PREDICTIVE EVALUATION OF LOGISTIC MODELS

by

Françoise Seillier-Moiseiwitsch

Department of Biostatistics, University of
North Carolina at Chapel Hill, NC.

Institute of Statistics Mimeo Series No. 2101

June 1992

PREDICTIVE EVALUATION OF LOGISTIC MODELS

Françoise Seillier-Moiseiwitsch

Department of Biostatistics
University of North Carolina at Chapel Hill

ABSTRACT

Logistic models are assessed for their ability to produce valid forecasts for new observables rather than for their goodness of fit to past observations. The diagnostics described here are based on scoring rules and, being updated easily, are particularly well-suited to a sequential context. Test statistics measuring different aspects of empirical validity are described. Simulations for moderate sample sizes compliment results pertaining to their asymptotic behaviour.

Key Words: deviance, logistic model, prediction, scoring rule.

1. Introduction

In the work of Guttman (1967), Akaike (1974), Stone (1974), Geisser (1975) and Geisser & Eddy (1979), attention has turned, in model selection, from the capability to explain past data to the ability to predict future out-turns. Putting philosophical standpoints aside, predictive model assessment offers sound evaluations: the data are not used simultaneously for estimation and validation. This approach also provides a safeguard against overparametrization. Furthermore, these papers rightly emphasize the importance of sound model evaluation once the model building phase is completed.

Here, the evaluation of logistic models is cast in a sequential framework. This often reflects constrains imposed by the data collection process. It also mimicks the way the model will be used in the future. So, at instance i, the forecast relies on the previous i-1 observations (as well as on the covariate information associated with all observations up to and including instance i) and once the outcome of the ith event becomes known, it will in turn be used to forecast the next event.

Therefore, neither will estimated parameter values be compared to posited ones nor will expectations be taken over unrealized values (as is done when calculating the mean squared error of prediction). The model evaluation is based solely on probabilities for future events generated by the model. These will be measured against the outcomes via a *scoring rule*. The training set, of size n, consists of the observations from which the first parameter estimates are generated. All other data points serve, in turn, two purposes: validation and estimation.

What is described here is an adaptation of crossvalidation to sequential set-ups (Dawid,1984). Each new data point involves a single re-estimation of the parameters while crossvalidation would require, in its one-out-at-a-time version, as many new fits as there are data points. As evidenced later, this set-up is more amenable to formal testing than crossvalidation. Earlier papers have looked into the well-foundedness of the proposed diagnostics from a general view-point (Seillier-Moiseiwitsch & Dawid,1993; Seillier-Moiseiwitsch et al.,1992).

The next section reviews scoring rules and their use. The proposed diagnostic tests are described in section 3. Proofs of their distributional properties are relegated to the Appendix. Section 4 summarizes some empirical convergence results. In section 5, several bootstrap procedures are described and used to improve the relevance of the reference distribution of the tests. Some empirical power studies are presented in section 6.

2. Scoring Rules

To be of use to a decision-maker, model-based probabilistic statements should exhibit a high degree of realism. Their realism can be quantified via measures of *calibration* and *resolution* (DeGroot & Fienberg,1983). In *well-calibrated* forecasting systems, the frequency of occurence of the event of interest among those instances which were assigned a probability p ($\bar{a}(p)$ say) should be close to p. *Perfect resolution* is attained when all the events with the same likelihood of occurence are given the same probability. In the context of simple two-decision situations, the performance of a model can always be improved by its calibrated version (i.e. quoting $\bar{a}(p)$ rather than p) (Schervish, 1983). Hence, high resolution is the more desirable of these two attributes. Indeed, given enough data, recalibration is

always possible while improving resolution would involve a reassessment of the information at hand (and hence the formulation of a new model).

Scoring rules were devised to help in determining the suitability of probability forecasts. Among the class of all possible such functions, those which encourage honesty, i.e. which are optimized whenever the forecasters report their actual beliefs, are termed *proper*. These have been shown to be an agregate measure of both calibration and resolution (DeGroot & Fienberg,1983). Let $A_1, \cdots, A_n$ denote the sequence of events of interest and p_i the model's probability that A_i occurs (calculated on the basis of $A_1, \cdots, A_{i-1}$). For simplicity, asssume that A_i's are binary events. A generic bounded scoring rule can be characterized as follows :

$$S(A_i,p_i) = J(p_i) + (A_i - p_i)\frac{dJ(p_i)}{dp_i}$$

where $J(p_i)$ is a strictly convex, differentiable and bounded function on [0,1]. Then the average score after n instances can be partitioned as follows:

$$S(N) = \frac{1}{N}\sum_{i=1}^{N} S(a_i, p_i) = S_1(N) + S_2(N)$$

$$\text{with } S_1(N) = \sum_{p\in P} f(p)\,\bar{a}(p)\left[\left(S(1,p) - S(1,\bar{a}(p))\right) + (1-\bar{a}(p))\left(S(0,p) - S(0,\bar{a}(p))\right)\right] \text{ and } S_2(N) = \sum_{p\in P} f(p)\,\Phi(\bar{a}(p))$$

where $\Phi(t) = tS(1,t) + (1-t)S(0,t) \quad 0 \le t \le 1$, P is the set of allowable probabilities and f(p) the frequency of prediction p.

Example 1: The Brier score

$$BS = \frac{1}{m}\sum_{i=n+1}^{N} (a_i - p_i)^2$$

where m=N–n and n is the number of observations that ensures that the first estimates of the parameters are relatively stable.

Example 2: The overall calibration score

$$OC = \frac{1}{m}\sum_{i=n+1}^{N} (a_i - p_i)$$

quantifies the overall bias in the forecasts.

Example 3: The logarithmic score

$$LO = \frac{1}{m}\sum_{i=n+1}^{N} \left(a_i \log p_i + (1-a_i)\log(1-p_i)\right)$$

is simply the average loglikelihood under the predictive distribution.

When enough data are available, an in-depth study of predictive performance would involve partitioning the original sequence into a number of subsequences and computing a score on each subsequence separately. These subsequences can be generated in a number of ways. For instance, one can divide the probability range into subintervals and consider instances which were assigned probability

forecasts in the the same subinterval. One can also stratify the set of events according to their covariates: similar covarites values would put the evnets into the same subset.

For a general survey of scoring rules and their merits, see Murphy & Epstein (1967a,b), Winkler & Murphy (1968), Murphy & Winkler (1970), Dawid (1986).

3. Diagnostic Tests for Predictive Validity

Starting with the case where A_i's are binary events, we assume that

$$A_i \sim \text{Bernoulli}(p_i(\underline{\theta})) \quad \text{with} \quad \text{logit}\, p_i(\underline{\theta}) = \underline{\theta}^T \underline{x}_i \quad \text{for } i \in \{1,...,N\}$$

where both $\underline{\theta}$ and $\underline{x}_i$ are $k \times 1$ vectors. Denote by $P_{\underline{\theta}}$ the probability distribution stemming from this model.

As in classical hypothesis testing, the diagnostics take the form of Z-statistics. The raw scores are therefore standardized. But, here, the standardization is performed with respect to the forecasting model, that is, assuming that the estimated probability of the event A_i (based on $\underline{a}^{(i-1)}$) is in fact the sampling model. As a result, after subtracting the expectation under the predictive model, the scoring functions considered fit the following expression:

$$S(N) = \frac{1}{N}\sum_{i=1}^{N}(A_i - p_i)\, g(p_i) \quad \text{where } g(.) \text{ is an arbitrary function and } p_i = P_{\hat{\theta}_{i-1}}(A_i = 1) \ .$$

Examples:

$g(p_i) = 1$ leads to the normalized overall calibration score, $g(p_i) = 1 - 2p_i$ to the normalized Brier score and $g(p_i) = \text{logit}\, p_i$ to the normalized logarithmic score.

