AUGUST 2011

MOODY'S ANALYTICS

MODELING METHODOLOGY
FROM MOODY'S KMV

# The Effect of Imperfect Data on Default Prediction Validation Tests[1]


## Authors

Heather Russell
Qing Kang Tang
Douglas W. Dwyer

## Contact Us

Americas
+1-212-553-5160
clientservices@moodys.com

Europe
+44.20.7772.5454
clientservices.emea@moodys.com

Asia (Excluding Japan)
+85 2 2916 1121
clientservices.asia@moodys.co

Japan
+81 3 5408 4100
clientservices.japan@moodys.com



## Abstract

Analysts often find themselves working with less than perfect development and/or validation samples, and data issues typically affect the interpretation of default prediction validation tests. Discriminatory power and calibration of default probabilities are two key aspects of validating default probability models. Both are susceptible to data issues. This paper considers how data issues affect three important power tests: the accuracy ratio, the Kolmogorov–Smirnov test, and the conditional information entropy ratio. The effect of data issues upon the Hosmer–Lemeshow test, a default probability calibration test, is also considered. A simulation approach is employed that allows the impact of data issues on model performance, when the exact nature of the data issue is known, to be assessed.

We obtain several results from the tests of discriminatory power. For example, we find that random missing defaults have little impact on model power, while false defaults have a large impact on power. We also see that random misclassification errors are essentially a combination of missing and false defaults, and they have little impact on power if the percentage of misclassification is limited. As with other common level calibration test statistics, the Hosmer–Lemeshow test statistic simply indicates to what degree the level calibration passes or fails. We find that the presence of any data issue tends to cause this test to fail, and, thus, we introduce additional statistics to describe how realized default probabilities differ from those expected. In particular, we introduce statistics to compare overall default probability level with the realized default rate, and to compare the sensitivity of the default rate to changes in the predicted default probability. We find that development (respectively, validation) sample data issues tend to cause calibrated default probabilities to be closer together (respectively, further apart) than default rates computed for risk buckets of the validation sample.


[1] Published in *The Journal of Risk Model Validation* (1–20) Volume 6/Number 1, Spring 2012. 

## Table of Contents

# 1 Introduction

Default risk prediction models utilize market- and firm-specific information, such as equity prices and financial ratios, to provide a forecast of the probability of a firm defaulting over a future time period. Banks typically use default risk prediction models in their assessment and approval processes for loan applications and, increasingly, for the pricing of credits and capital allocation purposes. Utilizing a better default risk model can lead to economically significant differences in portfolio performance (see, for example, Stein (2005)). Thus, the ability to distinguish between a "good" versus a "bad" default risk model is of tremendous economic importance to a bank. Further, faced with a quickly changing external environment, a model that worked well in the past may not necessarily continue to work well in the future. Ongoing performance validation of existing models enables a bank to use them appropriately and to update them as necessary.

What constitutes a good default risk prediction model? The performance of default risk prediction models includes two components: the ability to correctly rank order firms by their relative default probabilities, and the ability of the model to generate default probabilities that closely match the realized default probabilities. We look at several statistical tests that measure the performance of default risk prediction models along one or more of these two dimensions. Along the level calibration dimension, we study the Hosmer–Lemeshow (HL) test. Along the discriminatory power dimension, we study the accuracy ratio (AR), the Kolmogorov–Smirnov (KS) statistic, and the entropy ratio, each of which provides a slightly different measure of the first performance dimension, which is the ability of a model to distinguish between defaulting and non-defaulting firms. Although the objectives of these validation tests are similar, their construction is quite different. Depending on the objectives, one may choose one test over another.[2]

Despite the wide array of available statistical tests, validation of model performance continues to challenge. Generally, model validation tests are designed under the assumption that the data is clean (i.e., that a default is a clearly defined and meaningful concept, that all defaults are captured in the data, and that each default can be properly linked to the relevant information regarding the firm (e.g., financial statements)). In practice, the characteristics of the data often differ across various dimensions of the data. For example, in the early period of the data, the default coverage may be less comprehensive than during the latter period of the data. Moreover, the quality and consistency of the financial statements of small firms may be lower than that of larger firms. Consequently, the interpretation of validation test results may need to be adjusted for these data deficiencies. Dwyer (2007) provides an overview of data considerations encountered in validating private firm default risk models such as sample selection biases, data collection issues, and differences in definition of default. Analysts are generally aware of the existence of such data issues in their validation and/or development samples, but they do not have an exact measure of the extent of such issues.

While many studies have focused on issues pertaining to the details of individual validation tests and methodologies, there has only been limited work done thus far on interpreting the impact of data issues on validation test results. Dwyer and Stein (2006) outline an approach for adjusting default rates to reflect sampling bias (see also Bohn and Stein (2009)). We investigate the impact of a variety of data issues on common validation tests within a simulation framework. The simulation approach allows us to isolate the effects of each data issue in order to examine its impact on validation results individually. In practice, multiple data issues may occur simultaneously within a data sample. A better understanding of the implications of data issues can help guide future model development and validation efforts.

This paper is organized as follows. Section 2 introduces common validations tests for assessing the performance of default risk prediction models and highlights their similarities and differences. Section 3 describes the simulation framework for generating defaults and estimated probabilities of default (PDs) and introduces the models used to simulate data issues such as missing and false defaults. Section 4 summarizes our results and findings.

