On the Choice-Based Sample Bias in Probabilistic Business Failure Prediction

Kenth Skogsvik*)**)

KPMG Professor in Financial Accounting and Corporate Finance

(December 2005)


**Key words:** Business Failure Prediction, Choice-Based Sample Bias, Financial Analysis, Probabilistic Prediction Model, Probit/Logit Analysis

**JEL code:** C11, G33, M41

*) The comments of Peter Jennergren, Beate Juettner-Nauroth, Erik Ruist, Stina Skogsvik and seminar participants at the Center for Financial Analysis and Managerial Economics in Accounting at Stockholm School of Economics, are gratefully appreciated. The author is also grateful for financial support from the Torsten and Ragnar Söderberg Foundation.

**) Address for correspondence: Stockholm School of Economics, Box 6501, S-113 83 Stockholm, Sweden. Phone: +46-8-736 9307. E-mail: kenth.skogsvik@hhs.se.
On the Choice-Based Sample Bias in Probabilistic Business Failure Prediction

Abstract

Probabilistic failure prediction models are commonly estimated from non-random samples of business companies. The proportion of failure companies in such samples is often much larger than the proportion of failure companies in most real-world decision contexts. This so-called “choice-based sample bias” implies that calculated failure probabilities will be (more or less) biased. The purpose of the paper is to analyse this bias and its consequences for standard applications of probabilistic failure prediction models (for example probit/logit analysis) and in particular to investigate whether the bias can be eliminated without having to re-estimate the underlying statistical model. It is shown that there is a straightforward linkage between sample-based probabilities of failure and the corresponding population-based probabilities. Knowing this linkage, sample-based probabilities can be adjusted for the “choice-based sample bias”, provided that sufficiently large samples of randomly selected failure companies and randomly selected survival companies have been used in the estimation of the underlying model. Empirical observations in previous research are in line with the theoretical results of the paper.
1. **Introduction**

Over the years a considerable number of empirical investigations of the association between financial statement numbers and the event of business failure have been carried out. Various types of statistical methods have been used in order to explore this association, ranging from fairly ad hoc applications of regression analysis to more sophisticated variants of discriminant or probit/logit analysis. Previous research has shown that the statistical assumptions associated with regression analysis and discriminant analysis typically are not well fulfilled in the context of business failure prediction. Also, these methods are somewhat awkward in the sense that they do not directly provide for the assessment of failure probabilities. In this regard probit/logit analysis is superior, as this type of method presupposes an underlying probabilistic association between some set of independent variables (e.g. accounting numbers) and the outcome variable (“failure” versus “survival”).

Failure probabilities are important in various decision contexts. The relevance of probabilistic failure predictions in discounted cash flow modelling (bond valuation as well as the valuation of owners’ equity) is corroborated in for example Skogsvik (2005). Most decision models involving the calculation of “expected values” presume that unbiased – population-based – probabilities exist. In practice such probabilities are rarely exogenously given, but have to be estimated. In the context of business failure prediction, probit/logit analysis would then appear to be quite suitable. However, the issue at hand requires careful attention to the distorting impact of non-random sampling in the estimation of such models.

Failure prediction models have commonly been estimated from non-random samples of failure and survival companies in previous research. Often a “matched-pairs” design has been

---


2 Inconsistent with the underlying assumptions in regression analysis, the dependent variable is dichotomous in a prediction of business failure. Multivariate discriminant analysis hinges on the idea that the independent variables are multivariate normally distributed in the subpopulations of failure and survival companies, and that the variance /covariance matrices for the independent variables are the same for both populations. As observed in for example Foster (1986), pp.107-111, and Skogsvik (1987), pp. 210-214, financial statement numbers rarely fulfill these assumptions.
used, resulting in a sample proportion of failure companies of 0.50.\textsuperscript{3} This type of sampling implies typically that the proportion of failure companies in the sample is much larger than the proportion of such companies in the grand population of (failure and survival) companies. This causes a “choice-based sample bias” of the constant and the coefficients in estimated standard probit/logit models, in turn meaning that the probabilities being assessed in such models are more or less biased.

As noted in previous literature, there are statistical techniques that will generate unbiased parameter estimates in probabilistic prediction models even if non-random samples of failure and survival companies are used.\textsuperscript{4} These techniques imply that the estimation procedure is calibrated for some a priori probability of business failure. As long as an estimated model is used in contexts where the same a priori probability of failure applies, this is obviously unproblematic. However, if the a priori probability changes (due to, for example, changes in population characteristics over time), the estimated model will no longer be unbiased. A cumbersome way of dealing with such changes would be to re-estimate the model with a new assessment of the a priori probability. As a user of some previously estimated model this is not likely to be a viable alternative – not having access to the original empirical data would obviously be an effective impediment. A better course of action would rather be to make corrective adjustments to the model-based probabilities.\textsuperscript{5}

The purpose of the paper is to analyze the relationship between sample-biased probabilities and the corresponding unbiased – population-based – probabilities. Knowing this relationship would enable users of probabilistic prediction models to estimate unbiased probabilities, i.e. probabilities that are corrected for the choice-based sample bias.

