www.joim.com

Journal Of Investment Management, Vol. 4, No. 2, (2006), pp. 61–71

ARE THE PROBABILITIES RIGHT? DEPENDENT DEFAULTS AND THE NUMBER OF OBSERVATIONS REQUIRED TO TEST FOR DEFAULT RATE ACCURACY

Roger M. Stein[a,1]

Users of default prediction models often desire to know how accurate the estimated probabilities are. There are a number of mechanisms for testing this, but one that has found favor due to its intuitive appeal is the examination of goodness of fit between expected and observed default rates. While large data sets are required to test these estimates, particularly when probabilities are small as in the case of higher credit quality borrowers, the question of how large *often arises. In this short note, we demonstrate, based on simple statistical relationships, how a lower bound on the size of a sample may be calculated for such experiments. Where we have a fixed sample size, this approach also provides a means for sizing the minimum difference between predicted and empirical default rates that should be observed in order to conclude that the assumed probability and the observed default rate differ. When firms are not independent (correlation is non-zero), adding more observations does not necessarily produce a confidence bound that narrows quickly. We show how a simple simulation approach can be used to characterize this behavior. To provide some guidance on how severely correlation may impact the confidence bounds for of an observed default rate, we suggest an approach that makes use of the limiting distribution of Vasicek (1991) for evaluating the degree to which we can reduce confidence bounds, even with infinite data availability. The main result of the paper is not so much that one can define a likely error bound on an estimate (one can), but that, in general,under realistic conditions, the error bound is necessarily large implying that it can be exceedingly difficult to validate the levels of default probability estimates using observed data.*

1 Introduction

Credit default models have now become ubiquitous in banking and investment processes. It is typical for

[a] Moody's Investors Service, New York.

such credit models to assign probabilities of default in addition to assigning specific credit grades. Given the high reliance sometimes placed on these probabilities, users of these models often wish to know the accuracy of the probabilities because the levels of the (real-world) probabilities can have direct implications for underwriting, portfolio construction, and the allocation of capital. The need to validate probabilities can be particularly acute when a model has been produced by an outside group; for example, the research department of a financial institution or an outside academic researcher or a vendor.

Recently, the issue of calibration has attracted the attention of both regulators and practitioners due to its central role in determining capital adequacy and other risk measures. For example, the New Basel Accord (BIS, 2004) stipulates that a bank must demonstrate in its analysis that the probability estimates produced by a default model are reflective of underwriting standards and of any differences in the rating system that generated the data and the current rating system. Where only limited data are available, or where underwriting standards or rating systems have changed, the bank must add a greater margin of conservatism in its estimate of PD (cf., paragraph 451 of "A Revised Framework").

In such validation exercises, it is typically the case that users focus on issues both of model power (the model's ability to distinguish between defaulting and non-defaulting firms) and on the accuracy of a model's calibration (the appropriateness of the probability levels that the model produces). While much literature on default model validation focuses on aspects of power through the use of power curves and their associated statistics, a powerful model can turn out to be poorly calibrated and the "most accurate" probability model may not be the most powerful (cf., Stein, 2002).

In this article, we focus on one approach for validating default probability levels, which, though perhaps not as rigorous as some alternatives, has gained popularity in industry. In this approach, a researcher examines whether the observed default rate for borrowers of a certain credit grade is within the expected range for that credit grade. For example, a bank using a model to produce probability of default predictions might wish to know whether the predicted default rate in the "Pass − 2" rating category were correct. The bank might test this by examining all borrowers that were graded "Pass − 2" by the model over a period of time. By calculating the actual number of defaults observed and comparing this with the predicted or target number for "Pass − 2" borrowers, the firm could try to assess the accuracy of the model.

While most researchers readily acknowledge that large data sets may be required to perform such tests, particularly when probabilities are small as in the case of higher credit quality borrowers, the question of *how large* often arises. In this short note, we review some of the statistical machinery that can be used to help determine the number of records required to perform such tests. As it turns out, the number of records required can be large. For example, when default rates are low, say 50 bps, in order to be comfortable at a 99% confidence level that a 10 bps (20%) difference between a default model and actual data were not just due to noise in the data, we would need over 33,000 independent firms.

