Research Report
Department of Statistics
Göteborg University
Sweden

Bayes prediction of binary outcomes based on correlated discrete predictors

Robert Jonsson
Anders Persson

Research Report 2002:3
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Dept of Statistics	Nat: 031-773 12 74	Nat: 031-773 10 00	http://www.stat.gu.se/stat
P.O. Box 660	Int: +46 31 773 12 74	Int: +46 31 773 10 00	
SE 405 30 Göteborg			
Sweden			

BAYES PREDICTION OF BINARY OUTCOMES BASED ON CORRELATED DISCRETE PREDICTORS

by Robert Jonsson and Anders Persson

Department of Statistics, Göteborg University, Sweden

ABSTRACT

An approach based on Bayes theorem is proposed for predicting the binary outcomes $X = 0, 1$, given that a vector of predictors $\mathbf{Z}$ has taken the value $\mathbf{z}$. It is assumed that $\mathbf{Z}$ can be decomposed into g independent vectors given $X = 1$ and h independent vectors given $X = 0$. First, point and interval estimators are derived for the target probability $\mathbb{P}(X = 1 \mid \mathbf{z})$. In a second step these estimators are used to predict the outcomes for new subjects chosen from the same population. Sample sizes needed to achieve reliable estimates of the target probability in the first step are suggested, as well as sample sizes needed to get stable estimates of the predictive values in the second step. It is also shown that the effects of ignoring correlations between the predictors can be serious. The results are illustrated on Swedish data of work resumption among long-term sick-listed individuals.

Key words: Conditional independence; Confidence intervals; Interactions; Multinomial probabilities; Prediction; Work resumption.

1 Introduction

In many situations there is a great need for predicting categorical outcomes at the individual level. For example, during recent years there has been an increasing rate of cases with long-term sickness in many countries, and in Sweden the increase has been about 30% per year during the period 1997-2001 (SOU (2002)). This has focused on the need for better individual predictions of future state of health, which in term would facilitate the proper rehabilitating interventions. Commonly used methods for such predictions have been logistic regression (Cox (1970)) or 'computer diagnosis' based on empirical Bayes weights (Afifi and Azen (1979), pp. 306-10). The latter two approaches give identical results, since they only differ in the way in which the predictor variables are represented. With a few exceptions, the two approaches have been used under the assumption that the predictors are independent. The reasons for such an assumption are seldom declared, except for the need for simplification, even if it has been pointed out that the assumption may be unrealistic in most applications (Afifi and Azen (1979), p. 307). The effects of assuming predictors to be independent, when they actually are dependent, upon bias and precision of the estimated parameters and on the prediction error seems to have been ignored.

In this paper we suggest an approach based on Bayes theorem for predicting the two outcomes 'healthy' ($X = 0$) and 'non-healthy' ($X = 1$). The vector of predictors $\mathbf{Z}$ have discrete elements and these are allowed to be dependent in such a way that there are dependency between some predictors and independency between some sets of predictors. Furthermore, the number of independent sets of predictors given $X = 0$ may be different from the corresponding number given $X = 1$. In a first step point and interval estimators are derived for the probability $\mathbb{P}(X = 1 \mid \mathbf{z})$, where $\mathbf{z}$ denotes an outcome of the vector $\mathbf{Z}$. The performance of the estimators are studied in simulations (Section 3 and Section 4). Then, in a second step the estimates are used to predict the outcomes for new subjects being sequentially chosen from the same population (Section 5). The success of the predictions is studied by simulations from which the agreement between

predicted and actual outcomes are summarized by the predictive values for the outcomes $X = 0$ and $X = 1$, as well as the proportion of correct predictions. Special attention is devoted to the sample size needed to get reliable estimates of $\mathbb{P}(X = 1 \mid \mathbf{z})$ in the first step, but also to the sample size needed to get stable estimates of the predictive values in the second step. In the simulation study data from a study, called the ISSA-project, will be used (Bergendorff et al. (1997), (2001) and Riksförsäkringsverket och Sahlgrenska Universitetssjukhuset (1997)). In the latter, work resumption among sick-listed men and women with lower back- and neck pain was considered. Here, 5-10 predictors were chosen from more than 200 variables. The extraction of predictors from the original list of variables was made by simply choosing those variables for which a change in the variable value caused the largest change in the empirical probability of work resumption. The variables selection process will not be considered in this paper. Instead attention will be paid to the problem of how to use a given number of predictors in an optimal way. These issues are further considered in (Persson (2002)). The paper finally ends with a discussion in Section 6.

2 Notations and Some Basic Results

Let the binary outcome variable X denote the health state for a given individual, 'non-healthy' $(X = 1)$ and 'healthy' $(X = 0)$, with probability $p^{(x)} = \mathbb{P}(X = x)$, $x = 0, 1$. Groups of predictors such that elements within groups are dependent and elements in different groups are independent will be called independent groups. In general, it will be assumed that the complete vector of predictors $\mathbf{Z}$ can be decomposed into g independent groups of predictors given $X = 1$, $\mathbf{Z}_1, ..., \mathbf{Z}_g$ and h independent groups given $X = 0$, $\mathbf{Z}_1, ..., \mathbf{Z}_h$. The conditional probabilities are defined as

$$\begin{aligned} \mathbb{P}(\mathbf{Z}_r = \mathbf{z}_r \mid X = x) &= q^{(x)}(\mathbf{z}_r) \text{ and} \\ \mathbb{P}(\mathbf{Z}_s = \mathbf{z}_s \mid X = x) &= q^{(x)}(\mathbf{z}_s), \end{aligned} \tag{1}$$

where $x = 0, 1$, $r = 1, ..., g$ and $s = 1, ..., h$. Thus,

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid X = x) = q^{(x)}(\mathbf{z}) = \begin{cases} \prod_{r=1}^{g} q^{(1)}(\mathbf{z}_r) \\ \prod_{s=1}^{h} q^{(0)}(\mathbf{z}_s) \end{cases}.$$

The observed frequencies corresponding to the outcomes in (1) are denoted by $N^{(x)}(\mathbf{z}_r)$ and $N^{(x)}(\mathbf{z}_s)$, respectively. Obviously, $\sum_{\mathbf{z}} N^{(x)}(\mathbf{z}) = N^{(x)}$, $x = 0, 1$ and $N^{(1)} + N^{(0)} = n$, the fixed total sample size. The above notations are illustrated in Table 1 for the case with two binary predictors.

		$Z_2 \mid X = x$		
		0	1	
$Z_1 \mid X = x$	0	$N^{(x)}(0,0), q^{(x)}(0,0)$	$N^{(x)}(0,1), q^{(x)}(0,1)$	$N_1^{(x)}(0), q_1^{(x)}(0)$
	1	$N^{(x)}(1,0), q^{(x)}(1,0)$	$N^{(x)}(1,1), q^{(x)}(1,1)$	$N_1^{(x)}(1), q_1^{(x)}(1)$
		$N_2^{(x)}(0), q_2^{(x)}(0)$	$N_2^{(x)}(1), q_2^{(x)}(1)$	$N^{(x)}, 1$

Table 1: Cell frequencies and probabilities with two predictor variables, where $x = 0, 1$.

The probability of interest is $\pi = \mathbb{P}(X = 1 \mid \mathbf{z})$, and from Bayes theorem it follows that

$$\pi = \frac{\mathbb{P}(X = 1) \cdot \mathbb{P}(\mathbf{Z} \mid X = 1)}{\sum_x \mathbb{P}(X = x) \cdot \mathbb{P}(\mathbf{Z} \mid X = x)} = \frac{A}{1 + A}, \text{ where } A = \frac{p^{(1)} q^{(1)}(\mathbf{z})}{p^{(0)} q^{(0)}(\mathbf{z})}. \quad (2)$$

Note that the quantities π and A in (2) are functions of $\mathbf{z}$ although this notation has been suppressed for convenience. Thus, with k binary predictors there are 2^k possible outcomes for π and A.

When all predictors are independent, both conditionally on $X = 1$ and on $X = 0$, then $q^{(x)}(\mathbf{z})$ is a product of the marginal probabilities. For practical reasons it is often a great advantage if conditional independency between predictors, or at least between sets of predictors, can be assumed. This is because empty individual cells are more likely to appear than empty marginal cells, and under independency the probability π can be estimated from marginal frequencies

with greater accuracy than from within-cell frequencies. For example, with 11 binary predictors there are $2^{11} = 2048$ individual cells, in contrast to $2 \cdot 11 = 22$ marginal cells. In addition to the case with no independent sets of predictors and the case with independent predictors, there are a variety of cases with partial independency.

The conditional variable $\left(N^{(x)}(\mathbf{z}) \mid N^{(x)} = n^{(x)}\right)$ is obviously multinomially distributed with parameters $n^{(x)}$ and $\mathbf{q}^{(x)}$, where $\mathbf{q}^{(x)}$ is vector of all possible probabilities which have been assigned to $\mathbf{Z}$. Thus, for binary predictors $\mathbf{q}^{(x)} = \left(q^{(x)}(1,...,1),...,q^{(x)}(0,...,0)\right)$. The probability generating function (pgf) of $M\left(n^{(x)}, \mathbf{q}^{(x)}\right)$ can be expressed as

$$\mathsf{E}\left[\prod_{z_1 \ldots z_k} \left(s^{(x)}_{z_1 \ldots z_k}\right)^{N^{(x)}(\mathbf{z})} \mid N^{(x)} = n^{(x)}\right] = \left[\mathbf{s}^{(x)} \left(\mathbf{q}^{(x)}\right)^T\right]^{n^{(x)}}$$

where $\mathbf{s}^{(x)} = \left(s^{(x)}_{1\ldots1}, ..., s^{(x)}_{z_1\ldots z_k}, ..., s^{(x)}_{0\ldots0}\right)$ and $\left(\mathbf{q}^{(x)}\right)^T$ is the transpose of $\mathbf{q}^{(x)}$.

Lemma 1 *The vector of all cell frequencies* $\left(N^{(1)}(\mathbf{z}) \vdots N^{(0)}(\mathbf{z})\right)$ *is multinomially distributed with parameters* $\left(n, p^{(1)}\mathbf{q}^{(1)} \vdots\, p^{(0)}\mathbf{q}^{(0)}\right)$.

