Validation techniques for PD and LGD in credit risk modeling

Free talks

Orphée Van Essche

Altaïr

22 avril 2025

Outline

Introduction

Probability of Default (PD)
- Statistical Background
- Discriminatory Power
- Calibration Accuracy

Loss Given Default (LGD)
- Statistical Background
- Discriminatory Power
- Calibration Accuracy

References

Outline

Introduction

Basel II introduced significant changes in credit risk modeling, putting a strong emphasis on the validation of internal models used to compute regulatory capital, such as Probability of Default (PD) and Loss Given Default (LGD) models.

Central to this framework is the use of **quantitative validation techniques**, ensuring that models not only fit historical data but also generalize well to new, unseen observations. The use of such techniques is also known as **backtesting**.

This presentation outlines key validation techniques for PD and LGD models commonly used in practice. The structure is as follows :

- We begin with a **theoretical framework** for rating systems, especially used in the context of PD modeling, and explore the importance of **monotonicity** in PD estimates and its connection to optimal classification.
- We then discuss tools to assess **discriminatory power**—how well a model distinguishes between different levels of credit risk—and the **calibration accuracy** of model outputs to observed outcomes, first in a PD validation context.
- Finally, we extend some of the tools presented in the context of PD to LGD validation and present new ones.

Evolution of Credit Risk Assessment and the Basel Accord

Traditional Distinctions : Rating vs. Scoring

	Rating	**Scoring**
Calculated by	Rating agencies	Credit institutions
Based on	Expert-based approach	Statistical models
Used for	Pricing of corporate bonds	Credit approval decisions

Modern Credit Risk Management : A Unified View

- Convergence of rating and scoring methodologies.
- Both systems now primarily determine Probability of Default (PD).
- Used for :
  - Credit approval and pricing ;
  - Regulatory and internal capital allocation.

A Word on the Basel Accords

- Basel Committee on Banking Supervision (BCBS) established in 1975 by G10 central banks to coordinate global banking supervision.
- With main objective to strengthen global banking stability through consistent, risk-based capital standards - known as the Basel Accords.

Regulatory Evolution : Basel I (1998) → Basel II (2004)

Basel I : Uniform Capital Requirements

Fixed 8% capital requirement for borrowers, regardless of creditworthiness.

Basel II : Risk-Sensitive Capital Requirements

- Introduction of risk-based approaches using ratings and scoring.
- Two main approaches :

1. **Standardized Approach (SA)**
   - Capital requirements based on **external ratings** (from rating agencies).
2. **Internal Ratings-Based (IRB) Approach**
   - PDs derived from **internal ratings** (developed by credit institutions).
   - Capital requirements based on :
     - Probability of Default (PD), under A-IRB and F-IRB ;
     - Loss Given Default (LGD), under A-IRB only ;
     - Exposure at Default (EAD), under A-IRB only.

Types of Internal Rating Systems

- **Expert-Based :** Rating grades assigned based on qualitative criteria.
- **Statistical Scoring :** PDs mapped from continuous/discrete score variables into at least seven non-default rating grades.
- **Hybrid Models :** Statistical models combined with expert adjustments.

Basel III (2017) : Key Implications for IRB Approach

- **Output Floor :** IRB capital requirements must be at least **72.5%** of Standardized Approach (SA) levels.
- **Restrictions on IRB Use :** Certain exposures (e.g., large corporates, banks) can no longer use IRB models.
- **Parameter Floors :**
  - Minimum PD of 0.05% for low-risk exposures.
  - Constraints on LGD and EAD to reduce variability.
- **Stronger Model Governance :** Increased regulatory scrutiny on risk parameter estimation and validation.

Validation of PD & LGD under Basel II/III

Regulatory Requirements (CRR Art. 185 (a))

1. Banks must have a robust system in place to validate the accuracy and consistency of the *estimation of all relevant risk components*.
2. The internal validation process must allow supervisors to meaningfully assess the *performance of internal rating* and risk estimation systems.

Key Aspects of Quantitative Validation

- **Calibration Accuracy (Risk Quantification) :** Ensures that estimated PDs correctly reflect the "true" risk levels of rating grades (linked to requirement 1).
- **Discriminatory Power (Risk Differentiation) :** Evaluates whether the rating system correctly ranks borrowers from "good" to "bad" (linked to requirement 2).

Backtesting Requirement (CRR Art. 185 (b))

- Banks must regularly compare realized default rates with estimated PDs for each grade and demonstrate that realized default rates remain within expected ranges.
- Banks using own estimates of LGDs and conversion factors shall also perform analogous analysis for these estimates.

