New Performance Measures for Credit Risk Models

Technical report delivered to Standard & Poor’s Rating Services following 2014 Workshop on Mathematical Problems in Industry, held at New Jersey Institute of Technology, Newark, NJ

Contributors: JiaMing Chen*, Sunil Dhar†, Dean Duffy‡, Yiheng Liu§, Richard O. Moore†, William Morokoff¶, Michael Pedneault†, Andrew Pole†, Yuzhou Qian*, David Rumschitski‖, Ting Wang**, Liming Yang¶, Maxim Zyskin††

June 23-27, 2014

1. Introduction

One of the most important functions of a ratings company is to provide the financial services industry with information they can use to assess the creditworthiness of a firm. This information is derived in part from probability of default (PD) models constructed from historical data specific to a company or class of companies, including debt levels, earnings, market capitalization, etc. The model then provides the likelihood of default over different time horizons for use by potential investors and creditors. Models range from purely statistical regressions over the available data to dynamic models based on geometric brownian motion of the firm’s total value.

---

*Rensselaer Polytechnic Institute
†New Jersey Institute of Technology
‡National Aeronautics and Space Administration (retired)
§Worcester Polytechnic Institute
§Standard & Poor’s Rating Services
‖City College of New York
**University of Maryland at Baltimore County
††Rutgers University
One of the key challenges associated with PD models is assessing their performance against real-world data. The ability of a performance measure to capture the true skill of a model is highly dependent on the data available for assessment. A dataset consisting of a broad range of a priori obligor risk will provide a reasonable test for the skill of a PD model; however, a dataset consisting of obligors that are either highly likely to default or very unlikely to default will not provide a meaningful test, or will simply provide a performance measure that does not reflect the model’s true skill.

Another challenge in developing both the PD model and a performance measure for its evaluation is to properly account for correlations in default events. Defaults experienced by different firms are often driven by similar underlying factors, rendering invalid the assumption of independent observations often required for the applicability of statistical tests.

This document summarizes the results from a week-long workshop whose goal was to devise a better testing criterion for model performance. The discussion below begins in Section 2 by considering data quality, as measured by its ability to provide a meaningful assessment of model performance. The incorporation of this data quality measure into standard model performance measures such as the accuracy ratio (AR) is discussed in Section 3. Section 4 then turns to the question of how to incorporate correlations in a regression PD model and in its performance assessment against real-world data.

2. Data Quality Analysis

This section considers the discriminating power of the observational data available for the performance assessment of a given PD model. In particular, it proposes the use of a measure of divergence between two conditional distributions as an indicator of the reliability of a performance measure such as the AR.

2.1 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence between two (absolutely continuous) distributions $P_1$ and $P_2$ is defined as

$$D(P_1||P_2) = \int P_1(x) \log \frac{P_1(x)}{P_2(x)} dx$$

(2.1)

for continuous variable $x$ and

$$D(P_1||P_2) = \sum_i \ln \left( \frac{P_1(x_i)}{P_2(x_i)} \right) P_1(x_i)$$

(2.2)

for discrete variable $x_i$. It is not, in general, symmetric, i.e.,

$$D(P_1||P_2) \neq D(P_2||P_1);$$

(2.3)
however, it is positive semidefinite, i.e., \( D(P_1||P_2) \geq 0 \), with equality if and only if \( P_1 = P_2 \) almost everywhere. This is seen by appealing to Jensen’s Inequality, which states that if a function \( f(x) \) is convex, then
\[
E[f(X)] \geq f(E[X]).
\]
Thus,
\[
D(P_1||P_2) = \int P_1(x) \log \frac{P_1(x)}{P_2(x)} \, dx = E_{P_1} [\log \frac{P_1(X)}{P_2(X)}] \\
\geq - \log E_{P_1} \left[ \frac{P_2(X)}{P_1(X)} \right] = - \log \int \frac{P_2(x)}{P_1(x)} P_1(x) \, dx = - \log 1 = 0. \tag{2.4}
\]
Although it is trivial to symmetrise the KL divergence by introducing
\[
D^*_{KL}(P_1||P_2) = \frac{1}{2} \left( D_{KL}(P_1||P_2) + D_{KL}(P_2||P_1) \right) \tag{2.5}
\]
\[
= \int \ln \left( \frac{P_1(x)}{P_2(x)} \right) (P_1(x) - P_2(x)) \, dx, \tag{2.6}
\]
it still cannot be considered a strict distance metric since it does not satisfy the triangle inequality. Nevertheless, it provides useful information on how closely one distribution resembles another, and this information can be used to inform the quality of a performance measure as derived from a particular dataset.

