

# The In-and-Out-of-Sample (IOS) Likelihood Ratio Test for Model Misspecification

Brett Presnell and Dennis D. Boos\*

## Abstract

A new test of model misspecification, based on the ratio of in-sample and out-of-sample likelihoods, is proposed. The test is broadly applicable, and in simple problems approximates well known, intuitive methods. Using jackknife influence curve approximations, it is shown that the test statistic can be viewed asymptotically as a multiplicative contrast between two estimates of the information matrix that are equal under correct model specification. This approximation is used to show that the statistic is asymptotically normally distributed, though it is suggested that p-values be computed using the parametric bootstrap. The resulting methodology is demonstrated with a variety of examples and simulations involving both discrete and continuous data.

KEY WORDS: Goodness of fit; lack of fit; cross-validation; jackknife; parametric bootstrap; generalized linear model; sandwich matrix.

Institute of Statistics Mimeo Series No. 2536

August 2002

---

\*Brett Presnell is Associate Professor, Department of Statistics, University of Florida, Gainesville, FL 32611-8545 (E-mail: [presnell@stat.ufl.edu](mailto:presnell@stat.ufl.edu)). Dennis D. Boos is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (E-mail: [boos@stat.ncsu.edu](mailto:boos@stat.ncsu.edu)). The authors thank Alan Agresti for helpful comments and suggestions, Jim Booth for suggesting the leukemia example, and Sujit Ghosh for directing us to related work in Bayesian model selection. The first author is also grateful to the Department of Statistics at North Carolina State University for their support and hospitality during a sabbatical visit when much of this work was accomplished.

# 1 Introduction

## 1.1 Motivation

Model misspecification is always a concern when statistical inferences are based on a parametric likelihood. Even when a model manages to capture the gross features of the data, misspecification of distributional forms, dependence structures, link functions, and other “structural” components of the model can cause estimates to be biased and their precision to be misrepresented. Other inferences based on likelihood are similarly affected, including tests of nested hypotheses within the model.

These concerns have led to a long and continuing history of “specification tests” in the econometrics literature, beginning with Hausman (1978) and White (1982). The statistics literature too has seen a renewed interest in general methods for testing model adequacy, as exemplified by Bayarri and Berger (2000) and Robins, van der Vaart and Ventura (2000).

Nevertheless, a useful general approach to testing for model misspecification is hard to find. A major difficulty is that there is often no obvious way to embed distributional assumptions and other structural components of the assumed model into a larger model, with the aim of treating this larger model as the “full model” and the assumed model as the “reduced model” in a comparison of nested models. With categorical data, the “saturated model” can sometimes play the role of the larger model (see Agresti, 2002, pp. 139–141), but not when there are continuous predictors or otherwise too many settings of explanatory variables. For continuous data, no such approach is generally available.

To overcome these difficulties, we suggest comparing the maximized model likelihood to a corresponding “likelihood” motivated by cross-validation. To set notation, let  $Y_1, \dots, Y_n$  be independent random variables with hypothesized densities  $f(y; x_i, \theta)$ , where  $x_i$  is a possible predictor for  $Y_i$  and  $\theta$  is an unknown  $p$ -dimensional parameter. Note that  $Y_i$  may be discrete (in which case its density might be called a probability mass function), that both  $Y_i$  and  $x_i$  may be vector valued, and that we take the  $x_i$  to be nonrandom, so that our analysis is conditional on the values of any predictors. Let  $\hat{\theta}$  be the maximum likelihood estimator (MLE) of  $\theta$ , and let  $\hat{\theta}_{(i)}$  be the MLE when the  $i$ th observation is deleted from the sample.

To motivate our test statistic, note that  $f(Y_i; x_i, \hat{\theta}_{(i)})$ , the estimated likelihood of  $Y_i$  given the model and the remaining data, is a measure of how well the model is able to predict the  $i$ th observation. In particular, if  $f(Y_i; x_i, \hat{\theta})$  is much larger than  $f(Y_i; x_i, \hat{\theta}_{(i)})$ , then the fitted model must shift appreciably in order to accommodate the  $i$ th observation, suggesting that the model is in some way inadequate to describe the data. This leads us to compare the ordinary maximized, or “in-sample” likelihood,  $\prod_{i=1}^n f(Y_i; x_i, \hat{\theta})$ , with the cross-validated, or “out-of-sample” likelihood,  $\prod_{i=1}^n f(Y_i; x_i, \hat{\theta}_{(i)})$ , as a global measure of model adequacy. Our proposed test statistic is thus the logarithm of the “in-and-out-of-sample” (IOS) likelihood

ratio,

$$\text{IOS} = \log \left( \frac{\prod_{i=1}^n f(Y_i; x_i, \hat{\theta})}{\prod_{i=1}^n f(Y_i; x_i, \hat{\theta}_{(i)})} \right) = \sum_{i=1}^n \{l(Y_i; x_i, \hat{\theta}) - l(Y_i; x_i, \hat{\theta}_{(i)})\}, \quad (1)$$

where  $l(Y_i; x_i, \theta) = \log f(Y_i; x_i, \theta)$ . Note that IOS is always nonnegative, because

$$\begin{aligned} \sum_{j=1}^n l(Y_j; x_j, \hat{\theta}_{(i)}) - l(Y_i; x_i, \hat{\theta}_{(i)}) &\geq \sum_{j=1}^n l(Y_j; x_j, \hat{\theta}) - l(Y_i; x_i, \hat{\theta}) && \text{by definition of } \hat{\theta}_{(i)} \\ &\geq \sum_{j=1}^n l(Y_j; x_j, \hat{\theta}_{(i)}) - l(Y_i; x_i, \hat{\theta}) && \text{by definition of } \hat{\theta}, \end{aligned}$$

which implies that  $l(Y_i; x_i, \hat{\theta}) \geq l(Y_i; x_i, \hat{\theta}_{(i)})$  for all  $i = 1, \dots, n$ .

## 1.2 Some Simple Examples

In simple cases, the IOS statistic often approximates well-known, intuitive statistics, as seen in the following examples.

**Example 1 (Poisson Distribution).** The assumed probability mass function is  $f(y; \lambda) = e^{-\lambda} \lambda^y / y!$  with log likelihood  $-n\lambda + \sum_{i=1}^n Y_i \log(\lambda) - \sum_{i=1}^n \log(Y_i!)$  maximized at  $\hat{\lambda} = \bar{Y}$ , the sample mean. Then

$$\text{IOS} = \sum_{i=1}^n (\bar{Y}_{(i)} - \bar{Y}) + \sum_{i=1}^n Y_i \log(\bar{Y} / \bar{Y}_{(i)}) \approx \frac{s^2}{\bar{Y}},$$

where  $s^2$  is  $\sum (Y_i - \bar{Y})^2 / (n - 1)$  and we have used the fact that  $\bar{Y}_{(i)} - \bar{Y} = (\bar{Y} - Y_i) / (n - 1)$  and the Taylor expansion  $\log(\bar{Y}_{(i)}) \approx \log(\bar{Y}) + (\bar{Y}_{(i)} - \bar{Y}) / \bar{Y}$ . Of course  $s^2 / \bar{Y}$  is the familiar Fisher (1973, p. 58) statistic (divided by  $n - 1$ ) for testing goodness of fit for the Poisson distribution.

**Example 2 (Exponential Distribution).** The assumed density is  $f(y; \sigma) = \exp(-y/\sigma) / \sigma$  with log likelihood  $-\sum_{i=1}^n Y_i / \sigma - n \log(\sigma)$  maximized at  $\hat{\sigma} = \bar{Y}$ . Then

$$\text{IOS} = \sum_{i=1}^n Y_i \left( \frac{1}{\bar{Y}_{(i)}} - \frac{1}{\bar{Y}} \right) - \sum_{i=1}^n \log(\bar{Y} / \bar{Y}_{(i)}) \approx \frac{s^2}{\bar{Y}^2},$$

using approximations as in the previous example. Here again we see an obvious comparison of sample moments motivated by the exponential moment relationship  $\text{var}(Y_i) = \{E(Y_i)\}^2$ .

**Example 3 (Normal Distribution).** The assumed density is  $f(y; \mu, \sigma) = \exp\{-(y - \mu)^2 / 2\sigma^2\} / \sqrt{2\pi}\sigma$  with MLE's  $\hat{\mu} = \bar{Y}$  and  $\hat{\sigma}^2 = \sum (Y_i - \bar{Y})^2 / n$ . Somewhat lengthy calculations lead to

$$\text{IOS} \approx \frac{1}{2} \left( \frac{\hat{\mu}_4}{\hat{\sigma}^4} + 1 \right),$$

where  $\hat{\mu}_4 = \sum_{i=1}^n (Y_i - \bar{Y})^4/n$ . It may seem surprising that, in the limit, the IOS statistic for normality depends only on sample kurtosis and not on sample skewness, but as we look at the in-probability limit of the statistic below, the technical reason becomes clear. Note that under normality, the sample kurtosis converges in probability to 3 and IOS  $\xrightarrow{P} 2$  as  $n \rightarrow \infty$ .

**Example 4 (Normal Regression through the Origin).** The assumed density is normal with mean  $\beta x_i$  and variance  $\sigma^2$ , where  $\sum_{i=1}^n x_i = 0$ . The approximate version of IOS given in Section 2 leads to

$$\text{IOS} \approx \frac{\sum_{i=1}^n \hat{e}_i^2 x_i^2}{\hat{\sigma}^2 \sum_{i=1}^n x_i^2} + \frac{1}{2} \left( \frac{\hat{\mu}_4}{\hat{\sigma}^4} + 1 \right),$$

where  $\hat{\beta} = \sum_{i=1}^n Y_i x_i / \sum_{i=1}^n x_i^2$ ,  $\hat{e}_i = Y_i - \hat{\beta} x_i$ ,  $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \hat{e}_i^2$ , and  $\hat{\mu}_4 = n^{-1} \sum_{i=1}^n \hat{e}_i^4$ . Thus IOS retains the kurtosis term but adds a term that is related to the linear model specification, especially the homogeneity of errors assumption (cf. White, 1980).

Under the null hypothesis of correct model specification, the IOS statistic in each of these examples converges in probability to  $p$ , the dimension of the parameter vector  $\theta$ . This is no accident. To see why, consider the case of independent and identically distributed (iid)  $Y_1, \dots, Y_n$  with no predictors. Let  $I(\theta) = E\{-\ddot{l}(Y_1; \theta)\}$  and  $B(\theta) = E\{\dot{l}(Y_1; \theta)\dot{l}(Y_1; \theta)^T\}$ , where  $\dot{l}(y; \theta) = \partial l(y; \theta)/\partial \theta$ , the  $p \times 1$  gradient vector of partial derivatives of  $l(y; \theta)$  with respect to the elements of  $\theta$ , and where  $\ddot{l}(y; \theta) = \partial^2 l(y; \theta)/\partial \theta \partial \theta^T$ , the  $p \times p$  Hessian matrix of second-order partial derivatives. Finally, let  $\theta_0$  represent the in-probability limit of  $\hat{\theta}$ , assumed to exist even when  $f(y; \theta)$  is not correctly specified. Then Theorems 2 and 4 of Section 4 imply that under suitable regularity conditions,

$$\text{IOS} \xrightarrow{P} \text{IOS}_\infty = E\{\dot{l}(Y_1; \theta_0)^T I(\theta_0)^{-1} \dot{l}(Y_1; \theta_0)\} = \text{tr}\{I(\theta_0)^{-1} B(\theta_0)\} \quad (2)$$

as  $n \rightarrow \infty$ , where  $\text{tr}(A)$  denotes the trace of a matrix  $A$ . Of course under correct specification,  $I(\theta_0) = B(\theta_0)$  and  $\text{IOS}_\infty = p$ , the trace of the  $p$ -dimensional identity matrix.

