ELSEVIER Computational Statistics & Data Analysis 39 (2002) 127–136

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

# Measures of relative model fit

Alan Agresti *, Brian Caffo

*Department of Statistics, University of Florida, 204 Griffin-Floyd Hall, Gainesville, FL 32611-8545, USA*

Received 1 October 2000

**Abstract**

Most software for statistical model fitting reports the value of the maximized log likelihood function, but the numerical value is difficult to interpret because of its log scale, its dependence on the sample size, and the possible omission of constants. A sample-size-scaled version of the likelihood function summarizes the model fit. A related index uses the mean of the contributions to the likelihood function. The ratio or difference of either index for two models is a summary measure of relative model fit. We discuss these measures and briefly consider interval estimation for them. 

*Keywords:* AIC; Deviance; Discrete data; Dissimilarity; Likelihood function; Non-nested models.

## 1. Introduction

Table 1, taken from the General Social Survey of 1990, provides responses of 1308 subjects in the US to the question, "Within the past 12 months, how many people have you known personally that were victims of homicide?" Table 1 shows responses for those who identified their race as white or as black. For a certain negative binomial model fitted to these data, software (SAS, using PROC GENMOD) reports the maximized log likelihood value of $-497.9$. It is increasingly common for researchers to report the maximized log likelihood in tables that present results of their model fitting. This provides information needed to perform likelihood-ratio tests comparing pairs of nested models, but how does one interpret the magnitude of a value such as $-497.9$? Moreover, how does one compare its value to that for a

* Corresponding author.
*E-mail address:* aa@stat.ufl.edu (A. Agresti).

0167-9473/02/$ - see front matter 
PII: S 0167-9473(01)00054-8

Table 1
Data from 1990 General Social Survey on number of victims of murder known in past year, by race

| | Race | |
|---|---|---|
| Response | Black | White |
| 0 | 119 | 1070 |
| 1 | 16 | 60 |
| 2 | 12 | 14 |
| 3 | 7 | 4 |
| 4 | 3 | 0 |
| 5 | 2 | 0 |
| 6 | 0 | 1 |

different type of model, such as the value of $-500.7$ for a Poisson loglinear mixed model containing a normal random effect?

One complication is that different software for fitting a model may provide different log likelihood values, since some software (such as SAS GENMOD) drops constants from the likelihood function that do not affect parameter estimation. The most common use of the likelihood is for inference, but our focus in this article is mainly on description. When the sample size is large, for instance, a difference in log likelihoods may provide strong evidence that one model fits better than another, but the models may provide a similar fit in practical terms. We discuss scalings of the maximized likelihood that estimate a parameter summarizing the model fit. A ratio of the index for two models summarizes the relative model fit. This can be helpful for choosing among models, including non-nested models such as the negative binomial model and the Poisson model with random effects for Table 1.

It is common in practice to use likelihood functions and their variants (conditional likelihood, marginal likelihood, profile likelihood, partial likelihood, ...) in a variety of ways, both formally for inference and informally for description and model selection in measures such as AIC, BIC, and Bayes factors. We do not claim any striking originality in the ideas that follow, but the measures presented may help practitioners get a feel for the magnitudes of reported log likelihoods.

## 2. Summarizing model fits

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be independent observations. Most applications have a variety of potential models for $\mathbf{y}$. Let $f_m(y_i; \theta_m)$ denote the probability density or mass function of $y_i$ for model $m$, where the parameters $\{\theta_m\}$ may refer to different types of models and different dimensions. Let $f(\cdot)$ refer to the true, but unknown, mass function of each $y_i$. Since in practice an unsaturated model $m$ only approximates reality, we treat $\theta_m$ as the probability limit of its maximum likelihood (ML) estimator $\hat{\theta}_m$. This is the value that minimizes the Kullback–Leibler information criterion

$$I(f, f_m) = E_f \log[f(Y)/f_m(Y; \theta_m)]$$

between $f_m$ and $f$, where $E_f$ denotes the expectation taken with respect to $f$ (White, 1982).