It is informative to consider the KL divergence restricted to two Gaussian distributions \( P_1 \) and \( P_2 \) with means \( \mu_1 \) and \( \mu_2 \) and standard deviations \( \sigma_1 \) and \( \sigma_2 \), respectively. Then,
\[
D(P_1||P_2) = \log \left( \frac{\sigma_2}{\sigma_1} \right) + \frac{\sigma_1^2 - \sigma_2^2 + (\mu_2 - \mu_1)^2}{2\sigma_2^2}, \tag{2.7}
\]
from which it is clear that the KL divergence defined this way is only symmetric if \( \sigma_2 = \sigma_1 \). In that case, the KL divergence is just the squared distance between the two means weighted by the variance.

### 2.2 Data Model

The context in which the KL divergence can be useful in characterising the data used in model performance is in distinguishing between the distribution of obligors conditioned on default and that of obligors conditioned on non-default. This can be demonstrated using a simple method for generating synthetic data.

Let \( x_i = \frac{i-1}{N-1} \) be the \textit{a priori} credit risk associated with obligor \( i \), with values from 0 (very creditworthy) to 1 (very uncreditworthy). Boolean data \( y_i \) indicate whether or not company
defaults, and are generated by:

\[ y_i = H(x_i + \sigma r_i - c), \]

(2.8)

where \( r_i \) is drawn from standard normal \( N(0, 1) \), \( c \) is a uniform threshold for default and \( H(x) \) is the Heaviside function defined as:

\[ H(x) = \begin{cases} 
1, & \text{for } x > 0 \\
0, & \text{for } x \leq 0.
\end{cases} \]

Figure 1 depicts a dataset of 100 obligors, showing defaults and non-defaults with a default threshold of \( c = 0.8 \).

![Figure 1: Defaults (x) and non-defaults (o) generated from 100 ranked obligors with a default threshold of c = 0.8.](image)

This simple data model allows an analytic expression to be derived for the probability of default of obligor with risk value \( x \):

\[ PD_i(x) = P(x + \sigma r - c > 0) = P(r > \frac{c - x}{\sigma}) = 1 - \Phi\left(\frac{c - x}{\sigma}\right), \]

(2.9)

where \( \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2} \, ds \) is the cumulative density function (cdf) for the standard
normal distribution. From this, the probability density function (pdf) of obligor with risk $x$ conditioned on default and on non-default can be computed using Bayes Theorem:

$$ P_D(x) = P(x|y = 1) = \frac{P(y = 1|x)P(x)}{P(y = 1)} = \frac{1 - \Phi\left(\frac{c-x}{\sigma}\right)}{\int_0^1 (1 - \Phi\left(\frac{c-x}{\sigma}\right))dx} \quad (2.10) $$

and

$$ P_N(x) = P(x|y = 0) = \frac{P(y = 0|x)P(x)}{P(y = 0)} = \frac{\Phi\left(\frac{c-x}{\sigma}\right)}{\int_0^1 \Phi\left(\frac{c-x}{\sigma}\right)dx}. \quad (2.11) $$

The KL divergence between these two distributions can now be assessed empirically for a given dataset or theoretically using the above formulas. In calculating the empirical divergence, it is tempting simply to compute the empirical mean and standard deviation, then use reduced formula 2.7. As figures 2 and 3 show, however, these distributions deviate significantly from Gaussians, suggesting that an alternative empirical formula based on a piecewise-constant approximation of the empirical cdf [1] be used instead. The analytical formula for the KL divergence, in its symmetric form, is given by

$$ D_{KL}^* = \int \ln \left( \frac{p(x|y = 1)}{p(x|y = 0)} \right) (p(x|y = 1) - p(x|y = 0)) dx $$

$$ = \int \left[ \ln \left( \frac{\Phi\left(\frac{c-x}{\sigma}\right)}{1 - \int_0^1 \Phi\left(\frac{c-x}{\sigma}\right) dx} \right) \left( \frac{\Phi\left(\frac{c-x}{\sigma}\right)}{\int_0^1 \Phi\left(\frac{c-x}{\sigma}\right) dx} - \frac{1 - \Phi\left(\frac{c-x}{\sigma}\right)}{1 - \int_0^1 \Phi\left(\frac{c-x}{\sigma}\right) dx} \right) \right] dx. $$

