, \mathbf{z}_i, \boldsymbol{\theta}) \}$$

We intend to maximize the expected complete-data log likelihood given the observed variables \mathbf{y} and \mathbf{x} . This is an iterative process in which we use the g th guess of the model parameters, denoted $\boldsymbol{\theta}_{(g)}$, then compute the next guess $\boldsymbol{\theta}_{(g+1)}$.

In the expectation (E) step, we derive the functional form of the expected complete-data log likelihood. The complete-data log likelihood is a linear function of the latent class indicator variables, so

$$E(c_i | \mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_{(g)}) = \frac{\pi_i f_i(\mathbf{y}, \mathbf{z}_i, \boldsymbol{\theta}_{(g)})}{\sum_{j=1}^k \pi_j f_j(\mathbf{y}, \mathbf{z}_j, \boldsymbol{\theta}_{(g)})}$$

We denote this posterior probability by p_i , so the expected complete-data log likelihood for a given observation is given by

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}_{(g)}) = \sum_{i=1}^k p_i \{ \log \pi_i + \log f_i(\mathbf{y}, \mathbf{z}_i, \boldsymbol{\theta}) \}$$

Note that $Q(\boldsymbol{\theta} | \boldsymbol{\theta}_{(g)})$ is a function of $\boldsymbol{\theta}_{(g)}$ solely through the posterior probabilities p_i .

Now that we have the conditional complete-data log likelihood, the maximization (M) step is to maximize $Q(\boldsymbol{\theta} | \boldsymbol{\theta}_{(g)})$ with respect to $\boldsymbol{\theta}$ to find $\boldsymbol{\theta}_{(g+1)}$.

Survey data

`gsem` with categorical latent variables also supports estimation with survey data; however, only the linearized variance estimator is supported. For details on VCEs with survey data, see [SVY] [variance estimation](#).

Predictions

The predicted mean for a given response within a latent class is computed by applying the associated link function to the linear prediction; see [Link functions](#) above. For ordinal and multinomial responses, the predicted mean for a given response level is the predicted probability for that level. For survival outcomes, the formulas for predicted means are provided in [Survival distributions](#) above.

Let \widehat{z}_i be the linear prediction for the i th latent class; then, the predicted probability for the i th latent class is given by

$$\widehat{\pi}_i = \frac{\exp(\widehat{z}_i)}{\sum_{a=1}^k \exp(\widehat{z}_a)}$$

The predicted posterior probability for the i th latent class is given by

$$\tilde{\pi}_i = \frac{\hat{\pi}_i f_i(\mathbf{y}, \hat{\mathbf{z}}_i, \hat{\boldsymbol{\theta}})}{\sum_{j=1}^k \hat{\pi}_j f_j(\mathbf{y}, \hat{\mathbf{z}}_j, \hat{\boldsymbol{\theta}})}$$

Let $\hat{\mu}_i$ be the predicted mean of response y in the i th latent class. The predicted overall mean of y , using the fitted latent class probabilities, is given by

$$\hat{\mu} = \sum_{i=1}^k \hat{\pi}_i \hat{\mu}_i$$

The predicted overall mean of y , using the posterior latent class probabilities, is given by

$$\tilde{\mu} = \sum_{i=1}^k \tilde{\pi}_i \hat{\mu}_i$$

References

- Goodman, L. A. 2002. Latent class analysis: The empirical study of latent types, latent variables, and latent structures. In *Applied Latent Class Analysis*, ed. J. A. Hagenaars and A. L. McCutcheon, 3–55. Cambridge: Cambridge University Press.
- Masyn, K. E. 2013. Latent class analysis and finite mixture modeling. In *The Oxford Handbook of Quantitative Methods*, ed. T. D. Little, vol. 2, 551–610. New York: Oxford University Press.
- Rabe-Hesketh, S., and A. Skrondal. 2006. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A* 169: 805–827.
- Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2005. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* 128: 301–323.
- Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.
- . 2009. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society, Series A* 172: 659–687.

Also see

[SEM] [gsem](#) — Generalized structural equation model estimation command