Outline

Outline

Basic Setting

Definition (Score Variable)

Let S be a *score* or *rating* assigned to a borrower based on observable covariates X associated with the borrower's creditworthiness, i.e. risk factors :

$$S = f(X),$$

where the function f is usually estimated using regression or other methods.

Mathematical Framework

Each borrower is associated with two random variables :

- **Score Variable** S : A continuous value (discrete rating grades can be treated similarly) that reflects the borrower's creditworthiness.
- **State Variable** Y : Represents the borrower's financial state at the end of a fixed period T (usually one year), which can take two values :

$$Y = \begin{cases} D, & \text{if borrower defaults by } T \\ N, & \text{otherwise} \end{cases}$$

Since Y is unobservable at time t, it is a *latent variable*.

The institution uses S to predict Y.

Joint Distribution of (S, Y)

Using Conditional Densities

The joint statistical distribution of the continuous variable S (score) and the dichotomous variable Y (state) can be described using :

1. The *marginal distribution* of Y given by :

$$P[Y = y], \quad y = D, N$$

   where $P[Y = D] = 1 - P[Y = N] = p$ represents the *unconditional* PD.

2. The *conditional distribution* of S given Y, which is given by :

$$F_y(s) = P[S \leq s \mid Y = y] = \int_{-\infty}^{s} f_y(u)\, du, \quad y = D, N$$

   where f_D and f_N represent the *conditional densities* of S.

Key Relationship

We can derive the following expression for the conditional PD given the score :

$$\pi(s) := P[Y = D \mid S = s] = \frac{p\, f_D(s)}{p\, f_D(s) + (1 - p) f_N(s)}$$

Key Insights on Densities and Conditional PDs

Information Carried by Densities

- If the score variable S carries no information, the densities will be identical :

$$S \perp\!\!\!\perp Y \iff f_D(s) = f_N(s) \quad \forall s \in \mathbb{R} \iff \pi(s) = p \quad \forall s \in \mathbb{R}$$

  The conditional PD given the score is the unconditional PD p.
- Conversely, if S perfectly discriminate between good and bad in the sense that the conditional densities f_D and f_N have disjoint supports :

$$\{s : f_D(s) > 0\} \cap \{s : f_N(s) > 0\} = \emptyset \Leftrightarrow \pi(s) = \begin{cases} 1 & \text{if} \quad s \in \{s : f_D(s) > 0\} \\ 0 & \text{if} \quad s \in \{s : f_N(s) > 0\} \end{cases}$$

  The conditional PD given the score is the default indicator $\mathbb{1}_D$.

Optimality of Conditional PD

- The conditional probability of default $\pi(S)$ is the best approximation (in the least-squares sense) of the state variable Y given the score variable S.
- To see that, let $D = 1$ and $N = 0$ such that $Y = \mathbb{1}_D$. In this case :

$$\pi(S) = E[Y \mid S] = \text{argmin}_{f(S)} E[(Y - f(S))^2]$$

Monotonicity of Conditional PDs (1/2)

Reminder from Neyman-Pearson Lemma

Let a random variable $X \sim P_\theta$ with density f_θ, where $\theta \in \Theta = \{\theta_0, \theta_1\}$. Consider the hypothesis testing problem :

$$\begin{cases} H_0 : & f_\theta = f_{\theta_0} \\ H_1 : & f_\theta = f_{\theta_1} \end{cases}$$

For a fixed significance level $\alpha \in (0,1)$, the test ϕ_α^* defined as :

$$\phi_\alpha^*(x) = \mathbb{1}\left(\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > q_\alpha\right),$$

where q_α satisfies $P_{\theta_0}\left[f_{\theta_1}(X)/f_{\theta_0}(X) \leq q_\alpha\right] = 1 - \alpha$, is the uniformly most powerful (UMP) test in the class of α-level tests, i.e. test satisfying $E_{\theta_0}[\phi_\alpha] \leq \alpha$.

Application to the Scoring Problem

By setting $f_{\theta_0} = f_D$, $f_{\theta_1} = f_N$, and $X = S$ (the score variable), the test simplifies to :

$$\phi_\alpha^*(s) = \mathbb{1}\left(\frac{f_N(s)}{f_D(s)} > q_\alpha\right),$$

where q_α is determined such that $P[f_N(S)/f_D(S) \leq q_\alpha \mid Y = D] = 1 - \alpha$.

Monotonicity of Conditional PDs (2/2)

Cut-off Decision Rules

If the likelihood ratio $s \mapsto \frac{f_N}{f_D}(s)$ is monotonically increasing, the test can be expressed as :

$$\phi_\alpha^*(s) = \mathbb{1}\left(s > r_\alpha\right),$$

where r_α satisfies $P\left[S \leq r_\alpha \mid Y = D\right] = 1 - \alpha$.