Limit theorems being unaffected by early realizations, the results are stated for N, rather than N-n, scores.

Theorem:

Let g denote a bounded function of the probability forecast and assume that its first derivative is also bounded.

$$\frac{\sum_{i=1}^{N}(A_i - p_i)\, g(p_i)}{\left\{\sum_{i=1}^{N} p_i(1-p_i)\,(g(p_i))^2\right\}^{1/2}} \xrightarrow{D} N(0,1) \quad \text{under } P_{\underline{\theta}} \quad \text{as } N \to \infty \ .$$

The proof is given in the Appendix. For the logarithmic score, though g is not bounded, this result still holds if one assumes either that the model does not generate categorical forecasts or that, if it does, these are correct. This is certainly the case under the null hypothesis that the assumed model is generating the observations.

The next result can be applied when one divides the original sequence into a number of subsequences in order to investigate the predictive performance of the model in more depth.

Corollary 1:

Consider the following partition into K subsequences $\{E_k, k=1,...,K\}$ and $\Delta(A)$ takes the value 1 if A occurs and 0 otherwise, then

$$\sum_{k=1}^{K} \frac{\{\sum_{i=1}^{N} \Delta(A_i \in E_k)\,(A_i - p_i)\,g(p_i)\}^2}{\sum_{i=1}^{N} \Delta(A_i \in E_k)\,p_i\,(1-p_i)\,(g(p_i))^2} \xrightarrow{\mathcal{D}} \chi^2_K \text{ under } P_{\underline{\theta}} \text{ as } N \to \infty .$$

This follows from applying the above theorem to each subinterval and invoking a multivariate martingale central limit theorem (Aalen,1977).

For discrete covariates taking a fixed number of possible values, the usual limit distribution for the deviance obtains for a predictive version of this statistic. Let m_j refer to the number of outcomes with covariate vector $\underline{x}_j$ and the number of distinct covariate vectors be J. The index i runs from 1 to N, the total number of binary events. Here the convergence is considered under the condition that each m_j goes to infinity.

Corollary 2: *Predictive Deviance*

Letting $A^j = \sum_i A_i \Delta(\underline{x}_i = \underline{x}_j)$,

$$2\sum_j \{A^j \log(A^j / \sum_i p_i \Delta(\underline{x}_i = \underline{x}_j)) + (m_j - A^j)\log((m_j - A^j)/(m_j - \sum_i p_i \Delta(\underline{x}_i = \underline{x}_j)))\} \xrightarrow{\mathcal{D}} \chi^2_{J-k} \text{ under } P_{\underline{\theta}} \text{ as } m_j \to \infty \text{ for all } j$$

An outline of the proof is written up in the appendix. If $m_j = 1$ for all j, McCullagh's distributional results (1986) for the deviance apply in this setting: the predictive deviance is degenerate conditionally on the sufficient statistic and thus provides no information regarding the predictive performance of the model.

The above theorem is easily generalized to the case where $A_i \sim \text{Binomial}(n_i, p_i(\underline{\theta}))$ with n_i fixed and $i \in \{1,...,N\}$. Let $p_{ik} = p_{\underline{\hat{\theta}}_{i-1}}(A_i = k)$ and $A_{ik} = \delta(A_i = k)$ with $k \in \{1,...,n_i\}$. Then,

$$S(A_i, p_i) = \sum_{k=1}^{n_i} S(A_{ik}, p_{ik})$$

Corollary 3:

Let E and Var denote the expectation and variance under the forecast distribution.

$$\frac{S(N) - \sum_{i=1}^{N}\sum_{k=1}^{n_i} E(S(A_{ik}, p_{ik})}{\{\sum_{i=1}^{N} \text{Var}(S(A_i, p_i))\}^{1/2}} \xrightarrow{\mathcal{D}} N(0,1) \quad \text{under } P_{\underline{\theta}} \text{ as } N \to \infty .$$

Examples:

For the Brier score,

$$S(A_i, p_i) = \sum_{k=1}^{n_i} (A_{ik} - p_{ik})^2 \quad \text{and} \quad E(S(A_i, p_i)) = \sum_{k=1}^{n_i} p_{ik}\,(1 - p_{ik})$$

$$\mathrm{Var}(S(A_i,p_i)) = \sum_{k=1}^{n_i} p_{ik}(1-p_{ik})(1-2p_{ik})^2 - \sum_{k,m=1;k\neq m}^{n_i} p_{ik}p_{im}(1-2p_{ik})(1-2p_{im})$$

As for the logarithmic rule,

$$S(A_i,p_i) = \sum_{k=1}^{n_i} A_{ik}\log p_{ik} \quad \text{and} \quad E(S(A_i,p_i)) = \sum_{k=1}^{n_i} p_{ik}\log p_{ik}$$

$$\mathrm{Var}(S(A_i,p_i)) = \sum_{k=1}^{n_i} p_{ik}(1-p_{ik})(\log p_{ik})^2 - \sum_{k,m=1;k\neq m}^{n_i} p_{im}\,p_{ik}\log p_{im}\log p_{ik}$$

This approach is unsuitable for the overall calibration score since $S(A_i,p_i)=0$. One can, however, compare actual outcomes with expected numbers under the predictive distribution:

Corollary 4:

$$\frac{\sum_{1}^{N}(A_i - n_i p_{i1})}{\{\sum_{i=1}^{N} n_i p_{i1}(1-p_{i1})\}^{1/2}} \xrightarrow{\mathcal{D}} N(0,1) \quad \text{under } P_{\underline{\theta}} \quad \text{as } N \to \infty .$$

redictive deviance is described in McCullagh (1986).

The next sections look into the behaviour of these diagnostics, with regards to rates of convergence and power, on simulated data.

4. Empirical Rate of Convergence

In the first instance, we investigate the rate of convergence of the limit theorem presented in the preceding section when the predictions are computed from the actual model. The latter contains 4 parameters:

$$\mathrm{logit}\, p_i(\underline{\theta}) = 1.96 - 0.35\,X1 + 0.25\,X2 - 2.58\,X3 \ .$$

X1 has 4 categories. Both X2 and X3 are continuous. A_i's are binary random variables.

Table 1 gives Kolmogorov-Smirnov test statistics calculated from 1000 scores. CA and BI denote the χ^2-functions based, respectively, on the overall calibration and the Brier scores. The probability range was divided into 11 subintervals, which entails that the asymptotic distribution is χ^2_{11}. Clearly, under the true model, the empirical distributions of OC, BS and LO do not depart significantly from that of a standard normal variable with as few as 75 data points, while the distributional convergence of CA and BI requires at least some 200 observations.

Sample Size	OC	BS	LO	CA	BI
50	0.03271	0.03585	0.04871 **	0.10100 ***	0.10539 ***
75	0.03209	0.03086	0.04293 *	0.08689 ***	0.07626 ***
100	0.01554	0.02617	0.03297	0.06621 ***	0.06140 ***
125	0.02546	0.03192	0.03180	0.05735 ***	0.05539 ***
150	0.02221	0.02046	0.02142	0.04864 **	0.05640 ***
175	0.01209	0.03085	0.03953 *	0.04336 **	0.04477 **
200	0.02595	0.02010	0.02973	0.03260	0.02934

Table 1: Significance levels: *=90% **=95% ***=99%
Kolmogorov-Smirnov test statistics for scores based on the true model (1000 simulations)

The reliability of the tests revolves around the percentiles of the score distribution being close to nominal ones. In table 2, are entered the number of scores which, under this scenario, fall below the 0.5, 1, 2.5 and 5 percentiles and above the 95, 97.5, 99 and 99.5 percentiles of the normal distribution. The results are again based on 1000 simulations. Similarly for aggregate scores in table 3. The asterisks indicate whether the entries are between 1 and 2 (*), between 2 and 3 (**) or further than 3 standard deviations (***) away from expected numbers. It appears that, for sample sizes greater than 75, the percentiles of OC approximate well those of the standard normal distribution, while for BS and LO only the 95 and 97.5 percentiles are reliably estimated. Clearly, when they fail to do so, their overall tendency is to err on the conservative side. Regarding CA and BI, for sample sizes of 125 and above, the numbers in the tail are as expected. Thus, if one is checking a fully specified model, all these test statistics can be computed on an original sequence of fairly moderate size.