Figure 2 demonstrates the impact of varying default threshold $c$ and noise standard deviation $\sigma$ on the empirical KL divergence between the distributions of obligors conditioned on default and obligors conditioned on non-default. As one might expect, the KL divergence

![Figure 2: Pdf conditioned on default (red) vs rescaled Gaussian (blue).](image-url)
appears to decrease monotonically as $\sigma$ is increased and the distributions become less distinguishable in the data. The impact of default threshold $c$ on KL divergence appears to be a modest increase with increasing $c$.

### 3. Incorporating KL Divergence in Performance Measure

In addition to characterising the quality of the data for use in assessing performance measures, the KL divergence can be used to augment the performance measure in a way that quantifies its reliability. This section proposes two strategies for attaching a KL value to the accuracy ratio (AR) value derived from a dataset.

The AR is derived from the cumulative accuracy profile (CAP), which is a plot of the cumulative number of obligors having actually defaulted versus their *a priori* inverse risk value as generated by the PD model (in this case ranked from riskiest to safest). A perfect model with noise-free data identifies all defaulters as the highest risk (the green line in Fig. 5), whereas an uninformative model identifies defaulters by random chance (the pink line in Fig. 5). The AR is calculated using a ratio of areas $A$ and $B$ identified on Fig. 5, as

$$AR = \frac{B}{A + B}.$$  

One can readily see that $AR \in [0, 1]$. A PD model with high AR value is generally considered to have performed well.

#### 3.1 Impact of data on $AR$

To illustrate the impact of data quality on AR, two datasets have been generated using the same constants $c$ and $\sigma$ from the simple data model described in Sec. 2.2. A PD model is tuned for each of these datasets by inferring a threshold value $c$ from the data by taking the average of the smallest $x$-value associated with a default and the largest $x$-value associated
Figure 4: Empirical KL divergence for data model with varying default threshold $c$ and noise standard deviation $\sigma$. Calculated over sample size of 1000.

Figure 5: Cumulative accuracy profile (CAP). The AR is given by $B/(A+B)$. 

7
Figure 6: CAP figures for the same data model but different realizations. Left: $AR = 0.9$, $KL = 2.4$. Right: $AR = 0.4$, $KL = 0.9$.

with a non-default (recall that $x$ increases with riskiness of obligor). Figure 6 illustrates the impact of differing datasets (corresponding to different samples from the same distribution) in assessing the performance of the same PD model using the same underlying data model. A dataset with high discrimination between defaulters and non-defaulters can yield an artificially high AR value, while the opposite is true for a dataset with low visibility between conditional distributions. In Fig. 6, the CAP on the left has achieved a high AR value but with a relatively benign dataset, whereas the CAP on the right has achieved a lower AR value but with a more challenging dataset.

This demonstration immediately suggests that a performance measure, the AR in this case, should always be accompanied by a statistic qualifying the data. The next section discusses how the two values might be combined into a single measure to assess model performance.

### 3.2 $AR$ interval

The demonstration above suggests that a high KL divergence artificially inflates an empirically determined AR value, which a low KL divergence artificially depresses it. This raises the possibility of treating the AR value as the upper or lower bound, respectively, on an interval whose length is determined by the KL divergence. Letting

$$
\epsilon = \epsilon_m \frac{KL - 1}{KL + 1},
$$

one can show that $\epsilon \in (-\epsilon_m, \epsilon_m)$, with the extremes representing $KL = 0$, i.e., completely indistinguishable distributions, and $KL = \infty$, i.e., completely separate distributions. Letting $\hat{AR} = AR - \epsilon$, a reasonable adjusted AR interval might then take the form $(\hat{AR}, AR)$ if $\epsilon < 0$ (relatively indistinguishable data) or $(\hat{AR}, AR)$ if $\epsilon > 0$ (relatively distinguishable data). The value of $\epsilon_m$ might be determined by identifying the adjusted AR interval with a confidence interval in the sense of an appropriate statistic; however, for the purposes of this discussion it is arbitrary. Analysis of multiple datasets will in general produce a different interval for each dataset; the intersection of these intervals, where it is not empty, could be used as the effective interval resulting from the multiple-dataset analysis.
4. Logistic Regression Model

Another important aspect of the data used to assess PD model performance is the existence of correlations that invalidate statistical tests derived under an assumption of independent observations. The remaining sections propose an appropriate statistical treatment of these correlations in the context of a regression model with a goodness-of-fit measure that is also adapted to accommodate correlations.