Now consider a realization (s, y) of (S, Y). The score s is observed, but the state y (either N or D) is unobserved. From the Neyman-Pearson lemma, there exists a half-line shape $R = (r_\alpha, \infty)$ such that the following *classification rule* is optimal :

$$\begin{cases} s \in R & \implies y = N, \\ s \notin R & \implies y = D. \end{cases}$$

Conclusion

This implies that the conditional PD is monotonically decreasing, as shown by the relation :

$$\pi(s) = \frac{p\, f_D(s)}{p\, f_D(s) + (1-p) f_N(s)}.$$

Thus, if the conditional PD as a function of s is observed to be non-monotonic, it casts doubt on the reliability of S in discriminating between good and bad borrowers.

Outline

How Can Discriminatory Power Be Measured and Tested ?

Definition (Discriminatory Power)

Discriminatory power can be described as the ability of a score or rating S to distinguish between defaulters and non-defaulters. Technically, it is a measure of :

- The discrepancy between the conditional densities f_D and f_N ; or equivalently
- The variation of the conditional PD, $\pi(S)$, with values as close as possible to 100% for defaulters and 0% for non-defaulters.

Commonly Used Tools

Two widely used tools to measure the discriminatory power of PD models are :

- **Cumulative Accuracy Profile** and related **Accuracy Ratio**
- **Receiver Operating Characteristic** and related **Area Under the Curve**

Remark

Tools can be categorized based on whether they require the estimation of p, the unconditional PD :

- p **involved :** These tools can only be applied to samples with the correct proportion of defaulters.
- p **not involved :** These tools can be applied even to non-representative samples.

Cumulative Accuracy Profile (CAP) : Concept and Definition

Definition (CAP function)

The *CAP function* is defined as :

$$CAP(u) = F_D(F^{-1}(u)), \quad u \in (0,1),$$

where :

- $F_D(s)$ is called *hit rate* and indicates the proportion of defaulters regarded as suspect of default when s is used as threshold.
- $F(s) = (1-p)F_N(s) + p\, F_D(s)$ is called *alarm rate* and represents the proportion of the entire population detected by the same threshold.

Two Extreme Situations

- **Random Score :** The score carries no information on the default state :

$$S \perp\!\!\!\perp Y \Leftrightarrow F_D(s) = F(s) \quad \forall s \in \mathbb{R} \Leftrightarrow CAP(u) = F(F^{-1}(u)) = u \quad \forall u \in (0,1)$$

- **Perfect Score :** The score perfectly disciminate between good and bad :

$$F(s) = p\, F_D(s)\; \forall s \in \{s : f_D(s) > 0\} \Leftrightarrow CAP(u) = \frac{F(F^{-1}(u))}{p} = \frac{u}{p}\; \forall u \in (0,p]$$

$$F_D(s) = 1\; \forall s \in \{s : f_N(s) > 0\} \Leftrightarrow CAP(u) = 1\; \forall u \in (p,1)$$

Cumulative Accuracy Profile (CAP) : Graphical Representation

Graphical Representation

The CAP function can be visualized using one of the following methods :

- Plotting all the points $(u, CAP(u))$, $u \in (0, 1)$; or
- Plotting all the points $(F(s), F_D(s))$, $s \in \mathbb{R}$.

Interpretation

$100 \times CAP(u)\%$ corresponds to the percentage of defaulters detected among the first $100 \times u\%$ of all borrowers (sorted by score).

Example on real data from two rating systems

Rating	Rating System 1			Rating System 2		
	Non-default	Default	LR	Non-default	Default	LR
AA	200	2	5.26	145	2	3.82
A	215	5	2.26	215	4	2.83
BBB	185	2	4.87	210	5	2.21
BB	200	14	0.75	200	11	0.96
B	150	27	0.29	180	28	0.34

Table 1 – Ratings and corresponding number of (non-)defaults for two rating systems.

