The Stata Journal (2015)
15, Number 2, pp. 512–522

Generalized maximum entropy estimation of discrete choice models

Paul Corral
American University
Washington, DC
paulcorral@gmail.com

Mungo Terbish
American University
Washington, DC
mungunsuvd@gmail.com

Abstract. In this article, we describe the `gmentropylogit` command, which implements the generalized maximum entropy estimation methodology for discrete choice models. This information theoretic procedure is preferred over its maximum likelihood counterparts because it is more efficient, avoids strong parametric assumptions, works well when the sample size is small, performs well when the covariates are highly correlated, and functions well when the matrix is ill conditioned. Here we introduce the generalized maximum entropy procedure and provide an example using the `gmentropylogit` command.

Keywords: st0390, gmentropylogit, generalized maximum entropy, maximum entropy, logit, discrete choice

1 Introduction

Maximum entropy (ME) estimation for discrete choice problems is an alternative to traditional maximum-likelihood (ML) estimation methods such as logit and probit. However, of all multinomial probability distributions, ME estimation is the most noncommittal option available to the econometrician looking to avoid parametric assumptions (Mittelhammer, Judge, and Miller 2000). Under the principle of ME, the econometrician selects the probabilities with the minimal information content for the multinomial problem (Mittelhammer, Judge, and Miller 2000). The ME procedure consists in maximizing Shannon's (1948) entropy measure,

$$H(p) \equiv -\sum_j p_j \ln p_j$$

where p_j is the probability of observing outcome j. Jaynes (1957a, 1957b) proposed applying the ME procedure to recover unknown probabilities. Denzau, Gibbons, and Greenberg (1989) and Soofi (1992) used the suggested procedure to solve the multinomial choice problem. Soofi (1992) established the link and equivalence between the ME procedure and the traditional logit and referred to the ME procedure as the ME logit.

Golan, Judge, and Perloff (1996) expanded the ME procedure by including the noise terms in the multinomial information constraints of the ME. This generalized maximum entropy (GME) estimation method accounts for unknown disturbances that impose challenges on the ME and ML methods (Golan, Judge, and Miller 1996). Adding the noise terms to the ME procedure results in an estimator with the same sampling properties as

 st0390

the ME counterpart for large samples (Golan, Judge, and Miller 1996). However, given finite samples, the GME outperforms its ME and ML counterparts. Specifically, the GME has the following preferred properties (Golan, Judge, and Perloff 1996):

1. It is more efficient.
2. It avoids strong parametric assumptions.
3. It works well when the sample size is small.
4. It works well when the covariates are highly correlated.
5. It works well when the matrix is ill conditioned.

We begin this article by familiarizing readers with the GME estimator using the methodology proposed by Golan, Judge, and Miller (1996) and Golan, Judge, and Perloff (1996). We then introduce the new command **gmentropylogit**, which can be used as an alternative to `logit` or `probit` to fit discrete choice models.

2 GME logit

Logit and probit models are widely used in the social sciences. These model the probability (p) that agent i will choose or face a certain outcome from the set of possible outcomes (Soofi 1992). These two estimation methods entail finding the probability that an outcome is observed, and they assume that the most likely outcome is the one observed (Soofi 1992). However, logit and probit impose a parametric structure on the probabilities. The underlying distribution for the probabilities is unknown, and the choice of logit and probit relies on this strong assumption (Golan, Judge, and Perloff 1996). Thus Golan, Judge, and Miller (1996) and Golan, Judge, and Perloff (1996) prefer to use estimation strategies that do not rely on such heavy assumptions: ME and GME.