Consider the logistic regression model described by

\[ E(y_i) = n_iP(x_i) = \frac{1}{1 + e^{-x_i^t\beta}}, \quad i = 1, 2, ..., m \]

where \( y_1, ..., y_m \) are the observed response values which are distributed as independent binomial random variables with parameters \( n_i, P(x_i), i = 1, ..., m \), respectively. Here \( y_i \) is the response variable "PD" for probability of default, with 0 representing non-default and 1 representing default. Note one could run a multinomial logistic regression model with more than two default classifications (e.g. default only, non-default only and a third: may or may not default). The SAS LOGISTIC Procedure with Fisher’s scoring optimization technique is used to obtain the nonlinear regression analysis for this model. Number of observations used in this simulated data is 10,000. The variable PD observes 132 defaults and 9868 non-defaults. The convergence criterion (GCONV=10\(^{-8}\)) is used consistently throughout by the SAS procedure, which is satisfied for this model fitting.

The SAS procedure first checks if the intercept is significant in the model, where

\[ a = -2 \log L = 1404.720. \quad (4.1) \]

Quantity \( a \) should ideally be as small as possible as a measure of goodness of fit. As shown in the table below, a Chi-squared test of the residual indicates that the intercept is indeed significant, with \( p\)-value \(< 0.0001\).

<table>
<thead>
<tr>
<th>Residual Chi-Square Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
</tr>
<tr>
<td>80.2819</td>
</tr>
</tbody>
</table>

The logistic procedure in SAS can also be used to check if the covariate \( x \) is significant in the model. As shown in the table below, introducing covariates \( x \) does lead to improvement in the model performance, as demonstrated by a decrease in each statistical criterion tabulated.

<table>
<thead>
<tr>
<th>Model Fitness Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
</tr>
<tr>
<td>AIC</td>
</tr>
<tr>
<td>SC</td>
</tr>
<tr>
<td>(-2 \log L)</td>
</tr>
</tbody>
</table>
The table below demonstrates significance of the slope term.

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>85.5715</td>
<td>1</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Score</td>
<td>80.2819</td>
<td>1</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Wald</td>
<td>69.7592</td>
<td>1</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

The maximum likelihood estimates of the intercept and slope, tabulated below, are somewhat surprising. Each was set at -4 to generate the data, whereas the estimates are quite different and presumably outside of any reasonable confidence interval.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-3.1171</td>
<td>0.1370</td>
<td>517.8136</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>-3.1187</td>
<td>0.3734</td>
<td>69.7592</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Despite the comments on the table above, the table below demonstrates through a high percentage concordance that the model is a good fit.

<table>
<thead>
<tr>
<th>Association of Predicted Probabilities and Observed Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Concordant</td>
</tr>
<tr>
<td>Percent Discordant</td>
</tr>
<tr>
<td>Percent Tied</td>
</tr>
<tr>
<td>Pairs</td>
</tr>
</tbody>
</table>