In typical situations where the model is misspecified,  $\text{IOS}_\infty > p$ , but this is not always the case. For example, in the iid Normal model of Example 3,  $\text{IOS}_\infty = (\mu_4/\mu_2^2 + 1)/2$ , where  $\mu_k = E[\{Y_1 - E(Y_1)\}^k]$ , so that  $\text{IOS}_\infty < p = 2$  whenever the true distribution has kurtosis  $\mu_4/\mu_2^2 < 3$ , i.e., whenever the tails of the distribution are lighter than those of the normal distribution.

### 1.3 Links with Other Methods

Equation (2) suggests a connection between IOS and the Information Matrix (IM) Test of White (1982). The IM test is based on a quadratic form in the vector of differences between elements of  $\hat{B}(\hat{\theta}) = n^{-1} \sum_{i=1}^n \dot{l}(Y_i; \hat{\theta}) \dot{l}(Y_i; \hat{\theta})^T$  and the (average) observed information matrix  $\hat{I}(\hat{\theta}) = n^{-1} \sum_{i=1}^n \{-\ddot{l}(Y_i; \hat{\theta})\}$ . Thus the IM test works by differencing two estimates of the Fisher information matrix, both of which are consistent under correct specification. On

the other hand, as we show in Section 4, IOS can be approximated for large samples by  $\text{tr}\{\widehat{I}(\widehat{\theta})^{-1}\widehat{B}(\widehat{\theta})\}$ , which is effectively a matrix ratio of two estimates of the Fisher information.

This approximate form provides another motivation for IOS. Recall that when the distributional form of the model is incorrectly specified, it still may be possible to carry out (asymptotically) valid inference using a sandwich estimator of the covariance matrix of  $\widehat{\theta}$ , though generally with some loss in efficiency if the parametric model happens to be correct. By contrasting the model dependent estimate,  $\widehat{I}(\widehat{\theta})$ , of  $\text{var}\{\dot{l}(Y_1; \theta_0)\}$  with the empirical estimate,  $\widehat{B}(\widehat{\theta})$ , IOS, like the IM Test, can indicate situations where model-based inference is invalid.

Of course the IOS test differs from the IM test in important ways. IOS is defined directly in terms of the likelihood, and is easily motivated and understood without any reference to asymptotic arguments. Calculation of IOS is very similar to the computation of a jackknife bias or variance estimator. This can be computationally intensive, particularly for large sample sizes and particularly after embedding in a bootstrap loop, but the computations are easy given a reliable routine for calculating the MLE. Moreover, if the routine uses numerical derivatives, then no analytical expressions are required for even first derivatives of the log-density. For large sample sizes, the approximate form of IOS can be used and is also relatively straightforward to calculate, though second derivatives of the log-likelihood are required.

Computation of the IM test statistic is less automatic. A number of different forms of the IM statistic have been suggested, all requiring at least second order partial derivatives, and for some forms third order derivatives of the log-likelihood (see Horowitz, 1994, for references and further details). One must also be careful to remove any structural zeros and linear dependencies in the elements of the differenced information matrix estimates. Thus, though there are undoubtedly situations in which the IM test will have greater power than the IOS test, in general IOS is simpler to use.

IOS has connections to a variety of model selection procedures as well. Stone (1977) gives a heuristic derivation of (2) and shows that when the model is correctly specified, the out-of-sample log-likelihood is asymptotically equivalent to Akaike's (1973) information criterion. Linhart and Zucchini (1986) also discuss use of the out-of-sample log-likelihood as a model selection criterion. The approximate form of IOS,  $\text{tr}\{\widehat{I}(\widehat{\theta})^{-1}\widehat{B}(\widehat{\theta})\}$ , arises as the "trace term" in the asymptotic model selection criterion of Linhart and Zucchini (1986, Section 4.1.1), and examples discussed there include the Poisson and normal examples above.

Geisser and Eddy (1979) discuss the use of out-of-sample likelihoods as model selection criteria, but focus primarily on Bayesian approaches. They refer to the out-of-sample likelihood as the *predictive sample reuse (PSR) quasi-likelihood* selection criterion and to its Bayesian analogue as the *PSR quasi-Bayes* criterion. The ratio of PSR quasi-Bayes criteria for two competing models is called a *pseudo-Bayes factor* by Gelfand and Dey (1994), who consider both the in-sample likelihood and the out-of-sample likelihood as approximations to

the PSR quasi-Bayes criterion.

In these references, the out-of-sample likelihood is considered solely for model selection; that is, for choosing between two or more competing models. It is not compared to the in-sample likelihood or employed in any other way to test for model misspecification. Geisser (1989, 1990) uses Bayesian analogues of the individual IOS factors,  $f(Y_i; x_i, \hat{\theta})/f(Y_i; x_i, \hat{\theta}_{(i)})$ , to identify outlying, or “discordant” observations. This is perhaps closer in spirit to the IOS test, but is distinguished by the Bayesian emphasis and more importantly by the focus on discerning individual discordant observations as opposed to measuring model adequacy in a global way.

We emphasize here that we do not propose to use IOS as a model selection tool in the usual sense of choosing variables for inclusion in a model, nor do we intend or expect IOS to usurp the role of the likelihood ratio test and its usual competitors in comparing nested models. IOS is useful for testing the fit of the of the final model selected in such a comparison, or perhaps of the largest model under consideration, but a more narrowly focused test should nearly always have greater power in a nested comparison. Indeed, in many situations IOS may have no power whatsoever to detect that an important variable has been excluded from the model.

In the remainder of this paper we investigate the IOS test from both a practical and a theoretical viewpoint. Section 2 provides an overview of technical results, with further details postponed to Section 4. In Section 3 we illustrate this procedure with a number of examples, and provide the results of several small simulation studies of the size and power of the test in settings motivated by the examples. Finally, in Section 5, we summarize our findings and discuss possible extensions of the IOS test to other settings, including dependent data models and partial likelihood.

## 2 Overview of Theoretical Results

In this section we provide an overview of theoretical results concerning IOS, with precise statements of theorems and their proofs postponed until Section 4. To simplify the statements of theorems and their proofs, we restrict attention to the case of iid random vectors,  $Y_1, \dots, Y_n$ .

The asymptotic form of the IOS statistic is

$$\text{IOS}_A = \frac{1}{n} \sum_{i=1}^n i(Y_i; \hat{\theta})^T \hat{I}(\hat{\theta})^{-1} i(Y_i; \hat{\theta}) = \text{tr}\{\hat{I}(\hat{\theta})^{-1} \hat{B}(\hat{\theta})\}.$$

In Section 4, we prove the following results:

**Theorem 1:** In location-scale families, the null distributions of IOS and  $\text{IOS}_A$  are independent of the true parameter values.

**Theorem 2:**  $\text{IOS}_A \xrightarrow{P} \text{IOS}_\infty$ , where  $\text{IOS}_\infty$  is defined in (2).

**Theorem 3:**  $\text{IOS}_A$  is asymptotically normally distributed.

**Theorem 4:**  $\text{IOS} - \text{IOS}_A = o_p(n^{-1/2})$ , and thus the conclusions of Theorems 2 and 3 also apply to  $\text{IOS}$ .

The form of the asymptotic variance of  $\text{IOS}$ , derived in the appendix, is fairly complicated for routine use. Moreover, examples like the sample kurtosis in Example 3 indicate that the convergence to normality is slow. Thus we suggest that p-values be obtained via the parametric bootstrap under the hypothesized model (Horowitz, 1994, makes the same suggestion for the IM test). Theorem 1 implies that in location-scale models these p-values are exact up to Monte Carlo error. More generally, the p-values are asymptotically uniform under the hypothesized model, and simulations indicate that they are quite accurate in small samples as well.

Clearly the approximation of  $\text{IOS}$  by  $\text{IOS}_A$  is critical to our results.  $\text{IOS}_A$  arises naturally from Taylor expansion of the log-density about  $\hat{\theta}$ ,

$$l(Y_i; \hat{\theta}_{(i)}) = l(Y_i; \hat{\theta}) + \dot{l}(Y_i; \hat{\theta})^T (\hat{\theta}_{(i)} - \hat{\theta}) + \frac{1}{2} (\hat{\theta}_{(i)} - \hat{\theta})^T \ddot{l}(Y_i; \tilde{\theta}_i) (\hat{\theta}_{(i)} - \hat{\theta}), \quad (3)$$

where  $\tilde{\theta}_i$  lies between  $\hat{\theta}_{(i)}$  and  $\hat{\theta}$ , and from the influence curve approximation

$$\hat{\theta}_{(i)} - \hat{\theta} = -\frac{1}{n} \hat{I}(\hat{\theta})^{-1} \dot{l}(Y_i; \hat{\theta}) + R_{ni}. \quad (4)$$

Substituting (3) and (4) into the definition (1) of the  $\text{IOS}$  statistic, we have

$$\begin{aligned} \text{IOS} &= -\sum_{i=1}^n \dot{l}(Y_i; \hat{\theta})^T (\hat{\theta}_{(i)} - \hat{\theta}) - \frac{1}{2} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta})^T \ddot{l}(Y_i; \tilde{\theta}_i) (\hat{\theta}_{(i)} - \hat{\theta}) \\ &= \text{IOS}_A - \sum_{i=1}^n \dot{l}(Y_i; \hat{\theta})^T R_{ni} - \frac{1}{2} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta})^T \ddot{l}(Y_i; \tilde{\theta}_i) (\hat{\theta}_{(i)} - \hat{\theta}). \end{aligned} \quad (5)$$

The influence curve approximation (4) is the key to developing the approximation of  $\text{IOS}$  by  $\text{IOS}_A$ , and deserves further discussion here. It is easily shown that whenever the MLE is the sample mean,  $\bar{Y}$ , then

$$\hat{\theta}_{(i)} - \hat{\theta} = \bar{Y}_{(i)} - \bar{Y} = -\frac{1}{n-1} (Y_i - \bar{Y}) = -\frac{1}{n-1} \hat{I}(\hat{\theta})^{-1} \dot{l}(Y_i; \hat{\theta}).$$

Thus (4) is exact in this case, with  $R_{ni} = 0$ , if we replace the factor  $1/n$  by  $1/(n-1)$ . This suggests the approximation  $\text{IOS} \approx \{n/(n-1)\} \text{IOS}_A$ , which may be more accurate than  $\text{IOS} \approx \text{IOS}_A$ , but because the difference in the approximations is asymptotically negligible, we have chosen to simplify expressions by replacing the factor  $1/(n-1)$  with  $1/n$ .