ME entails finding the probability distribution $(p_1, p_2, \ldots, p_n)$ for the set of values $(x_1, x_2, \ldots, x_n)$ given their moments of these values. GME builds upon ME by adding natural noise components. For the GME solution, we consider an experiment consisting of T trials. In each experiment with J unordered possible outcomes, a binary random variable $y_{i1}, \ldots, y_{tj}$ is observed, where y_{ij} for $i = 1, 2, \ldots, T$ takes one of the J unordered categories $j = 1, 2, \ldots, J$. On each trial i, each J alternative is observed in the form of a binary variable, y_{ij}, that equals unity if alternative j is observed and zero otherwise. Let p_{ij} be the probability of alternative j on trial i and be related to a set of variables through the model

$$p_{ij} = \text{Prob}(y_{ij} = 1|\mathbf{x}_i, \boldsymbol{\beta}_j) = F(\mathbf{x}_i'\boldsymbol{\beta}_j) > 0 \text{ for all } i \text{ and } j$$

where $\boldsymbol{\beta}_j$ is a $(K \times 1)$ vector of unknowns, $\mathbf{x}_i'$ is a $(1 \times K)$ vector of covariates, and $F(\cdot)$ is a function linking the probabilities p_{ij} with the covariates $\mathbf{x}_i'\boldsymbol{\beta}_j$, such that $\sum_j F(\mathbf{x}_i'\boldsymbol{\beta}_j) = 1$.

For the logit model, $F(\cdot)$ would be the logistic cumulative density function. For the probit model, $F(\cdot)$ would be the standard Gaussian cumulative density function. If we accommodate the model to have noisy data, then the model becomes

$$y_{ij} = F(\cdot) + e_{ij} = p_{ij} + e_{ij}$$

where p_{ij} denotes the unknown multinomial probabilities and e_{ij} denotes the natural noise components for each observation and is contained in the $[-1, 1]$ support space for each observation.

To recover the unknown and unobservable p and e, one must use the noisy observable, y_{ij}, and the known covariates, $\mathbf{x}_i$, to formulate the problem. In the GME formulation of the multinomial problem, this information is used as the cross-moments between the $\mathbf{x}$ matrix and all the other quantities

$$(I_j \otimes \mathbf{X}')\mathbf{y} = (I_j \otimes \mathbf{X}')\mathbf{p} + (I_j \otimes \mathbf{X}')\mathbf{e} \tag{1}$$

where $\mathbf{X}$ is a $(T \times K)$ matrix and there are TJ data points for y_{ij}. The y_{ij} data points are vectorized and thus are $(TJ \times 1)$ vectors. The same applies for $\mathbf{e}$ and $\mathbf{p}$, which are also vectors of dimension $(TJ \times 1)$. As can be seen from (1), it is an ill-posed problem where there are $\{T \times (J-1)\}$ unknown parameters but only $(K \times J) < \{T \times (J-1)\}$ data points.

To use Shannon's entropy measure to estimate the problem, we must reparameterize the noise components. Because $\mathbf{p}$ is already in a probability form, only the elements of $\mathbf{e}$ need to be reparameterized to proceed with GME, as suggested by Golan, Judge, and Perloff (1996). Because each e_{ij} will range between $[-1, 1]$, we can include a set of discrete points (v_{ij}) ranging between $[0, 1]$. The error terms are characterized by an H-dimensional support space, v, and an H-dimensional vector of weights, w, that correspond to each v. The unknown weights have the properties of probabilities with $\sum_h w_{ijh} = 1$. Thus the resulting reparameterization is

$$e_{ij} \equiv \sum_h v_{ijh} w_{ijh}$$

where the H-dimensional errors' support suggested by Golan, Judge, and Miller (1996, 253) as a conservative choice is $v = (-1/\sqrt{T}, 0, 1/\sqrt{T})$ for each e_{ij}. Hence, in this case, $H = 3$.

Under this reparameterization, (1) takes the form

$$(I_j \otimes \mathbf{X}')\mathbf{y} = (I_j \otimes \mathbf{X}')\mathbf{p} + (I_j \otimes \mathbf{X}')(\mathbf{wv})$$

where now both $\mathbf{p}$ and $\mathbf{w}$ are in the form of probabilities, $\mathbf{w}$ is a $TJ \times H$ matrix, and $\mathbf{v}$ is an $H \times 1$ matrix.