For model $m$, let $\ell_m(\hat{\theta}_m) = \Pi_i f_m(y_i; \hat{\theta}_m)$ be the maximized likelihood function, and let $L_m(\hat{\theta}_m) = \log[\ell_m(\hat{\theta}_m)]$. A difficulty in interpreting the values of $\ell_m(\cdot)$ and $L_m(\cdot)$ is that they depend on $n$. For models for discrete data, they decrease monotonically as $n$ increases. The scaled values

$$\hat{\gamma}_m = [\ell_m(\hat{\theta}_m)]^{1/n}, \quad \text{and} \quad \log(\hat{\gamma}_m) = L_m(\hat{\theta}_m)/n = \frac{1}{n}\sum_i \log f_m(y_i; \hat{\theta}_m) \tag{1}$$

remove the dependence on $n$. The geometric mean $[\ell_m(\hat{\theta}_m)]^{1/n}$ of the $n$ contributions to the maximized likelihood function for model $m$ is a sample version of a parameter

$$\gamma_m = \exp\{E_f[\log f_m(Y; \theta_m)]\}$$

summarizing model fit. Since $f$ is unknown, the sample version (1) takes this expectation with respect to the empirical distribution and at the value $\hat{\theta}_m$ for $\theta_m$. We refer to $\gamma_m$ as the *geometric mean likelihood* for model $m$. Being a re-scaling of the maximized likelihood, $\hat{\gamma}_m$ and $\gamma_m$ are non-decreasing as a particular model adds predictors.

In the discrete case, let $a_1, a_2, \ldots a_K$ denote the possible values for $Y$, with true probabilities $\{f(a_k)\}$. This true distribution has

$$\log(\gamma_f) = \sum_k f(a_k) \log f(a_k). \tag{2}$$

This is the maximum possible value for $\log(\gamma_m)$ and serves as a baseline. This distribution corresponds to an unspecified multinomial model, which is the saturated model. In the sample let $p_k$ denote the proportion of times that $a_k$ occurs, $k = 1, \ldots, K$. The sample value of $\log(\gamma_f)$ is $\sum_k p_k \log p_k$, a sample-size-standardized multinomial log likelihood used in measures of association (e.g., Theil, 1970). Similarly, $\log(\hat{\gamma}_m) = \sum_k p_k \log f_m(a_k; \hat{\theta}_m)$.

## 3. Comparing fits of two models

In practice, it is more informative to compare likelihood values for different models than to consider a model in isolation. A summary measure of the *relative model fit* for models $\ell$ and $m$ is

$$\rho_{\ell m} = \gamma_\ell / \gamma_m = \exp[E_f\{\log f_\ell(Y; \theta_\ell)] - E_f[\log f_m(Y; \theta_m)\}].$$

Its sample value is

$$\hat{\rho}_{\ell m} = \hat{\gamma}_\ell / \hat{\gamma}_m = \exp[\{L_\ell(\hat{\theta}_\ell) - L_m(\hat{\theta}_m)\}/n]$$

the ratio of geometric means of the contributions to the maximized likelihoods for the two models. The models do not need to be nested for $\rho_{\ell m}$ to be meaningful. In fact, the expectations in $\gamma_\ell$ and $\gamma_m$ could even refer to different true distributions, such as in comparing model fits for a particular model at different times based on

distinct samples. In the case of nested models, with model $\ell$ a special case of model $m$, $\log(\hat{\rho}_{\ell m})$ equals $(-1/2n)$ times the usual likelihood-ratio statistic.

In the discrete case, it can be informative to compare a model to the baseline of the saturated model, using $\hat{\rho}_{mf} = \hat{\gamma}_m/\hat{\gamma}_f$ for $\gamma_f$ in (2). This equals

$$\hat{\rho}_{mf} = \exp\left\{\sum_k p_k \log[f_m(a_k;\hat{\theta}_m)/p_k]\right\} = \exp(-D_m/2n),$$

where $D_m$ denotes the deviance for the model. Hence, the comparison of relative model fit using the saturated model as a baseline provides a sample-size-standardized interpretation of the deviance. A baseline model in the other direction is the null model having probability $K^{-1}$ at each $a_k$; for it, $\hat{\gamma}_0 = K^{-1}$ and $\hat{\rho}_{m0} = \exp\{\sum_k p_k \log[Kf_m(a_k;\hat{\theta}_m)]\}$.

