International Journal of
*Financial Studies*

MDPI

*Article*

# On Unbalanced Sampling in Bankruptcy Prediction

**Marek Gruszczyński**

Institute of Econometrics, SGH Warsaw School of Economics, 02-554 Warszawa, Poland; marek.gruszczynski@sgh.waw.pl

Received: 11 November 2018; Accepted: 28 May 2019; Published: 5 June 2019

**Abstract:** The paper discusses methodological topics of bankruptcy prediction modelling—unbalanced sampling, sample bias, and unbiased predictions of bankruptcy. Bankruptcy models are typically estimated with the use of non-random samples, which creates sample choice biases. We consider two types of unbalanced samples: (a) when bankrupt and non-bankrupt companies enter the sample in unequal numbers; and (b) when sample composition allows for different ratios of bankrupt and non-bankrupt companies than those in the population. An imbalance of type (b), being more general, is examined in several sections of the paper. We offer an extended view of the relationship between the biased and unbiased estimated probabilities of bankruptcy—probability of default (PD). A common error in applications is neglecting the possibility of calibrating the PD obtained from a bankruptcy model to the unbiased PD that is population adjusted. We show that Skogsviks' formula of 2013 coincides with prior correction known for the logit model. This, together with solutions for other binomial models, serves as practical advice for obtaining the calibration of unbiased PDs from popular bankruptcy models. In the final section, we explore sample bias effects on classification.

**Keywords:** bankruptcy prediction; choice-based sample; logit model; probability of default; financial microeconometrics

**JEL Classification:** C25; G33; M4

## 1. Introduction

Bankruptcy probability—or probability of default (PD)—is of interest to many participants in and observers of corporate financial markets. The purpose of this paper is to explore why bankruptcy models estimated with the use of unbalanced samples generate bankruptcy probabilities that are biased and to recommend techniques for their calibration to unbiased probabilities.

We begin with comments on how bankruptcy as a rare event is studied and predicted from samples of data on insolvencies, defaults, and going concern opinions. We believe that there is rising demand for reliable bankruptcy predictions from users of our models, including company managers, equity owners, lenders (banks), investors, and others. Therefore, there is a constant need to refine the various approaches in this area, including the use of binomial models.

Section 3 on bankruptcy prediction models and unbalanced samples turns to Edward Altman's legacy of bankruptcy prediction modelling and the influence of his works on the modern finance profession and academia. We also introduce two types of unbalancing in bankruptcy models: (a) when the ratio of bankrupt and non-bankrupt companies in the sample is different than 50:50; and (b) when the percentages of bankrupt and non-bankrupt companies in the sample are different than those in the population. The more general unbalanced sample type (b) is elaborated later in the paper. Unbalanced samples and their treatment in the past, as well as in new research, are discussed in Section 4.

Prior correction for the unbalanced samples in logit modelling is examined in Section 5. It is shown that the correction quoted by King and Zeng (2001) has been known in microeconometrics since

Anderson (1972) and Maddala (1983). The relevance of prior correction to calibrating the unbiased PD from estimated binomial models is shown in Section 7, which is preceded by Section 6 introducing the formula by Skogsvik and Skogsvik (2013). This formula represents the relationship between the biased probability of bankruptcy obtained from the model and the population-adjusted unbiased probability of bankruptcy.

Section 7 develops the equivalence of prior correction in the logit model with the Skogsvik formula, which provides additional justification to applying prior correction. Together with solutions for other binomial models, this sets out practical advice for the calibration of unbiased PDs based on biased predictions from binomial bankruptcy models. Section 7 also discusses the question of sample bias effects on the classification of companies. However, the entire paper puts aside the classification problem and emphasizes questions of PD prediction that are the subject of growing demand from analysts and practitioners in accounting and corporate finance.

## 2. Bankruptcy: A Rare Event

Bankruptcy or insolvency cases can be qualified as rare among the population of active companies. Table 1 presents the numbers of insolvencies and filings for bankruptcy in selected European countries for the years 2015–2017. Insolvencies in 2015 and 2016 represent the number of companies that were in such a state that prompted them to file for bankruptcy[1]. Failures in 2017 are the numbers of companies that filed for bankruptcy in 2017. From our point of view, the most important is the percentage of failed companies within the number of active companies. It is relatively low and rarely exceeds 1 percent.

**Table 1.** The number of insolvencies and failures in selected European countries in 2015–2017 and the percentage of failed companies in 2017.

| Country | 2015 [a] | 2016 [a] | 2017 [b] | Total Number of Companies in 2017 | Percentage of Failed Companies in 2017 |
|---|---|---|---|---|---|
| Austria | 5150 | 5226 | 5849 | 601,641 | 1.0 |
| Belgium | 9762 | 9170 | 9614 | 1,741,058 | 0.6 |
| Czech Republic | 2191 | 2115 | 854 | 1,186,539 | 0.1 |
| Denmark | 4029 | 6674 | 6497 | 570,496 | 1.1 |
| Finland | 3068 | 2848 | 1741 | 468,048 | 0.4 |
| France | 63,259 | 58,898 | 55,730 | 4,398,926 | 1.3 |
| Germany | 23,101 | 21,525 | 20,684 | 4,923,202 | 0.4 |
| Hungary | 9545 | 7528 | 7955 | 494,209 | 1.6 |
| Italy | 14,729 | 13,472 | 12,302 | 6,602,969 | 0.2 |
| The Netherlands | 6006 | 5012 | 5024 | 1,436,799 | 0.3 |
| Norway | 4462 | 4544 | 4927 | 758,531 | 0.6 |
| Poland | 747 | 805 | 1299 | 4,370,412 | 0.1 |
| Portugal | 4714 | 3616 | 2826 | 686,107 | 0.4 |
| Romania | 10,269 | 8371 | 8007 | 1,396,442 | 0.6 |
| Slovakia | 622 | 495 | 431 | 364,070 | 0.1 |
| Spain | 4729 | 4091 | 4059 | 3,561,348 | 0.1 |
| Sweden | 6426 | 6019 | 6544 | 1,307,362 | 0.5 |
| Switzerland | 4519 | 4648 | 6599 | 574,551 | 1.1 |
| UK | 19,825 | 19,825 | 16,920 | 6,054,041 | 0.3 |