The above odds ratio indicates that the odd of defaulting is 4.4% of the odds of not defaulting. Since the confidence interval does not contain 1 the odds ratios are significantly different from one another at 5% level of significance. The covariate \( x \) in the simulated data was Uniform on \([0, 1]\). Since the covariate \( x \) is a continuous random variable, the Hosmer and Lemeshow goodness-of-fit test is the most appropriate one to use. In this context, the null hypothesis is that the logistic model is the correct model versus the alternate hypothesis that the logistic model is not appropriate, based on the following table:
Partition for the Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th>Group</th>
<th>Total</th>
<th>dp = 1</th>
<th>dp = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Observed</td>
<td>Expected</td>
</tr>
<tr>
<td>1</td>
<td>1000</td>
<td>2</td>
<td>2.30</td>
</tr>
<tr>
<td>2</td>
<td>1001</td>
<td>3</td>
<td>3.13</td>
</tr>
<tr>
<td>3</td>
<td>1000</td>
<td>7</td>
<td>4.28</td>
</tr>
<tr>
<td>4</td>
<td>1001</td>
<td>7</td>
<td>5.79</td>
</tr>
<tr>
<td>5</td>
<td>1001</td>
<td>5</td>
<td>7.90</td>
</tr>
<tr>
<td>6</td>
<td>1000</td>
<td>8</td>
<td>10.75</td>
</tr>
<tr>
<td>7</td>
<td>1000</td>
<td>17</td>
<td>14.53</td>
</tr>
<tr>
<td>8</td>
<td>1000</td>
<td>17</td>
<td>19.77</td>
</tr>
<tr>
<td>9</td>
<td>1000</td>
<td>33</td>
<td>26.95</td>
</tr>
<tr>
<td>10</td>
<td>997</td>
<td>33</td>
<td>36.61</td>
</tr>
</tbody>
</table>

One gets a **Chi-Squared value** of 6.4121 with 8 degrees of freedom, which corresponds to the smallest type I error of wrongly rejecting the null hypothesis to be 60.12%. Hence the null hypothesis is not rejected and the logistic model could be a good fit as there is no strong evidence against it. In this modeling setting one could, for different probability (Prob) levels of definition of default, come up with the following classification table, where Sensitivity represents the percentage calculated among those who actually defaulted and for whom the model predicted default, and Specificity represents the percentage computed among those who actually did not default and for whom the model predicted no default.

<table>
<thead>
<tr>
<th>Prob</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Percentages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Event</td>
<td>Non-Event</td>
<td>Event</td>
</tr>
<tr>
<td>0.000</td>
<td>132</td>
<td>0</td>
<td>9868</td>
</tr>
<tr>
<td>0.010</td>
<td>107</td>
<td>5249</td>
<td>4619</td>
</tr>
<tr>
<td>0.020</td>
<td>74</td>
<td>7508</td>
<td>2360</td>
</tr>
<tr>
<td>0.030</td>
<td>39</td>
<td>8782</td>
<td>1086</td>
</tr>
<tr>
<td>0.040</td>
<td>5</td>
<td>9650</td>
<td>218</td>
</tr>
<tr>
<td>0.050</td>
<td>0</td>
<td>9868</td>
<td>0</td>
</tr>
<tr>
<td>0.060</td>
<td>0</td>
<td>9868</td>
<td>0</td>
</tr>
<tr>
<td>0.070</td>
<td>0</td>
<td>9868</td>
<td>0</td>
</tr>
<tr>
<td>0.080</td>
<td>0</td>
<td>9868</td>
<td>0</td>
</tr>
<tr>
<td>0.090</td>
<td>0</td>
<td>9868</td>
<td>0</td>
</tr>
<tr>
<td>0.100</td>
<td>0</td>
<td>9868</td>
<td>0</td>
</tr>
</tbody>
</table>

The corresponding receiver operating characteristic (ROC) is plotted in Fig. 7 to find the optimal probability level of cutoff, which is close to 0.02 level as is obvious from the above table as well.
Note that another link function other than logistic could be used to come up with other models. Some similar models that could easily be computed in this context are the generalized linear models with complementary log-log link or the probit link where \( y_1, \ldots, y_m \) are the observed values of independent binomial random variables. Also area under curve (AUC) from the information available in the graph could easily be computed.

5. **Accounting for Correlations in Goodness-of-Fit**

In the case of **logistic regression models** there are three types of Chi-squared tests used to evaluate the model. Pearson’s Chi-squared test is most suitable when all the covariates or the predictor variables are (I) **categorical or discrete** in nature. Pearson’s Chi-squared test does not correctly follow a Chi-squared distribution when continuous covariates are present. Alternately, Pulkstenis and Robinson (P-R) (2004) introduced a goodness-fit-model that is most powerful when (II) **both discrete and continuous** covariates are present, which is described for the sake of completeness with verbatim quotes from P-R. These authors also provided a SAS macro to implement both P-R tests, posted online at StatLib [2]. Two-level subgrouping within each covariate pattern of Table I is based on assigned ordinal scores for the observations within that covariate pattern. Ordinal scores are defined [3] by score \( \hat{\mu}_1 + 2\hat{\mu}_2 + \ldots + J\hat{\mu}_J \), where \( \hat{\mu}_j \) denotes the model-based \( Pr(Y = j) \). Covariate patterns are determined by the observed cross-classification of all levels of the categorical explanatory
variables in the model. Then, within each covariate pattern, data are sorted by assigned ordinal score and each covariate pattern is essentially split in two based on the median of the estimated scores within each of the I rows of Table I.