[2] Friedman and Sandow (2003) discuss some of the limitations of popular measures for two-state conditional probability models based upon ROC measures, probability rankings, or upon conditional entropy within the context of utility maximization.

# 2 Tests of Model Calibration and Discriminatory Power

In this study, we focus our attention on four commonly used validation tests: the AR, the KS statistic, the Conditional Information Entropy Ratio (CIER), and HL statistics. We do not directly address the area under the receiver operating characteristic (AUROC) and the average default position summary statistic, (see, for example, Mann (2011)), as they are equivalent to the Accuracy Ratio via a linear transformation (see, for example, Engelmann *et al* (2003)) .

The HL test differs from the AR, the KS statistic, and the entropy measures in an important way. The HL test is a test of statistical significance. It tests as a "null hypothesis," whether or not model calibration is "correct." With a sufficiently large validation sample, the null hypothesis will be rejected unless the model and data are perfect. Therefore, we propose two summary descriptive measures that show how the observed default rate differs from the model predictions.[3] The first is the ratio of the observed default rate to the predicted default rates, which shows whether or not the model over- or under-predicts defaults, on average, on the validation sample. The second measures the sensitivity of the observed default rate to the PD and describes whether or not the model is over-calibrated or under-calibrated on the validation sample, in a sense that we will make clear. To the best of our knowledge, these descriptive measures are new to the literature.

## 2.1 Accuracy Ratio

The most common test for measuring a model's discriminatory power is the cumulative accuracy profile (CAP) and its associated summary statistic, the AR (see, for example, Basel Committee on Banking Supervision (2005) for a detailed description of the CAP and the AR summary statistic). We construct the CAP by first sorting all debtors by their respective model scores, from the riskiest to the safest. For a given percentage of the total number of firms, the percentage of total defaulters is calculated and plotted on the vertical axis. The CAP in Figure 1 illustrates the construction of the CAP curve. The perfect model successfully assigns the actual defaulters the riskiest scores. Hence, the perfect model's CAP curve increases steeply initially to one and remains there. The random model, on the other hand, randomly assigns scores to all firms, thus resulting in a CAP curve that approximates a straight, diagonal line. In reality, the CAP curves of actual rating models fall somewhere in between these two extremes. The accuracy ratio is defined as the area between the CAP curve of the rating model under validation and that of the random model divided by the area between the CAP curve of the perfect model and that of the random model. Mathematically, the accuracy ratio, AR = B / (A+B).

Several summary statistics are equivalent to the AR. The AUROC and the average default position can both be derived from the AR via a linear transformation, and, thus, contain exactly the same information as the AR (see, for example, Engelmann *et al* (2003)). The Gini coefficient, though defined differently, is identical in value to the AR, and, hence, also carries the same information as the AR.[4]

---

[3] Of course, one can and should plot the realized against predicted default rates by bucket and observe. This method may reveal a pattern in the data that is not characterized by the summary descriptive measures that we propose.

[4] The Gini coefficient has been used as a measure of income inequality since the early twentieth century. It is defined with respect to a "Lorenz Curve" for a variable that need not be dichotomous. When applying this concept to credit risk, one could define it with respect to either a CAP curve (in which the percentage of firms is on the horizontal axis) or with respect to a ROC curve (in which the percentage of the "goods" is on the horizontal axis). Curtis (2009) shows that the Gini Coefficient is identical to the AR when the Gini Coefficient is defined with respect to an ROC curve. It also follows immediately from the result that the AUROC is equal to ½AR + ½, which is shown in Engelmann *et al* (2003).