Score	Sample Size	0.5	1	2.5	5	95	97.5	99	99.5
OC	50	6	13	33 *	52	36 **	9 ***	1 **	0 **
	75	8 *	11	27	53	42 *	21	12	4
	100	6	10	30 *	57 *	44	24	6 *	2 *
	125	7	15 *	26	50	50	23	4 *	1 *
	150	7	15 *	23	49	47	16 *	7	4
	175	8 *	13	25	47	47	21	9	7
	200	4	11	31 *	50	48	25	10	6
BS	50	0 **	1 **	6 ***	27 ***	58 *	34 *	18 **	8 *
	75	0 **	7	10 ***	37 *	52	28	15 *	11 **
	100	1 *	5 *	13 **	40 *	46	23	16 *	8 *
	125	0 **	6 *	17 *	36 **	48	27	18 **	11 **
	150	0 **	5 *	16 *	40 *	49	31 *	15 *	10 **
	175	0 **	2 **	14 **	46	55	33 *	11	5
	200	1 *	4 *	18 *	36 **	54	33 *	19 **	9 *
LO	50	0 **	0 ***	2 ***	15 ***	58 *	38 **	17 **	8 ***
	75	0 **	2 **	10 ***	36 *	58 *	28	17 **	11 **
	100	1 *	3 **	11 **	36 *	45	30	19 **	12 **
	125	0 **	4 *	16 *	34 **	50	29	17 **	9 ***
	150	0 **	3 **	15 **	38 *	44	26	13	10 ***
	175	0 **	1 **	13 **	37 *	61 *	29	14 *	7 ***
	200	1 *	3 **	14 **	31 **	65 **	34 *	16 *	10 ***

Table 2: Number of scores, based on true model, below 0.5, 1, 2.5 and 5 percentiles and above 95, 97.5, 99 and 99.5 percentiles of normal distribution (1000 simulations)

Score	CA		BI	
Sample Size	95	99	95	99
50	41 *	19 **	40 *	20 ***
75	66 **	25 ***	72 ***	25 ***
100	55	20 ***	61 **	18 **
125	51	9	47	11
150	50	10	52	13
175	48	17 **	48	11
200	59 *	12	50	16

Table 3: Number of scores, based on true model, above 95 and 99 percentiles of asymptotic distribution χ^2_{11} (1000 simulations)

Now, the parameters are estimated from an increasing training sample. In the tables below, the training size refers to the size of the first training set. Tables 4 displays Kolmogorov-Smirnov statistics

calculated from 1,000 overall calibration and Brier scores for various sample and training set sizes. For OC, distributional convergence is reached at moderate sample sizes: the two significant statistics (sample size of 200 with training sizes of 50 and 75) may well be flukes as both smaller and larger sample sizes, for similar training sets, did not produce significant results. For BS and LO, on other hand, none of the size combinations produced a non-significant statistic.

Sample	Training Size								
OC	50	75	100	125	150	200	250	275	300
100	0.041	0.029							
150		0.034	0.029						
200	0.046	0.045	0.024	0.020	0.023				
250	0.028	0.031	0.023	0.025	0.015	0.026			
300				0.035		0.018	0.034		
350							0.026	0.030	0.015
BS	50	75	100	125	150	200	250	275	300
100	0.313	0.169							
150		0.220	0.148						
200	0.333	0.267	0.217	0.134	0.111				
250	0.300	0.241	0.200	0.174	0.141	0.074			
300				0.157		0.089	0.041		
350							0.120	0.078	0.055

Table 4: Kolmogorov-Smirnov statistics for overall calibration and Brier scores based on true model when one estimates parameters (1000 simulations)

If one applies a somewhat less stringent criterion and considers the number of scores, out of 1,000, falling in the tails of the normal distribution, the outcome for OC and BS is shown in tables 5, 6 and 7. These summary statistics for aggregate scores, based on 11 subintervals, appear in table 8.

For OC, numbers are acceptable for training size 75 and above and sample size of 150 and above. The outcome of the simulations suggest using the majority of the observations to obtain accurate parameter estimates and setting aside no less than 75 data points for evaluation. The summary statistics indeed tend to move towards the critical region when the size of the training set is such that fewer than 75 points are left for assessment. These numbers will, of course, be dependent on the number of parameters in the underlying model.

For BS, the score distribution is highly asymmetric. For all sample and training sizes selected, the number of test statistics falling below or above the considered percentiles differed from the expected values by more than 3 standard deviations. However, when one looks at the numbers of scores outside central intervals with 90%, 95%, 98% and 99% nominal coverage (table 7), the picture changes. For the 95% and 98% intervals, acceptable numbers are attained with a sample of 250 observations , 175 of which are used as training set and a sample of 300 observations, 200 of which are set aside for

estimation. For the logarithmic score, similar results were obtained. As expected, the aggregate scores yield very conservative tests: the frequencies obtained by simulation grossly overestimate tail areas of the normal distribution.

Note that, in these tables, the entries tend to decrease as the training size increases, which seems to point to the longlasting effect of earlier unreliable predictions. On the other hand, for fixed training size, the entries, apart from a couple of exceptions, decrease as the sample size becomes large. This reflects the central limit property of the statistics.

Sample Size	Training Size	0.5	1	2.5	5	95	97.5	99	99.5
100	50	8 *	21 ***	44 ***	71 ***	65 **	31 *	8	3
	75	11 **	20 ***	43 ***	73 ***	58 **	35 **	11	6
150	75	11 **	13 *	34 *	64 **	72 ***	39 **	19 **	7
	100	3	17 **	32 *	53	73 ***	35 **	12	7
200	50	6	11	28	56	68 **	40 ***	19 **	8 *
	75	9 *	16 *	28	50	63 *	32 *	15 *	5
	100	9 *	14 *	30 *	57 *	62 *	28	11	8 *
	125	7	14 *	31 *	57 *	43 *	30 *	10	5
	150	11 **	12	30 *	52	43 *	31 *	9	7
250	50	2 *	11	27	65 **	59 *	29	10	6
	75	5	10	24	57 *	60 *	28	10	2 *
	100	3	8	28	56	53	26	7	4
	125	6	10	24	52	48	26	11	6
	150	6	10	32 *	54	48	27	13	7
	175	8 *	18 **	34 *	58 *	47	26	13	7
	200	4	8	28	65 **	58 *	29	8	5

Table 5: Numbers of overall calibration, based on true model, below 0.5, 1, 2.5 and 5 percentiles and above 95, 97.5, 99 and 99.5 percentiles of normal distribution, when one estimates parameters (1000 simulations)