As noted before, the difference between GME and ME estimation arises because of the inclusion of the noise terms, which give the GME the desired properties we enumerated before. The resulting solutions for ML multinomial logit and ME p are equivalent, even though their formulations are different (see Soofi [1992]). As Mittelhammer, Judge, and Miller (2000) discuss, this equivalence can be explained by the following: 1) the estimating equations or moment constraints in the ME formulation are the ML logit first-order conditions; and 2) the ME solution resulting from the optimization has the same form as the logistic multinomial probabilities. As we illustrate in the following section, the GME gains the advantage over the ME because the GME is a product of two logits for the p and for the w (Golan, Judge, and Perloff 1996).

2.1 GME formulation and solution

The objective function for the GME discrete choice problem is a dual objective function. The dual objective function is composed of the entropy of the probabilities (p) and the entropy for the weights (w) (Golan, Judge, and Perloff 1996; Golan, Judge, and Miller 1996). This implies the assumption of independence between the two. The objective function of the GME multinomial problem is the maximization of the Shannon entropy measure and takes the following form specified in Golan, Judge, and Perloff (1996),

$$\max_{p,w} H(p,w) = \max_{p,w}(-\mathbf{p}'\ln \mathbf{p} - \mathbf{w}'\ln \mathbf{w})$$

subject to the JK information-moment conditions,

$$(I_j \otimes \mathbf{X}')\mathbf{y} = (I_j \otimes \mathbf{X}')\mathbf{p} + (I_j \otimes \mathbf{X}')(\mathbf{wv}) \tag{2}$$

the normalization constraints,

$$(I_{T1} I_{T2} \dots I_{TJ})\mathbf{p} = 1 \quad \text{for} \quad i = 1, 2, \dots, T \tag{3}$$

and

$$\mathbf{1}'\mathbf{w}_{ij} = 1 \quad \text{for} \quad i = 1, 2, \dots, T \quad \text{and} \quad j = 1, 2, \dots, J$$

where $\mathbf{1}$ is a $(1 \times H)$ vector of 1s.

The corresponding Lagrangian is

$$\begin{aligned}
\mathcal{L} = & -\mathbf{p}' \ln \mathbf{p} - \mathbf{w}' \ln \mathbf{w} \\
& + \boldsymbol{\lambda}' \left\{ \left(I_J \otimes \mathbf{X}' \right) \mathbf{p} + \left(I_J \otimes \mathbf{X}' \right) V \mathbf{w} - \left(I_J \otimes \mathbf{X}' \right) \mathbf{y} \right\} \\
& + \boldsymbol{\mu}' \left\{ \mathbf{1} - \left(I_{1J} I_{2J} \dots I_{TJ} \right) \mathbf{p} \right\} + \boldsymbol{\rho}' \left(1 - \mathbf{1}'\mathbf{w} \right)
\end{aligned}$$

where λ, μ, and ρ are the corresponding Lagrange multipliers. Under this specification, the parameters of interest are $\boldsymbol{\lambda}$. Note that these provide the coefficients $-\lambda_j = \boldsymbol{\beta}_j$.

To simplify the math, we proceed with the Lagrangian in scalar form,

$$\max_{p,w} H(p,w) = \max_{p,w} \left(-\sum_{ij} p_{ij} \ln p_{ij} - \sum_{ijh} w_{ijh} \ln w_{ijh} \right)$$

subject to the JK information-moment conditions, where the JKth condition is

$$\sum_i y_{ij} x_{ik} = \sum_i x_{ik} p_{ij} + \sum_{ih} x_{ik} v_h w_{ijh} \tag{4}$$