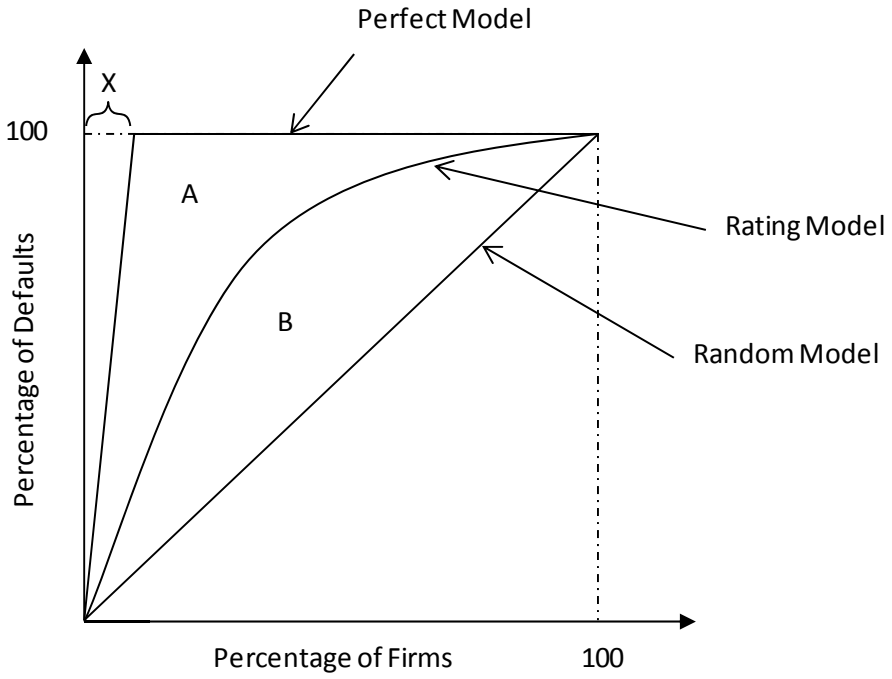


Figure 1 The Cumulative Accuracy Profile

## 2.2 Kolmogorov-Smirnov Statistic

The KS is closely related to the AR. Both are summary statistics derived from CAP plots. The KS statistic is a generic test to see whether or not two data samples come from different distributions. Within the context of default risk prediction model validation, the two data samples refer to the defaulter and non-defaulter data samples. Here, the KS statistic provides a measure of the maximum discrepancy or, conversely, the degree of overlap between the empirical defaulter and non-defaulter distributions. For each cutoff score[5] *C*, we can compute the difference between the hit rate and the false alarm rate, where hit rate is defined as the percentage of total defaulters with model scores higher than the cutoff score *C,* and the false alarm rate is defined as the percentage of defaulters with model scores lower than the cutoff score *C*.

The KS statistic is then defined as the maximum of these differences, as we enumerate through all possible values of cutoff score *C*.

$$\text{KS} = \max_{C} |\text{Hit Rate} - \text{False Alarm Rate}|$$

For a perfect model, the two distributions are completely segregated, resulting in a KS statistic of 1. Unlike the AR, the KS is not a global measure of discrepancy as it focuses only on the maximum discrepancy. As such, there is no direct one-to-one mapping between the AR and KS. Models with different ARs may have the same KS statistic and vice versa. KS is only sensitive to size and default rate above and below the cutoff, but not sensitive to the ability to differentiate between risk levels on either side of the cutoff.

## 2.3 Conditional Information Entropy Ratio

One way to compare the suitability of two default probability models is by looking at the likelihood of the realized defaults being produced under each of the two sets of probabilities. A comparison of quantities derived from these likelihoods (e.g. Akaike information criteria) reflects both differences in discriminatory power and calibration to default

[5] The score could be any credit risk measure, such as a default probability or rating, used to rank order credits from riskiest to safest.

probabilities. The CIER can be thought of as a likelihood-based test in which the credit risk measures are calibrated to realized defaults so that the final ratio measures only discriminatory power.

The information Entropy of an event with probability $p$ of occurring is defined as:

$$H(p) = -(p\log(p) + (1-p)\log(1-p))$$

Information entropy represents the negative of the expectation of the log of the likelihood of the realized outcome occurring. We extend the concept of informational entropy to a collection of events additively; the information entropy of a sample is the sum of the information entropies for each of the components.

We define the CIER for a specific credit risk measure on a data set for which we have default data as follows. First, we bucket the data into groups based upon the credit risk measure and assign default probabilities to each group based upon the empirical default rates. The CIER compares the information entropy $H_1$ of these bucketed default probabilities with the informational entropy $H_0$ of a single empirically determined default probability assigned to the entire set:

$$\text{CIER} = \frac{H_0 - H_1}{H_0}$$

The CIER reflects the drop in "uncertainty" when we change from a random model to a predictive model. The CIER construction ensures that it takes values between zero and one. If the model is completely random, we expect that the default rates computed on individual buckets is close to the default rate computed on the entire sample. Since $H_0$ is probably close to $H_1$ in this case, the CIER is probably close to zero. If the bucketed default rates perfectly predict default, then $H_1$ equals zero. Consequently the CIER equals one.

Unlike the AR and KS, the CIER is dependent upon the bucketing choice. If there is only one bucket, the CIER equals zero, regardless of the underlying measure. Similarly, if each observation is grouped into its own bucket, the CIER equals one. In our investigations with CIER, we use 10 buckets.

## 2.4 Hosmer-Lemeshow Test

An important aspect of validation is the verification of a model's ability to generate probability of defaults that track observed defaults, on average, over time. Known as level validation, it is related to how well a model has been calibrated. Level validation tests include the HL test, the binomial test, the traffic light approach, the Brier score, and the Spiegelhalter test. We focus on the HL Test, with the idea that our findings are similar for the other tests. The HL test groups firms into $k$ buckets by their predicted model scores and compares predicted and observed default rates in these bucket. In our investigations, we use 10 buckets. For each bucket $i$, we calculate the average predicted PD and note the number of firms. Next, we count the actual number of defaulters in each bucket, denoting it as $d_i$. The HL test statistic is then defined as:

$$\text{HL Statistic, } T = \sum_{i=0}^{k} \frac{(n_i PD_i - d_i)^2}{n_i PD_i(1-PD_i)}$$

The distribution of $T$ converges towards a $\chi^2_{k-1}$ distribution as $n_i \to \infty$.[6] The $p$-value of the HL test indicates whether or not the model's estimated probabilities of default are a good-fit relative to the observed probabilities of default. If the model calibration is inconsistent with the observed default rate, then the $p$-value will converge to one as the sample becomes large.

[6] This understanding assumes that the default events are independent. For the purposes of this paper, in all our simulation exercises, default events are independent. For discussion of the implications of correlation on tests of model calibration, see Kurbat and Korablev (2002) and Dwyer (2007).

## 2.5 Ratio of Means and Slope Measures

The HL test can reject a model calibration for a variety of reasons, even when the model calibration is useful.[7] Furthermore, the rejection of a hypothesis does not alone describe why the model calibration is inconsistent with the observed default rates of a particular validation sample.