Sources: For 2017, Global Bankruptcy Report (2017); for 2015 and 2016, Insolvency Outlook (2019). Note: [a] insolvencies as defined by Euler Hermes; [b] failures (i.e., companies that have filed for bankruptcy, as defined by Dun and Bradstreet).

Naturally, the percentages of bankrupt and failed companies differ across countries and across industries, as well as fluctuate over time. Countries differ in terms of bankruptcy regulations and their revisions (Entrepreneurship at a Glance, Organisation for Economic Co-operation and Development,

[1] The numbers for 2015 and 2016 are from Insolvency Outlook (2019) by Euler Hermes.

OECD 2017) and in terms of industrial structure. Also, business cycle swings affect companies differently across countries and industries. Therefore, detailed examination of past and present trends in bankruptcies should be performed on a country level. For example, in the Netherlands most bankruptcies in 2017 were recorded in trade, financial services, business services (specialized), and construction. Table 2 provides detailed figures for all sectors.

**Table 2.** Bankruptcies of businesses and institutions in the Netherlands in 2017 by sector.

| Sector | Proportion of Total Bankruptcies |
|---|---|
| Trade | 22% |
| Financial services | 15% |
| Specialized business services | 12% |
| Construction | 10% |
| Manufacturing | 8% |
| Renting, other business services | 8% |
| Accommodation and food services | 6% |
| Transport, storage | 5% |
| Information and communication | 5% |
| Care | 4% |
| Culture, sports, recreation | 2% |
| Real estate activities | 3% |
| Agriculture | 1% |

Source: Central Bureau voor de Statistiek (11 January 2019; https://www.cbs.nl/en-gb).

All in all, country statistics on bankruptcies that are available in central statistical offices and in specialized companies (e.g., Euler Hermes, Dun and Bradstreet, and Creditreform) indicate the rarity of the "bankruptcy" event. This is evidenced in the data on insolvencies, filings, and court resolutions as compared to the numbers of active companies.

Although instances of insolvency and bankruptcy are rare, the probability of bankruptcy (insolvency, default, etc.) is much talked about in contemporary accounting, corporate finance, and financial markets analysis—for obvious reasons. Bankruptcy is the state many parties would like to be forewarned of. Such parties include company management, equity owners, lenders, potential investors, and insurers. The rarity of actual events of insolvency creates questions of how to model such predictions for the entire population of companies.

## 3. Bankruptcy Prediction Models and Unbalanced Samples

The infrequent instances of company insolvency in large data sets of companies serve as the benchmark for assessing the probability of bankruptcy for all companies, including those outside the sample. Insolvency, bankruptcy, and default probabilities (PD—probability of default) are expected to be calculated for every firm. The modern methodology for assessing the PD began with the seminal paper by Edward Altman[2] on Z-Score in 1968 (Altman 1968). Altman's model uses multivariate discriminant analysis (MDA) to estimate the "score" that apprises stakeholders of a potential state of bankruptcy for the company in question.

Although a Z-Score as the outcome of multivariate discriminant analysis (MDA) is not expressed in terms of probabilities, Z-Scores may be inverted into probabilities with the use of logistic transformation. This conversion is "not strictly correct" (Hillegeist et al. 2004, p. 16) but may serve as a technique for comparison with other methods. A major complication is that independent variables in MDA must have a normal distribution, otherwise the MDA estimator is not consistent (Maddala 1983, p. 27). Nonetheless, if the Z-Score is taken as the argument in the logistic function, the result may be

[2] Professor Edward Altman holds a 2015 honorary doctorate awarded by my university—SGH Warsaw School of Economics. SGH holds an annual series of lectures dedicated to Professor Altman (http://www.sgh.waw.pl/pl/altman-lectures).

interpreted as the "default likelihood" and used as the "equity-implied probability of default" (Altman et al. 2011).

The Altman models have been challenged by approaches directly producing probabilities of bankruptcy, such as the logit model, as well as by more advanced machine-learning methods. Direct application of the Z-Score or its variants has proved problematic in other countries, under other legal regimes (accounting principles), and in other time frames. However, indirect applications (e.g., models with the same variables estimated for a new data set) are still acceptable. Let us cite here the paper by Altman et al. (2017) that shows the validity of the Z-Score approach internationally with large data sets, also compared to logit models that performed similarly or better. It is also worth referencing the paper by Barboza et al. (2017), which compares several machine-learning methods to discriminant analysis and logistic regression in predicting bankruptcy. It turns out that the Altman Z-Score variables fare relatively well in other setups and models.