The disposition of the paper is as follows. The choice-based sample bias of failure probabilities is derived in section 2. In section 3, the transformation of a biased (sample-based) probability into an unbiased (population-based) probability is derived. Consequences for the use of probabilistic prediction models are discussed in section 4, and some

\textsuperscript{3} In a survey of failure prediction studies, Zmijewski (1984) found that a ”matched-pairs” design had been used in 12 out of 17 previous studies.

\textsuperscript{4} According to Zmijewski (1984), p. 65, the choice-based sample bias can be handled in “weighted exogenous sample maximum likelihood”, “conditional maximum likelihood”, or “full information concentrated maximum likelihood”.

\textsuperscript{5} Some of the “methodological problems” being addressed in Zmijewski (1984) and Bergström et al (1999) appear to be non-issues if the bias of model-based probabilities can be corrected without having to re-estimate the probabilistic model (cf. section 4 of this paper).
implications in relation to previous empirical research are included in section 5. A short summary and concluding remarks follow in the last section.

2. The choice-based sample probability bias

As noted in the introduction, probabilistic failure prediction models have commonly been estimated from non-random samples in previous research. If unadjusted statistical methods have been used in this context, it can be feared that estimated coefficients have been (more or less) arbitrarily affected by the chosen sample compositions. Obviously this is quite disturbing if model-based probabilities are to be used in decision contexts where the expected proportion of failure companies strongly differs from the corresponding proportion in the estimation sample.

In order to analyze the effect of this choice-based sample bias, the following notation is introduced:

\[
\begin{align*}
\pi(t) & = \text{proportion of failure companies in year } t \text{ in the grand population} \\
& = \text{a priori probability of business failure in year } t; \text{ where} \\
& \quad 0 < \pi(t) < 1.00. \\
\end{align*}
\]

\[
\begin{align*}
p_{j,\text{fail}}^{(\pi)} & = \text{population-based probability of business failure in year } t \text{ conditioned on} \\
& \quad \text{survival at the end of year } t-1 \text{ for company } j. \\
\end{align*}
\]

\[
\begin{align*}
\text{prop} & = \text{proportion of failure companies in the estimation sample, where} \\
& \quad 0 < \text{prop} < 1.00. \\
\end{align*}
\]

\[
\begin{align*}
p_{j,\text{fail}}^{(\text{prop})} & = \text{estimated (unadjusted) probability of business failure in year } t \text{ conditioned} \\
& \quad \text{on survival at the end of year } t-1 \text{ for company } j, \text{ based on an estimation} \\
& \quad \text{sample with a proportion of failure companies equal to prop}. \\
\end{align*}
\]

\[
\begin{align*}
\{X_{j,t-1}\} & = \text{set of descriptors for company } j, \text{ observable at time } t-1. \\
\end{align*}
\]

In some decision context the idea is that the decision maker is armed with some (previously estimated) probabilistic failure prediction model, and:
• Numerical values of the descriptors \( \{X_{j,t-1}\} \) (commonly including accounting numbers as indicators of profitability, interest cost, operating and financial leverage, etc.)\(^6\) are obtained for the company.

• Based on the available prediction model and the numerical values of the descriptor variables, a probability of business failure \( \left( \hat{\pi}_{j,\text{fail}(t)}^{(\text{prop})} \right) \) is calculated.

If the prediction model has been estimated with a sample proportion of failure companies \( \text{prop} \) equal to \( \pi(t) \), the calculated probability can directly be used as an unbiased probability assessment. However, if \( \text{prop} \neq \pi(t) \) the calculated probability will no longer be an unbiased estimate of the corresponding population-based probability \( \pi_{j,\text{fail}(t)}^{(\pi)} \).