Sample	Training Set	0.5	1	2.5	5	95	97.5	99	99.5
100	50	0 **	0 ***	1 ***	4 ***	251 ***	175 ***	125 ***	93 ***
	75	0 **	0 ***	1 ***	9 ***	175 ***	126 ***	83 ***	57 ***
150	75	0 **	1 **	2 ***	10 ***	136 ***	114 ***	54 ***	45 ***
	100	0 **	1 **	4 ***	7 ***	183 ***	89 ***	69 ***	33 ***
200	50	1 *	2 **	4 ***	8 ***	265 ***	196 ***	118 ***	82 ***
	75	1 *	2 **	5 ***	8 ***	199 ***	142 ***	83 ***	56 ***
	100	1 *	1 **	3 ***	9 ***	162 ***	111 ***	61 ***	34 ***
	125	1 *	1 **	3 ***	10 ***	139 ***	95 ***	60 ***	41 ***
	150	0 **	0 ***	8 ***	20 ***	115 ***	65 ***	43 ***	28 ***
250	50	0 **	0 ***	2 ***	3 ***	265 ***	180 ***	119 ***	88 ***
	75	0 **	0 ***	3 ***	8 ***	201 ***	146 ***	80 ***	52 ***
	100	0 **	0 ***	2 ***	6 ***	163 ***	108 ***	54 ***	29 ***
	125	1 *	1 **	1 ***	11 ***	140 ***	84 ***	45 ***	29 ***
	150	0 **	0 ***	4 ***	10 ***	118 ***	67 ***	41 ***	25 ***
	175	1 *	1 **	2 ***	18 ***	104 ***	68 ***	36 ***	25 ***
	200	0 **	0 ***	3 ***	13 ***	105 ***	64 ***	35 ***	18 ***
300	125	0 **	0 ***	4 ***	18 ***	125 ***	77 ***	35 ***	21 ***
	175	0 **	0 ***	3 ***	20 ***	108 ***	70 ***	39 ***	26 ***
	200	0 **	2 **	6 ***	19 ***	109 ***	63 ***	35 ***	24 ***
	225	0 **	1 **	10 ***	27 ***	87 ***	54 ***	27 ***	21 ***
	250	0 **	0 ***	7 ***	28 ***	76 ***	43 ***	21 ***	14 ***

Table 6: Numbers of scores, based on true model, below 0.5, 1, 2.5 and 5 percentiles and above 95, 97.5, 99 and 99.5 percentiles of normal distribution, when one estimates parameters (1000 simulations)

Sample Size	Training Size	99%	98%	95%	90%
100	50	93 ***	125 ***	176 ***	255 ***
	75	57 ***	83 ***	127 ***	184 ***
150	75	45 ***	56 ***	116 ***	146 ***
	100	33 ***	70 ***	93 ***	190 ***
200	50	83 ***	120 ***	200 ***	273 ***
	75	57 ***	85 ***	147 ***	207 ***
	100	35 ***	62 ***	114 ***	171 ***
	125	42 ***	61 ***	98 ***	149 ***
	150	28 ***	43 ***	72 ***	135 ***
250	50	88 ***	119 ***	182 ***	268 ***
	75	52 ***	80 ***	149 ***	209 ***
	100	29 ***	54 ***	110 ***	169 ***
	125	30 ***	46 ***	85 ***	151 ***
	150	25 ***	41 ***	71 ***	128 **
	175	26 ***	37 ***	70 **	122 **
	200	18 **	35 ***	67 **	118 *
300	125	21 ***	35 ***	81 ***	143 ***
	175	26 ***	39 ***	73 ***	128 **
	200	24 ***	37 ***	69 **	128 **
	225	21 ***	28 *	64 **	114 *
	250	14 *	21	50	104

Table 7: Numbers of Brier scores, based on true model, in the middle 90%, 95%, 98% and 99% of normal distribution, when one estimates parameters (1000 simulations)

Scores		CA			BI		
Sample Size	Training Size	90	95	99	90	95	99
200	50	302	224	123	303	229	130
	75	245	196	86	262	191	90
	100	213	146	73	222	153	70
	125	187	133	67	191	130	68
	150	146	100	56	143	99	56
250	50	277	185	106	290	207	113
	75	239	150	70	243	164	71
	100	215	149	61	225	139	57
	125	181	136	58	183	133	61
	150	170	117	48	171	118	46
	175	167	113	54	163	117	49
	200	147	101	52	148	102	56
300	125	165	106	42	170	102	42
	175	172	116	40	162	106	43
	200	171	106	41	158	106	42
	225	152	100	46	159	105	49
	250	133	92	41	129	95	39

Table 8: Numbers of aggregate overall calibration and Brier scores (11 subintervals), based on true model, above 90, 95 and 99 percentiles of chi-square distribution, when one estimates parameters (1000 simulations)

5. Bootstrap Tests

These tests can be extended further using a predictive bootstrap approach. Such an approach would make the reference distribution more relevant to the data at hand when the sample size does not guarantee that normality holds. The simulation results from the preceding section indeed show that the convergence is somewhat slow. A description of possible bootstrap approaches follows.

The first two methods mimick the evaluation procedure. The parameter vector is first estimated from $\underline{a}^{(n)}$, which yields $\hat{\underline{\theta}}_n$. The probability that the next event A_{n+1} occurs is then computed using $\hat{\underline{\theta}}_n$ and the covariates associated with A_{n+1}. This probability generates a *bootstrap observation* a^*_{n+1}. It is this realization that enters the scoring function. For the first procedure, this whole process is repeated, in turn, on $\underline{a}^{(n+1)}, \underline{a}^{(n+2)}, \ldots, \underline{a}^{(N-1)}$. For the second procedure, the bootstrap observations become part of the training set: for event i, the bootstrap probability distribution is based on $\{\underline{a}^{(n)}, a^*_{n+1}, \ldots, a^*_{i-1}\}$. The second approach is therefore more likely to yield a bootstrap score distribution with large spread.

The last three procedures involve generating realizations from the model fitted on the full data set. The third procedure is a replica of the original evaluation process but now on the bootstrap outcomes. The last two procedures are similar to the first two above with a_i's replaced by data generated from the

fitted model.

Table 9 displays the numbers of overall calibration scores, out of 200 simulations based on the actual model, which fall below the 1st, 5th and above the 95th, 99th percentiles of the bootstrap distribution. The numbers outside the middle 90% and 98% of this distribution are also shown. Similarly for the Brier and logarithmic scores in table 10. For each iteration, 100 bootstrap samples were generated. Results for 200 bootstrap samples were not substantially different. Evidently, procedures 1 and 3 peformed best, the other three being overly conservative. For the overall calibration score, contrasting tables 5 and 9 reveals that a substantial improvement obtains with method 1 and particularly for method 3. For the Brier score (cf. table 7), method 1 yields coverage probabilities somewhat closer to the nominal ones. Method 3, on the other hand, achieves the nominal levels. For other combinations of sample and training sizes (data not shown), the overall features remain as these sizes increase. Also, as already observed, LO yields tail frequencies slightly worse than BR does.