Nonetheless, we find value in comparing realized and predicted default probabilities across buckets, as when computing the HL statistic. To see the discrepancy between the predicted and observed default rates, one could plot out the actual versus predicted default rates of each bucket and observe their relationship. We introduce summary statistics to quantitatively capture the key qualities likely to be observed in such a plot. We summarize the degree of level mis-calibration using the ratio of average actual default rates to average predicted default rates. Let $PD_0$ denote the initially predicted default probability, and let $D$ denote the default indicator. We define Ratio of Means as:

$$(1) \quad ratio\ of\ means = mean(PD_0)/mean(D)$$

There is more discretion involved in using a summary statistic to measure the sensitivity of calibrated default rates to predicted default rates. Since empirical default rates can have considerable sampling variability, we first use a probit model to smooth empirical default rates. We estimate the probit model:

$$(2) \quad D \sim \Phi[a + b\Phi^{-1}(PD_0)]$$

to obtain calibrated default rates:

$$(3) \quad PD = \Phi[\hat{a} + \hat{b}\Phi^{-1}(PD_0)]$$

where $\hat{a}$ and $\hat{b}$ are estimated from Equation (2). The slope $\hat{b}$ can be thought of as a measure of the smoothed empirical default rates sensitivity to the predicted default rates (see Appendix A).

Generally, a slope greater than one implies that the model over-estimates for firms with low predicted PDs and under-estimates for firms with high predicted PDs. In contrast, a slope of less than one may imply that the model over-estimates observed defaults for firms with high-PDs and under-estimates observed defaults for firms with low PDs. However, when the ratio of means is not close to one, the point at which calibrated and predicted default rates are equal may fall out of the data range. In this case, realized defaults tend to be either higher than predicted in all buckets or lower than predicted in all buckets. We illustrate this scenario in Section 3.1.1.

## 2.6 Comparison Between Validation Tests

Some tests are equivalent to the AR, in the sense that we can use linear transformations to convert between the three different summary statistics. However, we cannot map one-to-one between the AR, the KS statistic, and the CIER because of the subtle differences between their information content. Thus, in this section, we seek to provide a rough sense of the relationship between the three power tests.

First, we illustrate how the different statistics compare with one another when applied to portfolios of varying degrees of PD dispersion. We generate PDs using a simple one-factor model PD = Φ[X] where $X$ is normally distributed. We increase PD dispersion by increasing the standard deviation σ of $X$ while keeping the average sample default rate of each portfolio constant at 2% by adjusting the mean μ of $X$ accordingly. We find that AR, the KS statistic, and the CIER statistic generally increase for portfolios with higher PD dispersion (see Table 1). For example, we would expect the

[7] For example, when looking at the performance of a model in any one year, a "positive aggregate shock" can lead to an overstatement of default risk, and a "large negative aggregate shock" can lead to an underestimation of default risk.

statistics to be higher for a portfolio consisting of both investment and speculative grade credits than for the two subportfolios consisting of either investment-grade or speculative-grade credits alone.

Table 1 Comparison of validation test results for portfolios of different PD dispersions

| μ | σ | AR | KS | CIER |
|---|---|---|---|---|
| -2.054 | 0.001 | 0.004 | 0.014 | 0.000 |
| -2.054 | 0.003 | 0.008 | 0.019 | 0.000 |
| -2.054 | 0.006 | 0.010 | 0.019 | 0.000 |
| -2.054 | 0.011 | 0.014 | 0.018 | 0.000 |
| -2.054 | 0.023 | 0.030 | 0.032 | 0.000 |
| -2.056 | 0.045 | 0.070 | 0.059 | 0.002 |
| -2.062 | 0.091 | 0.128 | 0.097 | 0.005 |
| -2.087 | 0.181 | 0.243 | 0.177 | 0.017 |
| -2.178 | 0.354 | 0.450 | 0.330 | 0.052 |
| -2.515 | 0.707 | 0.731 | 0.571 | 0.134 |
| -2.693 | 0.849 | 0.797 | 0.637 | 0.178 |
| -2.890 | 0.990 | 0.843 | 0.689 | 0.224 |
| -3.101 | 1.131 | 0.880 | 0.736 | 0.278 |
| -3.324 | 1.273 | 0.907 | 0.774 | 0.326 |
| -3.557 | 1.414 | 0.925 | 0.803 | 0.369 |
| -3.798 | 1.556 | 0.938 | 0.825 | 0.406 |
| -6.161 | 2.828 | 0.983 | 0.926 | 0.641 |
| -8.952 | 4.243 | 0.993 | 0.958 | 0.756 |
| -29.117 | 14.142 | 0.999 | 0.991 | 0.923 |

# 3 Impact of Data Issues

One of the difficulties when studying data issues is the feasibility of isolating the issues one wants to analyze using real data. For this reason, we use simulation to produce data with the qualities we would expect if the data issues of concern were actually present.

Data issues can occur either in the data used to calibrate the model (the development data) or with separate data used to validate the existing model (the validation sample data). For the development sample, the issues we study focus on errors in the independent variable (i.e. the observed credit risk measure) and nonrandom missing defaults. For the validation sample, we study random missing defaults, random false defaults, name matching errors, and errors in the independent variable.