In order to simplify the notation somewhat, let henceforth the indices \( j \) and \( t \) be suppressed. Recognizing that both sample-based and population-based probabilities are conditioned on some set of descriptor variables \( \{X\} \), we can then write

\[ \hat{\pi}_{\text{fail}}^{(\text{prop})} = \hat{\pi}(\text{fail}|\{X\})^{(\text{prop})} \text{ and } \pi_{\text{fail}}^{(\pi)} = p(\text{fail}|\{X\})^{(\pi)}. \]

Given that \( 0 < \pi < 1.00 \) and assuming that \( p_{\text{fail}}^{(\pi)} > 0 \), the unbiased probability \( p_{\text{fail}}^{(\pi)} \) can be unfolded in accordance with Bayes’ theorem:\(^8\)

\[
(1) \quad p_{\text{fail}}^{(\pi)} = p(\text{fail}|\{X\})^{(\pi)} = \\
\quad = \frac{\pi \cdot p(\{X\}|\text{fail})^{(\pi)}}{\pi \cdot p(\{X\}|\text{fail})^{(\pi)} + (1 - \pi) \cdot p(\{X\}|\text{surv})^{(\pi)}} = \\
\quad = \left[ 1 + \frac{(1 - \pi)}{\pi} \cdot \frac{p(\{X\}|\text{surv})^{(\pi)}}{p(\{X\}|\text{fail})^{(\pi)}} \right]^{-1}
\]

where: \( p(\{X\}|\text{fail})^{(\pi)} \) = population-based probability of observing \( \{X\} \) (at time \( t \)) conditioned on business failure in the following year \( (t+1) \).

\( p(\{X\}|\text{surv})^{(\pi)} \) = population-based probability of observing \( \{X\} \) (at time \( t \)) conditioned on business survival in the following year \( (t+1) \).

---


\(^8\) Cf., for example, Chou (1984), p. 411.
The sample-based probability of failure \( \hat{p}_{\text{fail}}^{(\text{prop})} \) is implicitly affected by the proportion of failure companies in the estimation sample, with \( prop \) as the sample-based “a priori” probability of failure. Given that \( 0 < prop < 1.00 \) and assuming that \( \hat{p}_{\text{fail}}^{(\text{prop})} > 0 \), this sample-based probability can be unfolded in the same manner as above:

\[
(2) \quad \hat{p}_{\text{fail}}^{(\text{prop})} = \hat{p}(\text{fail} | \{X\})^{(\text{prop})} = \left[ 1 + \frac{(1 - \text{prop})}{\text{prop}} \cdot \frac{\hat{p}(\{X\} | \text{surv})^{(\text{prop})}}{\hat{p}(\{X\} | \text{fail})^{(\text{prop})}} \right]^{-1}
\]

where: \( \hat{p}(\{X\} | \text{fail})^{(\text{prop})} = \) probability of observing \( \{X\} \) (at time \( t \)) conditioned on business failure the following year \( (t + 1) \) in the estimation sample.

\( \hat{p}(\{X\} | \text{surv})^{(\text{prop})} = \) probability of observing \( \{X\} \) (at time \( t \)) conditioned on business survival the following year \( (t + 1) \) in the estimation sample.

Assume now that the sample of failure companies constitutes a random drawing from the sub-population of failure companies, and the sample of survival companies constitutes a random drawing from the sub-population of survival companies, in the sense that

\( \hat{p}(\{X\} | \text{fail})^{(\pi)} = p(\{X\} | \text{fail})^{(\pi)} \) and \( \hat{p}(\{X\} | \text{surv})^{(\pi)} = p(\{X\} | \text{surv})^{(\pi)} \). Given that \( p(\{X\} | \text{fail})^{(\pi)} > 0 \) and \( \hat{p}(\{X\} | \text{fail})^{(\text{prop})} > 0 \), this implies that

\( p(\{X\} | \text{surv})^{(\pi)} / p(\{X\} | \text{fail})^{(\pi)} = \hat{p}(\{X\} | \text{surv})^{(\text{prop})} / \hat{p}(\{X\} | \text{fail})^{(\text{prop})} \). Let this “odds ratio” henceforth be denoted \( \text{Or} \{X_0\} \).

The value of the “odds ratio” for the grand population of companies can simply be solved from (1):

\[
(3) \quad \text{Or} \{X\} = \left[ \frac{1}{p(\text{fail} | \{X\})^{(\pi)} - 1} \right] \cdot \frac{\pi}{(1 - \pi)}
\]
Inserting the above solution for \( \text{Or}(X_0) \) in (2) and recognizing that \( p(\text{fail} \mid X)^{(\pi)} = p^{(\pi)}_{\text{fail}} \)
and \( \hat{p}(\text{fail} \mid X)^{(\text{prop})} = \hat{p}^{(\text{prop})}_{\text{fail}} \), we get:

\[
\hat{p}^{(\text{prop})}_{\text{fail}} = \left[ 1 + \left( \frac{\pi}{1 - \pi} \right) \cdot \left( 1 - \hat{p}^{(\text{prop})}_{\text{fail}} \right) \cdot \left( \frac{1 - p^{(\pi)}_{\text{fail}}}{p^{(\pi)}_{\text{fail}}} \right) \right]^{-1}.
\]

The sample-based probability of failure is hence a function of the unbiased probability \( p^{(\pi)}_{\text{fail}} \), the proportion of failure companies in the population \((\pi)\), and the proportion of failure companies in the sample \((\text{prop})\).