In practice, observations are not typically *iid*, but the regression context is common, for which independent observations occur at different settings of predictor variables. The above sample estimators then refer to parameters that describe model fit of conditional distributions of the response, with expectations taken with respect to the joint distribution of all the variables. Specifically, if $X$ denotes a vector of explanatory variables, then for model $m$,

$$\gamma_m = \exp\{E[\log f_m(Y|X;\theta_m)]\}.$$

For instance, $\gamma_m$ is appropriate for surveys that randomly sample subjects and then classify those subjects on various predictors and response variables, since the empirical distribution estimates a true joint distribution of $(X, Y)$. Otherwise, when the expectation naturally applies only to $Y$, as in multi-center clinical trials, the indices apply to the empirical distribution on the predictor variables but the numerical value is not comparable to other samples with substantially different distributions of those predictor variables. Of course, this is true for most association measures, including correlation and R-squared measures.

## 4. A mean likelihood measure and related measures

Another index in the same spirit as $\gamma_m$ is $\mu_m = E[f_m(Y;\theta_m)]$. The sample version $\hat{\mu}_m = [\sum_i f_m(y_i;\hat{\theta}_m)]/n$ is the mean (rather than geometric mean) contribution to the maximized likelihood function. The measures of relative fit $\hat{\mu}_\ell/\hat{\mu}_m$ and $\hat{\mu}_\ell - \hat{\mu}_m$ compare models. These measures have a slightly simpler interpretation than ones based on $\hat{\gamma}_m$ involving geometric means, but they have the disadvantage of not being functionally related to the maximized likelihood. For instance, $\hat{\mu}_m$ need not necessarily increase as the model becomes more complex.

For discrete data, $\hat{\mu}_m$ relates to dissimilarity indices (Goodman and Kruskal, 1959). For observation $i$ that takes value $y_i$, the empirical distribution puts probability 1 at $y_i$ and 0 everywhere else. The amount of mass that needs to be moved from the fitted distribution for model $m$ to yield this empirical distribution equals $1 - f_m(y_i;\hat{\theta}_m)$. The mean of these in the sample, or $1 - \hat{\mu}_m$, is a dissimilarity measure of distance of $\mathbf{y}$ from the fitted distribution.

Although we have not seen $\gamma_m$, $\mu_m$, and related comparison measures directly used, other likelihood-based measures have been proposed for summarizing the fit of models. For instance, let $L_0$ denote the maximized log likelihood for some baseline model, such as a model having only an intercept parameter and no predictor effects. Let $L_s$ denote the log likelihood for the saturated model. Then, $D_m = -2(L_m - L_S)$ is the deviance for model $m$. The proportional reduction in deviance for the model of interest,

$$\frac{L_m - L_0}{L_s - L_0} = \frac{D_0 - D_m}{D_0}$$

is an alternative sample-size-adjusted index of model fit (Goodman, 1971). However, for different model forms, the null model differs, making comparison of values inappropriate. For instance, it is not appropriate to compare the value for a negative binomial model (in which the baseline model is negative binomial with only an intercept) to that for a Poisson generalized linear mixed model (GLMM); the measures are comparable only with the same baseline model, such as an ordinary Poisson generalized linear model (GLM) with only an intercept.

Other measures describe *predictive power* of a model, such as analogs of R-squared and the correlation between the observed response and the fitted value (e.g., Zheng and Agresti, 2000). Two models that provide the same fitted values necessarily have the same values of such as a measure, but they need not have the same values of measures such as $\hat{\gamma}$ because of the possible difference in likelihoods. In this sense, the measures in this article are indices of model fit and relative model fit as opposed to predictive power.

## 5. Example

We now return to Table 1 on $Y =$ annual numbers of homicide reports, for subjects classified by race. The sample mean was 0.52 for blacks and 0.09 for whites, with standard deviations of 1.07 and 0.39, respectively. A natural first choice for count data of this form is a Poisson GLM, such as a loglinear model with predictor a dummy variable for race. However, there is evidence of overdispersion, the sample variance being roughly double the mean for each race.

