GGT Documentation

Syntax

```
GGT, outcomevar(varname) orgchoice(varname) indID(varname) orgID(varname)
choicechar(varlist) [options]
```

Description

This program estimates the parameters of the Geweke, Gowrisankaran, and Town (2003) "GGT" model. The GGT model estimates the posterior distribution of organizational performance where there are many organizations from which individuals can choose to receive services. In this framework, individuals may select organizations based, in part, on information that is unobserved to the researcher and is correlated with the binary outcome. If this is the case, then standard approaches to inferring organization performance will yield biased estimates. The GGT model corrects for this unobserved selection allowing for flexible correlation in the error structure across the organizational choice and outcome equations. The estimation approach is Bayesian. **In sum, the model combines an organization choice multinomial probit model with an individual outcome binary probit model, allowing for correlation across equations for each individual[1].**

The parameters are estimated using Bayesian inference through Markov chain Monte Carlo techniques to simulate parameters and latent variables conditional on data to determine the posterior distribution of parameters. While we present the basics of the model below, we encourage all users of this Stata function to read the GGT paper to fully understand the model, assumptions underneath the model, and parameters used in the estimation.

Required Files

To speed up the computing process, the program code calls an included C plugin file. Thus, in addition to the Stata .ado files, the user additionally needs a .plugin file. The files necessary for the code to run are the following:

- GGT.ado
- callCcode.ado
- bayesqual12.plugin

Methods and Equations

The model presented below comes from the GGT application in which the authors estimate hospital quality measures. The authors assume that quality depends on patient mortality, a standard assumption for hospital quality calculations. However, GGT make the important note that patients may "select" into which hospital to attend, which would bias the hospital quality measures if not accurately controlled for. Thus, GGT define a model to allow for unobserved patient characteristics which are correlated with hospital choice and patient mortality.

We include a brief explanation of the model here to show which variables and parameters are referenced in the calling of the `GGT` Stata function.

The binary individual outcome equation: $m_i^* = c_i'\beta + x_i'\gamma + \varepsilon_i$ (equation (1) in GGT)

Here, m_i^* is the latent outcome variable for the observed binary variable m_i. This is patient mortality in the GGT application. The outcome variable depends on individual characteristics, x_i, and which organization the individual chooses, c_i (hospital choice in GGT).

The organization choice model is: $c_i^* = Z_i\alpha + \eta_i$ (equation (3) in GGT)

Here, c_i^* is the latent choice vector for the observed choice vector c_i where $c_{ij}=1$ if patient i chose organization j and 0 for all other vector entries. The individual choice is allowed to depend on individual-organization characteristic matrix, Z_i, such as distance to hospital in GGT.

[1] As noted in GGT, some possible applications for this model include: hospital quality based on mortality, school performance based on graduation rates, prison rehabilitation programs based on recidivism rates, and job training programs based on incidence of harassment complaints.

Selection is modeled as follows: $\varepsilon_i = \eta_i'\delta + \xi_i$ (equation (5) in GGT)

Here, GGT allow the error term in the organization choice equation to be correlated with the error term in the binary outcome equation with the parameter, δ.

The GGT Stata function estimates the latent observations and parameters (m_i^* c_i^*, α, β, γ, δ) through MCMC methods. The first 10,000 draws are eliminated as burn in. Then, using the remaining draws, the mean and variance of the organization quality measures (β-adjusted by the selection correction) are calculated and displayed as output on the Stata screen.

Because the process uses Bayesian inference, the estimation of the model depends not only on the necessary model variables, but also the prior distributions for each parameter. See section 2.2. of GGT for more information on prior distributions.

- Following GGT, the estimation method assumes independent prior distributions for α, γ, and δ. For these parameters, we assume mean 0 for each prior and allow user options for prior variances.
- For the parameter, β, we again follow GGT and use hyperpriors to allow correlation between organizations based on organization characteristics. β can be written as the sum of organization dummies and organization category dummies. For example, in GGT, there are four hospital ownership categories, k={1,2,3,4}, four hospital size categories, l={1,2,3,4}, and 144 unique hospitals, j. Thus, $\beta_j=p_k+s_l+u_j$ where $p_k=1$ if hospital j is in ownership category k, $s_l=1$ if hospital j is in size category l, and $u_j=1$ for hospital j. We assume p,s,u are jointly Normal with mean,0, and mutually independent, but allow dependence within each organization characteristic through definition of hyper-prior distributions. Specifically, assume p,s,and u have variance τ_p^2, τ_s^2, and τ_u^2 respectively, with a hyper-prior distribution defined as $\underline{s}^2 / \tau_{p,s,u}^2 \sim \chi^2(\underline{v})$, allowing user options for $\underline{s}^2$ and $\underline{v}$.