In order to better understand the association between the sample-based probability of failure and the proportion of failure companies in the estimation sample, the derivative of (4) with respect to \( \text{prop} \) can be calculated. Having previously assumed that \( 0 < \pi < 1.00 \), \( 0 < \text{prop} < 1.00 \), and \( 0 < p^{(\pi)}_{\text{fail}} < 1.00 \), we then get:

\[
\frac{\partial (\hat{p}^{(\text{prop})}_{\text{fail}})}{\partial \text{prop}} = - \left[ 1 + \left( \frac{\pi}{1 - \pi} \right) \cdot \left( 1 - \hat{p}^{(\text{prop})}_{\text{fail}} \right) \cdot \left( \frac{1 - p^{(\pi)}_{\text{fail}}}{p^{(\pi)}_{\text{fail}}} \right) \right]^{-2} \cdot \\
\left[ (1 - \pi) \cdot \hat{p}^{(\text{prop})}_{\text{fail}} \right] \cdot \left[ (1 - \pi) \cdot \text{prop} \cdot p^{(\pi)}_{\text{fail}} \right]^2 \cdot \\
\left[ (1 - \pi) \cdot \text{prop} \cdot p^{(\pi)}_{\text{fail}} \right]^2 \\
= \frac{\pi \cdot (1 - \text{prop}) \cdot (1 - p^{(\pi)}_{\text{fail}}) \cdot (1 - \pi) \cdot p^{(\pi)}_{\text{fail}}}{\pi \cdot (p^{(\pi)}_{\text{fail}} + \text{prop} - 1 - \text{prop} \cdot p^{(\pi)}_{\text{fail}})^2} > 0.
\]

The right-hand side (RHS) of (5) is positive, meaning that \( \hat{p}^{(\text{prop})}_{\text{fail}} \) is positively associated with the sample proportion of failure companies. Or, alternatively, the choice-based sample bias of \( \hat{p}^{(\text{prop})}_{\text{fail}} \) is – ceteris paribus – positively associated with the proportion of failure companies in the estimation sample.

---

\(^9\) The linkage between sample-based and population-based probabilities as formulated in (4), has been elaborated in Palepu (1986) in a methodological discussion of estimated probabilistic models for the prediction of takeover targets.
It is also interesting to know whether the sample-based probability $\hat{p}_{\text{fail}}^{(\text{prop})}$ – even though it can be more or less biased in relation to $p_{\text{fail}}^{(\pi)}$ – nevertheless would provide for a correct ranking with regard to the risk of business failure. That is, reintroducing the company index, if $p_{\text{fail,j}}^{(\pi)} > p_{\text{fail,k}}^{(\pi)}$ will then $\hat{p}_{\text{fail,j}}^{(\text{prop})} > \hat{p}_{\text{fail,k}}^{(\text{prop})}$, where $j$ and $k$ denote different companies? The question can be answered as follows:

\[
\frac{\partial (\hat{p}_{\text{fail}}^{(\text{prop})})}{\partial (p_{\text{fail}}^{(\pi)})} = - \left[ 1 + \left( \frac{\pi}{1-\pi} \right) \left( \frac{1 - \text{prop}}{\text{prop}} \right) \left( \frac{1 - p_{\text{fail}}^{(\pi)}}{p_{\text{fail}}^{(\pi)}} \right) \right]^{-2} \times \\
\left[ \frac{(-1) \cdot \pi \cdot (1 - \text{prop}) \cdot (1 - \pi) \cdot \text{prop} \cdot p_{\text{fail}}^{(\pi)} \cdot (1 - \pi) \cdot \text{prop} \cdot p_{\text{fail}}^{(\pi)}}{(1 - \pi) \cdot \text{prop} \cdot p_{\text{fail}}^{(\pi)}} \right] = \\
\frac{\pi \cdot (1 - \pi) \cdot \text{prop} \cdot (1 - \text{prop})}{\pi \cdot (p_{\text{fail}}^{(\pi)} + \text{prop} - 1) - \text{prop} \cdot p_{\text{fail}}^{(\pi)}} > 0
\]

There is hence a positive relationship between $\hat{p}_{\text{fail}}^{(\text{prop})}$ and $p_{\text{fail}}^{(\pi)}$, in turn meaning that the sample-based probability is a consistent indicator of risk in the following sense:

**Proposition:** If $p_{\text{fail,j}}^{(\pi)} > (\leq) p_{\text{fail,k}}^{(\pi)}$, where $j$ and $k$ denote different companies, the sample-based probability of failure $\hat{p}_{\text{fail,j}}^{(\text{prop})}$ is expected to be larger than (smaller than) the population-based probability $\hat{p}_{\text{fail,k}}^{(\text{prop})}$.