Proof of Lemma 1.

$$\begin{aligned}
&\mathsf{E}\left[\prod_{z_1 \ldots z_k} \left(s^{(1)}_{z_1 \ldots z_k}\right)^{N^{(1)}(\mathbf{z})} \prod_{z_1 \ldots z_k} \left(s^{(0)}_{z_1 \ldots z_k}\right)^{N^{(0)}(\mathbf{z})} \mid N^{(1)} = n^{(1)}\right] \\
&= \mathsf{E}\left[\prod_{z_1 \ldots z_k} \left(s^{(1)}_{z_1 \ldots z_k}\right)^{N^{(1)}(\mathbf{z})} \mid N^{(1)} = n^{(1)}\right] \cdot \mathsf{E}\left[\prod_{z_1 \ldots z_k} \left(s^{(0)}_{z_1 \ldots z_k}\right)^{N^{(0)}(\mathbf{z})} \mid N^{(0)} = n - n^{(1)}\right] \\
&= \left[\mathbf{s}^{(1)} \left(\mathbf{q}^{(1)}\right)^T\right]^{n^{(1)}} \cdot \left[\mathbf{s}^{(0)} \left(\mathbf{q}^{(0)}\right)^T\right]^{n - n^{(1)}}.
\end{aligned}$$

Now, $N^{(1)}$ is binomially distributed with parameters n and $p^{(1)}$. Thus, by taking the expectation of the last expression over $N^{(1)}$ we obtain the pgf of

$\left(N^{(1)}(\mathbf{z}) \vdots N^{(0)}(\mathbf{z})\right)$ as

$$\mathsf{E}\left[\left(\frac{\mathbf{s}^{(1)}\left(\mathbf{q}^{(1)}\right)^{T}}{\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^{T}}\right)^{N^{(1)}}\right] \cdot \left[\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^{T}\right]^{n}$$

$$= \left[\left(\frac{\mathbf{s}^{(1)}\left(\mathbf{q}^{(1)}\right)^{T}}{\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^{T}}\right) p^{(1)} + p^{(0)}\right]^{n} \cdot \left[\mathbf{s}^{(0)}\left(\mathbf{q}^{(0)}\right)^{T}\right]^{n}.$$

■

From Lemma 1 it follows that cell frequencies with equal as well as different values of x are negatively correlated. Consider for instance the data in Table 1. Here we obtain,

$$\begin{aligned}
\mathsf{Cov}\left(N^{(1)}(1,1), N^{(1)}(0,0)\right) &= -n\left(p^{(1)}\right)^{2} q^{(1)}(1,1)\, q^{(1)}(0,0) \\
\mathsf{Cov}\left(N^{(1)}(1,1), N^{(0)}(1,1)\right) &= -np^{(1)}\left(1-p^{(1)}\right) q^{(1)}(1,1)\, q^{(0)}(1,1).
\end{aligned}$$

When the predictors are dependent rather than independent, we may, for some combinations of the parameters of $p^{(x)}$ and $q^{(x)}(\mathbf{z})$ obtain extremely different results. To show this we calculate the difference between the probability π in the independent and dependent case. For simplicity and without loss of generality, we consider only the case with two predictors where $Z_1 = 1$ and $Z_2 = 1$. Figure 1 shows the differences for various values of $p^{(1)}/p^{(0)}$ with all possible 2×2 contingency tables with probabilities $.05\,(.1)\,.95$. The differences are symmetric when $p^{(1)}/p^{(0)} = 1$. Although, it is impossible from Figure 1 to identify the parameter values of $q^{(x)}(\mathbf{z})$, calculations show that the differences tends to zero when the parameter values are similar in both tables i.e. when $q^{(1)}(1,1) \approx q^{(0)}(1,1)$, for all values of $p^{(1)}/p^{(0)}$. The purpose of this illustration is to show that, in fact, it does matter if we assume that the predictors are independent or not.

Expression (2) seems to be the simplest way to express the dependency between π and the q-probabilities, but there are other ways. One is logistic regression.

Consider for example the case with two predictors which are dependent, both given $X = 1$ and $X = 0$. Then,

$$\begin{aligned}
\pi &= \mathbb{P}(X = 1 \mid z_1, z_2) = \frac{A}{1+A}, \text{ where} \\
A &= \frac{p^{(1)}}{p^{(0)}} \left(\frac{q^{(1)}(1,1)}{q^{(0)}(1,1)} \right)^{z_1 z_2} \left(\frac{q^{(1)}(1,0)}{q^{(0)}(1,0)} \right)^{z_1(1-z_2)} \\
&\quad \times \left(\frac{q^{(1)}(0,1)}{q^{(0)}(0,1)} \right)^{(1-z_1)z_2} \left(\frac{q^{(1)}(0,0)}{q^{(0)}(0,0)} \right)^{(1-z_1)(1-z_2)} \\
&= \exp\{\alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 z_2\}, \text{ where} \\
\alpha &= \log\left(\frac{p^{(1)} q^{(1)}(1,1)}{p^{(0)} q^{(0)}(1,1)} \right) \text{ is the intercept,} \\
\beta_1 &= \log\left(\frac{q^{(1)}(1,0)\, q^{(0)}(0,0)}{q^{(0)}(1,0)\, q^{(1)}(0,0)} \right), \beta_2 = \log\left(\frac{q^{(1)}(0,1)\, q^{(0)}(0,0)}{q^{(0)}(0,1)\, q^{(1)}(0,0)} \right) \text{ and} \\
\beta_3 &= \log\left(\frac{q^{(1)}(1,1)\, q^{(0)}(1,0)\, q^{(0)}(0,1)\, q^{(1)}(0,0)}{q^{(0)}(1,1)\, q^{(1)}(1,0)\, q^{(1)}(0,1)\, q^{(0)}(0,0)} \right) \text{ are regression parameters.}
\end{aligned}$$

In a similar way, it can be showed that in the case when the predictors are independent, both conditionally on $X = 1$ and $X = 0$ we obtain

$$\begin{aligned}
A &= \exp\{\alpha + \beta_1 z_1 + \beta_2 z_2\}, \text{ where} \\
\alpha &= \log\left(\frac{p^{(1)}}{p^{(0)}} \prod_{i=1}^{k} q_i^{(x)}(0) \right) \text{ and } \beta_i = \log\left(\frac{q_i^{(1)}(1)\, q_i^{(0)}(0)}{q_i^{(0)}(1)\, q_i^{(1)}(0)} \right)
\end{aligned}$$

for $i = 1, 2, ..., k$, where $q_i^{(x)}(z_i)$ denotes the marginal probabilities. With k dependent predictors there will be $2^k - 1$ β-coefficients, and this way of representing the q-probabilities will be extremely extensive. Notice also that omitting the interactions between the predictors in the logistic model is equivalent to assuming that the latter are independent.

Another approach for parametrization is the use of Bayes weights. Again, assume for simplicity that we have two predictors Z_1 and Z_1, then we may rewrite A in (2) as

$$\begin{aligned} & \exp\left\{\log\left(\frac{p^{(1)}}{p^{(0)}}\right)+\log\left(\frac{q^{(1)}(\mathbf{z})}{q^{(0)}(\mathbf{z})}\right)\right\} \\ = \quad & \exp\left\{\log\left(\frac{p^{(1)}}{p^{(0)}}\right)+\log\left(\frac{q^{(1)}(z_1)}{q^{(0)}(z_1)}\right)+\log\left(\frac{q^{(1)}(z_2)}{q^{(0)}(z_2)}\right)\right\}, \end{aligned}$$

where, $\log\left(q^{(1)}(z_i)/q^{(0)}(z_i)\right)$ are called Bayes weights (Afifi and Azen (1979), p. 306-10).

3 Point Estimation of π

The Maximum Likelihood (ML) estimator of the target probability in (2) is obtained as

$$\hat{\pi}=\frac{\hat{A}}{\hat{A}+1}, \text{ where } \hat{A}=\frac{\left(N^{(0)}\right)^{h-1}\prod_{r=1}^{g}N^{(1)}(\mathbf{z}_r)}{\left(N^{(1)}\right)^{g-1}\prod_{s=1}^{h}N^{(0)}(\mathbf{z}_s)}. \tag{3}$$

Some simple examples of (3) are:

$$(i)\ \hat{A}=\frac{N^{(1)}(\mathbf{z})}{N^{(0)}(\mathbf{z})},\ (ii)\ \hat{A}=\left(\frac{N^{(0)}}{N^{(1)}}\right)^{k-1}\prod_{i=1}^{k}\frac{N_i^{(1)}(z_i)}{N_i^{(0)}(z_i)},\ (iii)\ \hat{A}=\frac{N^{(0)}N_{13}^{(1)}(\mathbf{z})}{N_{12}^{(0)}(\mathbf{z})N_3^{(0)}(z_3)}(?!).$$

In (i) no sets of predictors are independent, and in (ii) all predictors are independent. In (iii), Z_1, Z_2 and Z_3 are dependent when $X=1$, while (Z_1, Z_2) and Z_3 are two independent groups of predictors when $X=0$.

The fact that (3) is the ML estimator is a direct consequence of Lemma 1. According to the latter, $N^{(x)}(\mathbf{z})/n$ and $N^{(x)}/n$ are the ML estimators of $p^{(x)}q^{(x)}(\mathbf{z})$ and $p^{(x)}$, respectively, so $N^{(x)}(\mathbf{z})/N^{(x)}$ is the ML estimator of $q^{(x)}(\mathbf{z})$ and from this the result in (3) follows.

Below some properties of the estimator in (3) are studied, and some expressions for the estimated variance are given. Results will be derived separately for the case when all predictors are dependent and for the more general case when g groups of predictors are dependent given $X = 1$ and h groups are dependent given $X = 0$. The reason for the separation of the two cases is that various degrees of approximations are used for deriving the results.

Case I. *No sets of predictors are independent*

The estimator in (3) is now obtained from the special case (i) above and an expression for the variance of the latter is given by

$$\mathsf{Var}\left[\hat{\pi}\right] = \frac{\pi\left(1-\pi\right)}{n}\left(\frac{1}{\pi'} + \frac{\left(1-\pi'\right)}{n\left(\pi'\right)^2}\right) = \frac{\pi\left(1-\pi\right)}{n} \cdot C, \text{ say,} \tag{4}$$

where $\pi' = p^{(1)}q^{(1)}\left(\mathbf{z}\right) + p^{(0)}q^{(0)}\left(\mathbf{z}\right)$. An estimator of the variance (4) is obtained from

$$\widehat{\mathsf{Var}}\left[\hat{\pi}\right] = \frac{\hat{\pi}\left(1-\hat{\pi}\right)}{\left(n-1\right)}\left(\frac{1}{\hat{\pi}'} + \frac{\left(1-\hat{\pi}'\right)}{n\left(\hat{\pi}'\right)^2}\right) = \frac{\hat{\pi}\left(1-\hat{\pi}\right)}{n} \cdot \hat{C}, \text{ say,} \tag{5}$$

where $\hat{\pi}' = n^{-1}\left(N^{(1)}\left(\mathbf{z}\right) + N^{(0)}\left(\mathbf{z}\right)\right)$.

In order to motivate these expressions, notice that according to Lemma 1 and the results (A1) and (A2) in the Appendix, it follows that, for a fixed value of $\mathbf{z}$, $N' = N^{(1)}\left(\mathbf{z}\right) + N^{(0)}\left(\mathbf{z}\right)$ is binomially distributed with parameters n and π' and also that $\left(N^{(1)}\left(\mathbf{z}\right)|N'\right)$ is binomially distributed with parameters N' and π. Thus, we obtain the expectation

$$\mathsf{E}\left[\hat{\pi}\right] = \underset{N'}{\mathsf{E}}\left[\mathsf{E}\left(\hat{\pi} \mid N'\right)\right] = \underset{N'}{\mathsf{E}}\left[\frac{N'\pi}{N'}\right] = \pi,$$

so $\hat{\pi}$ is unbiased. The variance is (Rao (1973), p. 97)

$$\mathsf{Var}\left[\hat{\pi}\right] = \underset{N'}{\mathsf{E}}\left[\mathsf{Var}\left(\hat{\pi} \mid N'\right)\right] + \underset{N'}{\mathsf{Var}}\left[\mathsf{E}\left(\hat{\pi} \mid N'\right)\right]$$

$$= \underset{N'}{\mathsf{E}}\left[\frac{N'\pi(1-\pi)}{(N')^2}\right] + \underset{N'}{\mathsf{Var}}[\pi] = \pi(1-\pi)\underset{N'}{\mathsf{E}}\left[(N')^{-1}\right] + 0. \tag{6}$$

Since there is a non-zero probability $\left[(1-\pi')^n\right]$ that N' takes the value 0, one should re-define the estimator of π either by adding 1 in the denominator or by conditioning on $N' > 0$. This would however make the estimator is unnecessary complicated in the large sample situation which is considered here. Instead a Taylor series expansion will be used. From Appendix (A4) it follows that

$$\underset{N'}{\mathsf{E}}\left[(N')^{-1}\right] \approx \frac{1}{n\pi'} + \frac{(1-\pi')}{(n\pi')^2}. \tag{7}$$

By inserting the approximate expectation (7) into (6) we obtain the variance in (4). The estimated variance in (5) is obtained by simply replacing the parameters π and π' by their obvious estimators. By using $n-1$ in the denominator rather than n, a slight improvement of the closeness to the true variance is obtained.