Today a large area of finance is dedicated to forecasting financial distress or bankruptcy, employing appropriate methodology. Nonetheless, it seems that the finance profession in academia still does not recognize this new methodology as staple content in core corporate finance and accounting courses. The notable exceptions are textbooks by Damodaran (*Applied Corporate Finance*, 5th ed., Damodaran 2015) and Berk and DeMarzo (*Corporate Finance*, 4th ed., Berk and DeMarzo 2017).

The methodology of bankruptcy modelling may be attributed to financial microeconometrics. and more recently, to advanced data analysis. Financial microeconometrics "emerges as a natural consequence of applying statistical and econometric methods to corporate finance, accounting, and other fields of finance; the applied edge of research in accounting and corporate finance is inevitably linked with the use of notions such as statistical sample, population, and the operation on sets of microdata" (Gruszczyński 2018).

Technically, bankruptcy prediction models aim at explaining the binomial outcome variable $y$ representing bankruptcy (insolvency) with two possible values: $y_i = 1$ for bankrupt companies; and $y_i = 0$ for non-bankrupt companies. Modelling involves the explanation of the y variable with the set of independent variables $X$ (*covariates*). Typical goals of modelling the binary $y$ variable are the following:

(a) prediction of $y_i$ values for given covariate values (i) within the sample and, possibly, (ii) outside the sample—serving to examine the accuracy of the classification of $y$ by the model;
(b) forecast (prediction) of bankruptcy probabilities $P(y_i = 1)$, which usually precedes (a);
(c) finding the best set of covariates $X$ for (a), (b), or both;
(d) prediction of the change in probability $P(y_i = 1)$ associated with a unit change of a particular covariate.

A correct bankruptcy prediction model has, possibly, good classification accuracy and supplies reliable predictions of bankruptcy probabilities. As indicated above, the emphasis of this discussion is on the rarity of bankruptcy cases within the population of all companies. This means that the number of cases with $y_i = 1$ is considerably smaller than cases with $y_i = 0$, and that this population's proportion is usually not represented in research samples.

There are two types of unbalanced samples (some researchers prefer "imbalanced") in bankruptcy prediction modelling. Let us consider the $n$-element sample for bankruptcy modelling that includes $n_1$ bankrupt companies and $n_2$ non-bankrupt companies.

(a) If the proportion of $n_1$ and $n_2$ in the sample is different than 50:50, then the sample is considered unbalanced.
(b) If the proportions $p_1 = n_1/n$ and $p_2 = n_2/n$ are different than the fractions of bankrupt and non-bankrupt companies in the population, then the sample is considered unbalanced in terms of $p_1$ and $p_2$.

The sample that is balanced in terms of definition (a) requires undersampling of healthy (survival) firms. This occurs because populations of companies are large—e.g., there are more than 4 million companies in Poland (see Table 1). It is, therefore, not feasible to sample healthy firms in the same manner as bankrupt ones. The "50:50 samples" appear in the studies that use matching techniques[3]—each bankrupt company is matched to a healthy company that is "similar" in terms of size, industry, etc. Such a sample is considered unbalanced from the point of view of definition (b).

Type (b) imbalance, as the more general situation, is the subject of further examination in this paper. In both cases, however, the question is whether the observation enters the sample randomly or not. If not, we have the problem of sampling bias, which is common in bankruptcy prediction models.

## 4. Sampling Bias, Weighting, Resampling

Since the paper by Zmijewski (1984), the question of unbalanced samples has been examined from many angles but is still far from being resolved (see also Platt and Platt 2002; Chen et al. 2006). Non-random samples in bankruptcy models are the source of two types of biases:

- Choice-based sample bias results when the probability of a company entering a sample depends on the dependent variable attributes (e.g., data on bankrupt companies is collected and then healthy companies are selected using a matching scheme);
- Sample selection bias results when only observations with complete data enter the sample.

Zmijewski (1984) has shown in several simulations with the probit model that choice-based sample bias declines if the ratio of bankrupt and non-bankrupt companies in the sample approaches that in the population. Also, neither bias appears to affect statistical inferences or overall classification rates. However, they were shown to have an impact on the estimates for single observations (e.g., on the estimates of the probability of the bankruptcy of a particular company).

Unbalanced samples are also sometimes handled with appropriate weights for observations in bankruptcy modelling. In their exercise of comparing models internationally, Altman et al. (2017) used weights of single units both for unequal numbers of bankrupt and non-bankrupt data and for unequal numbers of observations across countries. In a footnote, the authors state: "Although the score (logit) in principle has a probability interpretation, the 'probabilities' estimated using this weighting scheme in this study do not, however, represent empirical PDs (*sic*). It would still require calibration procedures for the models to obtain PDs that correspond to associated empirical PDs in the population. But this is not attempted in the study, as our focus is more general (the classification accuracies of the models across countries). It is also worth noting that the original Z-Score does not have a PD interpretation either".

Weighting is a technique not often utilized in "classical" bankruptcy studies worldwide, despite the known results generated, for example, for the binomial logit (Manski and Lerman 1977). In the case of logit models, the use of appropriate weights may be as effective as the application of simple correction that is discussed in the next section (see also Maalouf et al. 2018). On the other hand, as stated by Long and Freese (2014), "The use of weights is a complex topic, and it is easy to apply weights incorrectly". For the logit model, the choice of weights is not straightforward.