Our discussion is based on elementary statistical relationships. We characterize it as a lower bound because the analytic solutions shown assume no positive correlation among the data either in time or cross-sectionally, when in practice both of these assumptions are typically violated. We also show that effects of more realistic correlation structures can impact the calculation of such bounds significantly and we discuss this later in the article. For example, in the case of zero correlation, with sufficient data, the bound can be made arbitrarily small. However, when correlation is non-zero, adding more observations does not necessarily produce a confidence bound that narrows quickly.

Nonetheless, the analytic bound discussed here is useful in that it can be used to determine when the data at hand are *not* sufficient to draw rigorous conclusions about the probability estimates of a model. Conversely, where an experimenter has a fixed sample size, this approach can be used to size the minimum difference between an estimated and an empirical default rate that must be observed in order to not to conclude the rates are statistically indistinguishable.

Furthermore, in settings where the underlying assumptions are violated, simulation methods are available in many cases and we suggest approaches for this context. We also suggest a simple approach using the limiting distribution of Vasicek (1991) for evaluating the degree to which we can reduce the confidence bound even in the ideal case of infinite data availability. While this relies on a stylized analytic approximation, it provides useful insight.

It is important to note that while the approach of comparing expected to observed probabilities of default is a popular one, other more rigorous approaches can be used and we generally advocate these more sophisticated approaches. However, in industry, practitioners often use the expected default-rate approach and thus its properties, and, in particular its limitations, are of practical interest.

The remainder of this note proceeds as follows: in Section 2, we discuss the mathematical tools we use to calculate these bounds. In Section 3, we provide an example of how the approach can be applied. Section 4 presents some situations in which the analytic solution can break down. Section 5 discusses the limitations of the approach due to the simplifying assumption of no correlation in the data and presents a simulation example that demonstrates this. We also discuss the use of the limiting distribution of Vasicek (1991) for evaluating the degree to which we can reduce sampling error even in ideal cases.

2 Mathematical preliminaries

The mathematical machinery necessary for answering the questions set forth in this paper is well established in the form of the Law of Large Numbers and the Central Limit Theorem (CLT) and can be found in most textbooks on probability (cf., Papoulis, 1991; Grinstead and Snell, 1997).

In short, we take advantage of the limiting properties of the binomial distribution and assume it approaches normal distribution in the limit as the number of observations gets large. We can then formulate a standard hypothesis test, and with some algebra, solve for the sample size required to be sure that any difference between the true mean and the sample estimate will be small, with some level of confidence.

We start with an assumed predicted default probability, p, perhaps produced by a model, a rating system or as the target of some underwriting practice. We also have a data set containing n firms, d of which have defaulted. We wish to determine whether the assumed default rate is reasonably close to the true default rate.

We can define the empirical frequency of defaults (the default rate) as

$$f_d = d/n$$

We would like to be certain that

$$P(|f_d - p| < \varepsilon) \geq 1 - \alpha \tag{1}$$

where α is a significance level. For example, we might wish to be sure that the difference between the true default rate and our predicted default rate were less than 20 basis points.

In this case, the underlying distribution of f_d, the frequency of defaults, is binomial and we can appeal to the CLT and obtain a convenient (Gaussian) limit that facilitates calculations. Using the CLT, if $q \equiv (1 - p)$, we get the familiar result:

$$\begin{aligned} P(np_{\mathrm{L}} \leq np \leq np_{\mathrm{U}}) &\cong \frac{1}{\sqrt{2\pi}} \int_{\frac{n(p_{\mathrm{L}}-p)}{\sqrt{npq}}}^{\frac{n(p_{\mathrm{U}}-p)}{\sqrt{npq}}} e^{-x^2/2} dx \\ &= \Phi\left(\frac{n(p_{\mathrm{U}} - p)}{\sqrt{npq}}\right) - \Phi\left(\frac{n(p_{\mathrm{L}} - p)}{\sqrt{npq}}\right) \end{aligned}$$