The expression for the variance of $\hat{\pi}$ in (4) agreed well with the true variance determined from simulations. However, there were some deviations depending on the sample size n and the parameters $q^{(x)}(\mathbf{z})$. The best agreement was obtained with a uniform distribution of the q-probabilities. A simulation study with four cells as in Table 1, showed that with a uniform distribution, the absolute relative difference was below 1% even for a relatively small sample size $n = 50$, and declined rapidly for larger values of n. The agreement became worse when one of the cell probabilities was close to 1. For example, with the parameter setting $q^{(x)}(1,1) = 0.93$, $q^{(x)}(1,0) = 0.02 = q^{(x)}(0,1)$, $q^{(x)}(0,0) = 0.03$, $x = 0, 1$, the absolute relative difference was as large as 60% for $n = 50$. In the latter case one has to choose $n = 400$ to keep the absolute relative difference below 5% and to choose $n = 800$ in order to keep it below 0.5%. It was also found that similar conclusions could be drawn about the average performance of the estimated variance in (5) as for (4).

Even though the last example is a rather extreme one, it illustrates that some caution is needed when (4) and (5) are used in situations where the cell probabilities are close to 0 or 1.

By means of (4) it is possible to study analytically how the variance of $\hat{\pi}$ depends on the parameters $p^{(1)}$, $q^{(1)}(\mathbf{z})$ and $q^{(0)}(\mathbf{z})$. When $p^{(1)} = \frac{1}{2}$ the variance is a symmetric function of $q^{(1)}(\mathbf{z})$ and $q^{(0)}(\mathbf{z})$ which decreases as the latter of the two quantities increase, as can be seen in Figure 2. For $p^{(1)} \neq \frac{1}{2}$ the behavior of the variance is more complicated. When $p^{(1)} < \frac{1}{2}$ the variance decreases with increasing $q^{(0)}(\mathbf{z})$, but now the variance has a local maximum at some $q^{(1)}(\mathbf{z}) > 0$ (Figure 3). The value of $q^{(1)}(\mathbf{z})$ which gives this maximum will increase as $p^{(1)}$ tends to zero. When $p^{(1)} > \frac{1}{2}$ the same pattern is observed, but with $q^{(1)}(\mathbf{z})$ interchanged by $q^{(0)}(\mathbf{z})$ (Figure 4).

Case II. *g sets of predictors are independent given X=1 and h sets of predictors are independent given X=0*

An expression for the variance of $\hat{\pi}$ is given by

$$\mathsf{Var}[\hat{\pi}] = \frac{\pi^2(1-\pi)^2}{n}\left\{n\left[\prod_{r=1}^{g}\left(1+\frac{1-q^{(1)}(\mathbf{z}_r)}{np^{(1)}q^{(1)}(\mathbf{z}_r)}\right)\right.\right.$$

$$\left.\left.+\prod_{s=1}^{h}\left(1+\frac{1-q^{(0)}(\mathbf{z}_s)}{np^{(0)}q^{(0)}(\mathbf{z}_s)}\right)-2\right]+\frac{1}{p^{(1)}p^{(0)}}\right\} = \frac{\pi^2(1-\pi)^2}{n}\cdot D, \text{ say.} \quad (8)$$

An estimator of $\mathsf{Var}[\hat{\pi}]$ is

$$\begin{aligned}\widehat{\mathsf{Var}}[\hat{\pi}] &= \frac{\hat{\pi}^2(1-\hat{\pi})^2}{(n-1)}\cdot\hat{D}, \text{ where} \quad (9)\\ \hat{D} &= n\left[\prod_{r=1}^{g}\left(1+\frac{1}{N^{(1)}(\mathbf{z}_r)}-\frac{1}{N^{(1)}}\right)\right.\\ &\quad\left.+\prod_{s=1}^{h}\left(1+\frac{1}{N^{(0)}(\mathbf{z}_s)}-\frac{1}{N^{(0)}}\right)-2\right]+\frac{n^2}{N^{(1)}N^{(0)}}\end{aligned}$$

In contrast to Case I, the denominator of $\hat{\pi}$ now consists of a sum of products of multinomial variables and the exact distribution of this is very complicated. Instead all derivations will be based on Taylor approximations.

From Appendix (A4) it follows that

$$\begin{aligned} \mathsf{Var}\left[\hat{\pi}\right] &\approx \frac{\mathsf{Var}\left[\hat{A}\right]}{\left(\mathsf{E}\left[\hat{A}\right]+1\right)^4}, \text{ where} \\ \mathsf{Var}\left[\hat{A}\right] &= \underset{N^{(1)}}{\mathsf{E}}\left[\mathsf{Var}\left(\hat{A} \mid N^{(1)}\right)\right] + \underset{N^{(1)}}{\mathsf{Var}}\left[\mathsf{E}\left(\hat{A} \mid N^{(1)}\right)\right]. \end{aligned} \tag{10}$$

$$\text{Now, } \mathsf{Var}\left(\hat{A} \mid N^{(1)}\right) = \frac{\left(N^{(0)}\right)^{2(h-1)}}{\left(N^{(1)}\right)^{2(g-1)}} \cdot \mathsf{Var}\left(\frac{\prod_{r=1}^{g} N^{(1)}\left(\mathbf{z}_r\right)}{\prod_{s=1}^{h} N^{(0)}\left(\mathbf{z}_s\right)} \mid N^{(1)}\right),$$

where $\prod_{r=1}^{g} N^{(1)}\left(\mathbf{z}_r\right)$ and $\prod_{s=1}^{h} N^{(0)}\left(\mathbf{z}_s\right)$ are two independent products conditionally on $N^{(1)}$. These products consist of independent variables, which are distributed $M\left(N^{(1)}, q^{(1)}\left(\mathbf{z}_r\right)\right)$ and $M\left(N^{(0)}, q^{(0)}\left(\mathbf{z}_s\right)\right)$, respectively. From Appendix (A3) it follows that, for fixed values of $\mathbf{z}_r$ and $\mathbf{z}_s$,

$$\begin{aligned} \mathsf{E}\left(\textstyle\prod_{r=1}^{g} N^{(1)}\left(\mathbf{z}_r\right) \mid N^{(1)}\right) &= \left(N^{(1)}\right)^{g} \textstyle\prod_{r=1}^{g} q^{(1)}\left(\mathbf{z}_r\right), \text{ and} \\ \mathsf{E}\left(\textstyle\prod_{s=1}^{h} N^{(0)}\left(\mathbf{z}_s\right) \mid N^{(0)}\right) &= \left(N^{(0)}\right)^{h} \textstyle\prod_{s=1}^{h} q^{(0)}\left(\mathbf{z}_s\right), \text{ while} \end{aligned}$$

$$\begin{aligned} & \mathsf{Var}\left(\textstyle\prod_{r=1}^{g} N^{(1)}\left(\mathbf{z}_r\right) \mid N^{(1)}\right) \\ = & \left(N^{(1)}\right)^{2g} \left(\textstyle\prod_{r=1}^{g} q^{(1)}\left(\mathbf{z}_r\right)\right)^2 \left\{\textstyle\prod_{r=1}^{g}\left(1+\frac{1-q^{(1)}\left(\mathbf{z}_r\right)}{N^{(1)} q^{(1)}\left(\mathbf{z}_r\right)}\right)-1\right\}, \text{ and} \\ & \mathsf{Var}\left(\textstyle\prod_{s=1}^{h} N^{(0)}\left(\mathbf{z}_s\right) \mid N^{(1)}\right) \\ = & \left(N^{(0)}\right)^{2h} \left(\textstyle\prod_{s=1}^{h} q^{(0)}\left(\mathbf{z}_s\right)\right)^2 \left\{\textstyle\prod_{s=1}^{h}\left(1+\frac{1-q^{(0)}\left(\mathbf{z}_s\right)}{N^{(0)} q^{(0)}\left(\mathbf{z}_s\right)}\right)-1\right\}. \end{aligned}$$

By using the Taylor expansion in Appendix (A4) it is seen that the variance of any ratio of independent variables X and Y can be written

$$\mathsf{Var}\left(\frac{X}{Y}\right) \approx \left(\frac{\mathsf{E}(X)}{\mathsf{E}(Y)}\right)^2 \left(\frac{\mathsf{Var}(X)}{[\mathsf{E}(X)]^2} + \frac{\mathsf{Var}(Y)}{[\mathsf{E}(Y)]^2}\right).$$

From the last results and by taking the approximate expectation over $N^{(1)}$ it finally follows that

$$\underset{N^{(1)}}{\mathsf{E}}\left[\mathsf{Var}\left(\hat{A} \mid N^{(1)}\right)\right] \approx A^2\left[\prod_{r=1}^{g}\left(1+\frac{1-q^{(1)}(\mathbf{z}_r)}{np^{(1)}q^{(1)}(\mathbf{z}_r)}\right)+\prod_{s=1}^{h}\left(1+\frac{1-q^{(0)}(\mathbf{z}_s)}{np^{(0)}q^{(0)}(\mathbf{z}_s)}\right)\right].$$

In a similar way it can be shown that

$$\underset{N^{(1)}}{\mathsf{E}}\left[\hat{A} \mid N^{(1)}\right] \approx \frac{N^{(1)}\prod_{r=1}^{g} q^{(1)}(\mathbf{z}_r)}{N^{(0)}\prod_{s=1}^{h} q^{(0)}(\mathbf{z}_s)},$$

and by again using the Taylor approximation in Appendix (A3) one gets

$$\underset{N^{(1)}}{\mathsf{Var}}\left[\mathsf{E}\left(\hat{A} \mid N^{(1)}\right)\right] \approx \frac{A^2}{n}\frac{1}{p^{(1)}p^{(0)}}.$$

The expression for $\mathsf{Var}[\hat{\pi}]$ in (8) is finally obtained from (10) and by using the fact that $A^2/(A+1)^4 = \pi^2(1-\pi)^2$.

The estimator of the variance in (9) is simply obtained by inserting obvious estimators for parameters.

When $g = 1 = h$, the expression in (8) should reduces to (4). However, in this case it is easily shown that (8) can be written as

$$\mathsf{Var}[\hat{\pi}] = \frac{\pi(1-\pi)}{n}\frac{1}{\pi'}.$$

Thus, the two expressions in (4) and (8) are the same if

$$\frac{\pi(1-\pi)}{n^2}\cdot\frac{(1-\pi')}{(\pi')^2} \approx 0.$$

The agreement between the expressions for the variance of $\hat{\pi}$ in (8), the estimated variance in (9), and the true variance was determined from 100,000 simulations. In this case the comparison is complicated by the fact that there are many q-probabilities involved, and therefore we only consider the case with two independent sets of mutually dependent predictors $\mathbf{Z}_1 = (Z_1, Z_2)$ and $\mathbf{Z}_2 = (Z_3, Z_4)$,

both given $X = 1$ and $X = 0$. By varying the parameters $p^{(1)}$, $q_{12}^{(x)}(z_1, z_2)$ and $q_{34}^{(x)}(z_3, z_4)$, $x = 0, 1$, it was found that the absolute difference between the variance of $\hat{\pi}$ in the simulations and the variance given by (8) and (9) with a few exceptions were below .001 for $n \geq 200$. In no case the difference was larger than .0003 for $n \geq 400$. In the sequel we choose $n = 400$ and study how the variance of $\hat{\pi}$ in (8) depends on the magnitude of the q-probabilities and also on the number of independent sets of predictors

Figures 5-12 illustrate how the variance simultaneously depends on $q_{12}^{(1)}(z_1, z_2)$ and $q_{34}^{(1)}(z_3, z_4)$ for some values of $p^{(1)}$, $q_{12}^{(0)}(z_1, z_2)$ and $q_{34}^{(0)}(z_3, z_4)$. All variances are considered for a fixed set of (Z_1, Z_2, Z_3, Z_4), e.g. $(1, 1, 0, 1)$. Therefore, the z-arguments have been omitted in the legends to the figures. In Figure 5 it is seen that the variance is a symmetric function of its arguments when $p^{(1)} = \frac{1}{2}$ and $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot)$. For $p^{(1)} < \frac{1}{2}$ (see Figures 6-12), the pattern is more complex and in this case one can identify a saddle-point. The level of the latter increases as $q_{12}^{(0)}(z_1, z_2) = q_{34}^{(0)}(z_3, z_4)$ tends to zero, while at the same time the saddle becomes tighter. For $p^{(1)} > \frac{1}{2}$ this saddle-point pattern vanishes and the variance increases as $q_{12}^{(0)}(z_1, z_2)$ and $q_{34}^{(0)}(z_3, z_4)$ tends to zero (not shown in the figures).

