

Outline

Background

Definition

The gmss command

GMSS Library

Link functions

General distributions

Distributions for count outcomes

Distributions for binomial outcomes

Usage

Structures

predict and estat

Murphy–Topel example of development

BΣD

Version 0.8.0

GMSS is a (Stata version 19.0) library that facilitates estimation of individual-level regression models with support for user-specified link functions and standardized post-estimation facilities.

The project consists of a mata library and ado-file commands which access the library.

Class of Models

The library focuses on models for which the log-likelihood is given by $\ln L(Y|\theta_1, \dots, \theta_E) = \mathcal{L}$ where $\theta_i = \mathbf{x}_i \mathbf{b}_i$. Further, we have that $\mu_i = g(\theta_i)$ associating the linear predictor with the (possibly) range-restricted distribution parameter via a link function $g()$.

Parameters are organized such that the initial parameter represents a measure of central tendency. Subsequent parameters are associated with dispersion, scale, and/or shape.

Mata classes

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_i} = \mathbf{x}_i^T \left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_i} \right) \left\{ \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_i} \right\}$$

The basic idea of the library is that model calculations are carried out via the chain rule for which distributions handle the calculations in parentheses and link functions handle calculations in braces.

Methods

Distributions provide the following methods: `loglik()`,
`dldtheta(u)`, `d2ldtheta2(u, v)`.

Links provide the following methods: `g()`, `ginv()`, `dtheta()`,
and `d2theta()`.

Specifying a Regression Model

<code>gmss opt</code>	is used to specify an opt object containing options for Stata's optimizer.
<code>gmss link</code>	is used to specify a link object which associates variables with parameters.
<code>gmss dist</code>	is used to specify a dist object which associates a GMSS distribution with link objects in the order given in the help file (the mean parameter is always first).
<code>gmss init</code>	is used to specify a model object which associates a dist object with an opt object.
<code>gmss run</code>	is used to estimate a model.

Specifying Optimization Options

Options for Stata's optimizer are specified using the `gmss opt` statement. In addition to tolerances and iteration counts, you may specify

<code>evaluator(evalname)</code>	which of <code>d0</code> , <code>d1</code> , <code>d2</code> , <code>d1debug</code> , <code>d2debug</code> , <code>lf</code> , <code>lf0</code> , <code>lf1</code> , <code>lf2</code> , <code>lf1debug</code> , or <code>lf2debug</code> should be used as the evaluator (default: <code>d2</code>)
<code>technique(techstring)</code>	which value or string of values of <code>nr</code> , <code>dfp</code> , <code>bfgs</code> , <code>bhhh</code> , <code>nm</code> , or <code>gn</code> should be used (default: <code>nr</code>)
<code>tracelevel(levelname)</code>	which of <code>none</code> , <code>value</code> , <code>tolerance</code> , <code>step</code> , <code>coefdiffs</code> , <code>paramdifs</code> , <code>coefs</code> , <code>params</code> , <code>gradient</code> , or <code>hessian</code> to use (default: <code>value</code>)

Specifying Link Options

Options for the association of variables with parameters are specified using the `gmss link` statement.

<code>name(linkname)</code>	name of the GMSS link; default links are available
<code>label(label)</code>	label to use; the default is to use linkname and adding an underscore followed by mu, sigma, tau, delta, parm5, parm6, ...
<code>y(varname)</code>	name of the dependent variable (typically only required for the mean parameter)
<code>x(varlist)</code>	independent variable list for the parameter
<code>n(varname)</code>	variable containing the binomial information
<code>cons(off on)</code>	whether to include a constant (default: on)

Specifying Distribution Options

Options for distributions are specified using the `gmss dist` statement.