Method	Sample	Training Set	1	5	95	99	90	98
1	75	50	4 *	7	6 *	4 *	13 *	8 *
	100	50	2	8	8	4 *	16	6
		75	5 **	14 *	6 *	2	20	7 *
2	75	50	2	4 *	47 ***	20 ***	51 ***	22 ***
	100	50	0 *	2 **	62 ***	29 ***	64 ***	29 ***
		75	0 *	9	36 ***	9 ***	45 ***	9 **
3	75	50	1	9	8	4 *	17	5
	100	50	3	10	12	1	22	4
4	75	50	8 ***	20 ***	11	2	31 **	10 **
	100	75	8 ***	19 **	10	3	29 **	11 ***
5	75	50	5 **	14 *	17 **	7 ***	31 **	12 ***
	100	75	6 **	19 **	6 *	5 **	25 *	11 ***

Table 9: Numbers of overall calibration scores, based on true model, below 1, 5 , above 95, 99 percentiles and outside the middle 90%, 98% of the bootstrap distribution (200 simulations)

Score	Method	Sample	Training Set	1	5	95	99	90	98
BS	1	75	50	2	4 *	40 ***	11 ***	44 ***	13 ***
		100	50	0 *	2 **	45 ***	19 ***	47 ***	19 ***
			75	0 *	8	28 ***	7 ***	36 ***	7 *
	2	75	50	2	4 *	40 ***	15 ***	44 ***	17 ***
		100	50	0 *	2 **	46 ***	18 ***	48 ***	18 ***
			75	0 *	8	32 ***	8 ***	40 ***	8 *
	3	75	50	4 *	13	7	0 *	20	4
		100	50	4 *	12	10	3	22	7 *
	4	75	50	3	9	30 ***	11 ***	39 ***	14 ***
		100	75	4 *	8	32 ***	10 ***	40 ***	14 ***
	5	75	50	2	7	44 ***	15 ***	51 ***	17 ***
		100	75	4 *	11	33 ***	12 ***	44 ***	16 ***
LO	1	75	50	2	5 *	47 ***	20 ***	52 ***	22 ***
		100	50	0 *	2 **	56 ***	27 ***	58 ***	27 ***
			75	0 *	8	34 ***	14 ***	42 ***	14 ***
	2	75	50	2	4 *	47 ***	20 ***	51 ***	22 ***
		100	50	0 *	2 **	62 ***	29 ***	64 ***	29 ***
			75	0 *	9	36 ***	9 ***	45 ***	9 **
	3	75	50	4 *	12	7	1	19	5
		100	50	3	13	9	0 *	21	3
	4	75	50	3	10	34 ***	14 ***	44 ***	17 ***
		100	75	4 *	9	37 ***	10 ***	46 ***	14 ***
	5	75	50	2	8	41 ***	18 ***	49 ***	20 ***
		100	75	5 **	13	32 ***	16 ***	45 ***	21 ***

Table 10: Numbers of Brier and logarithmic scores, based on true model, below 1, 5 , above 95, 99 percentiles and outside the middle 90%, 98% of the bootstrap distribution (200 simulations)

6. Power Studies

In order to investigate the behaviour of these tests when the observations and the predictions are generated from different models, three types of departure from the underlying model are considered: ignoring one of the covariates, adding a redundant variable and substituting one of the covariates with a correlated one.

Table 11 gives the number of scores in the tails of the standard normal distribution if one were to use the coefficients of the actual model in computing the forecast probabilities. It therefore shows the best results attainable and hence the limitations of scoring-rule based tests. Numbers in parentheses

refer to the correlation between X3 and X5. Both are normally distributed covariables. X4 is a binary variable.

Even under these ideal circumstances the overall calibration score has almost no power to distinguish between correlated explanatory variables. On the other hand, the Brier score is sensitive to this type of model mispecification. Both scores are able to detect, with high probability, the omission of an explanatory variable. Neither could reliably identify redundant variables. Again, the results for the logarithmic score follow closely those of the Brier score. The performance of the overall calibration score is explained by its evaluation of average properties of the model rather than of the forecasts on an individual basis (as is the focus of the Brier score).

Model	Sample Size	Score	0.5	2.5	97.5	99.5
-X3	75	OC	256	473	1	0
		BR	0	0	596	414
	150	OC	503	710	0	0
		BR	0	0	854	722
+ .15 X4	75	OC	11	45	14	2
		BR	0	14	35	15
	150	OC	15	46	10	3
		BR	0	17	36	13
+ .55 X4	75	OC	43	117	3	1
		BR	0	10	65	25
	150	OC	72	194	2	0
		BR	0	4	74	30
+ .95 X4	75	OC	106	277	1	1
		BR	0	1	143	65
	150	OC	235	421	0	0
		BR	0	2	205	88
-X3+X5 (.25)	75	OC	57	99	119	53
		BR	0	0	1000	998
-X3+X5 (.45)	75	OC	44	90	97	36
		BR	0	0	995	980
-X3+X5 (.65)	75	OC	31	80	73	27
		BR	0	0	922	846
	150	OC	32	66	69	25
		BR	0	0	996	985
-X3+X5 (.85)	75	OC	18	49	43	12
		BR	0	0	457	292
	150	OC	22	51	48	16
		BR	0	0	728	547
-X3+X5 (.95)	75	OC	10	40	29	6
		BR	0	3	125	44
	150	OC	15	33	29	9
		BR	0	2	175	59

Table 11: Number of scores, based on models differing from true one (X1+X2+X3), below 0.5 and 2.5 percentiles and above 97.5 and 99.5 percentiles of normal distribution (1000 simulations)

When one estimates parameters, the power of these scores is drastically reduced, as evidenced by table 12. The entries in this table give the numbers of scores, resulting from 1,000 simulations, above and below percentiles of the normal distribution. The total sample size used is 150 and the training size 75. Again, the overall calibration score exhibits the least power. The logarithmic score is actually

the most likely, of these three rules, to detect departures from the actual model. By contrast with the previous scenario, the scores have highest power against the inclusion of a redundant covariate.

With bootstrap reference distributions, as generated by procedure 1, the chance of picking up model misspecifications, of the type investigated here, is higher than if one uses the asymptotic distribution. The entries of table 13 are based on 200 simulations and, for each of these, 100 bootstrap samples. The total sample size is 100 and the training size 50. Procedure 3 lacks ability to select the correct model. This is indeed expected as it makes use of bootstrap observations generated from a model estimated on the whole sample.

Model	Score	0.5	2.5	97.5	99.5
-X3	OC	7	30	25	7
	BR	0	2	110	38
	LO	0	2	116	39
+X4	OC	4	29	35	9
	BR	0	3	187	93
	LO	0	3	220	131
-X3+X5 (.25)	OC	2	28	31	5
	BR	0	2	181	64
	LO	0	2	192	68
-X3+X5 (.45)	OC	2	34	28	4
	BR	0	1	151	56
	LO	0	1	172	64
-X3+X5 (.65)	OC	3	31	25	7
	BR	0	3	125	54
	LO	0	3	144	66
-X3+X5 (.85)	OC	3	38	24	7
	BR	0	4	139	56
	LO	0	4	157	60

Table 12: Numbers of scores above and below percentiles of the normal distribution when the model is different from the actual one (1000 simulations).

Model	Method	Score	1	5	95	99
-X3	1	OC	3	8	10	4
		BR	0	0	54	17
		LO	0	0	56	17
	3	OC	1	8	9	4
		BR	3	9	7	3
		LO	3	9	7	3
+X4	1	OC	2	9	8	5
		BR	1	2	62	30
		LO	1	2	77	38
	3	OC	3	11	11	4
		BR	3	8	7	2
		LO	3	10	4	2
-X3+X5 (.25)	1	OC	3	8	11	3
		BR	0	1	65	23
		LO	0	1	74	28
	3	OC	2	11	13	4
		BR	1	4	12	5
		LO	1	6	9	3
-X3+X5 (.45)	1	OC	4	9	9	3
		BR	1	1	54	27
		LO	1	2	60	27
	3	OC	1	10	8	6
		BR	1	10	10	4
		LO	1	8	10	3

Table 13: Numbers of scores above and below percentiles of the bootstrap distribution when the model is different from the actual one (200 simulations)

7. Conclusion

The sequential test statistics for predictive performance, considered here, were shown to converge to their expected distribution. Simulation results demonstrate that this convergence is relatively slow, which leads to conservative tests. This can be remedied through bootstrap procedures. From these simulations, it also transpires that, for reasons of reliability, these tests should be employed in their two-sided version.