To study how the variance of $\hat{\pi}$ depends on the number of independent sets of predictors some simplifications have to be made. Put $g = h$, so there is an equal number of sub-groups of independent predictors both given $X = 1$ and given $X = 0$, and assume that all $q^{(1)}(\mathbf{z}) = q^{(1)}$ and $q^{(0)}(\mathbf{z}) = q^{(0)}$ while $p^{(1)} = \frac{1}{2}$. Then Figure 14 shows that the variance of $\hat{\pi}$ increases with increasing g as far as $q^{(1)} = q^{(0)}$, and that the increase is larger for small $q's$. When $q^{(1)} \neq q^{(0)}$ there is a different pattern. For large differences between the $q's$, the variance declines with increasing value of g, but for smaller differences the variance has a local maximum before it starts to decline. These findings suggest that much can be gained if it is possible to find (1) many predictors with the property that (2) the q-probabilities $q^{(1)}(\mathbf{z})$ differ much from $q^{(0)}(\mathbf{z})$. On the other hand, failure to identify predictors with different q-probabilities, or including such predictors for some reasons, will increase the variance of $\hat{\pi}$.

4 Interval Estimation of π

When the estimated value of π is used for predicting the state of an individual, it is customary to make the predictions '$X = 1$' if $\pi > \frac{1}{2}$ and '$X = 0$' if $\pi < \frac{1}{2}$ if the costs of misclassification are unknown. Such rigid classification rules may be useful if one wants to evaluate the prediction ability of certain predictors, but for practical purpose they can be risky. The predicted outcome of an individual sometimes calls for an intervention, by for instance offer the individual medical rehabilitation programs. Wrong predictions may then be very expensive. If the costs of misclassification are known, the rigid rule above can be replaced by generalized Bayes classification rules, which minimize the expected cost of misclassification (Afifi and Azen (1979), p. 292). However, the costs are seldom known, or may be hard to quantify. In such cases it may be wise to compute a confidence interval (CI) for π. CI's that are clearly outside $\frac{1}{2}$, can be considered to indicate that the corresponding predictions are more likely than CI's that cover $\frac{1}{2}$. In this section we consider various ways to construct a CI for π. As in the preceding section, two cases will be treated separately.

Case I. *No sets of predictors are independent*

We will compare the expected length and actual coverage probability of five different CI's. Let T d. as. $N(0,1)$ denote that a statistic T asymptotically has a standard normal distribution. The various CI's are derived from the following properties, where the same notations are used as in Section 3.

$$(i)\ \frac{\hat{\pi}-\pi}{\{\mathsf{Var}\,[\hat{\pi}]\}^{1/2}}\ \text{d. as. } N(0,1), \quad (ii)\ \frac{\hat{\pi}-\pi}{\left\{\frac{\pi(1-\pi)}{n}\cdot\hat{C}\right\}^{1/2}}\ \text{d. as. } N(0,1),$$

$$(iii)\ \frac{\hat{\pi}-\pi}{\left\{\frac{\pi(1-\pi)}{N'}\right\}^{1/2}}\ \text{d. as. } N(0,1),\ (iv)\ \left(N^{(1)}(\mathbf{z})\,|N'\right)\ \text{d. } B(N',\pi)\ \text{and}$$

$$(v)\ \frac{\log\left(\hat{A}\right)-\log(A)}{\left\{\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right]\right\}^{1/2}}\ \text{d. as. } N(0,1).$$

Here the statistics in (iii) and (iv) are conditional and based on the particular

outcome $N' = N^{(1)}(\mathbf{z}) + N^{(0)}(\mathbf{z})$, while $\log\left(\hat{A}\right)$ is an estimator of $\log(A)$ to be considered below. Let Z be the $100(1-\alpha/2)\%$ percentile of the standard normal distribution, and let $F(n_1, n_2)$ denote the $100(1-\alpha/2)\%$ percentile of the F-distribution with n_1 and n_2 degrees of freedom. Then the CI's derived from $(i)-(iv)$ are $\hat{\pi}_L < \pi < \hat{\pi}_U$, where $\hat{\pi}_L$ and $\hat{\pi}_U$ are obtained from:

$$(i)\ \ \hat{\pi} \pm Z \cdot \{\mathsf{Var}[\hat{\pi}]\}^{1/2}$$

$$(ii)\ \ \frac{2\hat{\pi}+\frac{Z^2\hat{C}}{n} \pm \left\{\left(2\hat{\pi}+\frac{Z^2\hat{C}}{n}\right)^2 - 4\hat{\pi}^2\left(1+\frac{Z^2\hat{C}}{n}\right)\right\}^{1/2}}{2\left(1+\frac{Z^2\hat{C}}{n}\right)}$$

$$(iii)\ \ \frac{2\hat{\pi}+\frac{Z^2}{N'} \pm \left\{\left(2\hat{\pi}+\frac{Z^2}{N'}\right)^2 - 4\hat{\pi}^2\left(1+\frac{Z^2}{N'}\right)\right\}^{1/2}}{2\left(1+\frac{Z^2}{N'}\right)} \qquad (11)$$

(iv)

$$\begin{aligned}
\hat{\pi}_L &= \frac{N^{(1)}(\mathbf{z})}{\left(N^{(0)}(\mathbf{z})+1\right)F\left[2\left(N^{(0)}(\mathbf{z})+1\right), 2N^{(1)}(\mathbf{z})\right]}, \\
\hat{\pi}_U &= \frac{\left(N^{(1)}(\mathbf{z})+1\right)F\left[2\left(N^{(1)}(\mathbf{z})+1\right), 2N^{(0)}(\mathbf{z})\right]}{N^{(0)}(\mathbf{z}) + \left(N^{(1)}(\mathbf{z})+1\right)F\left[2\left(N^{(1)}(\mathbf{z})+1\right), 2N^{(0)}(\mathbf{z})\right]}
\end{aligned}$$

$$(v)\ \ \frac{\exp\left\{\log\left(\hat{A}\right) \pm 1.96\left\{\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right]\right\}^{1/2}\right\}}{1+\exp\left\{\log\left(\hat{A}\right) \pm 1.96\left\{\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right]\right\}^{1/2}\right\}}, \text{ where}$$

$$\begin{aligned}
\log\left(\hat{A}\right) &= \log\left(N^{(1)}(\mathbf{z})\right) - \log\left(N^{(0)}(\mathbf{z})\right) \text{ and} \\
\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right] &= \frac{1}{N^{(1)}(\mathbf{z})} + \frac{1}{N^{(0)}(\mathbf{z})}.
\end{aligned}$$

The expressions (11) : $(i)-(iv)$ follows from well known results (Casella and Berger (1990), p. 444-49). (11) : (v) follows from very rough approximations (see Appendix (A4)) $\mathsf{E}\left[\log\left(\hat{A}\right)\right] \approx \log(A)$, and

$$\mathsf{Var}\left[\log\left(\hat{A}\right)\right] \approx \frac{\mathsf{Var}\left[N^{(1)}(\mathbf{z})\right]}{\left\{\mathsf{E}\left[N^{(1)}(\mathbf{z})\right]\right\}^2} + \frac{\mathsf{Var}\left[N^{(0)}(\mathbf{z})\right]}{\left\{\mathsf{E}\left[N^{(0)}(\mathbf{z})\right]\right\}^2} - 2\frac{\mathsf{Cov}\left[N^{(1)}(\mathbf{z}), N^{(0)}(\mathbf{z})\right]}{\mathsf{E}\left[N^{(1)}(\mathbf{z})\right]\mathsf{E}\left[N^{(0)}(\mathbf{z})\right]},$$

where

$$\begin{aligned}
\mathsf{E}\left[N^{(x)}(\mathbf{z})\right] &= np^{(x)}q^{(x)}(\mathbf{z}) \\
\mathsf{Var}\left[N^{(x)}(\mathbf{z})\right] &= np^{(x)}q^{(x)}(\mathbf{z})\left(1-p^{(x)}q^{(x)}(\mathbf{z})\right), x=0,1, \text{ and} \\
\mathsf{Cov}\left[N^{(1)}(\mathbf{z}), N^{(0)}(\mathbf{z})\right] &= -np^{(1)}p^{(0)}q^{(1)}(\mathbf{z})\,q^{(0)}(\mathbf{z}).
\end{aligned}$$

This implies that

$$\begin{aligned}
\mathsf{Var}\left[\log\left(\hat{A}\right)\right] &\approx \frac{1}{n}\left(\frac{1}{p^{(1)}q^{(1)}(\mathbf{z})}+\frac{1}{p^{(0)}q^{(0)}(\mathbf{z})}\right), \text{ and hence} \\
\widehat{\mathsf{Var}}\left[\log\left(\hat{A}\right)\right] &\approx \frac{1}{N^{(1)}(\mathbf{z})}+\frac{1}{N^{(0)}(\mathbf{z})}.
\end{aligned}$$

The simple expression in (11) : (v) is worth a comment. $\log\left(\hat{A}\right)$ is in fact a poor estimator of $\log(A)$. By instead using the alternative estimator

$$\log\left(\hat{A}\right)+\frac{1}{2}\left(\frac{1}{N^{(1)}(\mathbf{z})}+\frac{1}{N^{(0)}(\mathbf{z})}\right),$$

which follows by considering terms of the order n^{-1} in the Taylor expansion of $\mathsf{E}\left[\log\left(\hat{A}\right)\right]$, both bias and variance can be reduced substantially. The estimated variance of this alternative estimator is

$$\begin{aligned}
&\frac{1}{N^{(1)}(\mathbf{z})}+\frac{1}{N^{(0)}(\mathbf{z})}-\frac{1}{\left[N^{(1)}(\mathbf{z})\right]^2}-\frac{1}{\left[N^{(0)}(\mathbf{z})\right]^2} \\
&+\frac{1}{4}\left(\frac{1}{\left[N^{(1)}(\mathbf{z})\right]^3}+\frac{1}{\left[N^{(0)}(\mathbf{z})\right]^3}\right)-\frac{1}{4n}\left(\frac{1}{N^{(1)}(\mathbf{z})}+\frac{1}{N^{(0)}(\mathbf{z})}\right)^2.
\end{aligned}$$

To illustrate the difference between the two estimators of $\log(A)$, consider the case when there are 2 dependent predictors Z_1 and Z_2, given $X=1$ and given $X=0$, and with the parameter setting

$$
\begin{aligned}
q_{12}^{(1)}(1,1) &= .24, q_{12}^{(1)}(1,0) = .38, q_{12}^{(1)}(0,1) = .11, q_{12}^{(1)}(0,0) = .27, \\
q_{12}^{(0)}(1,1) &= .71, q_{12}^{(0)}(1,0) = .25, q_{12}^{(0)}(0,1) = .02, q_{12}^{(0)}(0,0) = .02.
\end{aligned}
$$

A simulation study using the relatively large sample size of $n = 400$, showed that the alternative estimator had a relative bias which was more than 50% smaller than the original estimator. The variance was reduced by 35% and the expression above for the estimated variance of the alternative estimator was very close to the actual variance. However, when the alternative estimator was used for making CI's, the distribution of the pivotal statistic for (v) was slightly skew, and for this reason the coverage rate of 95% was not maintained. The actual coverage rate could in fact drop down to 91%. This illustrates that a CI based on a crude estimator may perform better than a CI based on a more sophisticated estimator.