Models having an additional parameter to allow extra variability include the negative binomial and a Poisson GLMM. The Poisson GLMM is a mixture model whereby the Poisson applies, given the mean for a particular subject, and the log means for each race have a normal distribution. That is, the response for subject $i$ follows a Poisson distribution with an unknown mean $u_i$, where $\{\log(u_i)\}$ are independent $N(\mu_1, \sigma^2)$ for blacks and $N(\mu_2, \sigma^2)$ for whites. The negative binomial model results from a gamma mixture for $u_i$; this is a GLMM in which the log of the mean has a log gamma distribution, for each race. Such mixture models seem plausible here. Due to various unmeasured demographic factors, heterogeneity likely occurs among subjects of a given race in the distribution of $Y$.

Table 2 summarizes model fit for six models: the ordinary Poisson GLM, the Poisson GLMM, and the negative binomial model, each with and without an effect

Table 2
Indices of model fit and corresponding 95% confidence interval (CI) for Table 1

| | Model | | | | | |
|---|---|---|---|---|---|---|
| Index | Poisson GLM | | Poisson GLMM | | Negative binomial | |
| $L_m(\hat{\theta}_m)$ | −618.0 | −559.0 | −529.0 | −500.7 | −523.7 | −497.9 |
| $\hat{\gamma}_m$ | 0.623 | 0.652 | 0.667 | 0.682 | 0.670 | 0.683 |
| CI for $\gamma_m$ | (0.579,0.671) | (0.611,0.696) | (0.631,0.705) | (0.647,0.719) | (0.634,0.708) | (0.649,0.720) |
| $\hat{\mu}_m$ | 0.794 | 0.808 | 0.828 | 0.833 | 0.830 | 0.833 |
| CI for $\mu_m$ | (0.761,0.827) | (0.780,0.838) | (0.801,0.856) | (0.807,0.859) | (0.803,0.857) | (0.807,0.860)* |

*Note: For each model type, the first model is the null model and the second model has an effect for race. The Poisson models use the log link, with a normal random effect in the generalized linear mixed model (GLMM).

Table 3
Indices of relative model fit (compared to negative binomial model with race effect) and corresponding 95% confidence intervals, for Table 1

| | Model | | | | | |
|---|---|---|---|---|---|---|
| Index | Poisson GLM | | Poisson GLMM | | Negative binomial | |
| $\hat{\rho}_{m6} = \hat{\gamma}_m/\hat{\gamma}_6$ | 0.912 | 0.954 | 0.976 | 0.998 | 0.980 | 1.0 |
| CI for $\rho_{m6}$ | (0.879,0.946) | (0.932,0.977) | (0.964,0.989) | (0.994,1.002) | (0.969,0.992) | — |
| $\hat{\mu}_6 - \hat{\mu}_m$ | 0.039 | 0.025 | 0.006 | 0.001 | 0.003 | 0.0 |
| CI for $\mu_6 - \mu_m$ | (0.026,0.053) | (0.016,0.035) | (0.002,0.008) | (−0.002,0.002) | (0.001,0.006)* | — |

*Note: For each model type, the first model is the null model and the second model has an effect for race.

for race. It reports the maximized log likelihood function $L_m(\hat{\theta}_m)$ and the estimated geometric mean likelihood $\hat{\gamma}_m$. (It also reports a 95% confidence interval for $\gamma_m$ discussed in Section 6). Under the negative binomial model with race effect, for instance, the estimated geometric mean probability of the response was $\hat{\gamma}_6 = 0.683$. From $\{\hat{\gamma}_m\}$, the need to accommodate the overdispersion is clear. For instance, $\hat{\gamma}_m$ is higher for a negative binomial model with no effect than for a Poisson GLM with the race effect.

The two models suggested by $\{\hat{\gamma}_m\}$ are the Poisson GLMM and the negative binomial model with a race effect. The values $\hat{\gamma}_4 = 0.682$ and $\hat{\gamma}_6 = 0.683$ show no practical difference between the two in terms of this criterion. That both models fit relatively well is highlighted by noting that the general multinomial model for these data has a log likelihood of −489.5 and $\hat{\gamma}_f = 0.688$, barely larger than $\hat{\gamma}_4$ and $\hat{\gamma}_6$. Table 2 also contains the mean likelihood values $\{\hat{\mu}_m\}$ for the six models. They suggest the same conclusions.