and

$$\sum_j p_{ij} = 1$$

and

$$\sum_h w_{ijh} = 1$$

The resulting first-order conditions are

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_{ij}} &= -\ln\, p_{ij} - 1 - \sum_k \lambda_{jk} x_{ik} - \mu_i = 0 \\
\frac{\partial \mathcal{L}}{\partial w_{ijh}} &= -\ln\, w_{ijh} - 1 - \sum_k \lambda_{jk} x_{ik} v_h - \rho_i = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda_{jk}} &= \sum_i y_{ij} x_{ik} - \sum_i x_{ik} p_{ij} - \sum_i x_{ik} v_h w_{ijh} = 0 \\
\frac{\partial \mathcal{L}}{\partial \mu_i} &= 1 - \sum_j p_{ij} = 0 \\
\frac{\partial \mathcal{L}}{\partial \rho_i} &= 1 - \sum_h w_{ijh} = 0
\end{aligned}$$

Note that for the Lagrange parameters $\beta_{jk} = -\lambda_{jk}$,

$$p_{ij} = \exp\left(-1 - \mu_i - \sum_k \lambda_{jk} x_{ik}\right)$$

and

$$w_{ijh} = \exp\left(-1 - \rho_i - \sum_k \lambda_{jk} x_{ik} v_h\right)$$

We sum the probability equation over j and sum the noise component weights over h and have

$$\sum_j p_{ij} = \exp(-1 - \mu_i) \sum_j \exp\left(-\sum_k \lambda_{jk} x_{ik}\right) = 1$$

and

$$\sum_h w_{ijh} = \exp(-1 - \rho_i) \sum_h \exp\left(-\sum_k \lambda_{jk} x_{ik} v_h\right) = 1$$

Then, by setting $\lambda_1 = \mathbf{0}$, where $\mathbf{0}$ is a $(K \times 1)$ zero vector, we obtain

$$\widehat{p}_{ij} = \frac{\exp\left(-\sum_k \widehat{\lambda}_{jk} x_{ik}\right)}{\sum_j \exp\left(-\sum_k \widehat{\lambda}_{jk} x_{ik}\right)} = \frac{\exp\left(-\sum_k \widehat{\lambda}_{jk} x_{ik}\right)}{1 + \sum_{j=2}^{J} \exp\left(-\sum_k \widehat{\lambda}_{jk} x_{ik}\right)} \equiv \frac{\exp\left(-\sum_k \widehat{\lambda}_{jk} x_{ik}\right)}{\Omega_i}$$

and

$$\widehat{w}_{ijh} = \frac{\exp\left(-\sum_k x_{ik}\widehat{\lambda}_{jk}v_h\right)}{\sum_h \exp\left(-\sum_k x_{ik}\widehat{\lambda}_{jk}v_h\right)} \equiv \frac{\exp\left(-\sum_k x_{ik}\widehat{\lambda}_{jk}v_h\right)}{\Psi_{ij}\left(\widehat{\lambda}\right)}$$

As we can with the traditional logit, we can also define log odds-ratios, which we calculate as follows (Mittelhammer, Judge, and Miller 2000):

$$\ln\left(\frac{p_{ij}}{p_{i1}}\right) = -\mathbf{x}_i\boldsymbol{\lambda}_{\mathbf{j}}$$

Concurrently, the exponentiated coefficients may be interpreted as odds ratios.

The dual unconstrained formulation of the GME problem demonstrates that the GME is related to a class of generalized logit formulations (Golan, Judge, and Perloff 1996; Golan, Judge, and Miller 1996). Building on the Lagrangian, the dual unconstrained GME formulation as a function of the Lagrangian multipliers, λ, is, according to Golan, Judge, and Perloff (1996) and Golan, Judge, and Miller (1996),