## 3.1 Errors in Validation Data set

We construct the development data by simulating observed credit risk measures as 1,000,000 independent variables $X$ drawn from a standard normal distribution. We simulate defaults based on $X$ by assigning the default probability $\Phi[-2.7 + 0.8X]$. We construct validation data by slightly changing the development data set according to the data issue we wish to isolate. Specifically, we draw a random number from a standard uniform distribution and flag the observation as a default if this number is less than the default probability.

### 3.1.1 Random Missing Defaults (Validation Data set)

When building data sets to estimate default probabilities, capturing the default events is often challenging. If the validation data set spans a long time period, as we look further back into the history, the default information may be incomplete. If the sample contains very recent data, the most recent defaults may not be captured immediately. It is important to know how missing defaults impact validation tests. Here, we focus on the case where the validation sample is missing the defaults, but not the development sample.

We construct the validation data set by retaining the 1,000,000 independent variables $X$ representing the observed credit risk measures in the development data and randomly selecting half of the defaults in the development data set as the observed defaults in the validation data. We observe that the introduction of missing defaults in the validation data set does not have a significant impact on AR and KS test results (see Table 2). Relatively, the CIER is affected more, while the HL test rightfully rejects the goodness-of-fit of the model's level calibration.

Table 2 Impact of missing defaults in validation data set

| | Development Sample | Validation Sample |
|---|---|---|
| Accuracy Ratio | 0.79 | 0.78 |
| KS | 0.62 | 0.62 |
| CIER | 0.21 | 0.17 |
| Ratio of Observed Defaults Rate to Mean PD | | 0.50 |
| Intercept | | -0.54 |
| Slope | | 0.87 |
| HL *p*-value | | 1.00 |

### Intuition Behind the Results

We see only a minimal decline in the Accuracy Ratio, even though 50 percent of the actual defaults are misclassified as non-defaults. We can explain the Accuracy Ratio statistic's robustness to missing defaults in the validation data set by considering the CAP curve in Figure 2. Note, that with missing defaults, the CAP curve for the model being validated remains nearly unchanged. To provide some intuition as to why the CAP curve remains nearly unaffected by missing defaults, we consider a simplified example. Assume that a model maps all firms to four distinct rating scores. When defaults are randomly "missed," the distribution of the percentage of total defaults across the four rating scores remains unchanged. Furthermore, since the percentage of total firms in each rating category remains unchanged as well, the CAP curve should not be affected. We can attribute the slight decrease in AR to the increase in the area marked A, which accounts for the smaller percentage of total firms observed as in default. Without missing defaults, the AR is given by $\mathrm{AR} = \frac{C}{B+C}$. With missing defaults $\mathrm{AR} = \frac{C}{A+B+C}$.

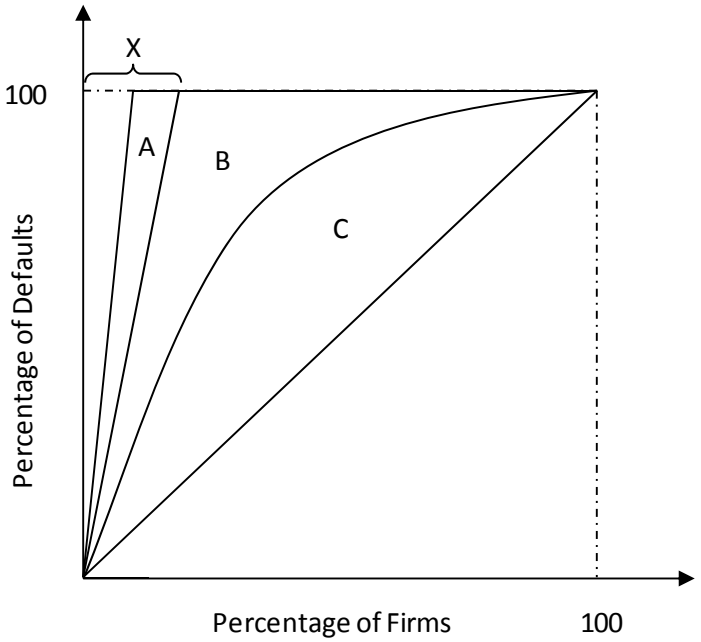


Figure 2 Shift in CAP curve as a result of random missing defaults

Furthermore, the fact that the KS statistic remains relatively unaffected can be explained by missing defaults. Recall, we define the KS statistic as |hit rate – false alarm rate|, where hit rate refers to the percentage of total defaulters with model scores lower than a cutoff score, and false alarm rate refers to the percentage of total non-defaulters with model scores lower than the same cutoff score. When defaults are randomly missing, the default distribution is reduced approximately proportionately, and these missing defaults are added to the left tail of the non-defaulter distribution. Since the defaulter distribution decreases approximately proportionately, for each cutoff score, *C*, the hit rate should remain approximately the same for all values of *C*. For the non-defaulter distribution, missing defaults increase the false alarm rate for low values of the cutoff score *C*. This reduces the KS statistic because the number of non-defaults (~98%) is typically much larger than the number of missing defaults (< 2%); the impact of the fatter left tail on the KS statistic is likely to be very small.

Unlike the AR and the KS statistic, the CIER is sensitive to random missing defaults. This trait demonstrates that CIER, besides being sensitive to the heterogeneity of the validation sample, is also dependent on its average default rate.