Table 3 summarizes the estimated relative model fit using $\hat{\rho}_{m6}$ and $\hat{\mu}_6 - \hat{\mu}_m$ with the negative binomial model as the baseline. This also shows the Poisson GLMM and negative binomial models as roughly comparable in this summary sense.

## 6. Confidence intervals for measures of relative fit

The measures $\hat{\gamma}_m$, $\hat{\mu}_m$, $\hat{\rho}_{\ell m}$, and $\hat{\mu}_\ell - \hat{\mu}_m$ are descriptive summaries of model fit that help us interpret the numerical value of the maximized likelihood. We intend them as descriptive rather than inferential tools. However, in some cases it may be useful to construct a confidence interval for corresponding true values in the sampled population.

A straightforward way to construct a large-sample confidence interval for $\gamma_m$ exploits the sample mean representation for $\log \hat{\gamma}_m$. By the central limit theorem, the asymptotic distribution of

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\log f_m(y_i;\theta_m) - E_f[\log f_m(Y;\theta_m)]\right]$$

is normal with mean 0 and variance having unbiased estimator

$$\left[\sum_i \{\log f_m(y_i;\theta_m) - L_{\mathrm{m}}(\theta_m)/n\}^2\right] \Big/ (n-1).$$

A consistent estimator of this is

$$s_m^2 = \left[\sum_i \{\log f_m(y_i;\hat{\theta}_m) - L_{\mathrm{m}}(\hat{\theta}_m)/n\}^2\right] \Big/ (n-1).$$

By standard arguments (e.g., Linhart, 1988), this is also the asymptotic distribution of

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\log f_m(y_i;\hat{\theta}_m) - E_f\{\log f_m(Y;\theta_m)\}\right] = \sqrt{n}[\log(\hat{\gamma}_m) - \log(\gamma_m)].$$

Thus, a large-sample $100(1-\alpha)\%$ confidence interval for $\gamma_m$ exponentiates endpoints of $\log(\hat{\gamma}_m) \pm z_{\alpha/2}s_m/\sqrt{n}$.

For a pair of models $\ell$ and $m$, standard methods can also generate a confidence interval for $\rho_{\ell m}$. Let $s_{\ell m}$ denote the sample standard deviation of $\log f_\ell(y_i;\hat{\theta}_\ell) - \log f_m(y_i;\hat{\theta}_m)$, $i=1,\ldots,n$. Then, a large-sample confidence interval for $\rho_{\ell m}$ exponentiates endpoints of $\log(\hat{\rho}_{\ell m}) \pm z_{\alpha/2}s_{\ell m}/\sqrt{n}$.

One could judge model $\ell$ to be "better" than model $m$ when the confidence interval for $\rho_{\ell m}$ contains only numbers larger than 1.0. In practice, however, other considerations may be more relevant. For instance, if $\rho_{\ell m} < 1$ but is very close to 1 and model $\ell$ is simpler and easier to interpret, it might be preferred. Also, if model $\ell$ is a special case of model $m$, then $\rho_{\ell m} \leqslant 1$ yet model $\ell$ may be preferred for finite $n$ because of the usual advantages of model parsimony, such as having better estimates of $\{f(a_k),\ k=1,\ldots,K\}$. Hence, although $\rho_{\ell m}$ helps us interpret the relative fits of two models, we do not propose basing model selection on inferential results for $\rho_{\ell m}$.

Table 2 shows 95% confidence intervals for $\gamma_m$ for the various models for Table 1. Table 3 shows intervals for $\rho_{m6}$ for comparing models to the negative binomial model. The interval for $\rho_{46}$ comparing the Poisson GLMM to the negative binomial model is (0.994, 1.002). All values in this interval being very close to 1.0 suggests

the models give comparable fits in the population in terms of the geometric mean likelihood summary.