<code>name(distname)</code>	name of the GMSS distribution (required)
<code>link(gmssnames)</code>	gmssname(s) of the associated links for the distribution's parameters (required)

Mata Library

GMSS includes a mata library of distributions, links, and functions with which one can estimate regression models. The `gmss` command is an ado-language front-end allowing users to specify information associated with a model. For example,

```
gmss link mean, y(price) x(foreign displ) cons(on)
gmss link sdev, cons(on)
gmss dist gauss, name(normal) link(mean sdev)
gmss opt d2, evaluator(d2) search(on) random(on)
gmss init regmodel, dist(gauss) opt(d2)
gmss run regmodel, nolog noheader
```

GMSS Links

The allowable linknames for `gms` include

<code>atanh</code>	<code>logcomplement</code>	<code>nbinoomial</code>
<code>cauchit</code>	<code>logcorr</code>	<code>power2</code>
<code>cloglog</code>	<code>logm1</code>	<code>powerm2</code>
<code>halflogit</code>	<code>logm2</code>	<code>probit</code>
<code>identity</code>	<code>logit</code>	<code>reciprocal</code>
<code>log</code>	<code>loglog</code>	

We emphasize that each of the GMSS distributions define default linknames, so specifying a particular link is optional.

GMSS Distributions

Bernoulli	Inv. normal
Beta	Logistic
Exponential	Normal
Gamma	Ordered
Inv. gamma	Ordered (heteroskedastic)

We also developed count and binomial distributions.

Modified Count Distributions

We developed helper distributions for zero-truncated, zero-altered, zero-inflated, zero-inflated(τ), zero-inflated marginalized, and heaped models.

$$\mathcal{L}_H = \sum_{b=1}^{H+1} P_B(B=b) \exp\{\mathcal{L}_C(y/k_b, \mu_1/k_b, \mu_2 \dots)\}$$

$$\mathcal{L}_{ZA} = \log[\mu_0] I(y=0) + [\mathcal{L}_C - \log(1 - P_0) + \log(1 - \mu_0)] I(y > 0)$$

$$\mathcal{L}_{ZI} = \log[\mu_0 + (1 - \mu_0)P_0] I(y=0) + [\mathcal{L}_C + \log(1 - \mu_0)] I(y > 0)$$

$$\mathcal{L}_{ZI(\tau)} = \log[\mu_0^* + (1 - \mu_0^*)P_0] I(y=0) + [\mathcal{L}_C + \log(1 - \mu_0^*)] I(y > 0)$$

$$\mathcal{L}_{ZM} = \log[\mu_0 + (1 - \mu_0)P_0] I(y=0) + [\mathcal{L}_{CM} + \log(1 - \mu_0)] I(y > 0)$$

$$\mathcal{L}_{ZT} = \mathcal{L}_C - \log(1 - P_0)$$

BΣD

GMSS Count Distributions

Bern. Poisson–Lindley	Gen. Poisson–Lindley 3	Poisson–Lindley
Beta neg. binomial	Gen. Poisson 1	Poisson–Loai
Borel	Gen. Poisson 2	Poisson–Mirra
Delaporte	Gen. Waring	Poisson
Double Poisson	Neg. Binomial 1	Poisson Trans. Exp.
Flory–Schulz	Neg. Binomial 2	Poisson Wt. Exp.
Gen. Binomial	Neg. Binomial P	Poisson Xgamma
Geometric	Poisson inv. Gaussian	Unif. Poisson–Ailamujia
Gen. Neg. Binomial 2	Planck	Waring
Gen. Poisson–Lindley 2	Poisson–Ailamujia	Yule

Each count distribution can be used in models: as-is,
zero-inflated, zero-inflated(τ), zero-inflated marginalized,
zero-altered, zero-truncated, and heaped.