The new generation of bankruptcy studies that has emerged with the use of machine learning techniques also propose new solutions for handling unbalanced samples. Zhou (2013) describes the use of oversampling and undersampling algorithms applied to 1981–2009 data on US bankruptcies and to 1989–2009 data on Japanese bankruptcies. Oversampling means sampling "the minority class over and over to achieve the balanced distribution of the two classes". Undersampling means "to select a portion of the majority class to achieve the distribution balance of the two classes" (Zhou 2013). The sampling techniques are: for oversampling, ROWR (random oversampling with replication)

[3] This is called matched-pairs sample design. Skogsvik and Skogsvik (2013) indicate that 70% of early studies on bankruptcy use this design (Zmijewski 1984). An example of more recent bankruptcy research with matched pairs is the study by Bodle et al. (2016).

and SMOTE (synthetic minority oversampling technique); and for undersampling, RU (random undersampling), UBOCFNN (undersampling based on clustering from the nearest neighbor), and UBOCFMGD (undersampling based on clustering from a Gaussian mixture distribution). These techniques may generate samples with a 50:50 composition of bankrupt and healthy companies and then may be used to verify various bankruptcy prediction models. What is important is that the major goal in studies using such techniques lies in finding the model that performs best in terms of classification accuracy. Other examples of such an approach are Choi et al. (2018) and Wagenmans (2017).

How does the unbalanced sampling in bankruptcy models interfere with bankruptcy probabilities? To answer this question, we will concentrate on the binomial logit model.

## 5. Prior Correction in the Logit Model

Let us consider the binomial logit model of bankruptcy and the consequences of unbalanced samples for prediction of bankruptcy probability. One method for overcoming the effects of unbalancing is weighting, as explained in the previous section. We advocate the use of a simple correction, sometimes called "prior correction" (King and Zeng 2001) or the "Anderson-Maddala correction" (Gruszczyński 2017). King and Zeng (2001) state that although econometricians attribute the correction to Manski and Lerman (1977), in fact, the correction has been well known since 1975 (Bishop et al. [1975] 2007). We challenge this finding by noting that the paper by Anderson (1972) first introduced this result, which was later restated by Maddala (1983, 1991).

Prior correction allows the analyst to convert the binomial logit model estimated based on an unbalanced sample to the model for the population. The condition is that the "fraction of ones" (i.e., bankrupt companies) in the population is known. As before, $y_i = 1$ means a bankrupt company, and $y_i = 0$, a non-bankrupt one. The subject of the modelling is the probability $P(y_i = 1)$. Let us assume that the fraction of ones in the population is equal to $\pi$. King and Zeng (2001) state that knowledge of $\pi$ "can come from census data, a random sample from the population measuring $y$ only". In the case of bankruptcy modelling, the fraction may be established from official data on bankruptcies and companies for a particular country, region, time period, etc. Assume that $\pi = N_1/N$, where $N_1$ is the number of bankrupt companies in the population and $N = N_1 + N_2$ is the population size (with the number of non-bankrupt companies equal to $N_2$).

Now, consider the $n$-size sample for bankruptcy modelling that includes $n_1$ bankrupt companies and $n_2$ non-bankrupt companies. The fraction of ones in the sample is $\overline{y} = n_1/n$. The proportion of bankrupt companies selected for the sample is $p_1 = n_1/N_1$ and the analogous proportion of non-bankrupt companies is $p_2 = n_2/N_2$.

Consider also the following binomial logit model with $k$ covariates $X$ and $k + 1$ parameters $\beta_0, \beta_1, \ldots, \beta_k$:

$$P(y_i = 1) = \frac{1}{1 + \exp\left(-\boldsymbol{x}_i'\boldsymbol{\beta}\right)} \tag{1}$$

where $\boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \ldots + \beta_k X_{ki}$. Logit model (1) can be written as:

$$logit\ P = ln\frac{P(y_i = 1)}{1 - P(y_i = 1)} = \boldsymbol{x}_i'\boldsymbol{\beta} \tag{2}$$

The maximum likelihood estimation of (2) in the $n$-element sample gives the estimate of intercept $\beta_0$ that needs to be corrected—if estimated (2) is to represent the population (not only the sample).

The correction, known as "prior correction" or the "Anderson-Maddala correction", amounts to subtracting the estimate of $\beta_0$ by:

$$prior\ correction = \ln\left[\left(\frac{1-\pi}{\pi}\right)\left(\frac{\overline{y}}{1-\overline{y}}\right)\right] \quad \text{[King and Zeng 2001]} \tag{3}$$

or

$$\delta = ln\left(\frac{p_1}{p_2}\right) \qquad \text{[Maddala 1983]} \qquad (4)$$

Corrections (3) and (4) are equal, which can be shown using the definitions of $p_1, p_2$, and $\pi$, $\overline{y}$. For the randomly selected sample, we have $p_1 = p_2$ and $\pi = \overline{y}$ and the prior correction is equal to zero. Thus, the non-random samples inherently imply the need to correct the model in order to have it represent the population. However, if the population is not precisely known, the fractions $p_1, p_2$, and $\pi$ can only be estimates or calibrations. Then, the correction (3) or (4) should be applied carefully and with a relevant comment. In any case, we apply the correction when we make inferences in the context of the entire population and not the sample itself. Later, we use the correction $\delta$ from (4). The considerations that follow in this section are a major extension of the paper Gruszczyński (2017).

We give a simple explanatory example based on Gruszczyński (2012). Let us consider a population of 100,000 companies, of which 60 are bankrupt and 99,940 are non-bankrupt. Typically, all bankrupt companies (i.e., 60) are selected for the sample. Then, from the 99,940 non-bankrupt companies, 60 companies are selected (e.g., at random). As a result, we have a sample of 120 companies with 50 percent of the companies from each group. After running the model based on the sample, we can calculate the estimates (theoretical values) of bankruptcy probabilities for companies in the sample of 120 companies. These are estimates with a sample selection bias. The unbiased estimates are obtained when we consider the entire population from which we did the sampling.