The power of these tests is drastically reduced when estimating parameters and using the asymptotic reference distribution. The bootstrap procedures actually improve power. Adding a redundant variable is the type of departure most often picked up when the model parameters need to be estimated. Removal of a covariate and replacement by a correlated one seem difficult to detect.

Of the three scoring rules studied here, the overall calibration score exhibits almost no power though it converges most quickly to the theoretical asymptotic distribution. The logarithmic score, with convergence similar in rate to that of the Brier score, is the most powerful.

Of the bootstrap approaches considered, the procedure that is most similar to the evaluation process and calls on the actual observations as training sample is to be preferred. Procedures based on bootstrap samples generated from distributions estimated from the whole sample lack the predictive appeal.

In sum, the statistics described here test different aspects of model-based predictions. The predictive deviance allows the comparison of nested models as in the usual goodness-of-fit setting. For models involving a moderate number of parameters, as employed in the simulation work presented in sections 4, 5 and 6, practical guidelines would state that the overall calibration score can be reliably referred to the standard normal distribution with 150 observations, 50 of which being left aside for model evaluation while the reference distribution for the other statistics should be obtained by bootstrap procedures.

References

Aalen, O.O. (1977) Weak convergence of stochastic integrals related to counting processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **38**, 261-77.

Akaike, H. (1974) A new look at statistical model identification. *I.E.E.E. Transactions on Automatic Control*, **AT-19**, 716-23.

Dawid, A.P. (1984) Statistical theory: The prequential approach (with discussion). *Journal of the royal statistical Society*, A, **147**, 278-92.

Dawid, A.P. (1986) Probability forecasting. *Encyclopedia of Statistical Sciences*, vol. 7, edited by S. Kotz, N.L. Johnson & C.B. Read. Wiley-Interscience, 210-218.

DeGroot, M.H. & Fienberg, S.E. (1983) The comparison and evaluation of forecasters. *The Statistician*, **32**, 14-22.

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320-8.

Geisser, S. & Eddy, W.F. (1979) A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153-60.

Guttman, I. (1967) The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society*, B, **29**, 83-100.

McCullagh, P. (1986) The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association*, **81**, 104-7.

McCullagh, P. & Nelder, J.A. (1989) Generalized Linear Models, Second Edition. Chapman and Hall, London.

Murphy, A.H. & Epstein, E.S. (1967a) Verification of probabilistic predictions: A briefreview. *Journal of Applied meteorology*, **6**, 748-55.

Murphy, A.H. & Epstein, E.S. (1967b) A note on probability forecasts and "hedging". *Journal of Applied meteorology*, **6**, 1002-4.

Murphy, A.H. & Winkler, R.L. (1970) Scoring rules in probability assessment and evaluation. *Acta Psychologica*, **34**, 273-86.

Schervish, M.J. (1989) A general method for comparing probability assessors. *The Annals of Statistics*, **17**, 1856-79.

Seillier-Moiseiwitsch, F. & Dawid, A.P. (1993) On testing the validity of probability forecasts. To appear in the *Journal of the American Statististical Association.*

Seillier-Moiseiwitsch, F., Sweeting, T.W. & Dawid, A.P. (1992) Prequential tests of model fit. To appear in the *Scandinavian Journal of Statistics.*

Stone, M. (1974) Cross-validatory choice and assessment of statistical models (with discussion). *Journal of the royal statistical Society*, B, **36**, 111-47.

Winker, R.L. & Murphy, A.H. (1968) "Good" probability assessors. *Journal of Applied meteorology*, **7**, 751-8.

Appendix

Proof of Theorem:

Let $p_i(\underline{\theta})$ be the probability that A_i occurs under the actual model i.e. $p_i(\underline{\theta}) = e^{\underline{\theta}^T \underline{x}_i} / (1+e^{\underline{\theta}^T \underline{x}_i})$. Denote by ' differentiation with respect to $\underline{\theta}$ and by $H(f(\underline{\theta}))$ the Hessian of the function f(.) with respect to $\underline{\theta}$. Let $L_n(\underline{\theta})$ be the loglikelihood for the first n instances i.e.

$$L_n(\underline{\theta}) = \sum_{i=1}^{n} \{ A_i \log p_i(\underline{\theta}) + (1-A_i) \log(1-p_i(\underline{\theta})) \} = \sum_{i=1}^{n} \{ A_i \underline{\theta}^T \underline{x}_i + \log(1-p_i(\underline{\theta})) \}$$

Let $X_i = (\underline{x}_1\ \underline{x}_2 \cdots \underline{x}_i)^T$, then $\underline{p}_i'(\underline{\theta}) = p_i(\underline{\theta})(1-p_i(\underline{\theta}))\ \underline{x}_i$ and $\underline{L}_n'(\underline{\theta}) = X_n^T\ (A_1 - p_1 \cdots A_n - p_n)^T$

Let $V_n = \mathrm{diag}(p_1(\underline{\theta})(1-p_1(\underline{\theta})), \ldots, p_n(\underline{\theta})(1-p_n(\underline{\theta})))$ and $\delta_n^2 = -L_n''(\underline{\theta}) = X_n^T V_n X_n$.

Consider the following Taylor expansion:

$$p_i(\hat{\underline{\theta}}_{i-1}) - p_i(\underline{\theta}) = (\hat{\underline{\theta}}_{i-1} - \underline{\theta})^T \underline{p}_i'(\underline{\theta}) + O_p((i-1)^{-1})$$

$$= \{ \underline{L}_{i-1}^T(\underline{\theta})\, \delta_{i-1}^{-2} + \tfrac{1}{2}((\hat{\underline{\theta}}_{i-1} - \underline{\theta})^T H_{i-1}^1 (\hat{\underline{\theta}}_{i-1} - \underline{\theta}) \cdots (\hat{\underline{\theta}}_{i-1} - \underline{\theta})^T H_{i-1}^k (\hat{\underline{\theta}}_{i-1} - \underline{\theta}) \}\, \delta_{i-1}^{-2}\, \underline{p}_i'(\underline{\theta}) + O_p((i-1)^{-1})$$

where $H_n^i = H(\partial L_n(\underline{\theta}_n^*) / \partial \theta_i)$ and the entries of $\underline{\theta}_n^*$ fall between the corresponding ones of $\underline{\theta}$ and $\hat{\underline{\theta}}_n$.

Let $R_i = \tfrac{1}{2}((\hat{\underline{\theta}}_{i-1} - \underline{\theta})^T H_{i-1}^1 (\hat{\underline{\theta}}_{i-1} - \underline{\theta}) \cdots (\hat{\underline{\theta}}_{i-1} - \underline{\theta})^T H_{i-1}^k (\hat{\underline{\theta}}_{i-1} - \underline{\theta}))\delta_{i-1}^{-2}\, \underline{p}_i'(\underline{\theta}) - \tfrac{1}{2}(\hat{\underline{\theta}}_{i-1} - \underline{\theta})^T p_i''(\underline{\theta}_{i-1}^+)(\hat{\underline{\theta}}_{i-1} - \underline{\theta})$

where $p_i''(\underline{\theta}) = H(p_i(\underline{\theta}))$ and the entries of $\underline{\theta}_{i-1}^+$ fall between the corresponding ones of $\underline{\theta}$ and $\hat{\underline{\theta}}_{i-1}$. Let p_i^* be between p_i and $p_i(\underline{\theta})$ and let g_i^* stand for the first derivative of g(.) evaluated at p_i^*. Then

$$\sum_{i=1}^{N} (A_i - p_i)\, g(p_i)$$

$$= \sum_{i=1}^{N} (A_i - p_i(\underline{\theta}))\, g(p_i(\underline{\theta})) + \sum_{i=2}^{N} \underline{L}_{i-1}^T(\underline{\theta})\, \delta_{i-1}^{-2}\, \underline{p}_i'(\underline{\theta})\, g(p_i^*)\, (A_i - p_i(\underline{\theta})) + (A_1 - p_1)\, g(p_1) - (A_1 - p_1(\underline{\theta}))\, g(p_1(\underline{\theta})) + \sum_{i=2}^{N} R_i\, g_i^*\, (A_i - p_i(\underline{\theta}))$$

$$-\sum_{i=2}^{N} g(p_i(\underline{\theta}))\,\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta}) - \sum_{i=2}^{N} p_i^2(\underline{\theta})\,(1-p_i(\underline{\theta}))^2 (\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i)^2\, g'(p_i^*) - \sum_{i=2}^{N} \underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,R_i\,g_i^* + \sum_{i=2}^{N} R_i\,g(p_i)$$