Last, we find that the validation data set with missing defaults causes the model to fail the HL test for level calibration. The ratio of means equals 0.5 and reflects the fact that expected default probabilities in the validation sample were off by a factor of two due to only half of the development sample defaults being observed. As the ratio of means moves further from 1, the slope may no longer reliably indicate either over-calibration or under-calibration of the model. In this example, the validation default probabilities are uniformly lower by a factor of two, and, thus, never cross one another.

Since the slope represents elasticity at the point where the probit model approximation to validation PD crosses the calibration PD, the slope of 0.87 reflects the limitations of the probit model functional form in this example.

### 3.1.2 Random False Defaults (Validation Data set)

In the past decade, banks have attempted to build default detection methods that collect default events consistent with the Basel definition of default. One of the challenges in doing so is in distinguishing between true defaults and technical defaults. For example, a borrower that became 90 days past due on a material credit obligation for reasons of financial distress would be considered a true default according to the Basel definition of default. However, some may argue that a borrower that becomes 90 days past due on a small amount for reasons not related to financial distress should not be viewed as a default under the Basel definition of default. Distinguishing between the two is often challenging. Supervisory guidance on the topic varies from country to country. It is important to assess the implications of technical defaults on the validation tests.

We construct the validation data set by retaining the observed credit risk measures in the development data and randomly setting 1% of non-defaults in the development data set so that the defaults in the development data set, together with these non-defaults, comprise the observed defaults in the validation data. Once again, we calibrate the PD model with a development data set that contains actual defaults. Unlike missing defaults, false defaults significantly impact all four validation tests, as shown in the Table 3.

Table 3 Impact of false defaults on validation test results

| | Development Sample | Validation Sample |
|---|---|---|
| Accuracy Ratio | 0.79 | 0.50 |
| KS | 0.62 | 0.40 |
| CIER | 0.21 | 0.09 |
| Ratio of Observed Default Rate to Mean PD | | 1.56 |
| Intercept | | -0.65 |
| Slope | | 0.53 |
| HL p-value | | 1.00 |

#### Intuition Behind the Results

In contrast to random missing defaults, a small percentage of random false defaults can significantly flatten the CAP plot. Figure 3 illustrates the changes to the CAP curve and defaulter/non-defaulter distributions when false defaults are found within the validation data set.

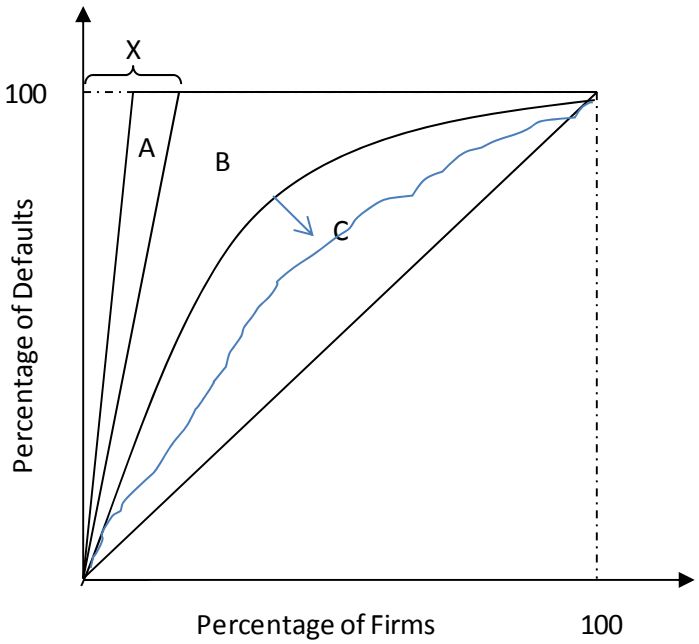


Figure 3 Shift in CAP curve as a result of false defaults

The model inevitably fails the HL test under this scenario, as we tested it on a validation data set that contained a significant number of false defaults. The ratio of means of two indicates that there are half as many defaults in the validation data set as predicted by the PDs calibrated to development data.

### 3.1.3 Name Changing Errors (Validation Data set)

In some cases, classification errors may not affect the observed default rate. For example, a subset of the IDs associated with the financial statement may be jumbled. We simulate such errors in the validation data set by starting with the development data and permuting the default flags for 10% of observations. Note that this permutation of default flags does not change the number of defaults or the default rate. Table 4 reports the results of this test.

Table 4 Impact of classification errors on validation test results

| | Development Sample | Validation Sample |
|---|---|---|
| Accuracy Ratio | 0.79 | 0.70 |
| KS | 0.62 | 0.56 |
| CIER | 0.21 | 0.16 |
| Ratio of Means | | 1.00 |
| Intercept | | -0.32 |
| Slope | | 0.81 |
| HL p-value | | 1.00 |

#### Intuition Behind the Results

This error is a combination of the first two cases (missing defaults and false defaults), with the results being driven mostly by the random false defaults case. The ratio of means remains one, because we randomly misclassify default status without changing the total number of defaults and total number of non-defaults in the data set. The slope is less than one, because the model appears "over-calibrated" on the validation sample. It appears that default rates are lower than anticipated for high-risk names and higher than predicted for low-risk names: a result of misclassified defaults.