where $\Phi(\cdot)$ is the standard cumulative normal distribution. Since here we are assuming that $p_{\mathrm{U}} - p = p - p_{\mathrm{L}} = \varepsilon$, this simplifies to

$$2\Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right) - 1 \geq 1 - \alpha$$

or, more conveniently,

$$\Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right) \geq 1 - \alpha/2$$

yielding

$$\frac{n\varepsilon}{\sqrt{npq}} \geq \Phi^{-1}(1 - \alpha/2)$$

Rearranging terms gives

$$n \geq \frac{pq}{\varepsilon^2}[\Phi^{-1}(1 - \alpha/2)]^2 \qquad (2)$$

Equation (2) gives the minimum required number of independent firms n, in the case that we wish to be certain that we will have enough data to determine whether a probability p is accurate to within ε at the α level.

Conversely, given that we have n independent firms, we can ask how big a difference between p and f_d we would need to observe in order to conclude at the α level that the assumed probability and the observed default rate differ. Rearranging terms in (2), we get

$$\varepsilon \geq \sqrt{\frac{pq}{n}} \Phi^{-1}(1 - \alpha/2) \qquad (3)$$

which gives the desired quantity.

Thus far we have assumed that we observed a predicted probability and wished to determine how well it matched an empirical default rate.[2] If, on the other hand, we were unsure of the true default rate and wished to estimate it, how many firms would we need? We can calculate this by observing that the quantity pq is maximized when $p = q = 0.5$. Setting $p = 0.5$ ensures that irrespective of what the true default rate is, we can measure it within ε with $100 * (1 - \alpha)\%$ confidence as long as we have at least n firms. Similarly, for a fixed number of firms, the estimate of the default rate that we obtain will be within ε $100 * s(1 - \alpha)\%$ of the time. This is just the standard confidence bound for a probability estimate.

To the extent that the sample size (e.g., the database of firms) we are using to test the accuracy of p is very much smaller than the full population (on the order of 5–10% of the full population), (2) and (3) will generally suffice. However, to the extent that the sample is much larger relative to the overall population, it is often advisable to make a *finite population correction* (fpc) (cf., Cochran, 1977). The fcp serves to adjust for the heightened variance observed in the sample of empirical defaults. The recommended adjustment to the variance is $(N - n)/(N - 1)$, where n is the sample size and the population size is N.

3 An example

Assume that a bank is using a rating system that produces scores in a number of predefined rating buckets and that it has assigned a probability of default to each bucket. The bank would like to determine whether the predicted default rate in the "Pass − 2" rating category is correct. Assume that firms in this category, according to the model, have a default rate in the range of 25–75 bps. How many companies would the firm need in this category to determine if the empirical default rate was within this range?[3]

Table 1 Required levels of ε for various sample sizes when $p = 0.005$ and $\alpha = 0.05$.

n	ε
1000	0.0044
2500	0.0028
5000	0.0020
10,000	0.0014

To evaluate this, we use (2), setting $p = 0.0050$ (the mid-point between 25 and 75 bps) and $\varepsilon = 0.0025$ (since 50 ± 25 bps $=$ [25 bps, 75 bps], the range of the "Pass $-$ 2" category). This yields an estimated sample size n of 3058 to achieve a 95% confidence level or 5281 to achieve a 99% confidence level.

If the firm only had 1000 firms available to test on, it could try to determine how big ε would have to be (i.e., how big a difference from 0.005 would the observed default rate, f_d, have to yield to indicate that the model's predictions were incorrect). To do so, the firm would use (3) and find that ε would be 0.0044 and 0.0057, for the 95% and 99% confidence bounds. In other words, given the number of firms, it would not be possible to conclude that the bucket probability was incorrect unless it was well outside the 25–75 bps range. Since $\varepsilon = 44$ bps in this case, any empirical default rate observed in the range of 6–94 bps would still not provide evidence that the probabilities were misspecified for the bucket. Table 1 gives some additional levels of ε for different sample sizes at the 95% confidence level.