For calculating correction $\delta$ we find that $p_1 = 1$ (considering all bankrupt companies) and $p_2 = 60/99,940$. The model for all 100,000 companies is obtained from the model estimated for the sample by reducing the intercept $\beta_0$ in (2) by the value of $\delta$; that is, by ln(1)–ln(60/99,940), or 7.417981.

To sum up, a model estimated on a sample not representing the population's proportion of bankrupt companies gives estimates of bankruptcy probability, which are biased with regard to the entire population. Unbiased probabilities of bankruptcy can be obtained after adjusting the model.

Note that this discussion leaves aside the question of classification and classification accuracy. We concentrate here on estimating the PD, the probability of bankruptcy and insolvency, especially when the model is used for companies outside the sample.

## 6. Predicting Bankruptcy Probabilities from Non-Random Samples

Prior correction in the form of (3) or (4) coincides with the findings by Skogsvik and Skogsvik (2013). They also emphasize that the bankruptcy probabilities obtained from the bankruptcy prediction models depend on the proportion of bankrupt companies in the sample, and they are, therefore, biased (if the proportion of bankrupt companies in the sample is not the same as in the population). According to the authors' findings, there is an algebraic relationship between the biased bankruptcy probability of a given company (from the sample-based model), and the unbiased probability, which results from the proportion of bankrupts in the population. This proportion of bankrupts in the population (denoted by $\pi$) is treated as the a priori probability of bankruptcy. The probability of bankruptcy of a single company calculated from the model (sample-based) denoted by the Skogsviks as $p_{fail}^{prop}$ is, therefore, biased. It is the function of:

- unbiased probability $p_{fail}^{\pi}$;
- proportion (*prop*) of bankrupt companies in the sample;
- proportion $\pi$ of bankrupt companies in the population.

The formula (derived from the Bayes theorem (Skogsvik and Skogsvik 2013) is as follows[4]:

[4] The "Skogsviks' formula" may be also found in Appendix B of the paper on prior correction by King and Zeng (2001) (Equation (28) in that paper). It is derived from Bayes theorem, as in Skogsviks' case.

$$p_{fail}^{prop} = \left[1 + \left(\frac{\pi}{1-\pi}\right)\left(\frac{1-prop}{prop}\right)\left(\frac{1-p_{fail}^{\pi}}{p_{fail}^{\pi}}\right)\right]^{-1} \tag{5}$$

It follows that if $prop > \pi$, then $p_{fail}^{prop} > p_{fail}^{\pi}$, and vice versa. This means that in the typical situation of bankruptcy modelling (i.e., when $prop > \pi$), the PD for a given company calculated from the model is higher than the "population-adjusted" PD for the same company. This leaves the question of how much higher? The authors give the example with $prop = 0.5$, as in matched-pairs modelling and $\pi = 0.02$. If the unbiased probabilities are $p_{fail}^{\pi} = 0.01,\ 0.02,\ 0.10$, they correspond to the (biased) predictions from the model $p_{fail}^{prop} = 0.33,\ 0.55,\ 0.84$, respectively. Thus, the model-predicted probabilities are considerably higher than the unbiased probabilities. It should be noted that Equation (5) has been derived assuming random sampling from the population of bankrupt companies, as well as (separately) from the population of non-bankrupt companies.

Factor $\left(\frac{\pi}{1-\pi}\right)\left(\frac{1-prop}{prop}\right)$ in Equation (5) can also be written in terms of $\pi,\ \overline{y}$ and $p_1, p_2$ from Equations (3) and (4) as:

$$\left(\frac{\pi}{1-\pi}\right)\left(\frac{1-prop}{prop}\right) = \left(\frac{\pi}{1-\pi}\right)\left(\frac{1-\overline{y}}{\overline{y}}\right)$$

and

$$\left(\frac{\pi}{1-\pi}\right)\left(\frac{1-prop}{prop}\right) = \frac{p_2}{p_1}.$$

The factor is equal to 1 when proportions in the sample are the same as in the population (i.e., for a randomly selected sample). In such a case, the probabilities $p_{fail}^{prop}$ and $p_{fail}^{\pi}$ are equal, so the model produces unbiased estimates of bankruptcy probability. Obviously, this result has little practical value, since random samples from the population in the case of bankruptcy studies are not feasible. To simplify, in further formulas we use $p_1$ and $p_2$, where Equation (5) now has the form of:

$$p_{fail}^{prop} = \left[1 + \frac{p_2}{p_1}\left(\frac{1-p_{fail}^{\pi}}{p_{fail}^{\pi}}\right)\right]^{-1} \tag{6}$$

The Skogsviks' equation seems to be important for calculating a specific unbiased probability $p_{fail}^{\pi}$ for a company, needed in financial risk management and in the valuation of a company's equity or its bonds. From Equation (6) it follows that the adjustment of the unbiased PD is equal to:

$$p_{fail}^{\pi} = \left[1 + \frac{p_1}{p_2}\left(\frac{1-p_{fail}^{prop}}{p_{fail}^{prop}}\right)\right]^{-1} \tag{7}$$

This result allows the analyst to translate the model's predicted PD for a single company into the PD "calibrated for the fraction of failure companies in the population". Figure 1 shows how the model PD is related to the population PD for $\pi = 0.02$ and $prop = 0.5$ (i.e., for $\frac{p_1}{p_2} = 49$). This is roughly a "real" case, with 2% solvencies annually and with the "balanced" sample composed of 50% bankrupts and 50% non-bankrupts.