$$\begin{aligned} M(\lambda_j) &= -\mathbf{p}(\lambda)' \ln \mathbf{p}(\lambda) - \mathbf{w}(\lambda)' \ln \mathbf{w}(\lambda) \\ &\quad + \lambda'\{(I_j \otimes \mathbf{X}')\mathbf{y} - (I_j \otimes \mathbf{X}')\mathbf{p} - (I_j \otimes \mathbf{X}')(\mathbf{wv})\} \\ &= -\mathbf{p}(\lambda)'(-\mathbf{x}_i'\lambda_j - \ln\,\Omega_i) - \mathbf{w}(\lambda)'(-\mathbf{x}_i'\lambda_j v_j - \ln\,\Psi_{ij}) \\ &\quad + \lambda'\{(I_j \otimes \mathbf{X}')\mathbf{y} - (I_j \otimes \mathbf{X}')\mathbf{p} - (I_j \otimes \mathbf{X}')(\mathbf{wv})\} \\ &= \mathbf{y}'(I_j \otimes \mathbf{X})\lambda + \sum_i \ln\{\Omega_i(\lambda_j)\} + \sum_i \sum_j \ln\{\Psi_i(\lambda_j)\} \end{aligned} \tag{5}$$

By minimizing the dual objective function with respect to λ, we obtain the multipliers $\widehat{\lambda}$. From these, we obtain the values of $\widehat{p}_{ij}$ and $\widehat{w}_{ijh}$. Compared with the primal GME problem, the dual unconstrained estimation is computationally more efficient because it requires estimation of KJ $\widehat{\lambda}$ to estimate the same $T \times J$ p as in the primal problem. The objective function constitutes a generalized log-likelihood function where the right-hand side involves the error terms introduced in the GME formulation. As the sample size increases, the error terms converge to zero, and the GME estimations approach the ME and ML solutions because these have zero moment conditions.

The average marginal effects (AMEs) are considered the more informative parameters, as in a probit or logit specification. The AME for a continuous variable is (Bartus 2005)

$$\text{AME}_k = \beta_k \frac{1}{T}\sum_{i=1}^{T} p_i(1-p_i)$$

If x_k is a dummy variable, then the equation for the AME is (Bartus 2005)

$$\text{AME}_k = \frac{1}{T}\sum_{i=1}^{T} [\{x_{ki}p_i(1-p_i)|_{x_k=1}\} - \{x_{ki}p_i(1-p_i)|_{x_k=0}\}]$$

3 The gmentropylogit command, output, and statistics

3.1 gmentropylogit

`gmentropylogit` is implemented via the `optimize()` function in Mata, where we optimize the dual unconstrained model (5). It is executed as a `d0` evaluator, and Stata obtains the numerical first and second derivatives for the objective function. The command also allows estimation of the marginal effects, which provide the impact of each x on the probability of a positive outcome. The syntax is as follows:

`gmentropylogit` *depvar* [*indepvars*] [*if*] [*in*] [, `mfx` `generate(`*varname*`)`]

Options

`mfx` displays, instead of the coefficients, the impact of each x on the probability of a positive outcome. `mfx` considers dummy variables and provides their estimates accordingly.

`generate(`*varname*`)` creates a new variable with a user-selected name that will contain the predicted probability of the fitted model.

Example

Here we illustrate the use of the `gmentropylogit` command to estimate the probability of a car being foreign using Stata's `auto.dta`.

```
. sysuse auto
(1978 Automobile Data)
. gmentropylogit foreign price mpg weight trunk
Iteration 0:   log likelihood = -213.88751
Iteration 1:   log likelihood = -184.47517
Iteration 2:   log likelihood = -184.17173
Iteration 3:   log likelihood = -184.16896
Iteration 4:   log likelihood = -184.16896
Generalized Maximum Entropy (Logit)               Number of obs      =     74
                                                  Degrees of freedom =      4
                                                  Entropy for probs. =   24.4
                                                  Normalized entropy = 0.4750
                                                  Ent. ratio stat.   =   53.9
                                                  P Val for LR       = 0.0000
Criterion F (log L) = -184.16896                  Pseudo R2          = 0.5250

------------------------------------------------------------------------------
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   .0005633   .0001571     3.59   0.000     .0002554    .0008713
         mpg |  -.0878192   .0831957    -1.06   0.291    -.2508797    .0752413
      weight |  -.0043673   .0010259    -4.26   0.000     -.006378   -.0023566
       trunk |      .0367   .1159334     0.32   0.752    -.1905252    .2639253
       _cons |   9.373785   3.919642     2.39   0.017     1.691428    17.05614
------------------------------------------------------------------------------
```