Technical Notes: Users may notice some slight differences in the above description of prior distributions from that in GGT Section 2.2. These do not change the model but do make the Stata code more tractable. We describe these changes below.

- We remove the constant term, β_1, from the linear equation defining β_j. Instead, we combine this constant term with constant term in γ. Thus, users should not specify a different prior variance for β_1 and should keep in mind that this term will be included γ.
- Within each parameter, we request users to specify a single value for prior variances. For example, suppose γ is the coefficient for two variables, illness severity and age. This Stata code allows users to specify σ_γ^2 in the prior distribution: $\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_\gamma^2 & 0\\ 0 & \sigma_\gamma^2\end{pmatrix}\right]$. Since we do not allow σ_γ^2 to differ by elements γ, users should modify/rescale variables if desired to fit into this framework.
- We do not allow non-zero elements in the off-diagonal prior variance specifications.

The remainder of this documentation explains the different function options and presents an example.

Options

Required Model Variables

outcomevar(*varname*) is required. It is the name of the variable that indicates the individual outcomes in the binary probit model. This variable needs to be 0 or 1 for each individual.

orgchoice(*varname*) is required. It is the name of the variable that indicates the organization that each individual selects/chooses. This variable should be 0,1 and should sum to 1 for each individual.

indID(*varname*) is required. It provides a unique identifier for each individual.

orgID(*varname*) is required. It provides a unique identifier for each organization.

choicechar(*varlist*) is required. It specifies the name of the variables that should be included in the choice equation. These are the Z variables in the Methods and Equations section above.

Optional Model Variables

orgchar(*varlist*) specifies the name of the variables that hold the different organization characteristics. These are the k and l variables in the Methods and Equations section above. The maximum number of variables in this varlist is 10. The variables must be categorical in nature and can either be string or factor type in Stata dataset.

indchar(*varlist*) specifies the name of the variables that should be included in the individual outcome probit equation. These are the X variables in the Methods and Equations section above. The maximum number of variables in this varlist is 20.

Optional Model Parameters

niter(*integer*) is the number of iterations for Gibbs sampling. The default is 100000.

alphapriorar(*real*) is the diagonal elements of the α prior variance-covariance matrix. The default is 1.

gammapriorvar(*real*) is the diagonal elements of the γ prior variance-covariance matrix. default is 1.

deltapriorvar(*real*) is the $\sigma_\delta{}^2$ term in the prior distribution: $\delta \sim N(0, \sigma_\delta{}^2\Sigma^{-1})$ where as in GGT, $\Sigma = I_{J-1} + e_{J-1}e'_{J-1}$ for identity matrix I, number of organizations, J, and vector of units, e. See footnote 17 in GGT for information on choosing $\sigma_\delta{}^2$. In `GGT`, the default is 0.038416.

priortau(*real*, *integer*) is the hyper-parameters for the organization characteristic variance hierarchical prior distributions. From the Methods and Equations section above, `GGT` allows users to specify $\underline{s}^2$ and $\underline{v}$ in the hierarchical prior, $\underline{s}^2 / \tau_o{}^2 \sim \chi^2(\underline{v})$ for organization characteristic, o. The first number in priortau refers $\underline{s}^2$, and the second number refers to $\underline{v}$. The default is priortau(1.25,5). Users must specify both elements if choosing to use this option.

noselection – This option should be specified if the user does not want to apply the selection correction. In this case, the program will simply estimate the parameters in GGT equation (1). Note: The code will also estimate α solely for the purpose of comparison.

noconstant- This option should be specified if the user does not want to include a constant in the outcome probit equation, i.e. γ will not include a constant term.

Reporting

savedraws- This option will save a .csv file in the directory which holds every 100 draws of each parameter via the MCMC Gibbs Sampling routine.

Remarks and Examples

In this section, we present an example to show how the data should be arranged in order to use `GGT`.

Assume we are interested in hospital quality. We have data on 300 patients and 8 hospitals. The individual patient variables include the following: the mortality measure ("mortality"), the hospital choice variable ("hosp_choice), and an illness severity measure ("severity"). Additionally, we have two variables "dist" and "dist2" representing the distance from each patient to each hospital along with its square (normalized to have similar scales, necessary since the priors are the same). We also have hospital characteristic variables, "hosp_size" and "hosp_ownership". The categories for hosp_size are "small" and "large", and the categories for "hosp_ownership" are "public" and "private".