The performance of the CI's in (11) : $(i) - (v)$ was found to depend on the q-probabilities. As for the expressions (4) and (5) in Section 3, the worst case was obtained when one of the cell probabilities are close to 1. This is illustrated in Table 2, where the 5 CI's are compared regarding expected length and coverage probability. First of all one may notice that none of the CI's keeps the stipulated level of 95% if the sample size, n, is 100 or less. For $n = 200$ the 95%-level is only maintained by (11) : (ii) and possibly by (11) : (iii). However, the expected lengths of the latter are too large to be accepted. When the q-probabilities tend to be more uniformly distributed, the probability that the 95% level is maintained increases, also for smaller samples. The overall conclusion is that (11) : (ii) performs best, even if the CI's may be somewhat conservative. When n is large the computational simple expression in (11) : (v) may be an alternative. (11) : (i) should be avoided. The latter CI's did not even maintain the 95% level in the most favorable case with uniformly distributed q-probabilities and $n = 1600$.

Case II. *g predictors are independent given X=1 and h predictors are independent given X=0*

Now the CI's are derived from the following properties, where the same notations are used as in Section 3 for Case I:

$$(i)\ \frac{\hat{\pi}-\pi}{\left\{\frac{\pi^2(1-\pi)^2}{n}\cdot\hat{D}\right\}^{1/2}}\ \text{d. as. } N(0,1),\ (ii)\ \frac{\log\left(\hat{A}\right)-\log(A)}{\left\{\hat{D}/n\right\}^{1/2}}\ \text{d. as. } N(0,1).$$

Due to the complexity of the statistic A in this case, we do not consider any conditional statistics, as in Case I. The CI's of π derived from (i) and (ii) above now are $\hat{\pi}_L < \pi < \hat{\pi}_U$, where $\hat{\pi}_L$ and $\hat{\pi}_U$ are the solutions of

$$(i)\ \frac{\left(z\sqrt{\hat{D}/n}\pm 1\right)\mp\sqrt{\left(z\sqrt{\hat{D}/n}\pm 1\right)^2\mp 4z\hat{\pi}\sqrt{\hat{D}/n}}}{2z\sqrt{\hat{D}/n}} \tag{12}$$

$$(ii)\ \frac{\exp\left\{\log\left(\hat{A}\right)\pm 1.96\left\{\hat{D}/n\right\}^{1/2}\right\}}{1+\exp\left\{\log\left(\hat{A}\right)\pm 1.96\left\{\hat{D}/n\right\}^{1/2}\right\}}$$

In (i) the upper part of the two signs $\pm$ and $\mp$ refers to $\hat{\pi}_L$ and the lower part to $\hat{\pi}_U$. In (ii) the upper part of $\pm$ refers to $\hat{\pi}_U$ and the lower part to $\hat{\pi}_L$.

(i) follows from the following arguments. Put $f(\pi) = (\hat{\pi}-\pi)/(\pi-\pi^2)$. Then the statement $-z < (\hat{\pi}-\pi)/\sqrt{\mathsf{Var}[\hat{\pi}]} < z$ is equivalent to $-z\sqrt{D/n} < f(\pi) < z\sqrt{D/n}$, where the meaning of D is clear from (8). Here $f(\pi)$ is a monotonously decreasing function of $\pi \in (0,1)$ for all $\hat{\pi} \in (0,1)$ with the inverse

$$\pi = \frac{f+1-\sqrt{(f+1)^2-4f\hat{\pi}}}{2f},$$

which gives the CI in (i).

Now, $\log\left(\hat{A}\right)$ can be written $\log(\hat{\pi}) - \log(1-\hat{\pi})$, and by using (8) together with Appendix (A4) one gets

$$\mathsf{E}\left[\log\left(\hat{A}\right)\right] \approx \log(A) + \left(\pi - \frac{1}{2}\right) \cdot \frac{D}{n} \text{ and } \mathsf{Var}\left[\log\left(\hat{A}\right)\right] \approx \frac{D}{n},$$

which motivates the use of the statistic in (ii). The expression for the CI in (12) follows easily by noticing that

$$c_L < \log(A) < c_U \text{ implies that } \frac{\exp\{c_L\}}{1+\exp\{c_L\}} < \pi < \frac{\exp\{c_U\}}{1+\exp\{c_U\}}.$$

When $\hat{D}$ in (12) is used for constructing a confidence interval for π, $N^{(1)}(\mathbf{z}_r)$ and $N^{(0)}(\mathbf{z}_s)$ in (12) should be replaced by $N^{(1)}(\mathbf{z}_r)+1$ and $N^{(0)}(\mathbf{z}_s)+1$, respectively. This will make the confidence interval less conservative.

Tables 3 and 4 show expected lengths and coverage probabilities for the two CI's in (12), the latter being determined from simulations. The differences between the two are very small. (12) : (i) tends to give somewhat shorter CI's, but (12) : (ii) tends to give CI's which agree better with the stipulated level of 95%. Again we point out that, although $\log\left(\hat{A}\right)$ is a poor estimator of $\log(A)$, CI's constructed from $\log\left(\hat{A}\right)$ perform well.

5 Prediction

In this section we consider the possibility to predict the outcomes $X = 1$ and $X = 0$ based on $\hat{\pi}$, the estimates of π. The outcome $X = 1$ will be predicted whenever $\hat{\pi} > \frac{1}{2}$ and otherwise the outcome $X = 0$ will be predicted. This rather strict classification rule is chosen merely for simplicity. In practical work it would perhaps be better to use a less rigid classification rule and take the CI's for π into consideration. The predictions will be performed in a two-step approach, where in the first step π is estimated from a sample of a certain population, and then in a second step this estimate is used to predict the outcomes for new subjects being chosen from the same population. If the predicted outcome is denoted by XP, the success of the predictions will be measured by the predictive

values $\mathbb{P}(X = 1 \mid XP = 1)$ and $\mathbb{P}(X = 0 \mid XP = 0)$, and the probability of a correct prediction $\mathbb{P}(Correct)$ (see Ch. 3 in Campbell and Machin (1990)). Of special interest will be to study how the predicting ability depends on the sample size, which is used in the first step to estimate π, and also to determine the sample size, which is needed in the second step for reaching stable estimates of the measures of predicting ability. Attention will also be paid to study how miss specification of the dependency structure of the predictors may affect the predicting ability.

5.1 A Simulation Example

In this section we consider the ability to predict work resumption for long-termed sick-listed subjects. The sample considered here is a part of a larger sample within the ISSA-study that has previously been described in detail (Bergendorff et al. (1997), (2001) and Riksförsäkringsverket och Sahlgrenska Universitetssjukhuset (1997)), and consisted of 545 full-time working employed men sick-listed for at least 28 days because of a lower back pain diagnosis. After 28 days the values on the following predictor variables were obtained: (1) Age, (2) Complete rehabilitation plan, (3) Comorbidity, (4) Working ability, (5) Sick-listing in family, (6) Suitable working tasks, (7) Ethnicity, (8) Heavy lifts. Here, Comorbidity means that the subjects has other diseases than lower back pain. Working ability was subjectively assessed on a scale ranking from 1 (low) to 10 (high). Suitable working tasks means that the employer was willing to adjust the working tasks in agreement with the subject's state of health. In a previous study, these variables were found to be the most important ones for predicting work resumption among men with lower back pain (Bergendorff et al. (2001)).

The outcomes to predict at 90 days are $X = 1$, if there is no work resumption and $X = 0$ otherwise. The predictor variables were dichotomized in the following way. Age $= Z_1 = 1$, if age > 30 years and 0 otherwise, Complete rehabilitation plan $(Z_2) = 1$, if yes and 0 otherwise, Comorbidity $(Z_3) = 1$, if yes and 0

otherwise, Working ability $(Z_4) = 1$, if scale value < 5 and 0 otherwise, Sick-listening in family $(Z_5) = 1$, if yes and 0 otherwise, Suitable working tasks $(Z_6) = 1$, if no and 0 otherwise, Ethnicity $(Z_7) = 1$, if Swedish and 0 otherwise and Heavy lift $(Z_8) = 1$, if yes and 0 otherwise.

Notice that all binary predictors have been defined in such a way that the outcome 1 of a predictor favors the outcome $X = 1$. The reasons for dichotomizing the variables Age and Working ability have given previously (Bergendorff et al. (2001)). Although the variable Age has been found to be continuously negatively related to the probability of work resumption in other studies (Jonsson (2001)), this was not the case in the present study where the selected subjects differed from the test of the population in several aspects. E.g. all were full-time working employed.

In this example the first task is to estimate

$$\pi = \mathbb{P}(X = 1 \mid \mathbf{z}) = \frac{p^{(1)} q^{(1)}(\mathbf{z})}{p^{(1)} q^{(1)}(\mathbf{z}) + p^{(0)} q^{(0)}(\mathbf{z})}.$$

A hierarchical cluster analysis (Anderberg (1973) and Jobson (1992)) suggested the following independent sets of vectors

$$\begin{aligned}
(\mathbf{Z} \mid X = 1) &= \{(Z_1, Z_2, Z_3 \mid X = 1), (Z_4, Z_5 \mid X = 1), (Z_6, Z_7, Z_8 \mid X = 1)\} \\
(\mathbf{Z} \mid X = 0) &= \{(Z_1, Z_8 \mid X = 0), (Z_3, Z_4, Z_6 \mid X = 0), (Z_2, Z_5, Z_7 \mid X = 0)\}
\end{aligned}$$

Thus, e.g. Z_1 (age) and Z_3 (comorbidity) were correlated among those who did not return to work after 90 days, but uncorrelated among those who returned to work. For a more detailed description of the dependency structures the reader is referred to the paper by Persson (2002). The corresponding q-probabilities were

$$\begin{aligned}
q^{(1)}(\mathbf{z}) &= q^{(1)}(z_1, z_2, z_3) \cdot q^{(1)}(z_4, z_5) \cdot q^{(1)}(z_6, z_7, z_8) \\
q^{(0)}(\mathbf{z}) &= q^{(0)}(z_1, z_8) \cdot q^{(0)}(z_3, z_4, z_6) \cdot q^{(0)}(z_2, z_5, z_7)
\end{aligned}$$

where,

z_1, z_2, z_3	$q^{(1)}(z_1, z_2, z_3)$	z_4, z_5	$q^{(1)}(z_4, z_5)$	z_6, z_7, z_8	$q^{(1)}(z_6, z_7, z_8)$
111	.09	11	.02	111	.22
110	.02	10	.17	110	.02
101	.50	01	.21	101	.06
011	.01	00	.60	011	.35
100	.27	–	–	100	.01
010	.01	–	–	010	.01
001	.07	–	–	001	.23
000	.03	–	–	000	.10

z_1, z_8	$q^{(0)}(z_1, z_8)$	z_3, z_4, z_6	$q^{(0)}(z_3, z_4, z_6)$	z_2, z_5, z_7	$q^{(0)}(z_2, z_5, z_7)$
11	.64	111	.01	111	.02
10	.24	110	.02	110	.02
01	.11	101	.01	101	.01
00	.01	011	.03	011	.05
–	–	100	.01	100	.01
–	–	010	.24	010	.30
–	–	001	.03	001	.05
–	–	000	.65	000	.54

These q-probabilities were estimated from the data set, and will be used as fixed probabilities for generating samples in the simulation study. The prevalence $p^{(1)}$ was 0.54. This figure was also taken from the empirical study.