4 Regions of breakdown for the analytic results

It is also useful to consider that when p and/or n is very small, the limiting properties on which this analysis relies may not be present. In such cases, it is not prudent to use the analytic results.

Table 2 Analytic vs. simulated levels of ε for various sample sizes when $\alpha = 0.05$.

n	p	Analytic ε	Simulated ε	% Diff.
100	0.001	0.0062	0.009	45
250	0.001	0.0039	0.003	−23
500	0.001	0.0028	0.003	8
1000	0.001	0.0020	0.002	2
50	0.025	0.0433	0.035	−19
100	0.025	0.0306	0.025	−18
250	0.025	0.0194	0.019	−2
500	0.025	0.0137	0.013	−5

Table 2 gives examples of the simulated and analytic results for several selected cases. In the simulations, for each of S simulation iterations, we generate n Bernoulli random variables with a known probability p and then calculate f_d based on the simulated sample. We then calculate the desired quantile (e.g., $\alpha = 0.05$) of the distribution of $|f_d - p|$ (over all S results) to produce the simulated value of ε. We compare this with the value calculated using (3). The results in Table 2 were generated using $S = 10{,}000$.

From Table 2, it is clear that the analytic result provides a reasonable estimate in cases where n and p are fairly large, or, more appropriately, not very small. However, the relative difference in predicted values ($[\varepsilon_{\text{simulated}} - \varepsilon_{\text{analytic}}]/\varepsilon_{\text{simulated}}$) can be quite large in cases where the values are too small. For example, even for moderately high default probabilities (e.g., $p = 2.5\%$), the difference between the analytic approximation to the error bound ε and the simulation result is almost 1% in default probability (83 bps) for small samples ($n = 50$).

This result is generally consistent with a common heuristic that recommends avoiding the approximation unless the quantity npq is a good deal larger than 2. These are also often the cases in which the distribution can become significantly skewed, which complicates the interpretation of the results.[4] From our informal experiments, we recommend using simulation in cases where npq is less than about 4. In the experiments, this resulted in relative errors of less than about 10% when estimating ε.

5 Complications and caveats: why a lower bound?

We have described the measures in (2) and (3) as lower bounds. The statistical theory provides that under the assumptions of the CLT, the limiting values should be *upper* bounds, given assumptions of independence and appropriate values of n, p, and ε.

In practice, however, it is rare that databases of corporate obligors and defaults strictly meet the assumptions of the CLT, as we have presented it, particularly as they relate to the independence of observations. This is due to additional sources of variance[5] and correlation. We discuss the effects of correlation on the estimates of ε and n in the following section.

5.1 Correlation among data

Estimates of ε and n become understated in the presence of correlation among the data. The analysis above assumes independent (i.i.d.) observations. However, in the case of correlated observations (e.g., if the firms are affected by similar economic factors, etc.), this assumption does not hold. This is because the financial statements (in the first case) and the credit quality of the firms (in the second case) may be correlated from observation to observation.

Note that the binomial distribution only approaches a normal distribution in the limit *when the observations are independent*. Under independence, the results of (2) and (3) allow us to solve analytically for the quantities of interest in many cases. On the other hand, there is no analytic solution to this problem in general when the observations are not independent. Furthermore, in the non-zero correlation case, we have no guarantee that the bound (ε) goes to zero as the number of observations (n) gets large.

This result is sometimes surprising to researchers. However, intuitively, if we have a portfolio containing a single firm the distribution of defaults is clearly binomial since the firm either defaults or does not. If we now add to it another firm that is 100% correlated with the first, we do not get any benefit in terms of reducing the variability of the portfolio losses since the second firm will always default when the first one does and thus, conceptually, represents just a larger position in the same firm. If we add a third and a fourth, the variance similarly does not decline and the distribution remains binomial. In fact, we can add an infinite number of 100% correlated firms, and still not reduce the variance or approach a Gausian distribution.