The adjustment of the probability of bankruptcy (default) is only possible when the number representing $\pi$ (i.e., the fraction of bankruptcies in a recognised population of companies for a given year) is known or may be feasibly approximated. In the case shown in Figure 1, the bankruptcy probability estimated from the model as 0.7 corresponds to the "population adjusted" probability of 0.045. This exaggeration of unbiased PDs that is inherent in bankruptcy models should be considered in practical uses.

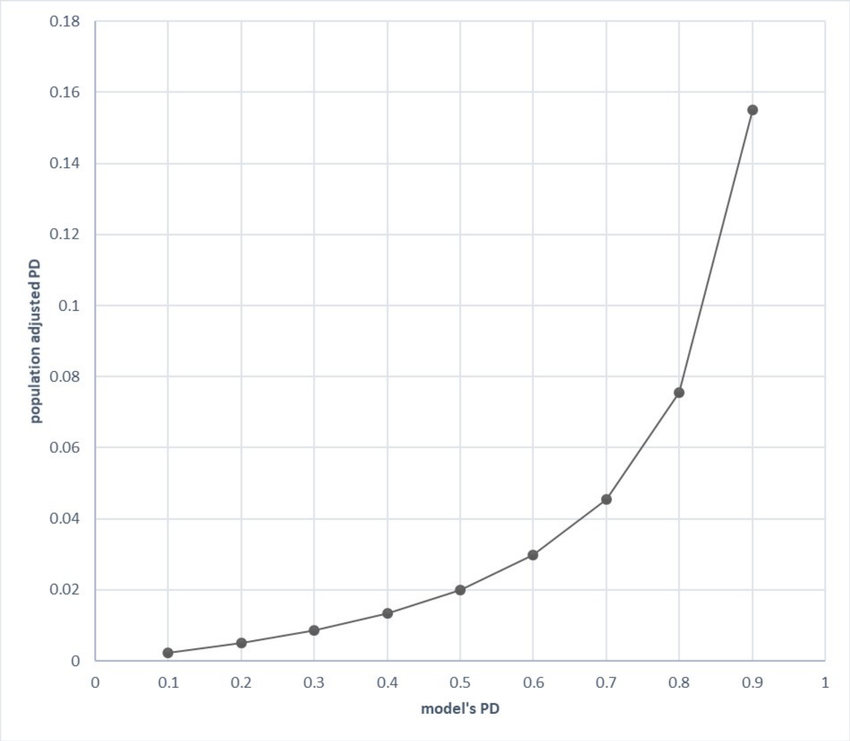

**Figure 1.** Plot of population adjusted PD (probability of default; $p^{\pi}_{fail}$) against the model's PD ($p^{prop}_{fail}$) for $\pi = 0.02$ and $prop = 0.5$.

## 7. Prior Correction: Formulae for Binomial Models and Classification Issues

The transformation of biased into unbiased probabilities can be further specified for various binomial models. We return to prior corrections (3) and (4) for the logit model (2):

$$logit\ P = \ln \frac{P(y_i = 1)}{1 - P(y_i = 1)} = \boldsymbol{x}'_i \boldsymbol{\beta} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \ldots + \beta_k X_{ki}$$

The correction $\delta$ of constant term $\beta_0$ estimate defined in (4) can be shown to coincide with the Skogsviks' equation. From (6) we have:

$$\frac{p_2}{p_1}\left(\frac{1 - p^{\pi}_{fail}}{p^{\pi}_{fail}}\right) = \left(\frac{1 - p^{prop}_{fail}}{p^{prop}_{fail}}\right) \tag{8}$$

and

$$ln\left(\frac{p_2}{p_1}\right) - ln\left(\frac{p^{\pi}_{fail}}{1 - p^{\pi}_{fail}}\right) = -ln\left(\frac{p^{prop}_{fail}}{1 - p^{prop}_{fail}}\right)$$

or

$$- ln\left(\frac{p_1}{p_2}\right) - logit\ p^{\pi}_{fail} = -logit\ p^{prop}_{fail}$$

and

$$\delta + logit\ p^{\pi}_{fail} = logit\ p^{prop}_{fail} \tag{9}$$

where $\delta$ is defined in (4). Therefore, the logit for the biased bankruptcy probability $p^{prop}_{fail}$ (i.e., what we receive from the estimated logit model) must be reduced by the value of $\delta$ in order to obtain the logit of unbiased bankruptcy probability $p^{\pi}_{fail}$. The reduction is contained in the constant term.

For example, an estimated logit model with seven explanatory variables is as follows[5]:

$$\widehat{logit\ P} = 0.79 + 0.26X_1 - 4.48X_2 + 0.48X_3 - 0.20X_4 - 0.01X_5 - 2.14X_6 - 7.60X_7 \qquad (10)$$

Equation (10) has been estimated for 40 bankrupt and 40 non-bankrupt companies. It means that the proportion of bankrupt companies in the sample is $prop = 0.5$. For the values of explanatory variables corresponding to one specified case (firm), the probability of bankruptcy resulting from this model is $p_{fail}^{prop} = 0.6$.