The various outcomes $(z_1, ..., z_8)$ give rise to 256 values of the estimated posterior probability π. The 5 smallest and largest of these are

$z_1z_2z_3z_4z_5z_6z_7z_8$	$\mathbb{P}(X = 1 \mid \mathbf{z})$	$z_1z_2z_3z_4z_5z_6z_7z_8$	$\mathbb{P}(X = 1 \mid \mathbf{z})$
11001010	.0156	01010000	.9852
11011010	.0316	01110000	.9852
10000010	.0391	01000100	.9860
11000010	.0432	01000110	.9860
10001010	.0786	01000111	.9860

Here one may notice that $z_1 = 1$ (age > 30 years) in all cases giving the smallest probability, while $z_1 = 0$ in all cases giving the largest probabilities.

The simulation experiment was performed in the following way: First, one sample was selected, each being based on the sample sizes $n = 25, 50, 100, 200, ..., 1000$, and from each sample π was estimated. The latter quantity was then used to

predict the outcome at 90 days for new subjects being selected from the same population. The number of new sampled subjects was $m = 1000, ..., 100000$, and for each of these, the outcome $X = 1$ was predicted ($XP = 1$) if $\hat{\pi} > \frac{1}{2}$, and the outcome $X = 0$ was predicted ($XP = 0$) if $\hat{\pi} < \frac{1}{2}$. The predicted outcomes were then compared with the actual outcomes, and the predictive values were computed as well as the proportion of correct predictions. Here it was found that the predictive values had stabilized already at $m = 1000$.

Figure 14 shows how the predictive values depend on the sample size n in the first sample. It is seen that the predictive values starts to stabilize when n is larger than 400 and that this stabilization process goes faster for ($XP = 1$) than for ($XP = 0$). The final values were 0.74 for ($XP = 1$), 0.73 for ($XP = 0$) and 0.73 for $\mathbb{P}(correct)$. The similarity between the latter values is merely a coincidence.

6 Discussion

When predicting the future state of health based on estimated probabilities, the choice of good predictors is of major importance, like in all areas of prediction. If very little is known about which variables that will serve as good predictors, a first step may be to perform preliminary study where as many variables as possible are included as candidates. This was made in the ISSA-study mentioned in Section 1 and 5.1. Here, 5-10 variables were chosen as predictors among a total of more than 200 variables. In this paper we have considered the situation where a first sample is taken in order to estimate π and where the prediction ability is evaluated in a second sample from the same population. Then the questions arise of how to extract the predictors from a larger list of candidates, how many to use and how to identify the dependency structure between them, if necessary. The dependency structure can be created by hierarchical cluster methods (Anderberg (1973) and Jobson (1992)). Simulations show that the procedure works very well with dichotomous variables. Since a correct specification of independent

clusters has been showed to be of such great importance this issue should be further investigated.

Throughout the paper it has been assumed that the dependency structures between sets of predictors are correctly specified. This is a matter of crucial importance, since by assuming sets of predictors to be conditionally independent when they in fact are dependent may have serious effects on bias and variance of the estimator of π. An illustrative example is the following one with two predictors. Let the cell probabilities in Table 1 be $q_{12}^{(1)}(1,1) = 0.10$, $q_{12}^{(1)}(1,0) = 0.40 = q_{12}^{(1)}(0,1)$, $q_{12}^{(0)}(1,1) = 0.20$, $q_{12}^{(0)}(1,0) = 0.10 = q_{12}^{(0)}(0,1)$, so that the correlation between Z_1 and Z_2 is -0.60 given $X = 1$ and $+0.52$ given $X = 0$. From (2) it follows that the target probability to estimate when $(z_1, z_2) = (1,1)$ is $\pi = 0.33$, and according to (4) $\mathsf{Var}[\hat{\pi}] = 0.0148$ when $n = 100$. On the other hand, by assuming independency between Z_1 and Z_2 the target probability becomes $\pi = 0.74$, while the variance of the estimator is 0.0067 when $n = 100$. Thus, both bias and variance will in this case differ with about 120%. This was just a counter example, but in practice the effects of ignoring correlations between the predictors can be serious and give rise to large differences between the estimated $\pi's$ (see the discussion in Persson (2002)).

The results in Section 3 support the idea to include as many predictors as possible in the model, provided that the difference between the q-probabilities $q^{(1)}(\mathbf{z})$ and $q^{(0)}(\mathbf{z})$ is large. When the latter difference is small, it may result in a local increase in the variance of $\hat{\pi}$ (see Figure 14). This argues against using predictors in the model with only slight differences between the q-probabilities. For $p^{(1)} = \frac{1}{2}$ and when both $q^{(1)}(\mathbf{z})$ and $q^{(0)}(\mathbf{z})$ are small, the variance of $\hat{\pi}$ in (4) will be large, as shown in Figure 2. When there are two independent groups of predictors and $p^{(1)} = \frac{1}{2}$, Figure 5 suggests that the variance of $\hat{\pi}$ will be large if both $q_{12}^{(1)}(z_1, z_2)$ and $q_{34}^{(1)}(z_3, z_4)$ are small. These results should apply to the example in Section 5.1 where $p^{(1)}$ was close to $\frac{1}{2}$. Notice that many of the q-probabilities were small. For $p^{(1)} < \frac{1}{2}$ there is a different pattern. Now, Figures 6-12 suggests that the variance will be large when there is a large difference between the $q^{(1)}$-probabilities.

There are also questions about sample sizes needed to get reliable estimates of model parameters and of predictive values. The variance of $\hat{\pi}$ can be reduced by increasing the sample size, but due to the complicated dependencies on the parameters of the expression for the variance, it is not easy to give clear-cut recommendations for the choice of a proper sample size. The smallest sample size needed to reach an acceptable level of the variances of $\hat{\pi}$, for making reliable CI statements and also for getting reliable values of the predictive values was $n = 400$. The latter may be smaller when the q-probabilities are relatively large, but $n = 400$ may be recommended as a safe rule of thumb. Even with samples of 400 it is seen from Tables 2-4 that the lengths of the CI's can be somewhat large, and that sample sizes above 1000 would be needed in order to get CI's with reasonable lengths.

Although all results of the paper apply to predictors with an arbitrary number of outcomes, we have only been concerned with dichotomized predictors in the example of Section 5.1, and this needs an explanation. The reasons for only using binary predictors were that almost all of the variable values were subjectively assessed on an ordinal scale (exceptions were Age and Income), and that more or less pronounced threshold values could either be detected on probability plots (e.g. Working ability on a 10-point scale), or determined after consulting experts in the field (e.g. Complete rehabilitation plan on a 5-point scale). It was supposed that dichotomized predictors would behave more robustly than the original ordinal variables when predictions were made for new subjects. It may be argued that information is lost by the dichotomization. However, in the present study it was felt that this loss of information could be neglected. For instance, the variable 'Complete rehabilitation plan' got the maximal value 5 if the document was signed by the insured, but 4 if the same document was not signed. Here it seemed to be more relevant to know whether such a document existed or not. A further reason for dichotomizing is to reduce the possibility of getting zero cell frequencies. When there are enough many possible outcomes for a predictor it will be inevitable that this will occur. The problem with zero frequencies and missing values are further considered in Persson (2002).

Acknowledgments

The authors would like to thank Christian Sonesson for helpful comments and valuable suggestions on earlier versions of the manuscript.

References

[1] Afifi, A.A. and Azen, S.P. (1979) *Statistical Analysis - A Computer Oriented Approach* (2nd ed.). New York: Academic Press.

[2] Anderberg, M.R. (1973) *Cluster Analysis for Applications.* New York: Academic Press.

[3] Bergendorff, S., Hansson, E., Hansson, T., Palmer, E., Westin, M. and Zetterberg, C. (1997) (In Swedish) *Projektbeskrivning och undersökningsgrupp.* Rygg och Nacke 1. Stockholm: Riksförsäkringsverket och Sahlgrenska universitetssjukhuset.

[4] Bergendorff, S., Hansson, E., Hansson, T. and Jonsson, R. (2001) (In Swedish) *Vad kan förutsäga utfallet av en sjukskrivning?* Rygg och Nacke 8. Stockholm: Riksförsäkringsverket och Sahlgrenska universitetssjukhuset.

[5] Campbell, M.J. and Machin, D. (1990) *Medical Statistics.* New York: Wiley.

[6] Casella, G. and Berger, R.L. (1990) *Statistical Inference.* Belmont California: Duxbury Press.

[7] Cox, D.R. (1970) *Analysis of Binary Data.* London: Chapman and Hall.

[8] Jobson, J.D. (1992) *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods.* New York: Springer-Verlag.

[9] Jonsson, R. (2001) (In Swedish) *Faktorer som är väsentliga vid arbetslivsinriktad rehabilitering samt deras prognosvärde.* Seminar Paper 2001:4. Department of Statistics, Göteborg University.

[10] Kotz, S. and Johnson, N.L. (1985) In *Encyclopedia of Statistical Sciences,* Vol 8. New York: Wiley.

[11] Persson, A. (2002) *Prediction of Work Resumption Among Men and Women with Lower Back- and Neck Pain in a Swedish Population.* Research Report 2002:4. Department of Statistics, Göteborg University.

[12] Rao, C.R. (1973) *Linear Statistical Inference and Its Applications* (2nd ed.). New York: Wiley.

[13] Riksförsäkringsverket and Sahlgrenska Universitetssjukhuset (1997) (In Swedish) *Enkäter till undersökningsgruppen och försäkringskassan.* Rygg och Nacke 2. Stockholm: Riksförsäkringsverket och Sahlgrenska Universitetssjukhuset.

[14] SOU (2002) (In Swedish) *Handlingsplan för ökad hälsa i arbetslivet.* Statens Offentliga Utredningar, 2002:2. Stockholm: Fritzes.

Appendix

Some results for multinomial distributions.

Let $(X_1^{(1)}, ..., X_k^{(1)}, X_1^{(0)}, ..., X_k^{(0)})$ be a random vector with a multinomial distribution denoted by $M(n, p^{(1)}q_1^{(1)}, ..., p^{(1)}q_k^{(1)}, p^{(0)}q_1^{(0)}, ..., p^{(0)}q_k^{(0)})$, where $\sum_{i=1}^k q_i^{(1)} = 1 = \sum_{i=1}^k q_i^{(0)}$ and $p^{(1)} + p^{(0)} = 1$. A binomial distribution with parameters n and p is denoted by $B(n, p)$.