Although these methods for constructing confidence intervals are straightforward, our evidence is that sample sizes may need to be quite large for actual coverage probabilities to be near nominal confidence levels. For instance, although substituting $\hat{\theta}_m$ in $(1/n)\sum_{i=1}^{n} \log f_m(y_i;\theta_m)$ makes no ifference asymptotically, for small to moderate samples it can affect the coverage properties. To examine the coverage issue, we conducted various simulation studies, of which we describe one below. Coverage probabilities tended to be too low unless $n$ was large. Coverages tended to be more accurate for the single-model measure $\gamma_m$ than for $\rho_{\ell m}$ comparing two models.

To illustrate, one simulation study treated counts in Table 1 as representing a "true" distribution. We simulated 100,000 random samples from a joint distribution having probabilities equal to $\{\text{count}/1308\}$. We calculated the sample proportion of times that the 90%, 95%, 99% confidence interval for each measure contained the true value, for the models discussed in Section 5. With Monte Carlo error of no more than about 0.003, estimated coverage probabilities for 95% intervals with $n=(50,300,1308)$ were (0.865,0.939,0.949) for $\gamma_4$ with the Poisson GLMM and (0.862,0.938,0.948) for $\gamma_6$ with the negative binomial model (both with race effect) but only (0.818,0.925,0.933) for $\rho_{46}$ comparing the two models. Results were somewhat better for simple models without predictors and worse for pairs of models of different form with different predictors. In this example, for instance, the expected percentage of observations for the black sample was only 12%, and the imbalance in allocated sample sizes could adversely affect results. Without the predictor, estimated coverage probabilities with $n=(50,300,1308)$ were (0.896,0.942,0.949) for $\gamma_3$ with the Poisson GLMM, (0.884,0.941,0.949) for $\gamma_5$ with the negative binomial model, and (0.827,0.944,0.954) for $\rho_{35}$ comparing the models.

Alternatively, for interval estimation we have used bootstrap methods such as a non-parametric percentile bootstrap and the Efron BCa bootstrap (Efron and Tibshirani, 1993). Results for the example are similar to those reported in Tables 2 and 3. These methods are computationally intensive (especially when used with models that themselves require substantial computation, such as the Poisson GLMM), and the limited simulation studies we have done have insufficient precision to determine if they perform better than those based on standard asymptotics. The bootstrap is also the method we used for interval estimation of $\hat{\mu}_m$ and $\hat{\mu}_\ell - \hat{\mu}_m$, since the above asymptotic argument does not apply. Tables 2 and 3 show percentile bootstrap intervals for the $\mu_m$ measures and their comparisons using the negative binomial baseline model.

Incidentally, similar problems with achieving nominal error rates have been noted in the literature on tests comparing non-nested models $f_1$ and $f_2$. We summarize briefly this literature. The ordinary likelihood-ratio test does not apply for testing the null hypothesis that $f_1$ holds against the alternative that $f_2$ holds. Cox (1962) formed an approximately standard normal null test statistic that compares the log-likelihood ratio to its estimated expectation under the null model. He also mentioned the possibility (later pursued by Atkinson, 1970 and others) of imbedding

the two models into a family of models $[f_1(y;\theta_1)]^{\lambda} f_2(y;\theta_2)]^{1-\lambda}$ and making inferences about $\lambda$ with standard methods. Since Cox's paper, most related work is in the econometrics literature for normal error models. For instance, Davidson and MacKinnon (1981) compared normal but possibly non-linear regression models by using a model based on a weighted average of the two and conducting inference about the weight; their method relates to Atkinson's (1970) in certain special cases. Fisher and McAleer (1981) gave a related approach and Royston and Thompson (1995) considered adjustments to improve distributional properties. Godfrey and Pesaran (1983) proposed mean- and variance-adjusted Cox tests to reduce the small-sample bias of that statistic, and Victoria-Feser (1997) proposed robust versions of the Cox tests. McAleer (1995) presented several examples of non-nested models, summarized proposed methods of comparison, and discussed uses of the methods in various applications.

## 7. Related problems for future research work

Based on results in Section 6, scope exists for improving confidence intervals for the measures proposed in this paper. Standard methods are simple, but evidence exists from our simulations and from methods deriving from Cox's testing work that it is not easy to construct inferential methods for non-nested models that achieve close to nominal error rates when $n$ is not large (see, e.g., McAleer, 1995).