Now, let us assume that the proportion of bankrupt companies in the population is $\pi = 0.02$. Therefore, the ratio from Equation (4) is $\frac{p_1}{p_2} = 49$ (as in the Figure 1 example). In order to calculate the population adjusted (unbiased) probability from Equation (10), the intercept should be reduced by:

$$\delta = ln\left(\frac{p_1}{p_2}\right) = 3.89182$$

The new intercept is now equal to −3.10. The probability of bankruptcy obtained from Equation (10) with the new intercept is exactly equal to the probability $p_{fail}^{\pi}$ calculated from Equation (5) or Equation (6). In this case $p_{fail}^{\pi} = 0.0297$.

The Skogsviks' Equation (5) applies to the outcomes of all binomial models. For example, in the linear probability model (LPM) the probability $p_{fail}^{prop}$ of bankruptcy is equal to the estimate of the dependent variable for a specific company. From Equation (7), the unbiased $p_{fail}^{\pi}$ is calculated, provided that the proportions of $p_1$ and $p_2$ are known. In the probit model, the estimate of bankruptcy probability can also be calculated and inserted into Equation (7) as $p_{fail}^{prop}$. It should be noted that only in the case of the binomial logit model does there exist a simple correction for the estimated model that coincides with the Skogsviks' formula. King and Zeng (2001) point out that in the case of other binomial models like probit the only possibility is the use of a Skogsviks-like equation.

None of the foregoing considerations refer to the question of classification accuracy (classification of companies by the model). Since the rankings of the companies in terms of probabilities $p_{fail}^{prop}$ and $p_{fail}^{\pi}$ are identical (Skogsvik and Skogsvik 2013), the classifications of companies within the sample (into two groups, bankrupt and non-bankrupt) based on biased and unbiased predictions are the same, assuming the appropriate choice of the cut-off point.

The cut-off point $\alpha$ is the limit of probability for classification: if the estimated probability is less than $\alpha$, the company is classified as non-bankrupt; if not, it is classified as bankrupt. The default cut-off point in programs for estimating binomial models such as Stata is $\alpha = 0.5$. We advocate the use of Cramer's rule (Cramer 1999; Śmigielski et al. 2010), according to which the cut-off points $\alpha$ are:

- for the biased predictions from the bankruptcy model, $\alpha = prop = \overline{y} = \frac{n_1}{n}$;
- for the unbiased predictions from (7), $\alpha = \pi = \frac{N_1}{N}$.

Cramer's rule is based on the notion that the typical cut-off point of 0.5 applied for unbalanced samples does not allow one to reasonably predict less frequent cases. Cramer (1999) states: "(The) choice of 0.5 is usually defended by the argument that it is optimal if the predicted $y_i$ determine a course of action and if moreover the cost of misclassification is the same for either form that this may take. But if the cut-off point is optimal for the use of predictions in actual decisions it need not also be optimal for assessing the within-sample performance of the fitted model". Cramer (1999) proposes using a cut-off point $\alpha$ equal to the proportion of ones in the sample because it yields predictions that are optimal in the sense that they maximize the "index of performance" for each observation[6]. In

[5] Model estimated by Ciesielski (2005).

[6] "Index of performance" is defined as the probability of the observed outcome estimated from the model—related to the "null value" of this probability (i.e., estimated from the model containing only the constant term).

effect, the success rate for the unbalanced samples is better spread over the two alternatives, $y_i = 1$ and $y_i = 0$.

As has been noted in Section 3, models of multivariate discriminant analysis are not directly used to estimate bankruptcy probabilities. However, the MDA estimation results can be corrected by considering the population's proportion of bankrupt and non-bankrupt companies (Zmijewski 1984; Altman and Eisenbeis 1978).

The classification performance of traditional MDA and binomial models is challenged by new methods of data analysis, sometimes called "classifiers," applied to bankruptcy data. In Section 3 above, we mentioned the paper by Barboza et al. (2017) that compares several machine-learning methods to discriminant analysis and logistic regression in predicting bankruptcy. The paper by Jones et al. (2017) examines "classic classifiers", such as the logit and MDA, against neural networks, support vector machines, and statistical learning techniques, such as generalized boosting, AdaBoost, and random trees. However, the authors have not commented on the issues of sample selection bias, unlike in the papers cited in Section 4, in which machine-learning techniques are employed for resampling.

## 8. Conclusions

Bankruptcy models are typically estimated from non-random samples with the proportion of bankrupt companies differing from that in the population. This causes bias in the estimated bankruptcy probabilities for individual companies. The motivation of this survey-type paper was to explore how this bias may be assessed with the use of the estimation results in typical binomial bankruptcy models. Accurate, or at least calibrated, estimates of the probability of bankruptcy are required in risk assessment, discounted cash flow modelling, and for management and other parties interested in the financial fate of a company. Therefore, it is essential to properly (i.e., unbiasedly) evaluate such probabilities.

The paper introduces two types of unbalancing in the samples for bankruptcy prediction. We concentrate on samples with fractions of bankrupt and non-bankrupt companies differing from those in the population. We show the development of the Skogsvik and Skogsvik (2013) formula of the relationship between biased and unbiased estimated probabilities of bankruptcy. Skogsviks' formula coincides with prior correction (King and Zeng 2001; Anderson 1972; Maddala 1983) for the logit model. Similar solutions for other binomial models are also advocated for use in the calibration of unbiased PDs.