To explore the impact of correlation, we performed a second set of simulations. This time, we assumed that there was a hidden factor (e.g., asset value[6]) that generated the probabilities of default for each firm and that this variable followed a Gaussian distribution. The hidden factor for each firm is correlated across firms in our simulations and default now occurs when the value of this hidden factor for a specific firm falls below a specific threshold (e.g., a default point) for that firm. In the case of a Gaussian factor, the threshold is chosen so that the probability of default for the firm is consistent with the probability of default of the firm. Simulation under this model involves generating the joint distributions of the firm-level factors for the population and evaluating whether each firm's factor has fallen below the firm's threshold. In the simulations shown here, all default probabilities and correlations are identical, but the same approach can be used for heterogeneous populations.[7]

Table 3 Required 5% significance levels of ε when firms are correlated to various degrees.

Correlation	n	$p = 1\%$	$p = 3\%$	$p = 5\%$
0.0	500	0.008	0.011	0.018
0.1	500	0.020	0.048	0.070
0.2	500	0.030	0.063	0.108
0.3	500	0.036	0.083	0.142
0.0	1000	0.006	0.008	0.011
0.1	1000	0.020	0.046	0.067
0.2	1000	0.029	0.061	0.102
0.3	1000	0.034	0.081	0.135

Table 4 Required 5% significance levels of ε for various sample sizes when firms are correlated (corr $= 0.03, p = 0.01$).

n	Corr $= 0.3$	Corr $= 0.0$	Analytic (Corr $= 0.0$)
25	0.070	0.030	0.039
50	0.050	0.030	0.028
100	0.040	0.020	0.020
250	0.038	0.010	0.012
500	0.036	0.008	0.009
1000	0.034	0.006	0.006
5000	0.034	0.002	0.003

We then estimated ε assuming different levels of (uniform) correlation among the firms.[8] We present the results in Tables 3 and 4.

In Table 3, as expected, the estimate of ε with zero correlation is significantly smaller than in the cases in which the correlation is non-zero. We observe that ε increases with the degree of positive correlation. For example, in Table 4, we see that the estimate of a 95% confidence level for ε using 1000 firms with a probability of default of 1% and no correlation is about 60 basis points. In contrast, ε turns out to be about six times greater, 3.4%, when the correlation is 0.3. In Table 4, we also see that the reduction in ε is small even as the sample size increases.

However, it is important to note that as the correlation increases the distributions of ε resulting from the simulations becomes increasingly skewed. For example, the skewness of the zero correlation case is moderately low at about 0.48 for this set of simulations. In contrast, the skewness of the distributions for the cases of $\rho = 0.1$ and $\rho = 0.3$ are 2.2 and 6.3, respectively. As a result of this skewness, we observe two effects. First the values of ε increase considerably with the correlation as the right tails of the distributions lengthen and produce more extreme values. Secondly, as a result of the loss of symmetry, the values of ε become more difficult to interpret since they are mostly generated at the tail on the right side of the mean of the distribution. We show this graphically in Figure 1 as well which presents the distribution of ε as correlation increases.

Also note that even in the case of zero correlation, there is evidence in Table 4 that the distributions become quite skewed when n is small, thus making the symmetric interpretation of ε more involved. In this case, it is not until n reaches about 500 that either the theoretical or simulated estimates of ε get smaller than p itself (see footnote 4). Since negative values of p are not feasible, this implies that the distribution must be skewed and thus the largest ε are being generated on the right-hand side of the distribution.