Modified Binomial Distributions

We developed helper distributions for n -truncated, n -altered, n -inflated, n -inflated(τ), n -inflated marginalized, all of the helper distributions for count distributions, and the combination of zero and n .

$$\mathcal{L}_{NA} = \log[\mu_n]I(y = n) + [\mathcal{L}_C - \log(1 - P_n) + \log(1 - \mu_n)]I(y < n)$$

$$\mathcal{L}_{NI} = \log[\mu_n + (1 - \mu_n)P_n]I(y = n) + [\mathcal{L}_C + \log(1 - \mu_n)]I(y < n)$$

$$\mathcal{L}_{NI(\tau)} = \log[\mu_n^* + (1 - \mu_n^*)P_0]I(y = n) + [\mathcal{L}_C + \log(1 - \mu_n^*)]I(y < n)$$

$$\mathcal{L}_{NM} = \log[\mu_n + (1 - \mu_n)P_n]I(y = n) + [\mathcal{L}_{CM} + \log(1 - \mu_n)]I(y < n)$$

$$\mathcal{L}_{NT} = \mathcal{L}_C - \log(1 - P_n)$$

GMSS Binomial Distributions

Beta-binomial
Binomial

Double binomial
Gen. binomial

Each binomial distribution can be used in models: as-is, zero-inflated, zero-inflated(τ), zero-inflated marginalized, zero-altered, zero-truncated, heaped, n -inflated, n -inflated(τ), n -altered, n -truncated, zero/ n -inflated, zero/ n -inflated(τ), zero/ n -marginal, and zero/ n -truncated.

The GMSS struct

GMSS ultimately defines a named Mata structure that is left in the global space. This structure includes the Mata optimization structure necessary for model estimation as well as the distribution and link objects instantiated by the library. Users can directly utilize these objects (which `gmss` leaves behind in the `e()` space). For example,

```
mata: moptimize_result_coefs('e(moptname)')
```

See Stata's documentation for `moptimize` for usage.

Post-estimation

Post-estimation prediction access the GMSS struct (which contains the distribution and link objects) allowing the user to get any of the derivatives or scores. The `gmss(string)` allows specification of the GMSS model object meaning that predictions can be done for models that are not currently posted.

```
predict newvarname [if] [in] , [ xb mu dmu d2mu  
  dldeta dldmu d2ldeta2 d2ldmu2 p0 dp0dtheta dp0dmu  
  d2p0dtheta2 d2p0dmu2 score Residuals EQnum(string)  
  gmss(string) ]
```

User Development

Because `gmss` leaves behind the Mata structures, the `predict` command returns a rich collection of model-based calculations. That in turn allows further development of models.

For example, we can revisit one of our first Stata Journal papers in which the topic of two-stage models was addressed.

$$V_{\text{MT}} = V_2 + V_2(CV_1C' - RV_1C' - CV_1R')V_2$$

$$R = \left(\frac{\partial \mathcal{L}_2}{\partial b_2} \right) \left(\frac{\partial \mathcal{L}_2}{\partial b_2} \right)' \quad C = \left(\frac{\partial \mathcal{L}_2}{\partial b_2} \right) \left(\frac{\partial \mathcal{L}_2}{\partial b_1} \right)'$$

Because `predict` will return each of the derivatives, and each model will save the model structure, we can develop a general calculation of this matrix. To highlight such development, we created the `gmss2stage` command.

```
gmss link lacc, y(z) x(age income ownrent selfemp)
gmss dist logitacc, name(bernoulli) link(lacc)
gmss init stage1, dist(logitacc) opt(d2)
gmss run stage1
predict double muhat
```

To complete the estimation, we initialize the stage 2 model, and then use the `gmss2stage` command to run it and post the final result.

```
gmss link lreps, y(y) x(muhat age income expend)
gmss dist poiss, name(poisson) link(lreps)
gmss init stage2
gmss2stage stage1 stage2, zhat(muhat)
```

Summary

- ▶ Users can utilize the `gmss` and `gmss2stage` commands to estimate many new models.
- ▶ Users can take advantage of enhanced access to Stata's optimization features.
- ▶ There are many enhancements planned for the library (generalization of the definition of the linear predictor, allowing latent and additive terms, mixed effects models, etc).

Where to Look for Materials and More Information

A manuscript has been submitted to the Stata Journal which describes the material presented here.

A YouTube channel for video tutorials:

<https://www.youtube.com/@biostatdoc>

A github repository is available for the library plus example files and further documentation:

<https://www.github.com/jwhardin/BiostatDoc>