In a decision context where a probabilistic prediction model has been estimated with a sample proportion of failure companies $\text{prop} > \pi$, the probability $\hat{p}_{\text{fail}}^{(\text{prop})}$ is thus expected to be positively biased, but nevertheless the ranking of companies is expected to be the same for the sample-based and the population-based measures of probability.
3. Estimating an unbiased failure probability

According to (4) above, the sample-based probability \( \hat{p}_{\text{prop}}^{(\text{fail})} \) can be calculated as a function of the corresponding population-based probability, the proportion of failure companies in the grand population, and the proportion of failure companies in the sample. In a decision context however, one might primarily be interested in an assessment of the unbiased failure probability \( p_{\text{π}}^{(\text{fail})} \). It is easy to imagine a decision maker that has estimated a probit/logit model from some choice-based sample of companies, and who then wants to transform calculated sample-based failure probabilities into unbiased probability assessments.

A straight-forward rewriting of (4) results in the following expression for an estimate of the unbiased probability \( p_{\text{π}}^{(\text{fail})} \):

\[
\hat{p}_{\text{π}}^{(\text{prop})} = \left[ 1 + \left( \frac{1 - \pi}{\pi} \right) \left( \frac{\text{prop}}{1 - \text{prop}} \right) \left( 1 - \frac{\hat{p}_{\text{prop}}}{\hat{p}_{\text{fail}}} \right) \right]^{-1}
\]

where: \( \hat{p}_{\text{fail}}^{(\text{prop}=\pi)} \) = sample-based estimate of probability of business failure (in year \( t \) conditioned on survival at the end of year \( t-1 \)) when the proportion of failure companies in the grand population is \( \pi \).

Note that \( \hat{p}_{\text{fail}}^{(\text{prop}=\pi)} \) in (7) constitutes an estimate of the population-based probability of business failure \( p_{\text{π}}^{(\text{fail})} \). As standard probit/logit estimation techniques do not provide for the specification of sampling errors associated with \( \hat{p}_{\text{fail}}^{(\text{prop})} \), it is hard to make a statement on the sampling characteristics of \( \hat{p}_{\text{fail}}^{(\text{prop}=\pi)} \). However, as indicated in the derivation of (4) in the previous section, a necessary condition for \( \hat{p}_{\text{fail}}^{(\text{prop}=\pi)} \) to be an unbiased estimator of \( p_{\text{π}}^{(\text{fail})} \) is that the sample of failure companies constitutes a random drawing from the sub-population of failure companies and the sample of survival companies constitutes a random drawing from the sub-population of survival companies.

4. Implications for the use of probabilistic prediction models

Questions to be addressed in this section are concerned with some methodological consequences of the choice-based sample bias of \( \hat{p}_{\text{fail}}^{(\text{prop})} \). Throughout the discussion it is assumed that the sample of failure companies is representative for the sub-population of
failure companies and the sample of survival companies likewise is representative for the sub-
population of survival companies.\(^{10}\)

A first question to be addressed is how the size and statistical significance of the
coefficients in standard probit/logit models will be affected by the choice-based sample bias.
As explained in section 2, \(\hat{\pi}^{(\text{prop})}\) will be positively biased in relation to \(\pi\) when \(\text{prop} > \pi\). Given that the proportion of failure companies in the estimation sample is larger than the proportion of failure companies in the grand population, the estimated parameters can hence be expected to be “exaggerated”. However, as explained in Manski & Lerman (1977), it is difficult to more precisely pin-point the characteristics of this bias for individual descriptor variables. Empirical tests reported in Zmijewski (1984) and Bergström et al (1999) nevertheless indicate that the level of significance for estimated coefficients appear to be fairly robust to variations in \(\text{prop}\) (as long as there are at least about 40 failure companies in the estimation sample).

A second question is concerned with the impact of the choice-based sample bias on the
classification (referring to the estimation sample) or prediction (referring to a holdout sample)
accuracy of an estimated prediction model. A classification/prediction test involves the choice
of some cut-off value \(\bar{p}_{\text{fail}}\), such that companies with \(\hat{\pi}^{(\text{prop})} > (\leq) \bar{p}_{\text{fail}}\) are classified/predicted to be failure (survival) companies. Defining “error type I” as an erroneous
classification/prediction of a failure company and “error type II” as an erroneous
classification/prediction of a survival company, it is easily recognized that the choice of \(\bar{p}_{\text{fail}}\) typically involves a trade-off between the magnitude of errors type I and type II.\(^{11}\)

Basically two approaches have been suggested in previous research as to the choice of
\(\bar{p}_{\text{fail}}\) – an “empirical” and an “analytical” approach. In the former, a probability cut-off value \(\bar{p}_{\text{fail}}^{(\text{emp})}\) is determined as the cut-off probability being associated with the lowest “average error rate” or lowest “average error cost” for the estimation sample. Regarding the specification of the average error rate, the following definitions have been used in previous studies:

\(^{10}\) In the sense that \(\hat{p}(X|\text{fail})^{(\text{prop})} = p(X|\text{fail})^{(\pi)}\) and \(\hat{p}(X|\text{surv})^{(\text{emp})} = p(X|\text{surv})^{(\pi)}\), as discussed in section 2.