We believe this paper may be of help to researchers and analysts in corporate finance, as well as for company managers, investors, lenders, and auditors. Most issues discussed in this paper are often neglected in the application of bankruptcy models.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

Altman, Edward I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 23: 589–609. [CrossRef]

Altman, Edward I., and Robert A. Eisenbeis. 1978. Financial Applications of Discriminant Analysis: A Clarification. *Journal of Financial and Quantitative Analysis* 13: 185–95. [CrossRef]

Altman, Edward I., Neil Fargher, and Egon Kalotay. 2011. A Simple Empirical Model of Equity-Implied Probabilities of Default. *Journal of Fixed Income* 20: 71–85. [CrossRef]

Altman, Edward I., Małgorzata Iwanicz-Drozdowska, Erkki K. Latinen, and Arto Suvas. 2017. Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Models. *Journal of International Financial Management and Accounting* 28: 131–71. [CrossRef]

Anderson, James A. 1972. Separate sample logistic discrimination. *Biometrika* 59: 19–35. [CrossRef]

Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. Machine learning models and bankruptcy prediction. *Expert Systems with Applications* 83: 405–17. [CrossRef]

Berk, Jonathan, and Peter DeMarzo. 2017. *Corporate Finance*, 4th ed. London: Pearson.

Bishop, Yvonne M., Stephen E. Fienberg, and Paul W. Holland. 2007. *Discrete Multivariate Analysis: Theory and Practice*. New York: Springer Science & Business Media. First published 1975.

Bodle, Kerry A., Patti J. Cybinski, and Reza Monem. 2016. Effect of IFRS adoption on financial reporting quality: Evidence from bankruptcy prediction. *Accounting Research Journal* 29: 292–312. [CrossRef]

Chen, Jianguo, Ben R. Marshall, Jenny Zhang, and Siva Ganesh. 2006. Financial Distress Prediction in China. *Review of Pacific Basin Financial Markets and Policies* 9: 317–36. [CrossRef]

Choi, Hyunchul, Son Hyojoo, and Changwan Kim. 2018. Predicting financial distress of contractors in the construction industry using ensemble learning. *Expert Systems with Applications* 110: 1–10. [CrossRef]

Ciesielski, Piotr. 2005. Prognozowanie upadłości podmiotów gospodarczych w Polsce (Predicting bankruptcy of companies in Poland). *Rachunkowość (Accounting)* 2005: 28–42.

Cramer, Jan S. 1999. Predictive performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society* 48: 85–94. [CrossRef]

Damodaran, Aswath. 2015. *Applied Corporate Finance*, 5th ed. Hoboken: Wiley.

*Global Bankruptcy Report*. 2017. Short Hills: Dun & Bradstreet Worldwide Network.

Gruszczyński, Marek, ed. 2012. *Mikroekonometria*, 2nd ed. Warsaw: Wolters Kluwer.

Gruszczyński, Marek. 2017. Błędy doboru próby w badaniach bankructw przedsiębiorstw [Sample bias in the research on corporate bankruptcy]. *Kwartalnik Nauk o Przedsiębiorstwie* 44: 22–29. [CrossRef]

Gruszczyński, Marek. 2018. Financial Microeconometrics as Research Methodology in Corporate Finance and Accounting. In *Efficiency in Business and Economics, Springer Proceedings in Business and Economics*. Edited by Dudycz Tadeusz, Osbert-Pociecha Grażyna and Brycz Bogumiła. Cham: Springer, pp. 71–80.

Hillegeist, Stephen A., Elisabeth K. Keating, Donald P. Cram, and Kyle G. Lundstedt. 2004. Assessing the Probability of Bankruptcy. *Review of Accounting Studies* 9: 5–34. [CrossRef]

Insolvency Outlook. 2019. The View. *Euler Hermes Economic Research*, January 9.

Jones, Stewart, David Johnstone, and Roy Wilson. 2017. Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks. *Journal of Business Finance and Accounting* 44: 3–34. [CrossRef]

King, Gary, and Langche Zeng. 2001. Logistic Regression in Rare Events Data. *Political Analysis* 9: 137–63. [CrossRef]

Long, Scott J., and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*, 3rd ed. College Station: Stata Press.

Maalouf, Maher, Dirar Homouz, and Theodore B. Trafalis. 2018. Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Computational Intelligence* 34: 161–74. [CrossRef]

Maddala, Gangadharrao S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Maddala, Gangadharrao S. 1991. A Perspective on the Use of Limited-Dependent and Qualitative Variables Models in Accounting Research. *The Accounting Review* 66: 788–807.

Manski, Charles F., and Steven R. Lerman. 1977. The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica* 45: 1977–88. [CrossRef]

OECD. 2017. *Entrepreneurship at a Glance 2017*. Paris: OECD Publishing.

Platt, Harlan D., and Marjorie B. Platt. 2002. Predicting corporate financial distress: Reflections on choice-based sample bias. *Journal of Economics and Finance* 26: 184–99. [CrossRef]

Skogsvik, Kenth, and Stina Skogsvik. 2013. On the choice-based sample bias in probabilistic bankruptcy prediction. *Investment Management and Financial Innovations* 10: 29–37.

Śmigielski, Janusz, Anna Majdzińska, and Witold Śmigielski. 2010. Using ROC curves to find the cut-off point in logistic regression with unbalanced data samples. *Statistics in Transition* 11: 381–402.

Wagenmans, Frank. 2017. Machine Learning in Bankruptcy Prediction. Master's Thesis, Utrecht University Repository, Utrecht, The Netherlands.

Zhou, Ligang. 2013. Performance of corporate bankruptcy prediction models on imbalanced data set: The effect of sampling methods. *Knowledge-Based Systems* 41: 16–25. [CrossRef]

Zmijewski, Mark E. 1984. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research* 22: 59–82. [CrossRef]