### 3.1.4 Errors in an Independent Variable (Validation Data set)

In practice, the quality of the financial statements that financial institutions use to assess the risk is not uniform. Some financial statements may be audited, while others may be unaudited. In general, audit standards improve as firm size increase. There also can be data entry issues with how financial statements are captured in a bank's system. For example, unit issues (e.g. financial statements in thousands being recorded in units of millions) are common and difficult to fully eliminate. We look at the implications of such issues on the validation tests, when the data errors are in the validation data but not in the development data set.

We construct a validation data set with errors in the observed credit risk measure, by constructing a new credit risk measure $X^*$, observed for the validation data out of the credit risk measures $X$ observed from the development data set. We construct the credit measure $X^*$ with a standard normal distribution and correlation of 0.8 with $X$. The default flags are correct for both the validation data and the development data.

Simulation results summarized in Table 5 show that all four validation tests are negatively affected by the measurement errors in the validation data set. Power is reduced on the validation data set because the credit risk measure based upon data with measurement error is less informative.

Table 5 Impact of measurement errors in validation data set on various test results

| | Development Sample | Validation Sample |
|---|---|---|
| Accuracy Ratio | 0.79 | 0.65 |
| KS | 0.62 | 0.50 |
| CIER | 0.21 | 0.13 |
| Ratio of Means | | 1.00 |
| Intercept | | -0.51 |
| Slope | | 0.71 |
| HL p-value | | 1.00 |

The slope term is less than one. This result reflects that, with measurement error in the validation data, extreme PDs can be the result of measurement error and, as a result, the defaults rates are lower (respectively, higher) than predicted when the PDs are relatively high (respectively, low).

## 3.2 Errors in the Development Data

We construct the validation data by simulating observed credit risk measures as 1,000,000 independent variables $X$ drawn from a standard normal distribution. We simulate defaults based on $X$ according to the default probability $\Phi[-2.7 + 0.8X]$ assigned to each observation. We construct development data sets by slightly changing the validation data set according to the specific data issue we wish to isolate.

### 3.2.1 Errors in an Independent Variable (Development Data Set)

Here, we investigate the impact of introducing random measurement errors to the development data set. We validate the model using data in which the independent variable is observed without noise.

We construct the development data set with errors in the observed credit risk measure, by constructing a new credit risk measure $X^*$, observed for the development data out of the credit risk measures $X$ observed from the validation data. We construct the credit measure $X^*$ with a standard normal distribution and a correlation of 0.8 with X. The default data for the development data remains the same as that for the validation data. Results are summarized in Table 6.

Unlike the previous observation, we find that the AR, the KS statistic, and the CIER validation tests are insensitive to measurement errors in the development data set. The ratio of means also remains unaffected by random errors in the independent variable. However, we find the resulting slope is greater than one, suggesting that under this scenario, the model is under-calibrated relative to the validation data set. The calibration is based on a noisy measure of $X$, which results in the model under-predicting for high risk names and over-predicting low risk names when it encounters a validation data set with a clean measure of $X$.

Table 6 Impact of measurement errors in development data set on various test results

| | Development Sample | Validation Sample |
|---|---|---|
| **Accuracy Ratio** | 0.66 | 0.78 |
| **KS** | 0.50 | 0.62 |
| **CIER** | 0.13 | 0.20 |
| **Ratio of Means** | | 1.00 |
| **Intercept** | | 0.67 |
| **Slope** | | 1.38 |
| **HL p-value** | | 1.00 |

#### Intuition Behind the Results

Although such random measurement errors can cause incorrect calibration of a PD model, the relative rank ordering of firm PD estimates is preserved, thus explaining the unaffected AR, the KS statistic, and the CIER test results. Since the overall default rate does not change, the ratio of means remains equal to one. The slope is greater than one since the validation sample's independent variable has greater discriminatory power than the development sample's independent variable.

We note that, for multifactor models, the AR, the KS statistic, and the CIER statistics would typically be lowered by the model developed with the noisy development data. Relative rankings do change when the relative weights between the model covariates shift, and noisy development data may give us suboptimal weights. For example, if one variable was measured accurately in the development data set while another was measured with noise, the one measured with noise would typically be under-weighted relative to what the weights would be on a clean data set.

### 3.2.2 Non-Random Missing Defaults (Development Data set)

In this section, we examine the correlation implications of the likelihood of a missing default being correlated with the credit risk measure. Often, we can observe this scenario when there has been a change in the definition of default. For example, many parts of the world have only recently begun to collect defaults that are 90 days past due. A large firm with many facilities may be more likely to become 90 days past due on one facility for technical reasons than a small firm with fewer loan facilities. As large firms are generally viewed as safer than smaller firms, this issue would be more prevalent for safer firms.

We construct the development data set by constructing an independent standard normal variable $Y$ and relabeling defaults with $X+Y<0$ as non-defaults. This implies that firms with low levels of $X$ (and high credit risk) are more likely to be misclassified as non-defaults. Results are summarized in Table 7.