5.2 *Sizing the effect with Vasicek's limiting disribution*

It is interesting to ask whether one can ever have "enough" data to be comfortable that a particular sample has specified average default probability that is close to the predicted probability when the data are highly correlated. Vasicek (1991) derives

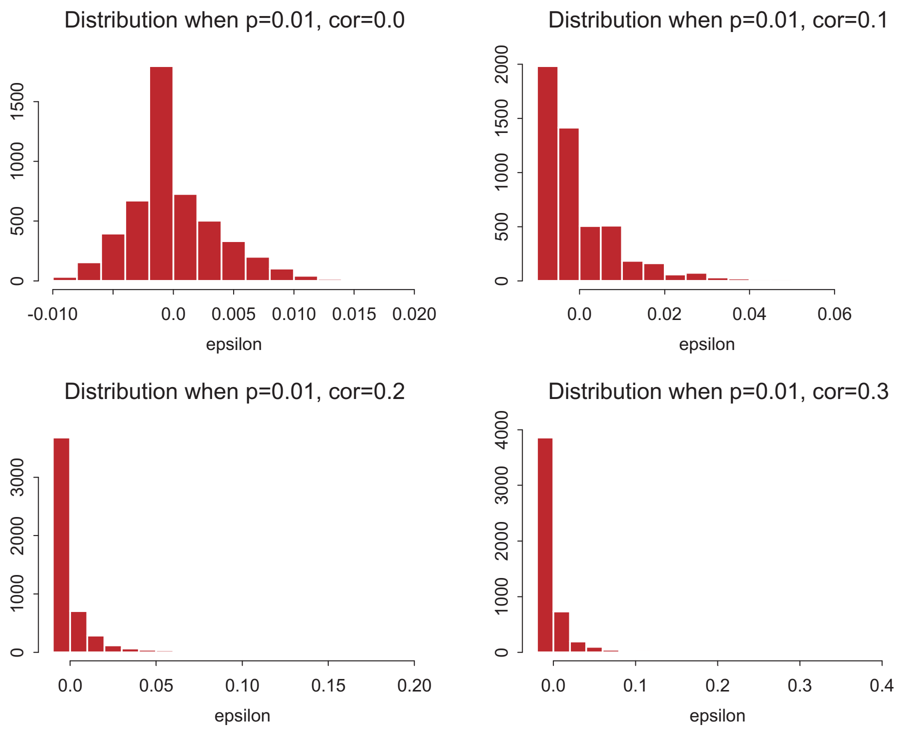

Figure 1 Distribution of ε at various levels of correlation when $p = 0.01$. This figure shows the distribution of ε when we assume uniform correlation at various levels and a uniform probability of default. We chose values of ρ at 0, 0.1, 0.2, and 0.3. Note that as the correlation increases the distribution becomes increasingly skewed. As a result we observe two effects. First, the values of ε increase significantly. Second, as a result of the loss of symmetry, these values become more difficult to interpret.

a limiting loss distribution for portfolios when all assets have identical asset correlation and probability of default. The limiting distribution holds for portfolios with infinite numbers of loans (or bonds) and is given as

$$F(\pi) = \Phi\left(\frac{\sqrt{1-\rho}\Phi^{-1}(\pi) - \Phi^{-1}(p)}{\sqrt{\rho}}\right) \quad (4)$$

where π is a an arbitrary default frequency and $F(.)$ is the cumulative probability. In words, (4) gives the probability that an infinitely large portfolio of loans with probability of default p and uniform correlation ρ will actually experience a realized default rate of π or less in any particular instance.

We can use (4) to get a quick sense of how much precision we might expect in testing default probabilities. Define $\pi_\alpha\ (p, \rho)$ as the value of π such that $F(\pi) = 1 - \alpha$ for a given probability p and correlation ρ. For example, $\pi_{05}(p, \rho)$ as the value of π such that $F(\pi) = 0.95$ for a given probability p and correlation ρ. Thus, for a given infinitely large portfolio with the probability of default p and correlation ρ, 95% of the time the realized frequency f_d will be less than $\pi_{05}(p, \rho)$. In statistical terms, this is the one-sided 95% confidence bound.