Table I. Data classification of observed counts for Pearson chi-square and deviance statistics.

<table>
<thead>
<tr>
<th>Covariate pattern</th>
<th>$Y = 1$</th>
<th>$Y = 2$</th>
<th>$\ldots$</th>
<th>$Y = J$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$O_{11}$</td>
<td>$O_{12}$</td>
<td>$\ldots$</td>
<td>$O_{1J}$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$O_{21}$</td>
<td>$O_{22}$</td>
<td>$\ldots$</td>
<td>$O_{2J}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$x_I$</td>
<td>$O_{I1}$</td>
<td>$O_{I2}$</td>
<td>$\ldots$</td>
<td>$O_{IJ}$</td>
</tr>
</tbody>
</table>

When the median of the estimated scores takes on an actual estimated score, one can arbitrarily place it in the first group of lower scores. Since the data include continuous covariates, ties among estimated scores should be rare. When they occur, tied observations may be grouped together into the first subgroup. This additional stratification (based on the medians), doubles the number of covariate patterns to incorporate information related to all continuous covariates in the model. Model-based expected counts are computed in an analogous fashion to the traditional Pearson chi-square and deviance statistics, and the proposed tests statistics are given by

$$
\chi^*^2 = \sum_{i=1}^{I} \sum_{h=1}^{2I} \sum_{j=1}^{J} \frac{(O_{ihj} - E_{ihj})^2}{E_{ihj}}, \\
D^*^2 = 2 \sum_{i=1}^{I} \sum_{h=1}^{2I} \sum_{j=1}^{J} O_{ihj} \log \frac{O_{ihj}}{E_{ihj}},
$$

where $i$ indexes covariate patterns, $h$ indexes the sub-stratification due to ordering by the ordinal scores, and $j$ indexes response categories. The degrees of freedom for these statistics are obtained by modifying the degrees of freedom for the traditional Pearson and deviance chi-square statistics given by $(I - 1)(J - 1) - k$, where $I$ is the number of rows in the cross-classification of Table I and $k$ is the number of covariates in the model. The degrees of freedom for $\chi^*^2$ and $D^*^2$ are given by $(2I - 1)(J - 1) - k - 1$. In this setting, $2I$ reflects the doubling of covariate patterns due to the contribution of continuous variables, $k$ is the number
of categorical terms in the model, and we subtract one additional degree of freedom due to
the contribution of continuous variables. This would be analogous to the regular deviance or
the Pearson chi-square test applied to a model with one continuous variable dichotomized at
its median and included as an additional indicator variable. In addition, the deviance form
of the test statistic is not defined if one of the observed cell counts equals zero. We adopt
the convention of assigning a contribution of zero for such cells. This adjustment is conservative
in that it produces a very slight reduction in power. Finally, if the covariates are all (III) continuous then the Hosmer-Lemeshow test should be used to achieve the maximum power
of the goodness-of-fit test as suggested by Myers et al. [4].

5.1 Taking Correlated Observations into Account with Repeated Time

Consider a data set consisting of $N$ subjects. For each subject $i (1 \leq i \leq N)$ $n_i$ observations
are available, i.e. repeated measures over time where typically $n_i = 10$ years. Given
the vector $b_i$ with the random effects for subject $i$, the repeated measurements $Y_{i1}, \ldots, Y_{in_i}$ are
assumed to be dependent. These dependencies will be captured in the correlation matrix
of $Y_{i1}, \ldots, Y_{in_i}$ under the generalized linear mixed model (GLMM) framework. Choices of
this correlation matrix would reflect the repeated measure over time (say 10 years) of the
individual companies and the various dependences between companies due to sharing the
same economic environment. For example, Stationary 10-dependent correlation structure
could be used as described in Kleinbaum et al. [5].