\(^{11}\) As an extreme, if \(\bar{p}_{\text{fail}} = 0\) there will be no errors type I but all survival companies will erroneously be classified/predicted as “failure companies” (errors type II). On the other hand, if \(\bar{p}_{\text{fail}} = 1.00\) all failure companies will be classified/predicted as “survival companies” (errors type I) but there will be no errors type II.
(8.a) \[ \text{rate}(e) = 0.5 \cdot \text{rate}(e_I) + 0.5 \cdot \text{rate}(e_{II}) \]

(8.b) \[ \text{rate}(e) = \text{prop} \cdot \text{rate}(e_I) + (1 - \text{prop}) \cdot \text{rate}(e_{II}) \]

(8.c) \[ \text{rate}(e) = \pi \cdot \text{rate}(e_I) + (1 - \pi) \cdot \text{rate}(e_{II}) \]

where: \( \text{rate}(e_I) \) = error rate type I = number of errors type I in relation to the number of failure companies.

\( \text{rate}(e_{II}) \) = error rate type II = number of errors type II in relation to the number of survival companies.

The average error rate in (8.a) is simply the arithmetic average of error rates type I and type II, while the average error rates in (8.b) and (8.c) are functions of the proportion of failure companies in the estimation sample or in the grand population. Note that if \( 0.5 = \text{prop} = \pi \), the average error rates in (8.a) to (8.c) will coincide, in turn implying that the corresponding cut-off values \( \overline{P}_{\text{fail}}^{(\text{emp})} \) will be the same. In other situations values of \( \overline{P}_{\text{fail}}^{(\text{emp})} \) will typically differ.

A probability cut-off value based on the average error cost provides for the lowest weighted error cost, in principle defined as:

(9) \[ \text{cost} = w_1 \cdot \text{cost}_I + w_2 \cdot \text{cost}_{II} \]

where: \( \text{cost}_I \) = cost associated with a classification/prediction error type I.

\( \text{cost}_{II} \) = cost associated with a classification/prediction error type II.

\( w_1 \) = relative weight of error cost type I.

\( w_2 \) = relative weight of error cost type II.

In previous research, \( w_1 \) and \( w_2 \) have often been assessed as a function of an a priori expected proportion of failure companies and empirically estimated values of \( \text{rate}(e_I) \) and \( \text{rate}(e_{II}) \). Using the expected proportion of failure companies in the grand population, one gets:
Values of $\overline{P}_{\text{fail}}^{(\text{emp})}$ according to (8.a), (8.b), (8.c) or (9) are affected by the choice-based sample bias in the same way as $\hat{p}_{\text{fail}}^{(\text{prop})}$ – i.e. $\overline{P}_{\text{fail}}^{(\text{emp})}$ will be positively biased if $\text{prop} > \pi$, and vice versa. However, this bias is not troublesome as long as the cut-off probabilities are only used with correspondingly biased values of $\hat{p}_{\text{fail}}^{(\text{prop})}$. Since rankings of companies based on $\hat{p}_{\text{fail}}^{(\text{prop})}$ and $p_{\text{fail}}^{(\pi)}$ are expected to coincide, there exists some biased value of $\overline{P}_{\text{fail}}^{(\text{emp})}$ that will generate the same average error rate, or average error cost, as an unbiased value of $\overline{P}_{\text{fail}}^{(\text{emp})}$ together with unbiased probabilities $p_{\text{fail}}^{(\pi)}$.

The “analytical” approach for determining a probability cut-off value – henceforth denoted $\overline{P}_{\text{fail}}^{(\text{ana})}$ – is founded in some decision context where unbiased probabilities are supposed to be used. In previous research, the choice of $\overline{P}_{\text{fail}}^{(\text{ana})}$ has often been guided by a simple trade-off between expected error costs in the following way:

\[(\#) \quad \boxed{\text{Expected error cost of classification/prediction} = \text{“survival”:} \quad p_{\text{fail}}^{(\pi)} \cdot \text{cost}_1}\]

\[(\#) \quad \boxed{\text{Expected error cost of classification/prediction} = \text{“failure”:} \quad (1 - p_{\text{fail}}^{(\pi)}) \cdot \text{cost}_\text{II}}\]

- **Decision rule**: Classification/prediction “failure” is made if $p_{\text{fail}}^{(\pi)} \cdot \text{cost}_1 > (1 - p_{\text{fail}}^{(\pi)}) \cdot \text{cost}_\text{II}$. Otherwise the classification/prediction “survival” is made.