Table 7 Impact of biased, missing defaults in the development data set

| | Development Sample | Validation Sample |
|---|---|---|
| Accuracy Ratio | 0.83 | 0.79 |
| KS | 0.67 | 0.62 |
| CIER | 0.23 | 0.20 |
| Ratio of Means | | 1.13 |
| Intercept | | -0.08 |
| Slope | | 0.91 |
| HL p-value | | 1.00 |

All four validation tests appear to be negatively affected by the introduction of biased default errors into the development data set. This finding reflects the fact that the defaults missing from the development sample are harder to predict. Note that, by construction, we expect that the default rates calibrated on the development sample will be too low for the development sample. However, if we rescale PDs derived from the development sample to get a ratio of means of one, we still expect a slope term of less than one, since defaults in the validation sample are more evenly distributed across different credit risk measure levels.

# 4 Conclusion

In credit risk research we generally want to use as much of the relevant data as possible. As a result, we inevitably find ourselves working with data of heterogeneous quality. More recent data is usually of higher quality than older data. The quality of information on large firms is usually higher than that of small firms. This paper simulates commonly seen data issues and investigates their implications for model performance validation tests. We employ a simulation approach that allows us to assess the impact of a data issue on model performance, when the exact nature of the data issue is known. We view this work as useful for providing guidance on how data issues in the real world may impact model performance validation tests, even though one cannot measure the extent of a data issue in the actual data precisely.

Results, summarized in Table 8, show that random missing defaults have little impact on model power, while false defaults have a large impact on power, and that random misclassification errors are essentially a combination of missing and false defaults and have little impact on power if the percentage of misclassification is limited. In addition, we find that random measurement errors in the independent variable of the validation sample causes power to decrease significantly, whereas, random measurement errors in the independent variable of the development sample only impacts calibration in a univariate context.

Interestingly, we find that development sample data issues tend to cause under-calibration, while data issues in the validation sample tend to cause over-calibration of the model levels in single-factor models. Furthermore, we find that the HL test tends to fail for large samples whenever there are any data issues. The two summary statistics (ratio of means and probit regression slope), complement the HL test by characterizing the nature of the calibration errors.

Table 8 Summary of Observations

| Data Issue | Observations |
|---|---|
| Random missing defaults in validation data | AR, KS minimally impacted<br>CIER is more sensitive. |
| Random false defaults in validation data | All tests are sensitive. |
| Name-changing errors in validation data | Model appears over-calibrated |
| Noise in independent variables in validation data set | Model appears over-calibrated |
| Noise in independent variables in development data set | Model is under-calibrated |
| Non-random missing defaults in development data (likelihood of a missing defaults correlated with one independent variable) | Discriminatory power appears high on development data. |

# Appendix A Probit Slopes and Elasticity

We show mathematically that the slope is the elasticity of the calibrated default rate to the predicted default rate at the point where the average and calibrated default rates are equal. Elasticity is defined as

$$(4) \quad Elasticity = \frac{d\,(PD)}{d\,(PD_0)}\,\frac{PD_0}{PD}$$

At the point where the predicted and calibrated default rates are equal, this becomes the derivative of the calibrated default rate with respect to the predicted default rate.

Implicit differentiation gives:

$$(5) \quad [1/\,\Phi'(\Phi^{-1}(PD))\,]\,d(PD_0) = \hat{b}\,[1/\Phi'(\Phi^{-1}(PD_0))\,]\,d(PD)$$

At the point where calibrated and predicted default rates are equal, this can be simplified to:

$$(6) \quad d(PD) = \hat{b}\,d(PD_0).$$

Thus, the slope of the probit regression is the elasticity of the calibrated default rate to the predicted default rate, at the point where calibrated and predicted default rates are equal.

## Acknowledgements

The authors would like to thank Uliana Makarov and Shisheng Qu for their comments.



## References

Basel Committee on Banking Supervision, 2005, "Studies on the Validation of Internal Rating Systems." Working Paper No. 14.

Bohn, J. R. and R. M. Stein, 2009, *Active Credit Portfolio Management in Practice*, John Wiley and Sons.

Curtis, Andrew, 2009, "Gini = Accuracy Ratio Proof," unpublished note, Barclays Commercial Bank.

Dwyer, Douglas W., 2007, "The Distribution of Defaults and Bayesian Model Validation." *Journal of Risk Model Validation*, Vol. 1, No. 1, Spring (2007), 23-53.

Dwyer, Douglas W., and Stein, Roger M., 2006, "Inferring the Default Rate in a Population by Comparing Two Incomplete Default Databases." *Journal of Banking & Finance*, Vol. 30, No.3, (March 2006), 797-810.

Engelmann, B., Hayden, E., and Tasche, D., 2003, "Testing Rating Accuracy."*Risk,* Vol. 16, No. 3, 82-86.

Friedman, Craig, and Sandow, Sven, 2003, "Model Performance Measures for Expected Utility Maximizing Investors." *International Journal of Theoretical and Applied Finance*, Vol. 6, No. 4 (2003), 355-401.

Kurbat, Matthew, and Korablev, Irina, 2002, "Methodology for Testing the Level of the EDF Credit Measure." Moody's KMV Technical Report #020729.

Mann, Christopher, 2011, "Measuring Ratings Accuracy Using the Average Default Position." Moody's Special Comment.

Stein, Roger. 2005. "The Relationship Between Default Prediction and Lending Profits: Integrating ROC Analysis with Lending Practices." *Journal of Banking and Finance*, Vol. 29, 1213-1236.

Stein, Roger M., 2007, " Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation." *Journal of Risk Model Validation*, Vol. 1, No. 1, (Spring 2007), 77-113.