From the probability generating function (pgf) it is easily verified that

$$X_i^{(1)} + X_i^{(0)} \text{ is distributed } B(n, p^{(1)}q_i^{(1)} + p^{(0)}q_i^{(0)}),\ i = 1, ..., k. \tag{A1}$$

Direct calculation yields that

$$(X_i^{(1)} \mid X_i^{(1)} + X_i^{(0)} = x) \text{ is distributed } B\left(x, \frac{p^{(1)}q_i^{(1)}}{p^{(1)}q_i^{(1)} + p^{(0)}q_i^{(0)}}\right),\ i = 1, ..., k. \tag{A2}$$

Let $N(\mathbf{z}_r)$, $r = 1, ..., g$, be independent vectors each being distributed $M(n, q(\mathbf{z}_r))$. For fixed $\mathbf{z}_r$, $r = 1, ..., g$, one may put $N_r = N(\mathbf{z}_r)$ and $q_r = q(\mathbf{z}_r)$. Then

$$\mathsf{Var}\left(\prod_{r=1}^g N_r\right) = n^{2g}\left(\prod_{r=1}^g q_r\right)^2 \left\{\prod_{r=1}^g \left(1 + \frac{1 - q_r}{nq_r}\right) - 1\right\} \tag{A3}$$

(A3) follows easily by repeated use of the expressions,

$$\begin{aligned}
\mathsf{Var}(N_1) &= nq_1(1 - q_1), \\
\mathsf{Var}(N_1N_2) &= \mathsf{Var}(N_1)\mathsf{Var}(N_2) + \mathsf{Var}(N_1)[\mathsf{E}(N_2)]^2 + [\mathsf{E}(N_1)]^2\mathsf{Var}(N_2) \\
&= n^4(q_1q_2)^2\left\{\left(1 + \frac{1 - q_1}{nq_1}\right)\left(1 + \frac{1 - q_2}{nq_2}\right) - 1\right\} \text{ and so on.}
\end{aligned}$$

Approximation of functions of moments

Let X_i, $i = 1, 2$ be two independent random variables with means μ_i and variances σ_i^2. Then it follows from a Taylor expansion that the function $g(X_1, X_2)$ has the approximate moments (Kotz and Jonsson (1985), p. 646)

$$\mathsf{E}[g(X_1, X_2)] \approx g(\mu_1, \mu_2) + \frac{1}{2}\left(\left[\frac{\partial^2 g}{\partial x_1^2}|\mu\right] \cdot \sigma_1^2 + \left[\frac{\partial^2 g}{\partial x_2^2}|\mu\right] \cdot \sigma_2^2 + 2\left[\frac{\partial^2 g}{\partial x_1 \partial x_2}|\mu\right] \cdot \sigma_{12}\right)$$

$$\mathsf{Var}[g(X_1, X_2)] \approx \left[\frac{\partial g}{\partial x_1}|\mu\right]^2 \cdot \sigma_1^2 + \left[\frac{\partial g}{\partial x_2}|\mu\right]^2 \cdot \sigma_2^2 + 2\left[\frac{\partial g}{\partial x_1}|\mu\right]\left[\frac{\partial g}{\partial x_2}|\mu\right] \cdot \sigma_{12} \tag{A4}$$

where all derivatives are evaluated at $\mu = (\mu_1, \mu_2)$. Also,

$$\mathsf{E}[g(X_i)] \approx g(\mu_i) + \frac{1}{2}\left[\frac{\partial^2 g}{\partial x_i^2}|\mu_i\right] \cdot \sigma_i^2 \text{ and } \mathsf{Var}[g(X_i)] \approx \left[\frac{\partial g}{\partial x_i}|\mu_i\right]^2 \cdot \sigma_i^2.$$

Legends to Figures

Figure 1: Calculation of the differences between the probability π with $Z_1 = 1$ and $Z_2 = 1$ in the independent and dependent case.

Figure 2: $\mathsf{Var}[\hat{\pi}]$ from (4) in the case with two dependent predictors (Z_1, Z_2), given that $n = 400$ and $p^{(1)} = \frac{1}{2}$.

Figure 3: $\mathsf{Var}[\hat{\pi}]$ from (4) in the case with two dependent predictors (Z_1, Z_2), given that $n = 400$ and $p^{(1)} = .10$.

Figure 4: $\mathsf{Var}[\hat{\pi}]$ from (4) in the case with two dependent predictors (Z_1, Z_2), given that $n = 400$ and $p^{(1)} = .90$.

Figure 5: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = \frac{1}{2}$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .05$.

Figure 6: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = .05$ and $q_{34}^{(0)}(\cdot) = .10$.

Figure 7: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = .05$ and $q_{34}^{(0)}(\cdot) = .20$.

Figure 8: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = .05$ and $q_{34}^{(0)}(\cdot) = .30$.

Figure 9: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .05$.

Figure 10: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .10$.

Figure 11: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .20$.

Figure 12: $\mathsf{Var}[\hat{\pi}]$ from (8) in the case with two independent groups of dependent predictors (Z_1, Z_2) and (Z_3, Z_4), given that $n = 400$, $p^{(1)} = .10$, $q_{12}^{(0)}(\cdot) = q_{34}^{(0)}(\cdot) = .30$.

Figure 13: Predictive values for healthy (solid line) and non-healthy (dotted line) for various sample sizes.

Figure 14: $\mathsf{Var}[\hat{\pi}]$ as a function of number of independent sets of predictors.