Now, $\sum_{i=2}^{N} R_i = o_p(N^{1/2})$

in view of the fact that

(i) all entries of $H_j^i(\underline{\theta})$ $(1 \le j \le k, 1 \le i \le N-1)$ and $p_k''(\underline{\theta})(1 \le k \le n)$ are bounded for all $\underline{\theta}$ and

(ii) $(\underline{\hat{\theta}}_{i-1} - \underline{\theta})$ is a vector the entries of which are $O_p((i-1)^{-1})$ (McCullagh & Nelder, 1989).

Hence, since we assumed that both g and its first derivative are bounded,

$$\sum_{i=1}^{N} (A_i - p_i)\,g(p_i) = \sum_{i=1}^{N} (A_i - p_i(\underline{\theta}))\,g(p_i(\underline{\theta})) + \sum_{j=1}^{N-1} \underline{l}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} \delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g_i^*\,(A_i - p_i(\underline{\theta})) - \sum_{j=1}^{N-1} \underline{l}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} g(p_i(\underline{\theta}))\,\delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})$$

$$-\sum_{i=2}^{N} p_i^2(\underline{\theta})\,(1-p_i(\underline{\theta}))^2\,(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i)^2\,g_i^* + o_p(N^{1/2})$$

One can show that

$$\mathrm{Var}_{\underline{\theta}}\Big(\sum_{j=1}^{N} \underline{l}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} \delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g_i^*\,(A_i - p_i(\underline{\theta}))\Big) = O(\log N)$$

$$\mathrm{Var}_{\underline{\theta}}\Big(\sum_{j=1}^{N} p_i^2(\underline{\theta})\,(1-p_i(\underline{\theta}))^2\,(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i)^2\,g_i^*\Big) = O(\log^2 N)$$

$$\mathrm{Cov}_{\underline{\theta}}\Big(\sum_{j=1}^{N-1} \underline{l}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} \delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g(p_i(\underline{\theta})),\ \sum_{i=2}^{N} p_i^2(\underline{\theta})\,(1-p_i(\underline{\theta}))^2\,(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i)^2\,g_i^*\Big) = O(\log N)$$

$$\mathrm{Cov}_{\underline{\theta}}\Big(\sum_{i=1}^{N} g(p_i(\underline{\theta}))\,(A_i - p_i(\underline{\theta})),\ \sum_{j=1}^{N-1} \underline{l}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} \delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g_i^*\,(A_i - p_i(\underline{\theta}))\Big) = O(N^{1/2})$$

$$\mathrm{Cov}_{\underline{\theta}}\Big(\sum_{i=1}^{N} g(p_i(\underline{\theta}))\,(A_i - p_i(\underline{\theta})),\ \sum_{i=2}^{N} p_i^2(\underline{\theta})\,(1-p_i(\underline{\theta}))^2\,(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i)^2\,g_i^*\Big) = O(N^{1/2})$$

$$\mathrm{Cov}_{\underline{\theta}}\Big(\sum_{j=1}^{N-1} \underline{l}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} \delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g_i^*\,(A_i - p_i(\underline{\theta})),\ \sum_{j=1}^{N-1} \underline{l}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} \delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g(p_i(\underline{\theta}))\Big) = O(N^{1/2})$$

$$\mathrm{Cov}_{\underline{\theta}}\Big(\sum_{j=1}^{N-1} \underline{l}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} \delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g_i^*\,(A_i - p_i(\underline{\theta})),\ \sum_{i=2}^{N} p_i^2(\underline{\theta})\,(1-p_i(\underline{\theta}))^2\,(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i)^2\,g_i^*\Big) = O(\log^{5/2} N)$$

Therefore,

$$\mathrm{Var}_{\underline{\theta}}\Big(\sum_{i=1}^{N} (A_i - p_i)\,g(p_i)\Big)$$

$$= \mathrm{Var}_{\underline{\theta}}\Big(\sum_{i=1}^{N} (A_i - p_i(\underline{\theta}))\,g(p_i(\underline{\theta}))\Big) + 2\sum_{i=2}^{N} g(p_i(\underline{\theta}))\,\underline{p}_i'^{T}(\underline{\theta}) \sum_{k=i+1}^{N} \delta_{k-1}^{-2}\,\underline{p}_k'(\underline{\theta})\,g(p_k(\underline{\theta})) - 2\sum_{j=1}^{N-1} g(p_j(\underline{\theta}))\,\underline{p}_j'^{T}(\underline{\theta}) \sum_{i=j+1}^{N} g(p_i(\underline{\theta}))\,\delta_{i-1}^{-2}\,\underline{p}_i'^{T}(\underline{\theta})$$

$$+ o(N) + O(\log N) + O(\log^2 N) + O(N^{1/2}) + O(\log^{5/2} N)$$

$$= \sum_{i=1}^{N} (g(p_i(\underline{\theta})))^2\,p_i(\underline{\theta})\,(1-p_i(\underline{\theta})) + o(N)$$

Also, since $p_i \to p_i(\underline{\theta})$ w.p. 1 as $i \to \infty$

$$\frac{1}{N}\sum_{i=1}^{N} p_i(1-p_i)\,g^2(p_i) \to \frac{1}{N}\sum_{i=1}^{N} p_i(\underline{\theta})(1-p_i(\underline{\theta}))\,g^2(p_i(\underline{\theta})) \quad \text{w.p. 1}$$

It remains to show that the terms which do not include R_i

$$\sum_{i=1}^{N}(A_i-p_i(\underline{\theta}))\left(g(p_i(\underline{\theta}))+\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g_i^*\right)-\sum_{j=1}^{N-1}\underline{l}_j'^{T}(\underline{\theta})\sum_{i=j+1}^{N} g(p_i(\underline{\theta}))\,\delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})-\sum_{i=2}^{N} p_i^2(\underline{\theta})(1-p_i(\underline{\theta}))^2\left(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\underline{x}_i\right)^2 g_i^*$$

satisfy the conditions for a central limit theorem. As one can show that

$$\sum_{i=2}^{N} p_i^2(\underline{\theta})(1-p_i(\underline{\theta}))^2\left(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\underline{x}_i\right)^2 g_i^* = o_p(N^{1/4}) \quad ,$$

one can disregard this term by invoking Slutzky's lemma. The remaining expressions can be written as follows:

$$\sum_{i=1}^{N}(A_i-p_i(\underline{\theta}))\left\{g(p_i(\underline{\theta}))+\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})\,g_i^*-\underline{x}_i^{T}\sum_{j=i+1}^{N} g(p_j(\underline{\theta}))\,\delta_{j-1}^{-2}\,\underline{p}_j'(\underline{\theta})\right\}$$

As these independent summands are bounded they satisfy the Lindeberg condition □

Proof of Corollary 2:

Let $\mathrm{Dev}_N(p_j)=2\sum_{j=1}^{J}\{A^j\log(A^j/m_j p_j)+(m_j-A^j)\log((m_j-A^j)/(m_j-m_j p_j))\}$ where $\sum_{j=1}^{J} m_j = N$.