More generally, (4) arises from Taylor expansion of the estimating equation for  $\widehat{\theta}_{(i)}$ :

$$\begin{aligned} 0 &= \sum_{j \neq i} \dot{l}(Y_j; \widehat{\theta}_{(i)}) = \sum_{j \neq i} \dot{l}(Y_j; \widehat{\theta}) + \sum_{j \neq i} \ddot{l}(Y_j; \check{\theta}_i) (\widehat{\theta}_{(i)} - \widehat{\theta}) \\ &= -\dot{l}(Y_i; \widehat{\theta}_{(i)}) - n\widehat{I}(\widehat{\theta})(\widehat{\theta}_{(i)} - \widehat{\theta}) - \left[ \sum_{j \neq i} \{-\ddot{l}(Y_j; \check{\theta}_i)\} - n\widehat{I}(\widehat{\theta}) \right] (\widehat{\theta}_{(i)} - \widehat{\theta}), \end{aligned} \quad (6)$$

where  $\check{\theta}_i$  lies between  $\widehat{\theta}_{(i)}$  and  $\widehat{\theta}$ . Here we have used the fact that  $\sum_{j \neq i} \dot{l}(Y_j; \widehat{\theta}) = -\dot{l}(Y_i; \widehat{\theta})$ , which follows from the estimating equation for  $\widehat{\theta}$ . Equation (6) implies that

$$\widehat{\theta}_{(i)} - \widehat{\theta} = -\frac{1}{n} \widehat{I}(\widehat{\theta})^{-1} \dot{l}(Y_i; \widehat{\theta}) - \widehat{I}(\widehat{\theta})^{-1} W_{ni} (\widehat{\theta}_{(i)} - \widehat{\theta}), \quad (7)$$

where

$$W_{ni} = \frac{1}{n} \sum_{j \neq i} [-\ddot{l}(Y_j; \check{\theta}_i)] - \widehat{I}(\widehat{\theta}), \quad (8)$$

and from this (4) follows with

$$R_{ni} = -\widehat{I}(\widehat{\theta})^{-1} W_{ni} (\widehat{\theta}_{(i)} - \widehat{\theta}). \quad (9)$$

A subtle but important detail is the use of  $\widehat{I}(\widehat{\theta})$  instead of  $I(\widehat{\theta})$  in the approximation. This is necessary in order to obtain  $\sqrt{n} \max_{1 \leq i \leq n} |(W_{ni})_{kl}| \xrightarrow{P} 0$ , which in turn is needed to prove in Theorem 4 that  $\text{IOS}$  and  $\text{IOS}_A$  differ only by  $o_p(n^{-1/2})$ .

### 3 Examples and Simulations

The examples in this section demonstrate application of the IOS test in a variety of settings. Simulation results are also given, investigating the size and power of the IOS test in settings motivated by the examples. All computations were done using the R system for statistical computation and graphics (Ihaka and Gentleman, 1996).

As discussed previously, we use parametric bootstrap p-values. In most of the examples we have used 4000 bootstrap samples to estimate bootstrap p-values, as this gives adequate accuracy for almost any purpose. In order to reduce computation times, we have generally used fewer bootstrap samples in our simulation studies. However, the minimum number of bootstrap samples used is 199, so that the accuracy of our estimates of test size and power is not too seriously affected (see Boos and Zhang, 2000, for details).

#### 3.1 A Simple Gamma Example

The data in Table 1 are taken from Larsen and Marx (2001, page 320) and record the maximum 24 hour precipitations for the 36 hurricanes that moved inland to the Appalachian

Table 1: Maximum 24 hour precipitation for 36 inland hurricanes (1900–1969).

31.00	2.82	3.98	4.02	9.50	4.50	11.40	10.71	6.31	4.95	5.64	5.51
13.40	9.72	6.47	10.16	4.21	11.60	4.75	6.85	6.25	3.42	11.80	0.80
3.69	3.10	22.22	7.43	5.00	4.58	4.46	8.00	3.73	3.50	6.20	0.67

Mountains in the period from 1900 to 1969. Larsen and Marx (2001) use the method of moments to fit a gamma distribution to these data.

Is the gamma model appropriate? The IOS test suggests not, with  $\text{IOS} = 3.60$  for this two parameter model, giving an (estimated) bootstrap p-value of .028 based on 4000 bootstrap samples. For this example the  $\text{IOS}_A$  statistic takes the value  $\text{IOS}_A = 2.84$ . Though this value does not seem particularly close to IOS, the bootstrap p-value is .022, in good agreement with the IOS p-value. For comparison, we also computed the Anderson-Darling statistic (.789) and the Kolmogorov-Smirnov statistic (.116), for which the bootstrap p-values were .044 and .274, respectively. Thus, of these tests, the IOS tests give the strongest evidence against the gamma model, with the Anderson-Darling test in reasonably close agreement. The Kolmogorov-Smirnov fails to find any evidence against the gamma model in this example.

The observations 31.00, 0.67, 22.22, and 0.80 make the largest contributions to IOS, 1.73, 0.49, 0.45, and 0.38 respectively. After deletion of the largest observation, 31.00, the IOS and  $\text{IOS}_A$  tests still yield mild evidence against a gamma model for the remaining data, with bootstrap p-values of .061 and .053, respectively, while the p-value for the Anderson-Darling test increases to .164.

### 3.1.1 Simulation

To examine the size of the IOS test in this context, we ran a small simulation with 4000 samples of size 36 drawn from a gamma distribution with shape parameter 2.187 equal to the maximum likelihood estimate for the rainfall data. For each sample, bootstrap p-values for the Anderson-Darling, the Kolmogorov-Smirnov, the IOS, and the  $\text{IOS}_A$  tests were estimated using 199 bootstrap samples. The estimated size of the nominal .05 level tests were .049, 0.055, 0.047, and 0.047, respectively, all well within simulation error of .05.

## 3.2 Leukemia Survival Times

Feigl and Zelen (1965) applied an exponential regression model (with identity link) to the survival times of thirty-three leukemia patients, using the logarithm of white blood count (WBC) and a binary factor (AG) as predictors. Patients were identified as AG positive by the presence of Auer rods and/or significant granulation of leukemic cells at diagnosis. There were 17 AG positive and 16 AG negative patients.

These data have since been analyzed by many authors. Aitkin, Anderson, Francis and Hinde (1989) fit a variety of generalized linear models to these data, modeling the conditional

survival distribution variously as gamma, Weibull, and lognormal (a Gaussian model fitted to  $\log(\text{time})$ ). They found that the exponential (with log link) was very nearly the best fitting model within both the gamma and the Weibull families. Using graphical methods, they found no evidence of lack of fit for any of these models. Comparing maximized likelihoods, they stated that the lognormal provided a slightly better fit than the exponential/gamma/Weibull model, but they did not judge the difference to be significant.

We applied the IOS test to the gamma model with log link and to the lognormal model with identity link. The linear predictor included a WBC:AG interaction term, so that both models have five parameters to be estimated. For both models we used 4000 bootstrap replications to estimate the p-value for the IOS test.

For the gamma model,  $\text{IOS} = 15.74$ . For 10 of the bootstrap samples, calculation of either the full MLE or at least one of the 33 leave-one-out estimators failed. Of the remaining 3990 bootstrap samples, 125 yielded a value of the IOS statistic greater than the observed value. Thus the estimated p-value is about .031 (or conservatively, .034), and there is appreciable evidence against the gamma model. For the lognormal model,  $\text{IOS} = 7.29$ , and the estimated bootstrap p-value is .22, providing no evidence against the model.

Does it matter which model we use for these data? We think that it does. The lognormal model provides more precise estimates in this example than the gamma model rejected by the IOS test. For example, in the gamma model, the Wald test yields a p-value of .25 for the AG:WBC interaction. Thus, the data analyst might be inclined to drop this term from the model and declare that the slopes for the two AG groups are not significantly different. Indeed, this is one of the models suggested by Aitkin et al. (1989) in the exponential case. In the lognormal model, however, the Wald test yields a p-value of .08 for the interaction, so that the data analyst might think more carefully before dropping it.

Recognizing the interaction is important in this example. The slope for the AG negative group is not significantly different from 0 in either model (the p-value for the Wald test is .48 in the gamma model and .34 in the lognormal model). Thus, for the AG negative group there appears to be at best a weak dependence of survival time on WBC, whereas in the AG positive group the relationship is significant. Indeed, this was perhaps the most important findings in the original analysis of Feigl and Zelen (1965). Though a careful data analyst might uncover this feature of the data with either the gamma or the lognormal model, there is a greater chance that it would go undetected if the gamma model were used.

### 3.2.1 Simulation

To study the size of the IOS test for the gamma model, we ran a simulation experiment in which 400 vectors of 33 responses were generated from the fitted gamma model, with the predictors fixed at the values observed in the data. For each simulated sample, bootstrap p-values were estimated using 199 parametric bootstrap resamples.

For 16 of the 400 simulated samples, we were unable to compute IOS either for the sample itself, or for at least one of the corresponding bootstrap samples. This may seem surprising, but for each simulated sample we must fit a gamma model both to the full data and to the 33 delete-one subsamples, and then repeat this procedure 199 times for the bootstrap samples. Thus we must compute maximum likelihood estimates for the gamma model for a total of  $34 \times 200 = 6800$  data sets for each simulated sample. Maximum likelihood estimation of the regression and shape parameters of a gamma model is not so stable that this process can always be counted on to succeed without any human intervention. In practice, one can simply keep track of the number of bootstrap samples for which the computation of IOS failed and report a suitably adjusted p-value as we have done in the example, but for this small simulation we decided to simply discard these samples. Of the remaining 384 simulated samples, 23 yielded an estimated bootstrap p-value less than .05. Thus the estimated size of the IOS test is .06, well within simulation error of the nominal .05 level.

For the lognormal model, maximum likelihood estimates are more easily computed, and here we used 1599 bootstrap resamples to estimate bootstrap p-values for each of 400 samples generated from the fitted normal model. In this case 22 of the estimated p-values were less than .05, so that the estimated size of the IOS test is .055, again well within simulation error of the nominal .05 level.

Finally, because the IOS test fails to reject the lognormal model in this example, the power of the IOS test of this model is of interest. For the alternative we chose to use a gamma model, and so for each of 400 samples generated from the fitted gamma model, we computed IOS for an assumed lognormal model. For each sample we again used 1599 bootstrap resamples to estimate the bootstrap p-value of the test. The power of the .05 level test was thus estimated as  $76/400 = .19$ . Of course it is difficult to say much about power on the basis of such a limited simulation, but with a five parameter regression model and a sample size of only 33, we judge this a reasonably positive performance, particularly given the dearth of competitors for the IOS test in this situation.

### 3.3 Drill Advance Rates

The data for this example are taken from Daniel (1976), and concern a  $2^4$  unreplicated factorial experiment to study the effect of four factors,  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$ , on drill advance rate. Box, Hunter and Hunter (1978) analyze these data using a lognormal model and find that only the main effects for  $x_2$ ,  $x_3$ , and  $x_4$  are important. Lewis, Montgomery and Myers (2001) fit a gamma model with a log link and also settle on  $x_2$ ,  $x_3$ , and  $x_4$  as predictors to be retained in the model. Lewis et al. (2001) prefer the gamma model to the lognormal because it yields shorter confidence intervals for the estimated mean at each of the 16 design points.