The individual patient ID variable is called "indnumber" and the hospital ID variable is called "hospnum". In the Stata dataset, there should be an observation for each individual-hospital pair, even if the individual did not choose that hospital. For example, with 300 patients and 8 hospitals, we have 300*8=2400 observations in the data. The table below shows the structure of the data for the first 2 patients. You can see that individual 1 went to hospital 7 and died while patient number 2 went to hospital 3 and did not die. The severity measure is constant within an individual while the distance and distance^2 measures differ for each patient-hospital pair. Additionally, notice the hospital characteristics are constant within hospitals, e.g., hosp_size and hosp_ownership is always "small", "public" for the row in which hospnum==1.

indnumber	hospnum	hosp_choice	mortality	severity	dist	dist2	hosp_size	hosp_ownership
1	1	0	1	1.549	0.015	0.000	small	public
1	2	0	1	1.549	0.250	0.013	large	public
1	3	0	1	1.549	0.259	0.014	large	public
1	4	0	1	1.549	0.080	0.001	large	private
1	5	0	1	1.549	0.097	0.002	large	public
1	6	0	1	1.549	0.160	0.005	large	public

1	7	1	1	1.549	0.459	0.042	small	private
1	8	0	1	1.549	0.491	0.048	small	public
2	1	0	0	0.723	0.052	0.001	small	public
2	2	0	0	0.723	0.162	0.005	large	public
2	3	1	0	0.723	0.067	0.001	large	public
2	4	0	0	0.723	0.097	0.002	large	private
2	5	0	0	0.723	0.187	0.007	large	public
2	6	0	0	0.723	0.019	0.000	large	public
2	7	0	0	0.723	0.110	0.002	small	private
2	8	0	0	0.723	0.058	0.001	small	public
3	1	0	1	2.684	0.142	0.004	small	public
3	2	0	1	2.684	0.070	0.001	large	public

Example 1:
If we want to see the hospital quality measures, β, using all the default settings, we would simply type the command:
```
GGT, outcomevar(mortality) orgchoice(hosp_choice) indID(indnumber) orgID(hospnum)
choicechar(dist dist2)
```

This will apply the selection model with using dist and dist2 as the choice characteristics. Since we did not specify indchar option, the code will assume only a constant and the hospital choice for the individual probit model. Additionally, since we did not specify orgchar, the code will assume no correlation across hospitals via hospital size or ownership. The sampling algorithm will assume the default prior variance options and number of iterations.

The output on the screen will be the summary statistics for the estimated β draws via the MCMC Gibbs sampler. The output for this example is shown below with "q_n" represented the quality for hospital ID, n. Notice that the number of observations is 900- this comes from the default 100,000 iterations, saving only every 100th draw, and deleting the first 10,000 draws as burn-in.

Variable	Obs	Mean	Std. Dev.	Min	Max
q_8	900	-.5500745	.3616526	-2.2093	.817451
q_7	900	.4810402	.2885439	-.361103	1.73498
q_6	900	.0461915	.2821159	-.803522	1.06977
q_5	900	.0610291	.2807342	-1.14918	1.13188
q_4	900	.0410323	.2765834	-.83994	1.01772
q_3	900	-.0477848	.2811626	-1.13624	.952603
q_2	900	.1456833	.3037255	-.930358	1.15258
q_1	900	-.2295427	.2927853	-1.42508	.574466

Note: The code may take several minutes to complete running due to its computational complexity. Once the code is complete, the word "complete" will display on the Stata screen. If after a couple minutes the code does not complete or Stata simply quits, this is likely due to an error with the prior variance specifications which are not compatible with the data. We suggest trying to call the program again using different prior variance values.

Example 2:
Now suppose we want to include the severity measure in the morality equation and we also want to allow hospital correlation based on size and ownership. Additionally, we want to rescale the prior variances based on the structure of the data. Specifically, we want the prior variance of alpha to be 5, the prior variance of gamma to be 3, selection term for delta to be 1, and the parameters for the hyperpriors to be 1 and 5. Finally, we want to save the draws for each of the parameters in a csv file to the directory.

To do this, we would type the command: `GGT, outcomevar(mortality) orgchoice(hosp_choice) indID(indnumber) orgID(hospnum) choicechar(dist dist2) indchar(severity) orgchar(hosp_size hosp_ownership) alphapriorvar(5) gammapriorvar(3) deltapriorvar(1) priortau(1,5) savedraws`

The output in this case is now:

Variable	Obs	Mean	Std. Dev.	Min	Max
q_8	900	-3.077418	1.962936	-10.19112	2.856249
q_7	900	1.453671	1.722678	-4.818444	7.110582
q_6	900	-.2475159	1.748768	-5.859746	5.017234
q_5	900	-.4813236	1.742171	-6.112561	5.860093
q_4	900	.1682982	1.676839	-5.128277	6.79628
q_3	900	-1.261563	1.753765	-6.99393	4.180546
q_2	900	-.5406034	1.740861	-6.037043	6.242453
q_1	900	-1.883937	1.795644	-7.425305	3.568303

Additionally, a file called "temp_GGT_output.csv" is saved in the directory. A screenshot of the first 11 rows and 6 columns is shown below.