The generalized linear mixed model (GLMM) framework could be used to develop these
models. To give a flavor of GLIMM one can consider the following: Given the vector $b_i$ with
the random effects for subject $i$, the repeated measurements $Y_{i1}, \ldots, Y_{in_i}$ are assumed to be independent. Using notation available in the literature [6] we define the GLIMM model in
this special case where the densities of $Y_{ij}$ follow the exponential family

$$f(y_{ij}|b_i, \beta, \phi) = \exp\left(\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right), \quad j = 1, \ldots, n_i. \quad (5.3)$$

The following (conditional) relations hold:

$$\mu_{ij} = E[Y_{ij}|b_i] = \psi'(\theta_{ij}) \quad \text{and} \quad \text{Var}[Y_{ij}|b_i] = \phi \psi''(\theta_{ij}) = \phi V(\mu_{ij}), \quad (5.4)$$

where the GLIMM is described by

$$g(\mu_{ij}) = x_{ij}^T\beta + z_{ij}^T b_i. \quad (5.5)$$

g(.) is called the link and $V(.)$ the variance function. $\beta (p \times 1)$ denotes the fixed effects
parameter vector and $b_i (q \times 1)$ the random effects vector. $x_{ij}(p \times 1)$ and $z_{ij}(q \times 1)$ contain
the subject’s covariate information for the fixed and random effects, respectively.
6. Conclusions and Recommendations

The conclusions of the Standard & Poor’s working group from the 2014 Mathematical Problems in Industry Workshop are presented below in the form of a list of recommendations. The working group recommends:

1. **Accounting for how the data set influences the ability of standard performance measures to accurately rate models of company default likelihood.** Statistical measures to compare historic data on company defaults with model predictions based on likelihood or goodness of fit, or graphical measures such as the area under curve (AUC) or accuracy ratio (AR), are sensitive to the distributions of defaulters and non-defaulters that are available in the data. In particular, if these distributions are well separated, all models look good and if they overlap significantly, all models look bad. The working group proposes quantifying this inherent observability of the data using the Kullback-Leibler divergence (KL) to measure an effective distance between the sample density conditioned on default and the sample density conditioned on non-default. To quantify each data sample the working group used both exact analytical expressions for its KL and a sample-based method that does not require computation of the underlying distributions. The KL distance can be used in one of the following ways:
   - to provide a separate statistic expressing confidence level in the model performance measure used (e.g., AR);
   - to provide an adjusted model performance measure that appropriately discounts AR when KL is large, indicating very observable data, or augments AR when KL is small, indicating indistinguishable data; or
   - to use the KL of the data set to construct an interval, e.g., $[AR - \epsilon, AR]$ with $\epsilon = \frac{1}{\kappa KL + 1} > 0$ (high observability) or $[AR, AR - \epsilon]$ with $\epsilon < 0$ (low observability), that replaces a single model performance rating with a range of performance ratings in which the true model performance lies. Analysis of multiple data sets would produce many intervals, the intersection of which could be taken as the final interval expressing model quality. Requiring that such an intersection be nonempty is one possible criterion for determining the empirical tuning parameter $\kappa$.

2. **Accounting for correlations in default probabilities.** Using regression to a logistic model with artificial default data, the working group showed that inclusion of covariates significantly improved agreement between model output and the data.

3. **Improving goodness-of-fit measure for regression models when discrete and continuous covariates are present.** A standard statistical test used to accept or reject a proposed discrete predictor variables regression model as appropriate to a given
data set is the scaled sum of squared differences between data and model output, known as the Pearson Chi-squared test. If the regression model contains only continuously varying explanatory variables, a related test known as the Hosmer-Lemeshow test is appropriate for model evaluation. Neither of these tests is appropriate to mixed discrete and continuous explanatory variables for regression evaluation, however, and the working group consequently proposes use of a recently developed mixed-parameter test developed by Pulkstenis and Robinson.

7. Acknowledgments

The working group would like to thank Standard and Poor’s Rating Services for supporting the Mathematical Problems in Industry Workshop and the work documented in this report. This material is based upon work supported by the National Science Foundation under Grant No. 1153954. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References