We tested the fit of both of these four parameter models using the IOS test. For the lognormal model,  $\text{IOS} = 8.32$  with an estimated bootstrap p-value of .67 based on 4000

Table 2: Free throws by basketball player Shaquille O’Neal in 23 playoff games.

Game	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Attempted	5	11	14	12	7	10	14	15	12	4	27	17	12	9	12	10	12	6	39	13	17	6	12
Made	4	5	5	5	2	7	6	9	4	1	3	5	6	9	7	3	8	1	8	3	0	1	3

bootstrap replications. For the gamma model,  $IOS = 8.37$  with an estimated bootstrap p-value of .64. Thus IOS finds no evidence against either of these models. This is perhaps not surprising given the small sample size, but it does suggest that the use of either model is justifiable.

### 3.4 Testing Multivariate Normality

Johnson and Wichern (1998) examine a data set consisting of four different measurements of stiffness taken on each of 30 wooden boards. Although the four variables taken individually appear to be marginally normal, a plot of Mahalanobis distances from the mean against quantiles of a chi-square distribution with 4 degrees of freedom suggests that at least two of the observations may be multivariate outliers.

We applied the IOS test and Mardia’s multivariate skewness and kurtosis tests (see Mardia, Kent and Bibby, 1979, section 5.7) to these data. Under multivariate normality, the distribution of all three of these statistics is free of the unknown mean and covariance matrix, so that parametric bootstrap p-values are exact up to simulation error. For the 14 parameter multivariate normal model,  $IOS = 30.7$ , with a p-value of .002 based on 4000 bootstrap replications. The multivariate skewness and kurtosis were 7.54 and 30.42 with bootstrap p-values of .008 and .001, respectively.

Johnson and Wichern (1998) identify observations 9 and 16 as outliers based on their Mahalanobis distances. This is supported by their large contributions, 5.1 and 13.5, respectively, to IOS. With these two observations deleted,  $IOS = 27.2$ , with a p-value of .006 based on 4000 bootstrap replications. Results for the skewness and kurtosis tests were similar, with bootstrap p-values of .004 and .006, respectively. Thus there is still strong evidence of non-normality, even after deletion of the two presumed outliers.

We have applied these three tests to a number of data sets, and generally p-values for the IOS test agrees very closely with the kurtosis test. For large sample sizes, IOS and the multivariate kurtosis appear to be nearly linearly related, and we speculate that  $IOS_A$  is indeed an affine function of Mardia’s multivariate kurtosis, although we have not verified this analytically.

### 3.5 A Simple Binomial Example

The data in Table 2, taken from Stefanski and Boos (2002), give the number of free throws attempted,  $n_i$ , and made,  $Y_i$ , in each of 23 consecutive games by professional basketball

Table 3: Size and power of nominal .05 level IOS and chi-squared tests in the free throw shooting example.

	Null	Random $p_i$	
	$p_i = .456$	Beta(2, 2.5)	Beta(9.5, 11.7)
IOS	0.049	0.982	0.478
$\chi^2$	0.049	0.991	0.478

player Shaquille O’Neal. We take as our null hypothesis that the  $Y_i$  are independent binomial random variables with common, unknown success probability. To test this hypothesis, Stefanski and Boos computed parametric bootstrap p-values for Pearson’s chi-squared statistic in the  $23 \times 2$  table (which is the score statistic in this example). We re-ran their analysis using 100,000 bootstrap samples, with the  $n_i$  fixed at their observed values for each game, and obtained a p-value of .028 for the chi-squared test. (Note that the p-value reported by Stefanski and Boos (2002) are incorrect because of a programming error discovered after publication.) The IOS statistic, on the other hand, takes the value 1.29, with a bootstrap p-value of .206. This suggests that the IOS test may have much lower power than the score test in the setting of this example, though we suspect that the difference in p-values is largely due to a greater sensitivity of the chi-square test to the extreme observation in game 14, in which O’Neal made 9 of 9 free throw attempts. To explore this issue further, we carried out a small simulation study.

### 3.5.1 Simulation

In each simulation the number of attempts were fixed at the observed values,  $n_1, \dots, n_{23}$ . Bootstrap samples were also generated with the same fixed binomial sample sizes, so that all analysis is conditional on the number of attempts. P-values were computed using 1000 parametric bootstrap resamples. All results are based on 20,000 simulated samples.

To study the size of the tests, we generated samples using a constant success probability,  $p = 0.456$ , equal to O’Neal’s overall free throw shooting percentage for the 23 games. The estimated sizes of the nominal .05 level tests, given in the first column of Table 3, are all well within probable simulation error of .05.

To study power, we let the success probabilities  $p_i$ ,  $1 \leq i \leq 23$ , vary independently from game to game according to a beta distribution, with the  $Y_i$  independent and binomially distributed conditional on the  $p_i$ . In the first simulation, we chose the parameters of the beta distribution to match the (unweighted) mean and variance of O’Neal’s observed success probabilities over the 23 games, 0.45 and 0.045, respectively. This corresponds to a beta distribution with parameters 2.0 and 2.5. Against this alternative, both the IOS and the chi-squared tests have power very close to 1, as shown in the fourth column of Table 3. To obtain a more informative comparison, we ran another simulation with beta distributed  $p_i$ , with the same mean as before, but with the variance reduced by a factor of four. This

Table 4: Beetle mortality data (Bliss, 1935).

Log Dose	Number of Beetles	Number Killed
1.691	59	6
1.724	60	13
1.755	62	18
1.784	56	28
1.811	63	52
1.837	59	53
1.861	62	61
1.884	60	60

corresponds to a beta distribution with parameters 9.5 and 11.7. The power of both the IOS and chi-squared tests against this alternative are moderate and nearly equal, as shown in the fifth column of Table 3. These results suggest that the power of the IOS test is comparable to that of the score test in this simple situation, though the latter may have greater power against a single extreme  $p_i$ .

### 3.6 Beetle Mortality Data

Table 4, taken from Agresti (2002, p. 247), reports the number of beetles killed after five hours of exposure to eight different concentrations of gaseous carbon disulphide (these data were originally reported by Bliss, 1935). Agresti finds that a binomial model with complementary log-log link fits these data very well, and that the model with logit link fits poorly. Specifically, the residual deviances, 3.5 for the model with complementary log-log link and 11.1 for the model with logit link, yield p-values of .74 and .085, respectively, when referred to the chi-square distribution with 6 degrees of freedom.

The IOS statistic is 1.45 for the model with complementary log-log link, and 4.07 for the model with logit link. The corresponding bootstrap p-values, estimated with 4000 bootstrap samples, are .71 and .136, respectively. The p-value for the logit model suggests that IOS may be somewhat less sensitive here than the deviance statistic. However, for the two largest dose levels the fitted counts in these models are very close to the number of beetles tested, so that the chi-square approximation to the distribution of the deviance statistic may not be adequate. The parametric bootstrap p-value for the deviance test of the logit model .111, in much closer agreement with the IOS p-value.

### 3.7 Horseshoe Crab Satellites

This example concerns data from Agresti (1996, p. 82–83) giving the number of satellite males for 173 nesting female horseshoe crabs. Agresti fits a Poisson model with log link to these data, using the carapace width of the female crab as predictor. To test the fit of this

model using the usual Pearson chi-square or residual deviance statistics, he is forced to pool the data over ranges of carapace width, and finds no evidence of lack-of-fit. Subsequently, however, he finds evidence of overdispersion, and prescribes adjustment of standard error estimates by an appropriate scaling factor.

Of course the continuous predictor is no hindrance to the IOS test. For this two parameter model,  $\text{IOS} = 5.55$ , suggesting a serious lack of fit, which is reinforced by the bootstrap: the estimated bootstrap p-value, based on 4000 bootstrap samples, was 0; the 99th percentile of the bootstrap IOS values was 3.20; only 5 of the 4000 bootstrap IOS values were greater than 4.0; and none were greater than 4.9.

For the negative binomial model with log link,  $\text{IOS} = 2.66$ . With 4000 bootstrap samples, the estimated p-value is .91, so there is no evidence against this model.

### 3.8 Testing Beta-Binomial Models

Binary data often exhibit overdispersion relative to a binomial model. Overdispersion may be caused, for example, by positive correlation of within litter responses in biological data. In such cases, beta-binomial models are often suggested as an alternative to the binomial (see Brooks, Morgan, Ridout and Pack, 1997; Slaton, Piegorsch and Durham, 2000).

Brooks et al. (1997) presented six data sets, labeled E1, E2, HS1, HS2, HS3, and AVSS, giving litters sizes and “success” counts for 205, 211, 524, 1328, 554, and 127 litters, respectively (Garren, Smith and Piegorsch, 2001, correct several errors in these data). Their analysis began by testing the goodness-of-fit of the beta-binomial distribution to each data set using the maximized likelihood as a test statistic. With this test they failed to find any evidence against the beta-binomial model for any of the six data sets. They then fitted several models to these data. In one particular model comparison, they found that for all but the AVSS data set, the beta-binomial model was significantly improved by a model mixing a beta-binomial distribution with a simple binomial component.

Garren, Smith and Piegorsch (2000) criticized Brooks, et al’s approach to testing goodness-of-fit for the beta-binomial model. In Garren et al. (2001) they proposed an alternative goodness-of-fit test based on pooling bootstrap p-values for Pearson chi-square statistics computed separately for each observed binomial sample size  $n$ . Garren et al. (2001) applied their test to each of the six data sets from Brooks et al. (1997) and to three similar data sets taken from Lockhart, Piegorsch and Bishop (1992), with counts from 50, 201, and 263 litters, respectively. In what follows these data sets are labeled LPB(a), LPB(b), and LPB(c).

We have applied the IOS and  $\text{IOS}_A$  tests to the nine data sets analyzed by Garren et al. (2001). The values of the IOS and  $\text{IOS}_A$  statistics are given in Table 5, together with estimated bootstrap p-values based on 400 bootstrap samples for IOS and 4000 bootstrap samples for  $\text{IOS}_A$ . For purposes of comparison, the bootstrap p-values computed by Garren et al. (2001) for their test procedure are also provided in Table 5 in the column labeled GSP.

Table 5: Comparison of IOS, IOS<sub>A</sub>, and the test of Garren et al. (2001) for nine data sets.

Data Set	No. of Litters	IOS	IOS <sub>A</sub>	p-value		
				IOS	IOS <sub>A</sub>	GSP
E1	205	2.58	2.45	.028	.027	.144
E2	211	2.36	2.26	.065	.076	.231
HS1	524	3.03	2.92	.000*	.000*	.000
HS2	1328	3.02	2.97	.000*	.000*	.000
HS3	554	3.12	3.05	.000*	.000*	.000
AVSS	127	2.04	1.94	.418	.433	.375
LPB(a)	50	2.12	1.91	.310	.283	.333
LPB(b)	201	2.09	2.02	.340	.309	.010
LPB(c)	263	4.73	4.38	.000*	.000*	.000

\* (No bootstrap IOS/IOS<sub>A</sub> exceeded observed value.)