However, the marginal effects are the more informative parameters because these inform us of the expected change in the probability.

```
. sysuse auto, clear
(1978 Automobile Data)
. gmentropylogit foreign price mpg weight trunk, mfx
Iteration 0:   log likelihood = -213.88751
Iteration 1:   log likelihood = -184.47517
Iteration 2:   log likelihood = -184.17173
Iteration 3:   log likelihood = -184.16896
Iteration 4:   log likelihood = -184.16896
Generalized Maximum Entropy (Logit), dF/dx        Number of obs      =     74
                                                  Degrees of freedom =      4
                                                  Entropy for probs. =   24.4
                                                  Normalized entropy = 0.4750
                                                  Ent. ratio stat.   =   53.9
                                                  P Val for LR       = 0.0000
Criterion F (log L) = -184.16896                  Pseudo R2          = 0.5250

     foreign        Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

       price     .0000584    .0000114    5.12   0.000     .0000361    .0000808
         mpg    -.0091075    .0083781   -1.09   0.277    -.0255283    .0073132
      weight    -.0004529    .0000473   -9.58   0.000    -.0005456   -.0003603
       trunk     .0038061     .011915    0.32   0.749    -.0195469    .0271591
       _cons     .9721323    .3524144    2.76   0.006     .2814127    1.662852

Partial effect for dummy is E[y|x,d=1] - E[y|x,d=0]
```

If we were to also specify the `generate()` option, the command would provide the following results:

```
. sysuse auto, clear
(1978 Automobile Data)
. gmentropylogit foreign price mpg weight trunk, generate(p_foreign)
Iteration 0:   log likelihood = -213.88751
Iteration 1:   log likelihood = -184.47517
Iteration 2:   log likelihood = -184.17173
Iteration 3:   log likelihood = -184.16896
Iteration 4:   log likelihood = -184.16896
Generalized Maximum Entropy (Logit)               Number of obs      =     74
                                                  Degrees of freedom =      4
                                                  Entropy for probs. =   24.4
                                                  Normalized entropy = 0.4750
                                                  Ent. ratio stat.   =   53.9
                                                  P Val for LR       = 0.0000
Criterion F (log L) = -184.16896                  Pseudo R2          = 0.5250

     foreign        Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

       price     .0005633    .0001571    3.59   0.000     .0002554    .0008713
         mpg    -.0878192    .0831957   -1.06   0.291    -.2508797    .0752413
      weight    -.0043673    .0010259   -4.26   0.000     -.006378   -.0023566
       trunk        .0367    .1159334    0.32   0.752    -.1905252    .2639253
       _cons     9.373785    3.919642    2.39   0.017     1.691428    17.05614

Percent correctly predicted:87.837838
```

Notice that the output report the percentage of correctly predicted results for the fitted model in the bottom left-hand corner.

3.2 Inference and diagnostic statistics

Along with the β coefficients, the marginal effects, and the traditional statistics, the `gmentropylogit` command returns inference and diagnostic statistics that are specific to the GME methodology. The maximum possible entropy in the model occurs when the distribution over the outcomes is uniform (Soofi 1992). Soofi (1992) explains that this occurs when the moment constraints (2) are not included or when the moment constraints are linearly dependent with the normalizing constraints (3). Once we add relevant data, we move away from the uniform distribution and reduce uncertainty (Golan, Judge, and Perloff 1996). The normalized entropy measures "the proportion of the remaining total uncertainty" (Golan, Judge, and Perloff 1996) for discrete choice models. Golan, Judge, and Perloff (1996) and Soofi (1992) present the following:

$$S(\widehat{p}) = \frac{\left(-\sum_i \sum_j \widehat{p}_{ij} \ln \widehat{p}_{ij}\right)}{(T \ln J)} = \frac{-\mathbf{p}' \ln \mathbf{p}}{(T \ln J)}$$