One must show that

(a) $\mathrm{Dev}_N(m_j^{-1}\sum_i p_i\,\Delta(\underline{x}_i=\underline{x}_j))\,/\,\mathrm{Dev}_N(p_j(\underline{\theta})) \xrightarrow{p} 1$

(b) $\mathrm{Dev}_N(p_j(\hat{\underline{\theta}}_N))\,/\,\mathrm{Dev}_N(p_j(\underline{\theta})) \xrightarrow{p} 1$

For (a), it is sufficient to show that

$$E_{\underline{\theta}}\left(\left\{\sum_{j=1}^{J} m_j^{-1}\sum_{i=1}^{N}\Delta(\underline{x}_i=\underline{x}_j)\left(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})+R_i\right)(1-p_j^{+})^{-1}\left(A^j p_i^{+-1}-m_j\right)\right\}^2\left\{\sum_{j=1}^{J}\left(A^j\log p_j(\underline{\theta})+(m_j-A^j)\log(1-p_j(\underline{\theta}))\right)\right\}^{-2}\right)$$

$\to 0$ as $m_j \to \infty$ for all j, where p_i^{+} falls between $p_j(\underline{\theta})$ and $\frac{1}{m_j}\sum_{i=1}^{N} p_i\,\Delta(\underline{x}_i=\underline{x}_j)$.

$$E_{\underline{\theta}}\left(\left\{\sum_{j=1}^{J} m_j^{-1}\sum_{i=1}^{N}\Delta(\underline{x}_i=\underline{x}_j)\left(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{p}_i'(\underline{\theta})+R_i\right)(1-p_j^{+})^{-1}\left(A^j p_i^{+-1}-m_j\right)\right\}^2\left\{\sum_{j=1}^{J}\left(A^j\log p_j(\underline{\theta})+(m_j-A^j)\log(1-p_j(\underline{\theta}))\right)\right\}^{-2}\right)$$

$$\le O(N^{-2})\,E_{\underline{\theta}}\left(\sum_{i=1}^{N}\left(\underline{L}_{i-1}'^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i+R_i\right)\right)^2$$

$$\le O(N^{-2})\left\{\sum_{i=1}^{N}\underline{x}_i^{T}\delta_{i-1}^{-2}\underline{x}_i+\sum_{i=1}^{N}E_{\underline{\theta}}^{1/2}(R_i^2)\sum_{i=1}^{N}E_{\underline{\theta}}^{1/2}\left(\left(\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i\right)^2\right)+E_{\underline{\theta}}\left(\sum_{i,k=1;i\neq k}\underline{x}_i^{T}\,\delta_{i-1}^{-2}\,\underline{L}_{i-1}'(\underline{\theta})\,\underline{L}_{k-1}^{T}(\underline{\theta})\,\delta_{k-1}^{-2}\,\underline{x}_k\right)\right.$$

$$\left.+E_{\underline{\theta}}\left(\sum_{i,k=1;i\neq k}^{N} R_i R_k\right)+E_{\underline{\theta}}\left(\sum_{i,k=1;i\neq k}^{N}\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i R_k\right)\right\}+o(N^{-1})$$

$$\le o(N^{-1})+o(N^{-1/4})+O(N^{-2})\left\{E_{\underline{\theta}}\left(\sum_{i,k=1;i<k}^{N}\underline{x}_i^{T}\,\delta_{i-1}^{-2}\,\underline{L}'_{i-1}(\underline{\theta})\,\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{k-1}^{-2}\,\underline{x}_k\right)+E_{\underline{\theta}}\left(\left(\sum_{i=1}^{N}R_i\right)^2\right)+E_{\underline{\theta}}\left(\left(\sum_{i=1}^{N}\underline{L}_{i-1}^{T}(\underline{\theta})\,\delta_{i-1}^{-2}\,\underline{x}_i\right)\left(\sum_{k=1}^{N}R_k\right)\right)\right\}$$

$$\leq o(N^{-1}) + o(N^{-1/4}) + O(N^{-2})\, E_{\underline{\theta}}^{1/2}\Big[\Big(\sum_{i=1}^{N} \underline{L}'^{T}_{i-1}(\underline{\theta})\, \delta_{i-1}^{-2}\, \underline{x}_i\Big)^2\Big]\; E_{\underline{\theta}}^{1/2}\Big[\Big(\sum_{k=1}^{N} R_k\Big)^2\Big]$$

$$\leq o(N^{-1}) + o(N^{-1/4})$$

For (b), it is sufficient to show that

$$E_{\underline{\theta}}\Big(\Big\{\sum_{j=1}^{K} (\underline{L}'^{T}_{N}(\underline{\theta})\, \delta_N^{-2}\, \underline{p}'_j(\underline{\theta}) + R_j^*)\,(1-\hat{p}_j)^{-1}\,(A^j \hat{p}_j^{-1} - m_j)\Big\}^2 \Big\{\sum_{j=1}^{K} (A^j \log p_j(\underline{\theta}) + (m_j - A^j)\log(1-p_j(\underline{\theta})))\Big\}^{-2}\Big)$$

$\to 0$ as $m_j \to \infty$ for all j where $\hat{p}_j$ falls between $p_j(\hat{\underline{\theta}}_N)$ and $p_j(\underline{\theta})$,

$$R_j^* = \tfrac{1}{2}\big((\hat{\underline{\theta}}_N - \underline{\theta})^T H_N^1 (\hat{\underline{\theta}}_N - \underline{\theta}) \cdots (\hat{\underline{\theta}}_N - \underline{\theta})^T H_N^k (\hat{\underline{\theta}}_N - \underline{\theta})\big)\delta_N^{-2}\, \underline{p}'_j(\underline{\theta}) - \tfrac{1}{2}(\hat{\underline{\theta}}_N - \underline{\theta})^T p_j''(\underline{\theta}_N^*)(\hat{\underline{\theta}}_N - \underline{\theta})$$

and the entries of $\underline{\theta}_N^*$ fall between those of $\underline{\theta}$ and $\hat{\underline{\theta}}_N$. Since J is finite, one only needs to demonstrate that each summand converges to 0.

$$E_{\underline{\theta}}\Big(\Big\{(\underline{L}'^{T}_{N}(\underline{\theta})\, \delta_N^{-2}\, \underline{p}'_j(\underline{\theta}) + R_j^*)\,(1-\hat{p}_j)^{-1}\,(A^j \hat{p}_j^{-1} - m_j)\Big\}^2 \Big\{\sum_{j=1}^{K} (A^j \log p_j(\underline{\theta}) + (m_j - A^j)\log(1-p_j(\underline{\theta})))\Big\}^{-2}\Big)$$

$$\leq O(1)\, E_{\underline{\theta}}\big(\underline{L}'^{T}_{N}(\underline{\theta})\, \delta_N^{-2}\, \underline{p}'_j(\underline{\theta}) + R_j^*\big)^2$$

$$= O(1)\big(\underline{p}'^{T}_{j}(\underline{\theta})\, \delta_N^{-2}\, \underline{p}'_j(\underline{\theta}) + E_{\underline{\theta}}(R_i^*) + E_{\underline{\theta}}\big((\underline{L}'^{T}_{N}(\underline{\theta})\, \delta_N^{-2}\, \underline{p}'_j(\underline{\theta})\, R_j^*\big)\big)$$

$$= O(N^{-1}) \qquad \square$$