The IOS and GSP tests agree except on all but the E1 and GSP(b) data sets, and perhaps the E2 data set. For the E1 data, the IOS test yields reasonably strong evidence against the beta-binomial model, whereas the GSP test does not. A more detailed examination of the terms of the IOS statistic reveals several observations that appear to contribute excessively to the observed value of IOS, including in particular a single observation,  $(n, y) = (14, 9)$ , that contributes 0.61. Recalculating IOS after removal of this observation yields an IOS value of 2.34, and a bootstrap p-value of .105 (again based on 400 bootstrap replications). For the LPB(b) data, the GSP test yields strong evidence against the beta-binomial model where IOS and IOS<sub>A</sub> find none. Finally, for the E2 data set, the IOS test perhaps casts some doubt on the beta-binomial model, while the GSP test certainly does not.

Thus it appears that IOS and the GSP test perform similarly, although each test may be sensitive to alternatives to which the other is not. Of course the IOS test is based on a general approach to testing for model misspecification, whereas the GSP test seems to be useful only for testing simple binomial or beta-binomial models. In particular, while IOS can be used in an automatic way in regression settings, the GSP test does not seem to generalize in this direction, except perhaps in situations with heavily replicated design points.

As an example, we have applied the IOS test to the Heckman-Willis model used by Slaton et al. (2000) to analyze data from a dose response experiment. In this model, it is assumed that, conditional on the litter size,  $N$ , the number of “successes,”  $Y$ , in a litter follows a beta-binomial distribution, with parameters depending on the value of the predictor  $x = \text{dose}$ . More precisely, it is assumed that  $Y$  has probability mass function

$$f(y|x) = \binom{N}{y} \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + y)\Gamma(\beta + N - y)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + N)}, \quad y = 0, \dots, N,$$

where

$$\alpha = \exp(a_0 + a_1x) \quad \text{and} \quad \beta = \exp(b_0 + b_1x),$$

Table 6: P-values for three tests of the adequacy of separate binomial models for the four dose groups in the boric acid data. P-values for the IOS and score tests are based on 4000 bootstrap samples. The likelihood ratio test takes the beta-binomial as the larger model and uses the chi-square distribution with one degree of freedom as its reference distribution.

Dose	p-values		
	IOS	Score	LR
0.0	.188	.205	.443
0.1	.479	.493	1.000
0.2	.074	.127	.310
0.4	.000	.000	.000

and  $a_0$ ,  $a_1$ ,  $b_0$ , and  $b_1$  are unknown parameters to be estimated.

The data, given in Table 4 of Slaton et al. (2000), are from an experiment in developmental toxicology and consist of observations on 107 litters at four different dose levels of boric acid. For the four-parameter Heckman-Willis model,  $IOS = 6.34$ , with a bootstrap p-value of .043 based on 400 bootstrap samples. For the asymptotic version of the test statistic, we observed  $IOS_A = 5.19$ . In one of 4000 bootstrap samples the observed Fisher information matrix was numerically singular preventing calculation of  $IOS_A$ . For 109 of the remaining 3999 bootstrap samples, the bootstrap value of  $IOS_A$  exceeded the observed value, yielding an estimated bootstrap p-value of .027 (a conservative estimate of the bootstrap p-value is  $110/4000 = .0275$ ). Thus there is fairly strong evidence against the Heckman-Willis model for these data.

Slaton et al. (2000) used a likelihood ratio test to compare the the Heckman-Willis model with a more general beta-binomial model which shared with the Heckman-Willis model the logistic form of the mean as a function of dose, but which allowed the intralitter correlation to vary freely among the four different dose levels (thus they compared the four parameter Heckman-Willis model to a larger six parameter beta-binomial model). The resulting p-value of .32 indicated no departure from the Heckman-Willis model, but of course this test assumes that the larger model is correctly specified. Our analysis suggests otherwise.

Considering the four dose groups separately, the p-values given in Table 6 show little or no evidence against a binomial model for each of the first three dose levels (0, 0.1, and 0.2 percent boric acid in feed), while for the fourth dose group there is very strong evidence of extra-binomial variation. However, even the beta-binomial does not adequately describe the fourth dose group, as for this model  $IOS = 4.46$ , with an estimated p-value of .013 based on 400 bootstrap samples.

The implied logistic mean function of the Heckman-Willis model also appears inappropriate for these data. In Figure 1 we show the overall proportions of dead embryos for the four groups along with the fitted mean curve from the Heckman-Willis model. The error bars represent deviations of two standard errors on either side of the first three dose groups' overall proportions, with the standard errors based on a simple binomial model for each

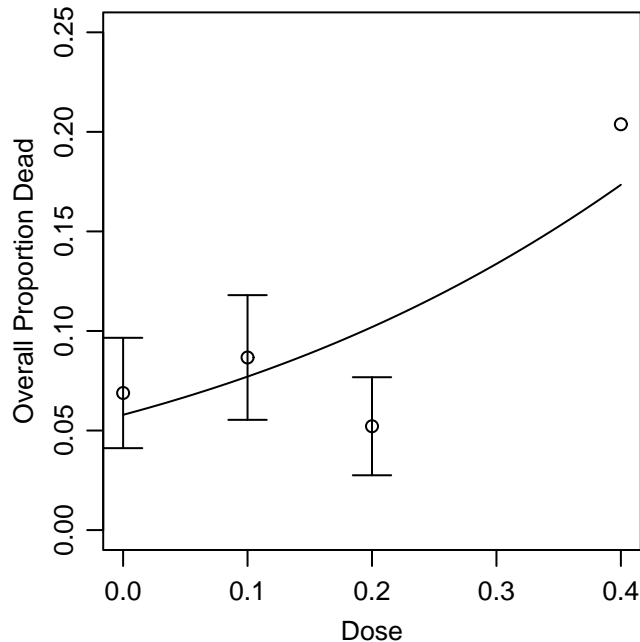


Figure 1: Overall group proportions and fitted mean curve from Heckman-Willis model for boric acid data. Error bars represent two standard errors to either side of the overall proportion.

group. It seems clear that the logistic mean function is unable to accommodate the observed proportions, particularly for dose level 0.2.

Judging from the overall proportions for the four dose groups, it appears that there is a threshold between dose levels 0.2 and 0.4 below which there is little or no discernible response to the treatment. To test this hypothesis, we applied the IOS test to a single binomial model for the combined data from the first three dose levels, obtaining  $\text{IOS} = 1.20$  with an estimated p-value of .13 based on 4000 bootstrap samples. Similarly the estimated bootstrap p-value for the score test is .19. The likelihood ratio statistic for testing the null hypothesis of a simple binomial model versus a beta-binomial alternative is 1.21 on one degree of freedom, with a chi-square p-value of .27. Finally, assuming a binomial model for each of the three dose groups, Pearson's chi-square statistic for testing the null hypothesis of no difference in death rates between the first three groups takes the value 2.62 on two degrees of freedom, yielding a p-value of .27. Thus, there seems to be no evidence against a single binomial model for the combined data from the first three dose levels, while neither the binomial nor the beta-binomial distribution adequately models the fourth dose group.

## 4 Technical Results

In this section we provide detailed statements and proofs of the theoretical results outlined in Section 2. Unless otherwise specified,  $\theta$  will be an unknown  $p$ -dimensional parameter vector belonging to a set  $\Theta \subset \mathbb{R}^p$ . The parametric models of interest have densities  $f(y; \theta)$  with respect to some dominating measure, but we will not emphasize measure-theoretic aspects. Recall that we take  $l(y; \theta) = \log f(y; \theta)$ ,  $\dot{l}(y; \theta) = \partial l(y; \theta) / \partial \theta$ , and  $\ddot{l}(y; \theta) = \partial^2 l(y; \theta) / \partial \theta \partial \theta^T$ . The maximum likelihood estimator is  $\hat{\theta}$ , and we assume that  $\hat{\theta}$  exists and solves the likelihood equations  $\sum_{i=1}^n \dot{l}(Y_i; \hat{\theta}) = 0$ . The matrix  $I(\theta) = E\{-\ddot{l}(Y_1; \theta)\}$  will be called the information matrix even in cases where the assumed density is not the true density (misspecification), though it is more common to define the information matrix to be  $B(\theta) = E\{\dot{l}(Y_1; \theta)\dot{l}(Y_1; \theta)^T\}$ . Here and throughout this paper, expectations are taken with respect to the true underlying distribution of the  $Y_i$ , which is not necessarily a member of the assumed model family  $\{f(y; \theta) : \theta \in \Theta\}$ . The (average) observed information matrix is  $\hat{I}(\hat{\theta}) = n^{-1} \sum_{i=1}^n \{-\ddot{l}(Y_i; \hat{\theta})\}$ , and we let  $\hat{B}(\hat{\theta}) = n^{-1} \sum_{i=1}^n \dot{l}(Y_i; \hat{\theta})\dot{l}(Y_i; \hat{\theta})^T$ . Finally,  $\theta_0$  represents the in-probability limit of  $\hat{\theta}$ , assumed to exist even when  $f(y; \theta)$  is not correctly specified.

For a vector  $x \in \mathbb{R}^p$ ,  $\|x\|_r$  represents the  $L_r$  norm, i.e.,  $\|x\|_r = (\sum_{i=1}^p |x_i|^r)^{1/r}$ . The  $L_r$  norm on  $\mathbb{R}^p$  induces the  $r$ -norm on the space of  $p \times p$  matrices, defined by  $\|A\|_r = \sup_{\|x\|_r=1} \|Ax\|_r$ ,  $A \in \mathbb{R}^{p \times p}$  (see Golub and van Loan, 1989, for details).

This first theorem shows that in location-scale models the distributions of IOS and IOS<sub>A</sub> do not depend on the true parameter values when the model is correctly specified. Thus the null distributions can be simulated exactly, and exact (up to simulation error) p-values can be obtained.

**Theorem 1.** *Suppose that  $Y_1, \dots, Y_n$  are iid with location-scale density  $f(y; \theta) = \sigma^{-1} f_0((y - \mu)/\sigma)$ , where  $f_0$  is a known density with derivative  $\dot{f}_0(y)$  existing and such that the maximum likelihood estimator  $\hat{\theta} = (\hat{\mu}, \hat{\sigma})^T$  exists and solves the likelihood equations. Then the distributions of IOS and IOS<sub>A</sub> do not depend on the value of  $\theta = (\mu, \sigma)^T$ .*

*Proof.* For arbitrary real numbers  $a > 0$  and  $b$ , and data  $\mathbf{Y} = (Y_1, \dots, Y_n)$ , it is easy to show that  $\hat{\mu}(a\mathbf{Y} + b) = a\hat{\mu}(\mathbf{Y}) + b$ ,  $\hat{\sigma}(a\mathbf{Y} + b) = a\hat{\sigma}(\mathbf{Y})$ ,  $l(ay + b; \hat{\theta}(a\mathbf{Y} + b)) = l(y; \hat{\theta}(\mathbf{Y})) - \log(a)$ ,  $\dot{l}(ay + b; \hat{\theta}(a\mathbf{Y} + b)) = a^{-1}\dot{l}(y; \hat{\theta}(\mathbf{Y}))$ , and  $\hat{I}_{a\mathbf{Y}+b}(\hat{\theta}(a\mathbf{Y} + b)) = a^{-2}\hat{I}_{\mathbf{Y}}(\hat{\theta}(\mathbf{Y}))$ . It then follows easily that  $\text{IOS}(a\mathbf{Y} + b) = \text{IOS}(\mathbf{Y})$  and  $\text{IOS}_A(a\mathbf{Y} + b) = \text{IOS}_A(\mathbf{Y})$ .  $\square$

The next theorem gives a consistency result for IOS<sub>A</sub> under both misspecified and correctly specified models. Recall that expectations are taken with respect to the true underlying distribution of the  $Y_i$ , which is not necessarily a member of the assumed model family  $\{f(y; \theta) : \theta \in \Theta\}$ .