The probability cut-off value can then be solved from the decision rule:12

\[(11) \quad \overline{P}_{\text{fail}}^{(\text{ana})} = \frac{1}{1 + (\text{cost}_1 / \text{cost}_\text{II})}\]

Analytically derived cut-off probabilities presuppose that unbiased probabilities of business failure are available. As stated previously, $\overline{P}_{\text{fail}}^{(\text{prop})}$ is positively (negatively) affected

---

12 The cut-off probability in (11) is implied by the decision rule in the sense that it is rational to classify/predict a company to be a “failure” company if $p_{\text{fail}}^{(\pi)} > \overline{P}_{\text{fail}}^{\text{ana}}$ and a “survivor” company otherwise.
by the choice-based sample bias when $\text{prop} > (\leq) \pi$ and hence not directly comparable with such cut-off probabilities. However, this problem can easily be handled through the relationship between $\hat{p}_{\text{fail}}^{(\text{prop})}$ and $\hat{p}_{\text{fail}}^{(\text{prop}=\pi)}$ in (7) above.

5. **Implications for the classification or prediction performance of probabilistic prediction models**

A large number of failure prediction models have been estimated and evaluated over the years, mostly without any specific attention to the importance of the choice-based sample bias. A pertinent question is thus to what extent this ignorance has distorted the “true” classification/prediction performance of the estimated models.

Addressing methodological issues related to probabilistic failure prediction models, the classification and prediction accuracy — in principle measured as $\text{rate}(\overline{e})'$ in (8.b) — for various proportions of failure companies in the estimation sample were calculated in Zmijewski (1984). In all tests a probability cut-off value of 0.50 was used, presumably viewed as an analytical cut-off value based on a symmetric loss function (i.e. $\text{cost}_1 / \text{cost}_2 = 1.00$).\(^{13}\) With regard to the observed classification results, it was noted:\(^{14}\)

“The results … generally indicate the existence of a bias and the overclassification of bankrupt firms when using unweighted probit.” (Zmijewski, 1984; p. 72.)

Observations of this kind are not surprising. As $\hat{p}_{\text{fail}}^{(\text{prop})}$ is positively associated with the sample proportion of failure companies, there will be more failure companies with calculated values of $\hat{p}_{\text{fail}}^{(\text{prop})}$ being larger than $\hat{p}_{\text{fail}}^{(\text{ana})} = 0.50$ when the sample proportion of failure companies is high. In turn a lower error rate type I — i.e. a lower fraction of misclassified failure companies — trivially follows.

The prediction accuracy was illustrated in Zmijewski (1984) with a holdout sample consisting of 41 failure and 800 survival companies, i.e. $\pi = 41/841 = 0.049$. The choice-based sample bias of $\hat{p}_{\text{fail}}^{(\text{prop})}$ was clearly observed in this holdout sample. With a proportion of failure companies in the estimation sample of 0.50, the average value of $\hat{p}_{\text{fail}}^{(\text{prop})}$ was found

\(^{13}\) Cf. Zmijewski (1984), note 16 on p. 72.

\(^{14}\) ”Unweighted probit” in the quotation refers to a standard application of probit analysis where no attention is paid to problems associated with the choice-based sample bias.
to be 0.19 in the holdout sample. When the proportion of failure companies in the estimation sample was reduced (from 0.50 to 0.286, 0.167, 0.091 and 0.048), the average value of $\hat{p}_{\text{fail}}^{(\text{prop})}$ decreased (from 0.19 to 0.11, 0.09, 0.07, 0.06 and 0.05, respectively).\footnote{From table 5 in Zmijewski (1984), p. 71.} Note that when the proportion of failure companies in the estimation sample was about equal to the proportion of failure companies in the holdout sample (0.048 $\approx$ 0.049), the average value of $\hat{p}_{\text{fail}}^{(\text{prop})}$ was very close to the a priori probability of failure in the holdout sample. This is consistent with $\hat{p}_{\text{fail}}^{(\text{prop})}$ being an unbiased estimate of $p_{\text{fail}}^{(\pi)}$ when $prop = \pi$ (cf. expression (4) in section 2).

Weighted average error rates in accordance with (8.c) were calculated with a probability cut-off value of 0.50 for the estimated probit models in Zmijewski (1984). The observed results were commented as follows:

“… the bankrupt firm correlation is positive … indicating an overclassification bias; the nonbankrupt firm correlation is negative … indicating an underclassification bias; and the overall correlation is negative, indicating that correct prediction rates increase when parameters which are less biased are used.” (Zmijewski, 1984, p.73.)