Now consider the following example. We have a real (i.e., finite) portfolio made up of loans with uniform asset correlation of 0.25 and that a true probability of default of 0.02. We cannot observe the true probability of default. Perhaps we believe that the probability "should be" 20 bps, maybe due to information from a model or other source. Using (4) we solve for $\pi_{05}(0.002, 0.25)$ to get a 95% confidence limit on how high the observed default rate might be while still not rejecting our assumption of a 20 bp probability of default. It turns out that $\pi_{05}(0.002, 0.25) \approx 87$ bps, since $F(0.0087) = 0.95$.

Now we can ask, even if we had extremely large amounts of data, how often we might falsely accept the 20 bps assumption at a 95% confidence level, given that the true (but unobservable) probability of default is 0.02 rather than 0.002 as we hypothesize. By using (4) again, we find that when the correlation is 0.25 and the probability of default is 0.02, $F(0.0087) = 0.5$. Thus, about half of the time, *even with infinite data*, we would mistake the default probability for 20 bps when it was actually 200 bps. If we were able to live with only 90% confidence, we will still accept the 20 bps hypothesis more than a third of the time.

This suggests that when correlations are higher it becomes more and more difficult to make strong statements about exact default probabilities, particularly when we observe only a single empirical realization of a data set or portfolio, regardless of how many loans it contains.

In other words, when correlation is high, ε is typically large.

It bears repeating, however, that while $F(.)$ is a limiting distribution, the limit only applies when asset correlations and default probabilities are both constant and uniform across the data. A similar result holds for finite populations, but again only in cases where correlations are constant. As discussed earlier, for the more general (and typical) case, no closed form solution exists.

5.3 Additional complications

In practice, with the exception of some classes of default model, it is sometimes difficult to estimate ρ, even when we might know something about the asset correlations. This is particularly true in overlapping samples such as those often used in testing default rates. In such samples, many of the firms will be present in the data set for more than one year and thus the sample overlaps to varying degrees from one year to the next. Covariance estimators for such situations have been developed for the study of census data but are less well known in the credit literature.[9]

Cantor and Falkenstein (2001), explore the case in which the correlation structure is more involved and the estimator of the variance becomes far more elaborate. These authors also show that failing to account for various correlation effects leads to significant underestimation of default rate variance. In a simulation study, Kurbat and Korablev (2002) make this point dramatically in a series of experiments in which they vary the levels of correlation among the firms at various risk levels. As we have also shown here, the authors' results show that the level of correlation among defaults can considerably affect the skewness of the resulting default distributions.

6 Conclusion

Users of default models that produce estimates of probabilities of default frequently desire to know the accuracy of the probabilities produced by the model. It is common for researchers to run experiments in which they attempt to estimate the goodness of fit between expected (under a model) and observed default rates. In this short note, we have reviewed some of the statistical machinery that can be used to help determine a lower bound on the number of records required to perform such tests.

The approach is based on simple statistical relationships. It is a lower bound as it assumes no correlation among the data in either in time or cross-sectionally, though both these assumptions are typically violated in reality. We provided a simple example of how even mild correlation among the firms in a sample can increase the required size of the effect that (and conversely the number of firms that must be observed).

That said, we feel that the bound is useful since it can be used to determine when the data at hand are not sufficient to draw rigorous conclusions about the probability estimates of a model. In addition, where an experimenter has a fixed sample size, this approach can be used to size the minimum difference between an estimated and an empirical default rate that must be observed in order to not to conclude that there is no evidence of a difference in the rates.

We show, both through simulation and the use of the results of Vasicek (1991), that the effects of correlation can dramatically increase the size of confidence bounds. Furthermore, even as data sets get very large, the error cannot be fully reduced.

There are a number of reasons to believe that estimates based on (2) and (3) will understate the actual required levels of the quantities ε and n in realistic settings. As a result, we feel that these formulae should be used as a hurdle test for examining default rate experiments.

While values that exceed the levels suggested by these formulae are necessary for experiments to have significance, they are almost certainly not sufficient. While they can be used to disqualify results in cases where an experiment produces an effect that is smaller than ε or that uses a sample smaller than n, they typically cannot be used to determine that a result is significant.