**Theorem 2.** *Let  $Y_1, \dots, Y_n$  be iid. Assume that  $l(y; \theta)$  has three partial derivatives with respect to  $\theta$  for all  $\theta \in \Theta$ . Assume further that:*

1. There exists  $\theta_0$  such that  $\widehat{\theta} \xrightarrow{P} \theta_0$  as  $n \rightarrow \infty$ .
2.  $I(\theta_0) = E\{-\ddot{l}(Y_1; \theta_0)\}$  is finite and nonsingular.
3.  $B(\theta_0) = E\{\dot{l}(Y_1; \theta_0)\dot{l}(Y_1; \theta_0)^T\}$  is finite.
4. There exists a function  $C(y)$  such that for all  $\theta$  in an open neighborhood of  $\theta_0$  and for all  $j, k, l \in \{1, \dots, p\}$ ,  $|\partial \ddot{l}(y; \theta)_{jk} / \partial \theta_l| \leq C(y)$ , and  $E\{C(Y_1)\} < \infty$ .
5. There exists a function  $D(y)$  such that for all  $\theta$  in an open neighborhood of  $\theta_0$  and for all  $j, k, l \in \{1, \dots, p\}$ ,  $|\ddot{l}(y; \theta)_{jk} \dot{l}(y; \theta)_l| \leq D(y)$ , and  $E\{D(Y_1)\} < \infty$ .

Under the above conditions,  $IOS_A \xrightarrow{P} IOS_\infty$  as  $n \rightarrow \infty$ .

*Proof.* Note that

$$\begin{aligned} & \left| \widehat{I}(\widehat{\theta})_{jk} - I(\theta_0)_{jk} \right| \\ & \leq \left| \frac{1}{n} \sum_{i=1}^n \{-\ddot{l}(Y_i; \widehat{\theta})_{jk}\} - \frac{1}{n} \sum_{i=1}^n \{-\ddot{l}(Y_i; \theta_0)_{jk}\} \right| + \left| \frac{1}{n} \sum_{i=1}^n \{-\ddot{l}(Y_i; \theta_0)_{jk}\} - I(\theta_0)_{jk} \right|. \end{aligned}$$

The second term is  $o_p(1)$  by the weak law of large numbers and Condition 2. By the mean value theorem and Condition 1, the first term can be bounded with probability converging to 1 by  $\{n^{-1} \sum_{i=1}^n C(Y_i)\} \|\widehat{\theta} - \theta_0\|_1$ , which, by Conditions 1 and 4 and the weak law of large numbers, is  $O_p(1) \cdot o_p(1) = o_p(1)$ . Thus  $\widehat{I}(\widehat{\theta}) \xrightarrow{P} I(\theta_0)$  as  $n \rightarrow \infty$ . Using Conditions 3 and 5 in place of 2 and 4, a similar argument shows that  $\widehat{B}(\widehat{\theta}) \xrightarrow{P} B(\theta_0)$  as  $n \rightarrow \infty$ . The theorem then follows from continuity of the trace and matrix inversion.  $\square$

*Remark 1.* There is no loss of generality in the requirement that the bounding functions  $C(y)$  and  $D(y)$  above do not depend on  $j, k$ , and  $l$ . If integrable bounding functions can be found for each  $j, k, l$  combination, we can then take  $C(y) = \max_{j,k,l} C_{jkl}(y)$ . The same remark applies to Lemma 1 and Theorem 4 below.

The following general result will be applied in Theorem 3 to prove asymptotic normality of  $IOS_A$ .

**Lemma 1.** *Suppose that  $Y_1, \dots, Y_n$  are iid and assume that*

1. The  $p \times 1$  estimator  $\widehat{\theta}$  has an influence curve approximating function  $h(y; \theta)$  such that

$$\widehat{\theta} - \theta_0 = \frac{1}{n} \sum_{i=1}^n h(Y_i; \theta_0) + R_{n1},$$

where  $\sqrt{n}R_{n1} \xrightarrow{P} 0$  as  $n \rightarrow \infty$ ,  $E\{h(Y_1; \theta_0)\} = 0$ , and  $\text{cov}\{h(Y_1; \theta_0)\}$  is finite.

2. The real-valued function  $q(Y_i; \theta)$  has two partial derivatives with respect to  $\theta$ , and

(a)  $\text{var}\{q(Y_1; \theta_0)\}$  and  $E\{\dot{q}(Y_1; \theta_0)\}$  are finite.

(b) there exists a function  $M(y)$  such that for all  $\theta$  in a neighborhood of  $\theta_0$  and all  $j, k \in \{1, \dots, p\}$ ,  $|\ddot{q}(y; \theta)_{jk}| \leq M(y)$ , where  $E\{M(Y_1)\} < \infty$ .

Then

$$\frac{1}{n} \sum_{i=1}^n q(Y_i; \hat{\theta}) - E\{q(Y_1; \theta_0)\} = \frac{1}{n} \sum_{i=1}^n Q(Y_i; \theta_0) + R_{n2},$$

where  $Q(y; \theta) = q(y; \theta) - E\{q(Y_1; \theta)\} + E\{\dot{q}(Y_1; \theta)\}^T h(y; \theta)$  and  $\sqrt{n}R_{n2} \xrightarrow{P} 0$  as  $n \rightarrow \infty$ , and it follows that  $n^{-1} \sum_{i=1}^n q(Y_i; \hat{\theta})$  is asymptotically normal with asymptotic mean  $E\{q(Y_1; \theta_0)\}$  and asymptotic variance  $\text{var}\{Q(Y_1; \theta_0)\}/n$ .

*Proof.* By Taylor expansion and adding and subtracting terms, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n q(Y_i; \hat{\theta}) &= \frac{1}{n} \sum_{i=1}^n [q(Y_i; \theta_0) + E\{\dot{q}(Y_1; \theta_0)\}^T h(Y_i; \theta_0)] \\ &\quad + \left[ \frac{1}{n} \sum_{i=1}^n \dot{q}(Y_i; \theta_0) - E\{\dot{q}(Y_1; \theta_0)\} \right]^T (\hat{\theta} - \theta_0) \\ &\quad + E\{\dot{q}(Y_1; \theta_0)\}^T \left\{ \hat{\theta} - \theta_0 - \frac{1}{n} \sum_{i=1}^n h(Y_i; \theta_0) \right\} \\ &\quad + \frac{1}{2} (\hat{\theta} - \theta_0)^T \left\{ \frac{1}{n} \sum_{i=1}^n \ddot{q}(Y_i; \tilde{\theta}) \right\} (\hat{\theta} - \theta_0), \end{aligned}$$

where  $\tilde{\theta}$  lies between  $\hat{\theta}$  and  $\theta_0$ . Using Conditions 1 and 2 and the asymptotic normality of  $\hat{\theta}$  that follows from Condition 1, it is straightforward to show that the last three terms are  $o_p(n^{-1/2})$  as  $n \rightarrow \infty$ . The result then follows immediately from the central limit theorem.  $\square$

The next theorem establishes the asymptotic normality of  $\text{IOS}_A$ . Here  $\text{vech}$  represents the usual column stacking operator for symmetric matrices (see Harville, 1997, section 16.4).

**Theorem 3.** Let  $Y_1, \dots, Y_n$  be iid. Suppose that Condition 1 of Lemma 1 holds for the maximum likelihood estimator  $\hat{\theta}$  with  $h(y; \theta_0) = I(\theta_0)^{-1} \dot{l}(y; \theta_0)$ , and that Condition 2 of Lemma 1 holds for both  $q(y; \theta) = -\ddot{l}(y; \theta)_{jk}$  and  $q(y; \theta) = \{\dot{l}(y; \theta) \dot{l}(y; \theta)^T\}_{jk}$  for each  $j, k \in \{1, \dots, p\}$ . Then  $\text{IOS}_A$  is asymptotically normal with asymptotic mean  $\text{IOS}_\infty$  and asymptotic variance  $D^T A D/n$ , where  $A/n$  is the asymptotic covariance matrix arising from the joint asymptotic normality of  $\text{vech}\{\hat{I}(\hat{\theta})\}$  and  $\text{vech}\{\hat{B}(\hat{\theta})\}$ , and  $D$  is the  $p(p+1)$  dimensional vector of partial derivatives of  $\text{IOS}_A$  taken with respect to the components of  $\text{vech}\{\hat{I}(\hat{\theta})\}$  and  $\text{vech}\{\hat{B}(\hat{\theta})\}$  and evaluated at their limits in probability.

*Proof.* Joint asymptotic normality of the elements of  $\text{vech}\{\widehat{I}(\widehat{\theta})\}$  and  $\text{vech}\{\widehat{B}(\widehat{\theta})\}$  follows from Lemma 1. The result then follows from a direct application of the delta method to  $\text{IOS}_A = \text{tr}\{\widehat{I}(\widehat{\theta})^{-1}\widehat{B}(\widehat{\theta})\}$ .  $\square$

The conditions on  $\widehat{\theta}$  in Theorem 3 may be satisfied by modifications of the usual conditions for asymptotic normality of  $\widehat{\theta}$  (e.g., pages 462–463 of Lehmann and Casella, 1998) to allow for misspecification. Since the form of the asymptotic variance  $D^T AD/n$  is fairly complicated, we defer further computations to the appendix.

Our last theorem establishes equivalence of IOS and  $\text{IOS}_A$  to order  $o_p(n^{-1/2})$ .

**Theorem 4.** *Suppose that the conditions of Theorem 2 hold. Assume further that:*

1.  $\sqrt{n} \max_{1 \leq i \leq n} \|\widehat{\theta}_{(i)} - \widehat{\theta}\|_2 \xrightarrow{P} 0$  as  $n \rightarrow \infty$ .
2. There exists a function  $G(y)$  such that for all  $\theta$  in an open neighborhood of  $\theta_0$  and for all  $j \in \{1, \dots, p\}$ ,  $|\dot{l}(y; \theta)_j| \leq G(y)$ , and  $E\{G(Y_1)^2\} < \infty$ .
3. There exists a function  $H(y)$  such that for all  $\theta$  in an open neighborhood of  $\theta_0$  and for all  $j, k \in \{1, \dots, p\}$ ,  $|\ddot{l}(y; \theta)_{jk}| \leq H(y)$ , and  $E\{H(Y_1)^2\} < \infty$ .

Then

$$\sqrt{n}(\text{IOS} - \text{IOS}_A) \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty.$$

If further, the conditions of Theorem 3 also hold, then IOS is asymptotically normal with asymptotic mean  $\text{IOS}_\infty$  and asymptotic variance  $D^T AD/n$ , where  $D$  and  $A$  are as defined in Theorem 3.