The overclassification bias for the failure companies and the underclassification bias for the survival are consistent with the choice-based sample bias of $\hat{p}_{\text{fail}}^{(\text{prop})}$ as previously noted. However, claiming that the overall error rates decrease as the model parameters are more unbiased, is more problematic. As the probability cut-off value (0.50) was distinctively different from the proportion of failure companies in the holdout sample (= 0.049), the cut-off value must rather have been optimal in some decision modeling context. Such a value should only be used when unbiased failure probabilities are available, i.e. values of $\hat{p}_{\text{fail}}^{(\text{prop})}$ based on estimation samples for which $prop = \pi$ or estimates of unbiased probabilities ($\hat{p}_{\text{fail}}^{(\text{prop}=\pi)}$). As the proportion of failure companies in the holdout sample was 0.049, the estimation sample with the lowest proportion of failure companies (prop = 0.048) should have been associated with almost unbiased values of $\hat{p}_{\text{fail}}^{(\text{prop})}$. It hence makes sense that the lowest weighted average error rate was observed for this estimation sample. On the other hand, roughly the same average error rates should have been possible to if the biased probabilities $\hat{p}_{\text{fail}}^{(\text{prop})}$ (based on the other estimation samples) had been adjusted in accordance with expression (7) above. No tests of this kind were reported in Zmijewski (1984), however.
6. Summary and concluding remarks

The purpose of the paper has been to enhance the usefulness of estimated probabilistic failure prediction models, in particular standard probit/logit model specifications. More specifically, the analysis has been limited to estimation problems associated with the “choice-based sample bias”, as the available literature (Zmijewski, 1984, and Bergström et al, 1999, in particular) is vague and somewhat misleading on this issue.

Probabilities of business failure are often statistically estimated in probabilistic prediction models, as in for example Ohlson (1980), Zavgren (1985) or Skogsvik (1990). The choice-based sample bias implies that such probabilities will be biased if the proportion of failure companies in the estimation sample (\( \text{prop} \)) deviates from the proportion of failure companies in the population (\( \pi \)). In previous empirical research, values of \( \text{prop} \) have in general been much larger than reasonable expected values of \( \pi \).

Through Bayes’ theorem it has been shown that a linkage between sample-based and population-based failure probabilities exists. Specifically, a sample-based probability \( \hat{p}_{\text{fail}}^{(\text{prop})} \) is a function of the corresponding unbiased probability \( p_{\text{fail}}^{(\pi)} \), the proportion of failure companies in the estimation sample, and the proportion of failure companies in the grand population (expression (4) in the paper). Characterizing this linkage between \( \hat{p}_{\text{fail}}^{(\text{prop})} \) and \( p_{\text{fail}}^{(\pi)} \), it has been found that \( \hat{p}_{\text{fail}}^{(\text{prop})} > p_{\text{fail}}^{(\pi)} \) if \( \text{prop} > \pi \) and vice versa, but that the expected ranking of companies is the same for \( \hat{p}_{\text{fail}}^{(\text{prop})} \) and \( p_{\text{fail}}^{(\pi)} \).

Knowing the linkage between \( \hat{p}_{\text{fail}}^{(\text{prop})} \) and \( p_{\text{fail}}^{(\pi)} \), the choice-based sample bias of \( \hat{p}_{\text{fail}}^{(\text{prop})} \) is less problematic. Estimated probabilities of failure can simply be adjusted for this bias (expression (7) in the paper) and the adjusted probabilities can be used in various decision models. There is hence really no need to re-estimate a probabilistic prediction model only because the sample proportion of failure companies differs from the proportion of failure companies in the population. Rather it is important to have sufficient numbers of randomly selected failure and survival companies.

Empirical observations in previous research – primarily Zmijewski (1984) – are in line with the analysis in the paper, i.e.:
Average values of \( \hat{p}_{\text{fail}}^{(\text{prop})} \) have been found to be larger than \( p_{\text{fail}}^{(\pi)} \) when \( \text{prop} > \pi \) and to increase as the value of \( \text{prop} \) goes up. This is consistent with the derived relationship between \( \hat{p}_{\text{fail}}^{(\text{prop})} \) and \( p_{\text{fail}}^{(\pi)} \) in (4) of the paper.

With a pre-determined probability cut-off value, an “overclassification” bias of failure companies and an “underclassification” bias of survival companies has been observed when \( \text{prop} > \pi \). This trivially follows from the choice-based sample bias of \( \hat{p}_{\text{fail}}^{(\text{prop})} \) being positive when \( \text{prop} > \pi \).

References