Notes

[1] I am grateful to Jeff Bohn, David Bren, Ahmet Kocagil, Matt Kurbat, Bill Morokoff, Richard Cantor, William Greene, and an anonymous referee for their generous comments. All remaining errors are of course my own.

[2] Importantly, an implicit assumption is that the data set accurately captures all relevant defaults. To the extent that this is not true (i.e., there are hidden defaults) the observed empirical probability may not reflect the true frequency of defaults. See Dwyer and Stein (2005) for a discussion.

[3] Here it is important to note that we are assuming that the true probabilities are distributed along the range of the upper and lower bounds of the bucket. An alternative assumption would be that the true probability is located somewhere in the range but that it could be represented as a single point (rather than a distribution). This latter assumption would require slightly different treatment. We chose the former as it seems more consistent with the formulation of the empirical problem.

[4] The skewness of the binomial distribution is given, after simplification as: $1 - 6pq/npq$. For theoretical binomial distributions the skewness becomes significant just below $p = 1\%$ and $p = 2\%$ for $n = 500$ and $n = 200$, respectively, using a variant of Fisher's test.

[5] The approach suggested in the example is a special case that does minimize to some degree the impact of estimation variability since the actual value of p that is being tested in the example is not an estimate. Rather it is the midpoint of the range of possible values of p associated with a particular rating class. We have not assumed that the classification of obligors into that class is done using any particular method so it is not necessary that the probabilities associated with the class be statistical estimates. They could, for example, be based on a particular underwriting target default rate. This is typically a special case. It is more often the case that users of a model are seeking to determine whether the probability produced by the model is consistent with the bank's experience in using the model. In this situation, this extra variability would likely increase the overall variance of the estimates of ε and n. Furthermore, if the estimate for the target default rate within the bucket were itself an estimate (say a mean or median of the PDs for borrowers within a particular rating class), the variance of this estimate would almost certainly affect the variance that (assuming

zero correlation) in turn will increase both the estimated values of ε and n.

[6] Note that correlation among asset values is not equivalent to correlation among defaults, although the two are related. See Gersback and Lipponer (2000) for a discussion.

[7] In the special case of uniform probabilities and correlations, Vasicek (1991) provides an analytic solution. We choose simulation here as it provides a more general solution in the case of more typical portfolios where correlation and probability of default are not uniform.

[8] I am grateful to Bill Morokoff for suggestions that greatly improved this section of the paper.

[9] The effects of overlap correlation effects can be significant and special methodologies have been developed to address the problem of "births and deaths" in the populations (e.g., new firms and firms that are delisted, dropped from the database, etc.). This phenomenon requires special attention as we observe that it is uncommon for the composition of the firms to be identical from year to year, rather some percentage of the sample overlaps of prior years from year to year with the addition of new firms and the loss of others.

References

Bank for International Settlements (BIS) (2004). *International Convergence of Capital Measurement and Capital Standards (A Revised Framework)*.

Cantor, R. and Falkenstein, E. (2001). "Testing for Rating Consistency in Annual Default Rates." *The Journal of Fixed Income*, September, 36–51.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

Dwyer, D. and Stein, R.M. (2005). "Inferring the Default Rate in a Population by Comparing Two Incomplete Databases." *Journal of Banking and Finance* **30**, 797–810.

Gersback, H. and Lipponer, A. (2000). *The Correlation Effect*. HeidelbergL: Alfred-Weber-Institut, University of Heidelberg.

Grinstead, C.M. and Snell, J.L. (1997). *Introduction to Probability*. Providence, RI, American Mathematical Society.

Kurbat, M. and Korablev, I. (2002). *Methodology for Testing the Level of the EDF™ Credit Measure*. San Francisco: Moody's KMV.

Papoulis, A. (1991). *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill.

Stein, R.M. (2002). *Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation*. New York, Moody's KMV.

Vasicek, O. (1991). *Limiting Loan Loss Probability Distributions*. San Francisco: Moody's KMV.

Keywords: Correlated default; default model; credit model; probability of default; model validation; calibration; capital allocation