*Proof.* The result will follow if we can show the two remainder terms in (5) are  $o_p(n^{-1/2})$ . Note that by Condition 1 of Theorem 2 and Condition 1 above, with probability tending to one,  $\widehat{\theta}$ ,  $\widehat{\theta}_{(i)}$ , and all values between  $\widehat{\theta}_{(i)}$  and  $\widehat{\theta}$ ,  $1 \leq i \leq n$ , lie within the neighborhoods of  $\theta_0$  where the bounds in Conditions 4 and 5 of Theorem 2 and in Conditions 2 and 3 of the present theorem apply. In what follows, we will use these bounds freely without restating the qualifier that the resulting inequalities hold only with probability tending to one. This is permissible because we are concerned only with convergence in probability.

As mentioned in Section 2, a key quantity in the proof is the matrix  $W_{ni}$  defined in (8). By Taylor expansion, we have

$$\begin{aligned} (W_{ni})_{kl} &= \frac{1}{n} \sum_{j=1}^n \ddot{l}(Y_j; \check{\theta}_i)_{kl} - \frac{1}{n} \sum_{j=1}^n \ddot{l}(Y_j; \widehat{\theta})_{kl} - \frac{1}{n} \sum_{j=1}^n \ddot{l}(Y_j; \check{\theta}_i)_{kl} \\ &= \frac{1}{n} \sum_{j=1}^n \sum_{m=1}^p \left( \frac{\partial \ddot{l}(Y_j; \theta)_{kl}}{\partial \theta_m} \Big|_{\theta=\check{\theta}_i} \right) (\check{\theta}_i - \widehat{\theta})_m - \frac{1}{n} \sum_{j=1}^n \ddot{l}(Y_j; \check{\theta}_i)_{kl}, \end{aligned}$$

where  $\tilde{\theta}_i$  is between  $\check{\theta}_i$  and  $\hat{\theta}$ . Thus,

$$|(W_{ni})_{kl}| \leq \frac{p}{n} \sum_{j=1}^n C(Y_j) \cdot \|\check{\theta}_i - \hat{\theta}\|_2 + \frac{1}{n} H(Y_i),$$

by Condition 4 of Theorem 2 and by Condition 3. Now by Condition 1, and the assumptions  $E\{C(Y_1)\} < \infty$  and  $E\{H(Y_1)^2\} < \infty$ , it follows that

$$\begin{aligned} \sqrt{n} \max_{1 \leq i \leq n} \|W_{ni}\|_2 &\leq \frac{p^2}{n} \sum_{j=1}^n C(Y_j) \cdot \sqrt{n} \max_{1 \leq i \leq n} \|\check{\theta}_i - \hat{\theta}\|_2 + \frac{1}{\sqrt{n}} \max_{1 \leq i \leq n} H(Y_i) \\ &= O_p(1) o_p(1) + o_p(1) \xrightarrow{P} 0 \end{aligned} \quad (10)$$

as  $n \rightarrow \infty$ . Referring back to (9), this also implies that

$$n \max_{1 \leq i \leq n} \|R_{ni}\|_2 \leq \|\hat{I}(\hat{\theta})^{-1}\|_2 \cdot \sqrt{n} \max_{1 \leq i \leq n} \|W_{ni}\|_2 \cdot \sqrt{n} \max_{1 \leq i \leq n} \|\hat{\theta}_{(i)} - \hat{\theta}\|_2 \xrightarrow{P} 0. \quad (11)$$

Addressing first the second remainder term in (5), recall from (7) and (9) that

$$\hat{\theta}_{(i)} - \hat{\theta} = -\frac{1}{n} \hat{I}(\hat{\theta})^{-1} \dot{l}(Y_i; \hat{\theta}) + R_{ni}. \quad (12)$$

Thus after some algebra, Conditions 1 and 3 and results (10) and (11) imply

$$\begin{aligned} &\left| \sqrt{n} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta})^T \ddot{l}(Y_i; \tilde{\theta}_i) (\hat{\theta}_{(i)} - \hat{\theta}) \right| \\ &\leq \|\hat{I}(\hat{\theta})^{-1}\|_2^2 \frac{1}{n^{3/2}} \sum_{i=1}^n \|\dot{l}(Y_i; \hat{\theta})\|_2^2 \|\ddot{l}(Y_i; \tilde{\theta}_i)\|_2 \\ &\quad + \|\hat{I}(\hat{\theta})^{-1}\|_2 \left( \max_{1 \leq i \leq n} \|R_{ni}\|_2 \right) \frac{2}{n^{1/2}} \sum_{i=1}^n \|\dot{l}(Y_i; \hat{\theta})\|_2 \|\ddot{l}(Y_i; \tilde{\theta}_i)\|_2 \\ &\quad + \left( \max_{1 \leq i \leq n} \|R_{ni}\|_2 \right)^2 n^{1/2} \sum_{i=1}^n \|\ddot{l}(Y_i; \tilde{\theta}_i)\|_2 \\ &\leq O_p(1) \frac{1}{n^{3/2}} \sum_{i=1}^n G(Y_i)^2 H(Y_i) + o_p(1) \frac{1}{n^{3/2}} \sum_{i=1}^n G(Y_i) H(Y_i) + o_p(1) \frac{1}{n^{3/2}} \sum_{i=1}^n H(Y_i). \end{aligned}$$

Now it follows immediately from Conditions 2 and 3 and the strong law of large numbers that  $n^{-3/2} \sum_{i=1}^n G(Y_i) H(Y_i) \rightarrow 0$  and  $n^{-3/2} \sum_{i=1}^n H(Y_i) \rightarrow 0$  with probability one. Also, by the Marcinkiewicz-Zygmund strong law of large numbers (see Loève, 1963, p. 243),  $n^{-3/2} \sum_{i=1}^n G(Y_i)^2 H(Y_i) \rightarrow 0$  with probability one as long as  $E\{[G(Y_i)^2 H(Y_i)]^{2/3}\} < \infty$ ,

which follows easily from Conditions 2 and 3 and Hölder's inequality:

$$\begin{aligned} E[\{G(Y_i)^2 H(Y_i)\}^{2/3}] &\leq (E[\{G(Y_i)^{4/3}\}^{3/2}])^{2/3} (E[\{H(Y_i)^{2/3}\}^3])^{1/3} \\ &= [E\{G(Y_i)^2\}]^{2/3} [E\{H(Y_i)^2\}]^{1/3} < \infty. \end{aligned}$$

For the first remainder term in (5) we require a somewhat finer analysis of  $R_{ni}$ . By (9) and (12) we have

$$R_{ni} = \frac{1}{n} \hat{I}(\hat{\theta})^{-1} W_{ni} \hat{I}(\hat{\theta})^{-1} \dot{l}(Y_i; \hat{\theta}) - \hat{I}(\hat{\theta})^{-1} W_{ni} R_{ni}.$$

Thus Condition 2 and results (10) and (11) imply

$$\begin{aligned} \sqrt{n} \left| \sum_{i=1}^n \dot{l}(Y_i; \hat{\theta})^T R_{ni} \right| &\leq \|\hat{I}(\hat{\theta})^{-1}\|_2 \left( \sqrt{n} \max_{1 \leq i \leq n} \|W_{ni}\|_2 \right) \left\{ \|\hat{I}(\hat{\theta})^{-1}\|_2 \frac{1}{n} \sum_{i=1}^n \|\dot{l}(Y_i; \hat{\theta})\|_2^2 \right. \\ &\quad \left. + \left( \sqrt{n} \max_{1 \leq i \leq n} \|R_{ni}\|_2 \right) \frac{1}{n} \sum_{i=1}^n \|\dot{l}(Y_i; \hat{\theta})\|_2 \right\} \\ &\leq o_p(1) \frac{1}{n} \sum_{i=1}^n G(Y_i)^2 + o_p(1) \frac{1}{n} \sum_{i=1}^n G(Y_i) \xrightarrow{P} 0. \quad \square \end{aligned}$$

Condition 1 in Theorem 4 seems restrictive, but it follows easily for  $\hat{\theta}$  that have either of the following two forms:

- (a)  $\hat{\theta} = g(\bar{h}_1, \dots, \bar{h}_k)$ , where  $\bar{h}_j = n^{-1} \sum_{i=1}^n h_j(Y_i)$ ,  $j = 1, \dots, k$ , and  $g$  has a derivative  $\dot{g}$  existing in a neighborhood of  $(E h_1(Y_1), \dots, E h_k(Y_1))$ .
- (b)  $\hat{\theta} = T(F_n)$ , where  $F_n$  is the empirical distribution function of the sample, and  $T(\cdot)$  is a Lipschitz continuous functional satisfying  $|T(G) - T(H)| \leq C \sup_y |G(y) - H(y)|$  for arbitrary distribution functions  $G$  and  $H$ .

For (b), Condition 1 of Theorem 4 follows simply because  $\sup_y |F_{ni}(y) - F_n(y)| \leq n^{-1}$ , where  $F_{ni}$  is the empirical distribution function for the sample with the  $i$ th observation left out. For (a), we follow an argument related to that found on p. 26 of Shao and Tu (1995). For simplicity we consider the case of  $k = 1$ ,  $h_1(y) = y$ , and  $g$  real-valued. If  $E[Y_1^2] < \infty$ , then, with probability one,  $n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$  converges and  $\max_{1 \leq i \leq n} (Y_i - \bar{Y})^2/n \rightarrow 0$ . Since  $|\bar{Y}_{(i)} - \bar{Y}| = (n-1)^{-1} |Y_i - \bar{Y}|$ , we thus have

$$(n-1) \max_{1 \leq i \leq n} (\bar{Y}_{(i)} - \bar{Y})^2 \rightarrow 0$$

with probability one. Taking square roots gives Condition 1 for  $\bar{Y}$ . Now, by the mean value theorem

$$|\hat{\theta}_{(i)} - \hat{\theta}| = |g(\bar{Y}_{(i)}) - g(\bar{Y})| = |\dot{g}(\tilde{Y}_i)| |\bar{Y}_{(i)} - \bar{Y}|,$$

where  $\tilde{Y}_i$  lies between  $\bar{Y}_{(i)}$  and  $\bar{Y}$ . By continuity of  $\dot{g}$ , we can bound  $|\dot{g}(\tilde{Y}_i)|$  uniformly by  $|\dot{g}(E[Y_1])| + \epsilon$  for all  $n$  sufficiently large with probability one.

## 5 Discussion

Perhaps the greatest strength of the IOS test is that it can be easily applied in a variety of situations without a great deal of analytic work. Thus, for example, though IOS appears to behave much like Mardia's kurtosis test when testing for multivariate normality, no special effort was required to devise a measure of multivariate kurtosis in order to use IOS, just some simple programming. Similarly, though the IM test could be applied in most if not all of our examples, it would typically require much more analytic work than IOS.

Testing for model misspecification using IOS can require considerable time for computations, but this is not always the case. In practice our approach is usually to first compute just IOS or  $\text{IOS}_A$ . The value of the statistic is often a reasonable guide to the outcome of the test, with values "close" to the number of parameters in the model generally yielding a large bootstrap p-value, as can be verified with, say, 100 bootstrap replicates, or even as few as 10 or 20 if computations are particularly time consuming. If the initial bootstrap replications indicate that the p-value may be relatively small, say .10 or less, then a more precise estimate is required and we increase the number of bootstrap replications accordingly. Of course testing with  $\text{IOS}_A$  is an alternative that is usually less computationally intensive, though writing the necessary routines requires more analytic work to compute the necessary derivatives. We have not explored whether numerical derivatives retain sufficient accuracy to replace analytic derivatives in the computation of  $\text{IOS}_A$ .