Here $S(\widehat{p}) \in (0,1)$ and $(T \ln J)$ is maximum uncertainty. $S(\widehat{p}) = 0$ indicates a complete lack of uncertainty, whereas $S(\widehat{p}) = 1$ indicates that the probability measure is uniform for i and j or that there is perfect uncertainty. This measure of uncertainty helps the user to compare different model specifications. For example, if eliminating an explanatory variable does not alter the $S(\widehat{p})$ measure, then, based on the data, the eliminated variable has no informational contribution and does not help explain the uncertainty of the unknown p_{ij}.

Because the normalized entropy measures the level of uncertainty in the model, its counterpart, the information index, measures the level of information in the model (Soofi 1992). The index provides a goodness-of-fit measure, and Soofi (1992) defines it as $I(\widehat{p}) = 1 - S(\widehat{p})$. Because $I(\widehat{p})$ measures the proportion of variation in the model's data, it is often referred to as pseudo-R^2 (Golan 2008).

Another diagnostic statistic that is characteristic of the discrete choice ME method is the entropy ratio statistic. The equivalence between ML and ME methods allows for deriving the relationship between the log-likelihood value and the entropy ratio statistic. Golan, Judge, and Perloff (1996) indicate that in the GME formulation, the ME will be attained if the information-moment conditions (4) are not enforced and will therefore have a uniform distribution of probabilities over each choice set. The log-likelihood value of the unconstrained problem where $\boldsymbol{\lambda} = 0$ is $\ln l_\omega = T \ln J$ (Golan 2008). The constrained likelihood is the value of $\max(H)$ and $\ln l_\Omega = -\mathbf{p} \ln \mathbf{p}$. Therefore, the entropy ratio statistic is

$$W = 2 \ln l_\omega - 2 \ln l_\Omega = 2T \ln J\{I(\widehat{p})\}$$

4 Conclusion

In this article, we describe the user-written command `gmentropylogit`, which provides users with an alternative to probit and logit estimation methods. The estimation method is preferable over logit and probit because of the aforementioned desired characteristics. The command also provides the user with a new set of information statistics that are relevant to ME estimation. Finally, the command includes the option `mfx` to provide users with the AMEs, which are usually the values of interest to the econometrician.

5 Acknowledgments

Amos Golan and Heath Henderson provided helpful comments in writing the command `gmentropylogit`.

6 References

Bartus, T. 2005. Estimation of marginal effects using margeff. *Stata Journal* 5: 309–329.

Denzau, A. T., P. C. Gibbons, and E. Greenberg. 1989. Bayesian estimation of proportions with a cross-entropy prior. *Communications in Statistics—Theory and Methods* 18: 1843–1861.

Golan, A. 2008. *Information and Entropy Econometrics—A Review and Synthesis*. New York: Now Publishers.

Golan, A., G. Judge, and J. M. Perloff. 1996. A maximum entropy approach to recovering information from multinomial response data. *Journal of the American Statistical Association* 91: 841–853.

Golan, A., G. G. Judge, and D. Miller. 1996. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Chichester, UK: Wiley.

Jaynes, E. T. 1957a. Information theory and statistical mechanics. *Physical Review* 106: 620–630.

———. 1957b. Information theory and statistical mechanics. II. *Physical Review* 108: 171–190.

Mittelhammer, R. C., G. G. Judge, and D. J. Miller. 2000. *Econometric Foundations*. New York: Cambridge University Press.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423.

Soofi, E. S. 1992. A generalizable formulation of conditional logit with diagnostics. *Journal of the American Statistical Association* 87: 812–816.

About the authors

Paul Corral is a third-year PhD student at the Economics department of American University, Washington, DC.

Mungo Terbish is a third-year PhD student at the Economics department of American University, Washington, DC.