The official column names are : iter, tau0, tau1, tau2, beta_orgatt1_type1, beta_orgatt1_type2, beta_orgatt2_type1, beta_orgatt2_type2, beta_orgatt3_type1, beta_orgatt3_type2, beta_orgatt3_type3, beta_orgatt3_type4, beta_orgatt3_type5, beta_orgatt3_type6, beta_orgatt3_type7, beta_orgatt3_type8, gamma1, gamma2, alpha1, alpha2, delta1, delta2, delta3, delta4, delta5, delta6, delta7.

-iter: indicates the Gibbs Sampler iteration.
-tau0, tau1, and tau2: the hyperprior draws for variances of hosp_size, hosp_ownership, and hospital organization dummies respectively.
-beta_orgattN_typeM: the β coefficient draws for the dummy variable indicating the Nth specified organization characteristic variable and the Mth category for that variable (where categories are ordered numerically in the case where the variable is string or factor).
Note: In this case, since we specified 2 organization characteristics, beta_orgatt3_typej corresponds to the β coefficient on the dummy variable for hospital j.
-gamma1, gamma2: the γ estimate draws for the coefficient on a constant and the severity measure (respectively) in the outcome probit equation.
-alpha1, alpha2: the α estimate draws for the coefficient on dist and dist2 (respectively) in the organization choice equation.
-delta1-delta7: the δ estimate draws in the selection equation.

iter	tau0	tau1	tau2	beta_orgatt1_type1	beta_orgatt1_type2	beta_orgatt2_type1	beta_orgatt2_type2
100	1.64005	0.947807	0.960714	-1.17202	-1.9744	-0.350753	-0.787612
200	0.791952	1.51049	0.740155	-0.746715	-0.166923	-0.435621	-0.919206
300	1.29789	1.61692	0.674868	0.448041	1.38033	1.1191	0.712085
400	3.16132	1.55191	1.25275	-1.90263	-2.499	0.917612	-1.0566
500	1.36342	1.74155	1.77559	0.299045	-0.347646	-1.10906	-1.28785
600	1.235	0.967284	1.2378	-0.576297	-0.39816	0.959383	-0.592853
700	1.76776	1.62961	2.20218	0.358753	-0.124939	-0.960883	-2.42629
800	0.972693	2.49559	1.04786	0.561438	0.402018	4.02357	0.979691
900	5.34903	1.51721	0.563776	3.56987	2.92047	-0.27793	-1.05311
1000	3.20846	2.10669	0.687972	0.895687	0.231262	1.41547	-0.130428

Example 3:

Finally, suppose we wish to compare the results to the case where we do not apply the selection correction. In this case, the program simply estimates equation (1) in GGT. We can still specify all the options, but the code will only use those that are necessary. e.g., since the nonselection model assumes that $\delta=0$, then specifying deltapriorvar is unnecessary. To run this model, we need to specify the "nonselection" option.

Note: Even though the equation we wish to estimate does not depend on patient-organization choice characteristics, the code will still require choice characteristics in its estimation of α.

```
GGT, outcomevar(mortality) orgchoice(hosp_choice) indID(indnumber) orgID(hospnum)
choicechar(dist dist2) indchar(severity) orgchar(hosp_size hosp_ownership)
alphapriorvar(5) gammapriorvar(3) priortau(1,5) noselection
```

The output for this scenario is as follows:

Variable	Obs	Mean	Std. Dev.	Min	Max
q_8	900	-.7867516	.5794054	-3.822582	1.01066
q_7	900	.2010573	.5644178	-2.832025	2.09774
q_6	900	-.0247391	.5788897	-2.722792	1.938399
q_5	900	-.1067564	.5616703	-3.02365	1.807913
q_4	900	.0623483	.5506349	-2.436699	1.650944
q_3	900	-.3221626	.5825404	-2.83032	1.668997
q_2	900	.0484907	.5699824	-2.7295	2.280766
q_1	900	-.6415367	.5895134	-3.490691	1.766314

References

Geweke, J., Gowrisankaran, G., & Town, R. J. (2003). Bayesian inference for hospital quality in a selection model. *Econometrica*, *71*(4), 1215-1238.