An alternative approach to large sample inference is to use the asymptotic normality of IOS (or  $\text{IOS}_A$ ) in conjunction with a jackknife estimate of its standard error, which is valid whether the model is correctly specified or not. Again, we have not explored this approach in any detail.

As presented here, the IOS test does require a fully specified likelihood and complete data. Thus, for example, IOS is not immediately useful for censored data unless we are prepared to specify a parametric model for the censoring distribution, which would not be typical. A version of the asymptotic form,  $\text{IOS}_A$ , might be used, but there remains the problem of how to carry out the parametric bootstrap, which would again seem to require a model for the censoring distribution. Similarly, it is not yet clear how IOS could be applied to dependent data, such as time series or spatial data. These problems remain to be solved.

## References

- Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. New York: Wiley.
- (2002) *Categorical Data Analysis*. New York: Wiley, 2 edn.
- Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM*. Oxford: Oxford University Press.
- Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Czaki), 267–281. Budapest: Akademiai Kiado.
- Bayarri, M. J. and Berger, J. O. (2000) P-values for composite null models. *J. Am. Statist. Ass.*, **95**, 1127–1142.
- Bliss, C. I. (1935) The calculation of the dosage-mortality curve. *Annals of Applied Biology*, **22**, 134–167.
- Boos, D. D. and Zhang, J. (2000) Monte Carlo evaluation of resampling-based hypothesis tests. *J. Am. Statist. Ass.*, **95**, 486–492.
- Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters*. New York: John Wiley & Sons, Inc.
- Brooks, S. P., Morgan, B. J. T., Ridout, M. S. and Pack, S. E. (1997) Finite mixture models for proportions. *Biometrics*, **53**, 1097–1115.
- Cramer, H. (1946) *Mathematical Methods of Statistics*. Princeton: Princeton University Press.
- Daniel, C. (1976) *Applications of Statistics to Industrial Experimentation*. New York: John Wiley & Sons.
- Feigl, P. and Zelen, M. (1965) Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**, 826–838.
- Fisher, R. A. (1973) *Statistical Methods for Research Workers*. New York: Hafner, 14 edn.
- Garren, S. T., Smith, R. L. and Piegorsch, W. W. (2000) On a likelihood-based goodness-of-fit test of the beta-binomial model. *Biometrics*, **56**, 947–949.
- (2001) Bootstrap goodness-of-fit test for the beta-binomial model. *J. Appl. Statist.*, **28**, 561–571.
- Geisser, S. (1989) Predictive discordancy tests for exponential observations. *Canad. J. Statist.*, **17**, 19–26.
- (1990) Predictive approaches to discordancy testing. In *Bayesian and likelihood methods in statistics and econometrics: Essays in honor of George A. Barnard* (eds. S. Geisser, J. S. Hodges, S. J. Press and A. Zellner), 321–335. Amsterdam: North-Holland Publishing Co.

- Geisser, S. and Eddy, W. F. (1979) A predictive approach to model selection. *J. Am. Statist. Ass.*, **74**, 153–160.
- Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: Asymptotics and exact calculations. *J. R. Statist. Soc. B*, **56**, 501–514.
- Golub, G. H. and van Loan, C. F. (1989) *Matrix Computations*. Johns Hopkins, second edn.
- Harville, D. A. (1997) *Matrix Algebra from a Statistician's Perspective*. New York: Springer-Verlag.
- Hausman, J. A. (1978) Specification tests in econometrics. *Econometrica*, **46**, 1251–1271.
- Horowitz, J. L. (1994) Bootstrap-based critical values for the information matrix test. *J. Econometrics*, **61**, 395–411.
- Ihaka, R. and Gentleman, R. (1996) R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.
- Johnson, R. A. and Wichern, D. W. (1998) *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall, 4 edn.
- Larsen, R. J. and Marx, M. L. (2001) *An Introduction to Mathematical Statistics and Its Applications*. Englewood Cliffs, NJ: Prentice-Hall Inc., 3 edn.
- Lehmann, E. L. and Casella, G. (1998) *Theory of point estimation*. New York: Springer-Verlag.
- Lewis, S. L., Montgomery, D. C. and Myers, R. H. (2001) Examples of designed experiments with nonnormal responses. *Journal of Quality Technology*, **33**, 265–278.
- Linhart, H. and Zucchini, W. (1986) *Model Selection*. New York: Wiley.
- Lockhart, A. M. C., Piegorsch, W. W. and Bishop, J. B. (1992) Assessing overdispersion and dose-response in the male dominant lethal assay. *Mutation Research*, **272**, 35–38.
- Loève, M. (1963) *Probability Theory*. Princeton: van Nostrand, 3 edn.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.
- Robins, J., van der Vaart, A. and Ventura, V. (2000) Asymptotic distribution of p values in composite null models. *J. Am. Statist. Ass.*, **95**, 1143–1156.
- Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. New York: Springer-Verlag.
- Slaton, T. L., Piegorsch, W. W. and Durham, S. D. (2000) Estimation and testing with overdispersed proportions using the beta-logistic regression model of Heckman and Willis. *Biometrics*, **56**, 125–133.
- Stefanski, L. A. and Boos, D. D. (2002) The calculus of m-estimation. *Amer. Statist.*, **56**, 29–38.

Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc. B*, **39**, 44–47.

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838.

— (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–26.

## Appendix: The Asymptotic Variance of IOS

As shown in Theorem 4,  $\text{IOS} = \text{IOS}_A + o_p(n^{-1/2})$ , where

$$\text{IOS}_A = n^{-1} \sum_{i=1}^n \left\{ \dot{l}(\hat{\theta}; Y_i)^T \hat{I}(\hat{\theta})^{-1} \dot{l}(\hat{\theta}; Y_i) \right\}.$$

The derivation of the asymptotic variance of  $\text{IOS}_A$  is easier if we use an estimating equations/M-estimation approach (see Stefanski and Boos, 2002). Let  $\hat{t} = \text{IOS}_A$  and  $\hat{J} = \hat{I}(\hat{\theta})^{-1}$ , and define

$$\psi(\theta, J, t; y) = \begin{pmatrix} \dot{l}(\theta; y) \\ \text{vech}\{\ddot{l}(\theta; y) + J^{-1}\} \\ \dot{l}(\theta; y)^T J \dot{l}(\theta; y) - t \end{pmatrix}.$$

Note that  $\hat{\theta}$ ,  $\hat{J}$ , and  $\hat{t} = \text{IOS}_A$  jointly satisfy the system of  $q = (p+1)(p+2)/2$  equations  $\sum_{i=1}^n \psi(\hat{\theta}, \hat{J}, \hat{t}; Y_i) = 0$ .

To simplify notation, let  $t_0 = \text{IOS}_\infty$ ,  $I_0 = I(\theta_0)$ ,  $J_0 = I_0^{-1}$ ,  $\dot{l}_0 = \dot{l}(\theta_0, Y_1)$ ,  $\ddot{l}_0 = \ddot{l}(\theta_0, Y_1)$ , and  $\psi_0 = \psi(\theta_0, Y_1)$ . Under the conditions of Theorem 3, the vector  $(\hat{\theta}^T, (\text{vech } \hat{J})^T, \hat{t})$  is asymptotically normally distributed with mean  $(\theta_0^T, (\text{vech } J_0)^T, t_0)$  and covariance matrix  $n^{-1} C^{-1} D (C^{-1})^T$ , where  $C = E\{\dot{\psi}_0\}$ , and  $D = E\{\psi_0 \psi_0^T\}$ . Here  $\dot{\psi}_0 = \dot{\psi}(\theta_0, J_0, t_0; Y_1)$ , with  $\dot{\psi}$  being the  $q \times q$  matrix-valued function

$$\begin{aligned} \dot{\psi}(\theta, J, t; y) &= \frac{\partial \psi(\theta, J, t; y)}{\partial \{\theta^T, (\text{vech } J)^T, t\}} \\ &= \begin{pmatrix} \ddot{l} & 0 & 0 \\ \frac{\partial}{\partial \theta^T} \text{vech}\{\ddot{l}\} & -H_p(J^{-1} \otimes J^{-1}) G_p & 0 \\ 2\dot{l}^T J \ddot{l} & \left[ \text{vech}\{2\dot{l}\dot{l}^T - \text{diag}(\dot{l}\dot{l}^T)\} \right]^T & -1 \end{pmatrix}, \end{aligned}$$

where  $\otimes$  is the Kronecker product,  $G_p$  is the duplication matrix (of dimension  $p^2 \times p(p+1)/2$ ),  $H_p$  (of dimension  $p(p+1)/2 \times p^2$ ) is any left inverse of  $G_p$  (see Harville, 1997, section 16.4),

and we have dropped the arguments  $\theta$  and  $y$  on the right hand side for brevity. Thus

$$C = \begin{pmatrix} -I_0 & 0 & 0 \\ E\left[\frac{\partial}{\partial\theta^T} \text{vech}\{\ddot{l}(\theta; Y_1)\}\Big|_{\theta=\theta_0}\right] & -H_p(I_0 \otimes I_0)G_p & 0 \\ 2E\{i_0^T I_0^{-1} \ddot{l}_0\} & \left(\text{vech}\left[2E(\dot{l}_0 \dot{l}_0^T) - \text{diag}\{E(\dot{l}_0 \dot{l}_0^T)\}\right]\right)^T & -1 \end{pmatrix}$$

and

$$D = \begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{12}^T & D_{22} & D_{23} \\ D_{13}^T & D_{23}^T & D_{33} \end{pmatrix}$$

where

$$\begin{aligned} D_{11} &= E\{\dot{l}_0 \dot{l}_0^T\}, \\ D_{12} &= E\{\dot{l}_0 (\text{vech } \ddot{l}_0)^T\}, \\ D_{13} &= E\{\dot{l}_0 \dot{l}_0^T I_0^{-1} \dot{l}_0\}, \\ D_{22} &= E\{(\text{vech } \ddot{l}_0)(\text{vech } \ddot{l}_0)^T\} - (\text{vech } I_0)(\text{vech } I_0)^T, \\ D_{23} &= E\{\dot{l}_0^T I_0^{-1} \dot{l}_0 (\text{vech } \ddot{l}_0)\} + t_0 \text{vech } I_0, \end{aligned}$$

and

$$D_{33} = E\{\dot{l}_0^T I_0^{-1} \dot{l}_0\}^2 - t_0^2.$$

Of course, under the null hypothesis of correct model specification,  $\theta_0$  is the “true” value of  $\theta$ ,  $I_0$  is the Fisher information matrix,  $E\{\dot{l}_0 \dot{l}_0^T\} = I_0$ , and  $t_0 = p$ , the dimension of  $\theta$ .

A convenient way to find the asymptotic variance of IOS in particular examples is to use a symbolic math program to compute the bottom right element of  $n^{-1}C^{-1}D(C^{-1})^T$ . Even in the simplest problems, however, the asymptotic variance can be fairly complicated. This is illustrated by the asymptotic variance given in equation (27.7.8) of Cramér (1946) for the sample kurtosis, which arises in the simple iid normal model in Example 3 of Section 1.2.