Not surprisingly, in our simulation studies confidence intervals much more often overestimated than underestimated the true parameter value. The sample value $\log(\hat{\gamma}_m) = L_{\mathrm{m}}(\hat{\theta}_m)/n \geqslant L_{\mathrm{m}}(\theta_m)/n$, and hence it tends to overestimate $\log(\gamma_m)$. A useful topic for future work is to find adjustments of $\hat{\gamma}_m$ and $\hat{\rho}_{\ell m}$ that reduce bias, in terms of estimating $\gamma_m$ and $\rho_{\ell m}$. The jackknife is one possibility, but again our simulation study had insufficient precision to determine if it has better performance.

Introducing a penalty for the number of parameters may also help with bias reduction. A related area for future work relates to AIC-type corrections involving the likelihood such as have traditionally been used to aid in model selection. Denote the number of parameters in model $m$ by $p_m$. By the same arguments that apply to Akaike's AIC index (see, e.g., Burnham and Anderson 1998), a corrected value for $\hat{\gamma}_m$ is

$$\hat{\gamma}_m^c = \exp\{[L_{\mathrm{m}}(\hat{\theta}_m) - p_m]/n\} = \exp(-p_m/n)\hat{\gamma}_m. \tag{3}$$

For the definition of AIC for model $m$ as $AIC_m = L_{\mathrm{m}}(\hat{\theta}_m) - p_m$ (Some sources define it as $-2[L_{\mathrm{m}}(\hat{\theta}_m) - p_m]$), $\hat{\gamma}_m^c = \exp(AIC_m/n)$ is a sample-size-scaled version of AIC. Similarly, a corrected estimate of the index of relative model fit is

$$\hat{\rho}_{\ell m}^c = \hat{\gamma}_{\ell}^c/\hat{\gamma}_m^c = \exp[(p_m - p_{\ell})/n]\hat{\rho}_{\ell m} = \exp[(AIC_{\ell} - AIC_m)/n].$$

For Table 1, $n$ is large and such AIC-type corrections have no substantive impact.

## Acknowledgements

This research was partially supported by grants from NSF and NIH. The authors acknowledge helpful comments from referees, such as pointing out the econometric literature on tests comparing models.

## References

Atkinson, A.C., 1970. A method of discriminating between models (with discussion). J. Roy. Statist. Soc. B 32, 323–353.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: A Practical Information-Theoretic Approach. New York: Springer.

Cox, D.R., 1962. Further results on tests of separate families of hypotheses. J. Roy. Statist. Soc. B 24, 406–424.

Davidson, R., MacKinnon, J.G., 1981. Several tests for model specification in the presence of alternative hypotheses. Econometrica 49, 781–794.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. London: Chapman & Hall.

Fisher, G.R., McAleer, M., 1981. Alternative procedures and associated tests of significance for non-nested hypotheses. J. Econometrics 16, 103–119.

Godfrey, L.G., Pesaran, M.H., 1983. Tests of non-nested regression models: Small sample adjustments and Monte Carlo evidence. J. Econometrics 21, 133–154.

Goodman, L.A., 1971. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics 13, 33–61.

Goodman, L.A., Kruskal, W.H., 1959. Measures of association for cross classifications, II: further discussion and references. J. Amer. Statist. Assoc. 54, 123–163.

Linhart, H., 1988. A test whether two AIC's differ significantly. South African Statist. J. 22, 153–161.

McAleer, M., 1995. The significance of testing empirical non-nested models. J. Econometrics 67, 149–171.

Royston, P., Thompson, S.G., 1995. Comparing non-nested regression models. Biometrics 51, 114–127.

Theil, H., 1970. On the estimation of relationships involving qualitative variables. Amer. J. Sociol. 76, 103–154.

Victoria-Feser, M.-P., 1997. A robust test for non-nested hypotheses. J. Roy. Statist. Soc. Ser. B 59, 715–727.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–26.

Zheng, B., Agresti, A., 2000. Summarizing the predictive power of a generalized linear model. Statist. Med. 19, 1771–1781.