Figure 1

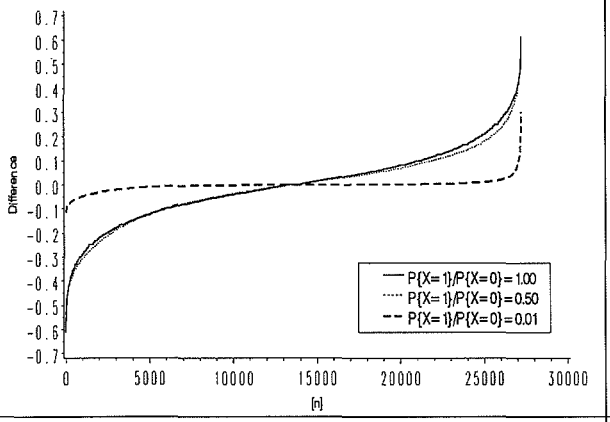


Figure 2

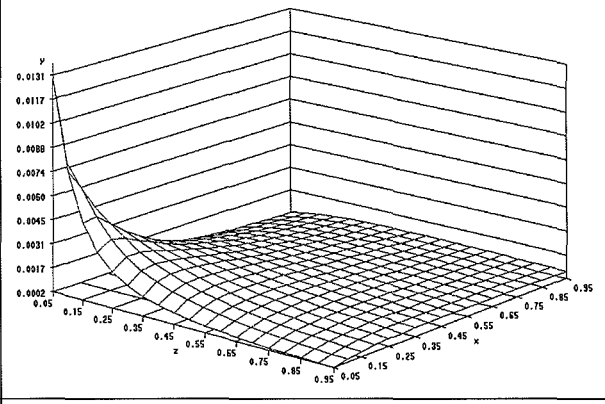


Figure 3

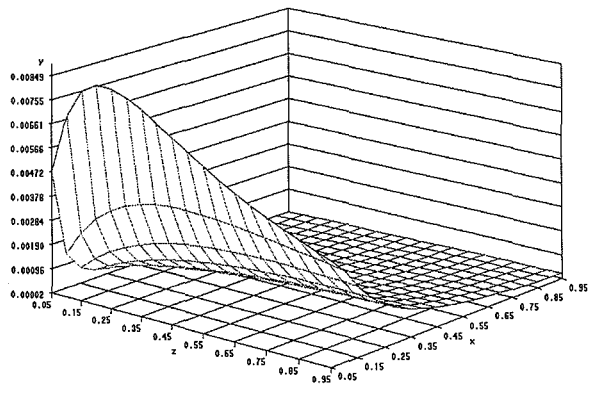


Figure 4

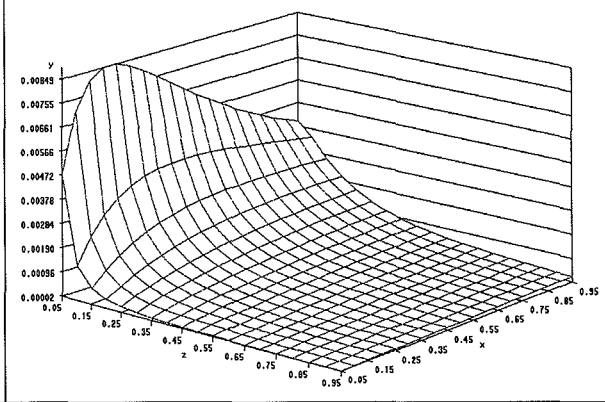


Figure 5

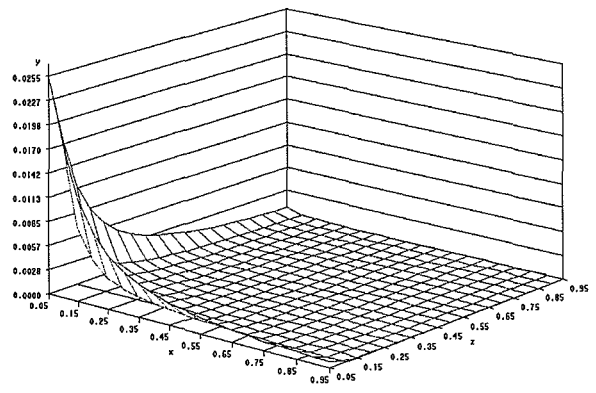


Figure 6

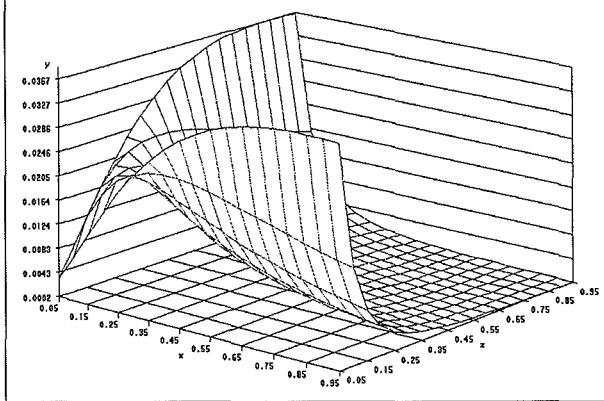


Figure 7

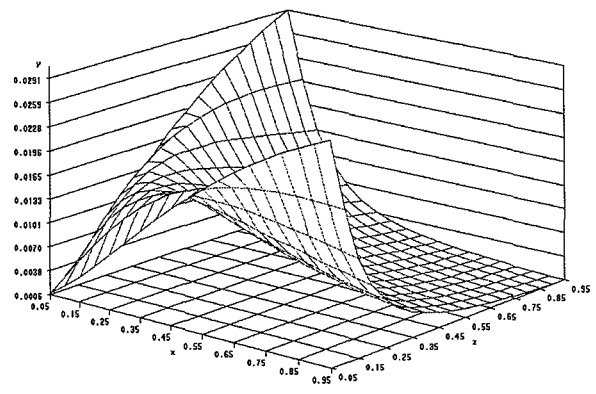


Figure 8

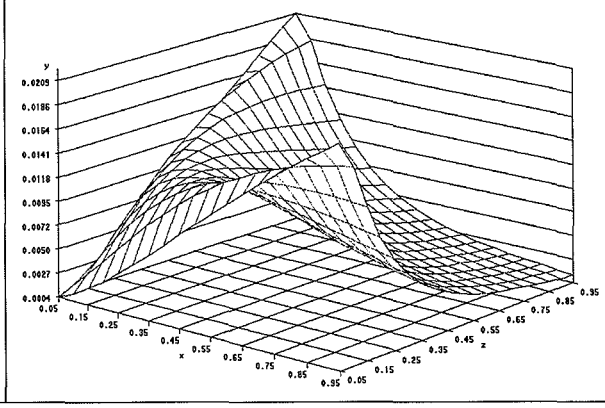


Figure 9

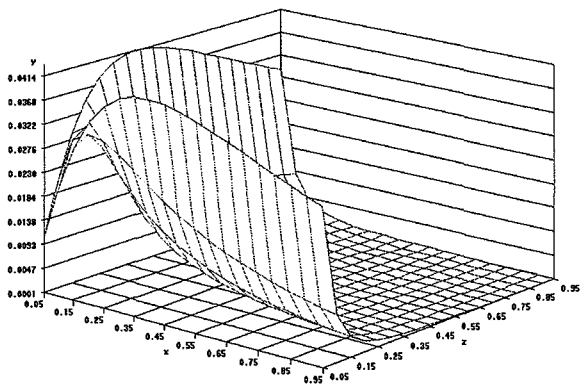


Figure 10

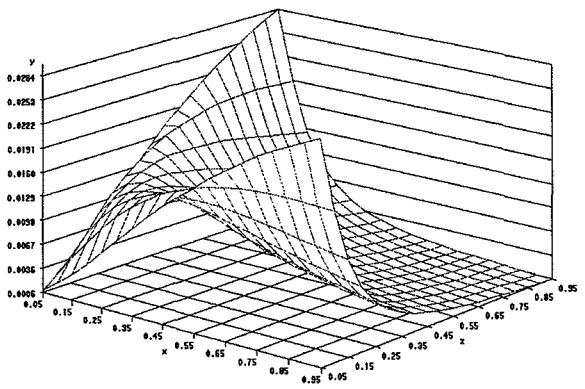


Figure 11

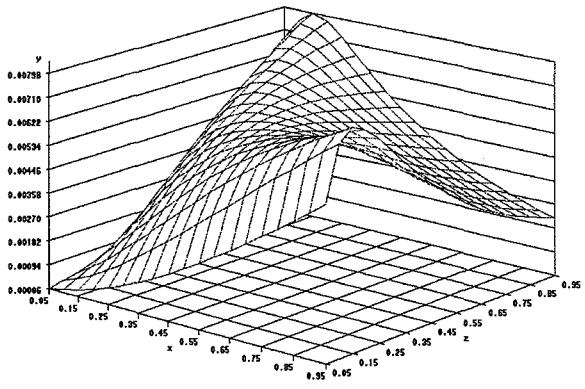


Figure 12

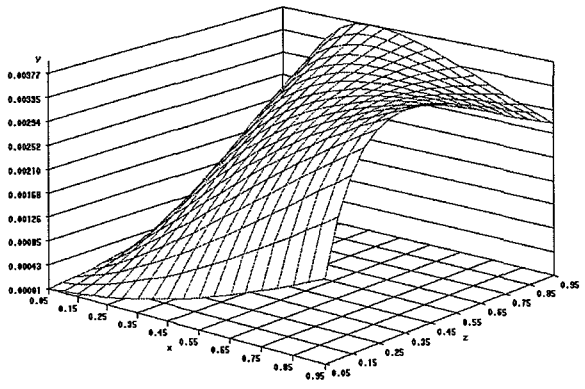


Figure 13

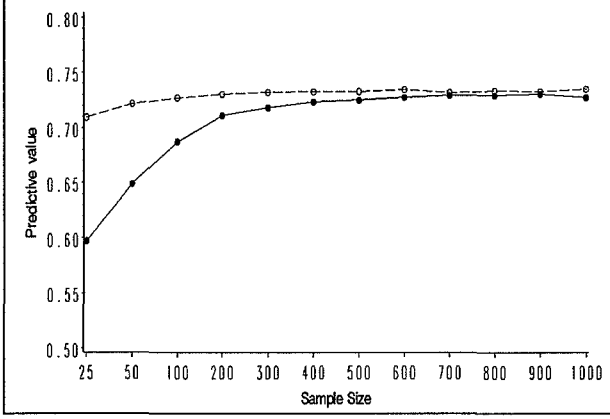
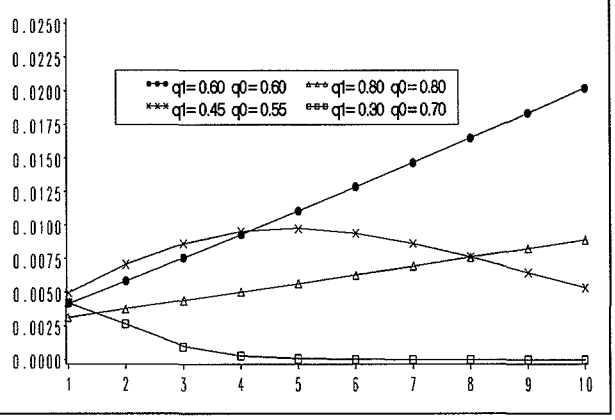


Figure 14



		Expected Length (z_1, z_2)				Coverage Probability (%) (z_1, z_2)			
CI	n	(1,1)	(1,0)	(0,1)	(0,0)	(1,1)	(1,0)	(0,1)	(0,0)
(11: *i*)	50	.29	.36	.36	.50	95	15	15	28
	100	.20	.60	.60	.73	95	40	40	59
	200	.14	.79	.78	.76	95	72	72	84
	400	.10	.70	.70	.58	95	89	89	92
	800	.07	.50	.50	.41	95	93	93	93
(11: *ii*)	50	.27	.84	.84	.82	95	64	64	78
	100	.20	.79	.79	.75	95	87	87	94
	200	.14	.71	.71	.63	95	97	97	97
	400	.10	.58	.58	.50	95	96	96	96
	800	.07	.45	.45	.38	95	96	96	96
(11: *iii*)	50	.29	1.69	1.69	1.55	95	63	63	77
	100	.20	1.44	1.43	1.23	95	85	85	92
	200	.14	1.08	1.08	.86	95	94	94	95
	400	.10	.73	.73	.59	95	96	95	96
	800	.07	.50	.50	.41	95	95	95	95
(11: *iv*)	50	.30	.94	.94	.92	97	39	39	53
	100	.21	.89	.89	.85	96	64	64	78
	200	.15	.81	.81	.73	96	87	86	94
	400	.10	.67	.67	.57	96	97	97	97
	800	.07	.50	.50	.42	95	97	97	97
(11: *v*)	50	.28	.84	1.06	.82	95	15	15	28
	100	.20	.80	.98	.76	95	40	40	60
	200	.14	.72	.84	.65	95	75	75	90
	400	.10	.59	.66	.51	95	95	95	97
	800	.07	.45	.48	.38	95	96	96	96

Table 2: Expected lengths and actual coverage probabilities (%) of the various CI's in (11): (*i*)-(*v*) for π, based on two dependent binary predictors. The q probabilities were $q^{(x)}(1,1) = .93$, $q^{(x)}(1,0) = .02$, $q^{(x)}(0,1) = .02$ and $q^{(x)}(0,0) = .03$, $x = 0,1$. The stipulated CI-level was 95%, and each figure was computed from 100,000 simulations.

	Expected Length: z_1, z_2, z_3, z_4															
Sample size, n	1,1,1,1	1,1,1,0	1,1,0,1	1,1,0,0	1,0,1,1	1,0,1,0	1,0,0,1	1,0,0,0	0,1,1,1	0,1,1,0	0,1,0,1	0,1,0,0	0,0,1,1	0,0,1,0	0,0,0,1	0,0,0,0
50	.42	.41	.68	.65	.48	.45	.69	.66	.66	.64	.76	.75	.61	.59	.74	.72
100	.29	.30	.64	.57	.37	.34	.65	.58	.59	.58	.72	.69	.54	.51	.70	.66
200	.20	.21	.58	.46	.28	.25	.57	.47	.51	.48	.65	.58	.44	.41	.62	.54
400	.13	.15	.48	.35	.21	.18	.46	.35	.39	.36	.51	.43	.30	.27	.47	.37
800	.10	.11	.37	.25	.15	.13	.33	.26	.27	.23	.33	.27	.17	.13	.26	.19
1600	.07	.08	.28	.18	.11	.09	.23	.18	.18	.14	.18	.17	.10	.07	.10	.09
Coverage Probability (%): z_1, z_2, z_3, z_4																
Sample size, n	1,1,1,1	1,1,1,0	1,1,0,1	1,1,0,0	1,0,1,1	1,0,1,0	1,0,0,1	1,0,0,0	0,1,1,1	0,1,1,0	0,1,0,1	0,1,0,0	0,0,1,1	0,0,1,0	0,0,0,1	0,0,0,0
50	94	96	25	59	99	97	25	59	31	29	07	18	28	28	06	17
100	95	96	55	89	97	97	53	88	56	54	28	48	53	53	28	47
200	95	96	85	97	96	96	81	96	80	79	65	78	79	78	64	77
400	95	95	98	97	95	95	93	96	92	91	88	91	91	91	87	91
800	95	95	98	96	95	95	95	96	95	94	93	94	94	94	92	93
1600	95	95	96	95	95	95	95	95	95	95	94	95	95	95	94	94

Table 3: Expected length and actual coverage probabilities (%) of the various CI's in (12): (*i*) for π, based on two independent groups of dependent binary predictors (Z_1,Z_2) and (Z_3,Z_4). The q probabilities were $q_{12}^{(1)}(1,1)=.24$, $q_{12}^{(1)}(1,0)=.38$, $q_{12}^{(1)}(0,1)=.11$, $q_{12}^{(1)}(0,0)=.27$, $q_{12}^{(0)}(1,1)=.71$, $q_{12}^{(0)}(1,0)=.25$, $q_{12}^{(0)}(0,1)=.02$, $q_{12}^{(0)}(0,0)=.02$, $q_{34}^{(1)}(1,1)=.34$, $q_{34}^{(1)}(1,0)=.55$, $q_{34}^{(1)}(0,1)=.04$, $q_{34}^{(1)}(0,0)=.07$, $q_{34}^{(0)}(1,1)=.45$, $q_{12}^{(0)}(1,0)=.48$, $q_{34}^{(0)}(0,1)=.02$, $q_{34}^{(0)}(0,0)=.05$ and $p^{(1)}=.50$. The stipulated CI-level was 95%, and each figure was computed from 100,000 simulations.

	Expected Length: z_1, z_2, z_3, z_4															
Sample size, n	1,1,1,1	1,1,1,0	1,1,0,1	1,1,0,0	1,0,1,1	1,0,1,0	1,0,0,1	1,0,0,0	0,1,1,1	0,1,1,0	0,1,0,1	0,1,0,0	0,0,1,1	0,0,1,0	0,0,0,1	0,0,0,0
50	.38	.41	.80	.74	.54	.48	.83	.78	.79	.74	.92	.89	.64	.55	.82	.78
100	.27	.30	.75	.63	.40	.36	.75	.65	.66	.59	.80	.74	.46	.36	.63	.54
200	.19	.21	.66	.49	.29	.26	.61	.49	.49	.41	.57	.50	.29	.22	.37	.30
400	.13	.15	.53	.36	.21	.19	.45	.36	.34	.27	.33	.31	.19	.13	.18	.16
800	.09	.11	.40	.26	.15	.13	.32	.26	.24	.19	.20	.21	.12	.09	.10	.10
1600	.07	.08	.29	.18	.11	.09	.23	.19	.17	.13	.13	.14	.09	.06	.6	.07
Coverage Probability (%): z_1, z_2, z_3, z_4																
Sample size, n	1,1,1,1	1,1,1,0	1,1,0,1	1,1,0,0	1,0,1,1	1,0,1,0	1,0,0,1	1,0,0,0	0,1,1,1	0,1,1,0	0,1,0,1	0,1,0,0	0,0,1,1	0,0,1,0	0,0,0,1	0,0,0,0
50	96	96	25	59	96	96	25	59	36	36	9	22	36	36	10	23
100	96	96	55	89	95	96	55	89	60	60	34	55	60	60	34	56
200	95	95	84	97	95	95	84	97	84	84	72	84	84	84	73	84
400	95	95	96	96	95	95	96	96	95	95	94	96	95	95	94	96
800	95	95	96	95	95	95	96	95	96	96	96	96	96	96	96	96
1600	95	95	95	95	95	95	95	95	95	95	95	95	96	95	95	95

Table 4: Expected length and actual coverage probabilities (%) of the various CI's in (12): (*ii*) for π, based on two independent groups of dependent binary predictors (Z_1,Z_2) and (Z_3,Z_4). The same q-probabilities as in Table 3 were used. The stipulated CI-level was 95%, and each figure was computed from 100,000 simulations.

Research Report

2001:1	Holgersson, H.E.T.:	On assessing multivariate normality.
2001:2	Sonesson, C. & Bock, D.:	Statistical issues in public health monitoring – A review and discussion.
2001:3	Andersson, E.:	Turning point detection using non-parametric statistical surveillance. Evaluation of some influential factors.
2001:4	Andersson, E. & Bock, D.:	On seasonal filters and monotonicity.
2001:5	Andersson, E., Bock, D. & Frisén, M.:	Likelihood based methods for detection of turning points in business cycles. A comparative study.
2001:6	Sonesson, C.:	Evaluations of some exponentially weighted moving average methods.
2001:7	Sonesson, C.:	Statistical surveillance. Exponentially weighted moving average methods and public health monitoring.
2002:1	Frisén, M. & Sonesson, C.:	Optimal surveillance based on exponentially weighted moving averages.
2002:2	Frisén, M.:	Statistical surveillance. Optimality and methods.