Examples of CAP curves

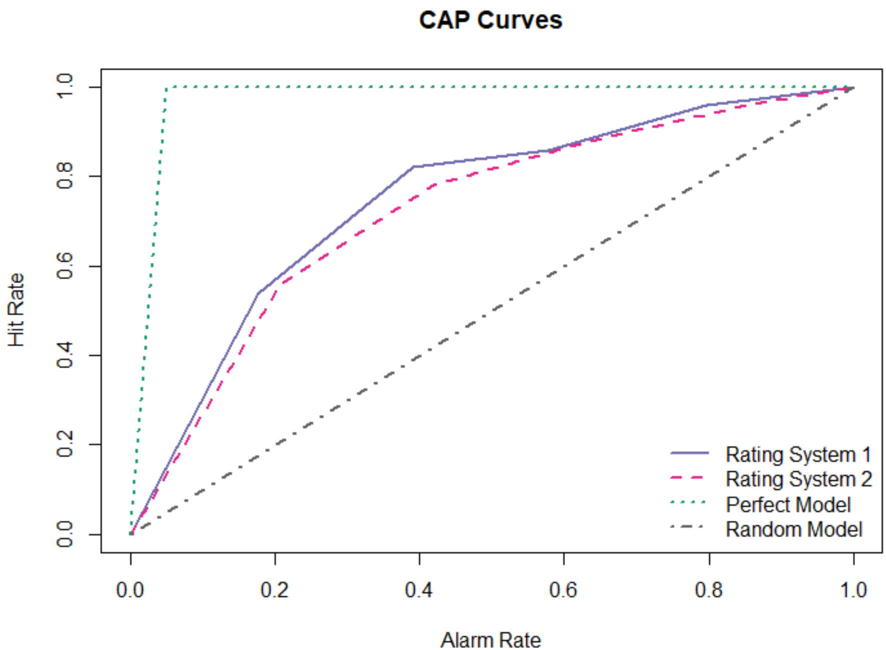


Figure 1 – CAP curves for two rating systems (cf. data in Table 1).

Cumulative Accuracy Profile (CAP) : Interpreting the Slope

Slope of the CAP Curve

It can be shown that :

$$CAP'(u) = \frac{P[Y = D \mid S = F^{-1}(u)]}{p} = \frac{\pi(F^{-1}(u))}{p}.$$

Coming back to the two extreme situations :

- **Random Score :**

$$CAP'(u) = 1 \quad \forall u \in (0,1) \iff \pi(s) = p \quad \forall s \in \mathbb{R}.$$

- **Perfect Score :**

$$CAP'(u) = \begin{cases} 1/p & \text{if } u \in (0,p] \\ 0 & \text{if } u \in (p,1) \end{cases} \iff \pi(s) = \begin{cases} 1 & \text{if } s \leq F^{-1}(p) \\ 0 & \text{if } s > F^{-1}(p) \end{cases}$$

Interpretation :

- Strong growth of $CAP(u)$ for $u \approx 0 \Rightarrow \pi(s) \approx 1$ for low scores s.
- Weak growth of $CAP(u)$ for $u \approx 1 \Rightarrow \pi(s) \approx 0$ for high scores s.

Accuracy Ratio (AR) as a Measure of Discriminatory Power

Definition (Accuracy Ratio (AR))

The *Accuracy Ratio* (AR) or *Gini-Coefficient* is defined as the ratio of the area between the CAP curve and the diagonal to the area between the perfect CAP curve and the diagonal :

$$AR = \frac{\int_0^1 CAP(u)\,du - 1/2}{1/2 - p/2} = \frac{2\int_0^1 CAP(u)\,du - 1}{1 - p} \in [-1, 1].$$

- The closer the AR is to 0, the lower the discriminatory power of the score.
- The closer the AR is to 1, the higher the discriminatory power of the score.

Alternative Definition

Alternatively, it can be shown that the AR can be expressed as the difference between two probabilities :

$$AR = P[S_D < S_N] - P[S_D > S_N],$$

where $S_N \sim F_N$, $S_D \sim F_D$ and $S_N \perp\!\!\!\perp S_D$.

Receiver Operating Characteristic (ROC) : Concept and Definition

Definition (ROC Function)

The equation of the *ROC function* is given by :

$$ROC(u) = F_D(F_N^{-1}(u)), \quad u \in (0, 1),$$

where $F_N(s)$ is called the *false alarm rate* and represents the proportion of the non-defaulters that will be regarded as defaulters when s is the threshold.

Note : Unlike CAP curves, constructing ROC curves does not require the unconditional PD p.

Slope of the ROC Curve

It can be shown that :

$$ROC'(u) = \frac{f_D(F_N^{-1}(u))}{f_N(F_N^{-1}(u))}, \quad u \in (0, 1).$$

Since a score variable is optimal if its (inverse) likelihood ratio is monotonically decreasing, the ROC curve must be concave given the form of its derivative.

Interpretation :

- Strong growth of $ROC(u)$ for $u \approx 0 \Rightarrow$ high $f_D(s)$ and low $f_N(s)$ for low s.
- Weak growth of $ROC(u)$ for $u \approx 0 \Rightarrow$ low $f_D(s)$ and high $f_N(s)$ for high s.

Receiver Operating Characteristic (ROC) : Graphical Representation

Graphical Representation

The ROC function can be visualized using one of the following methods :

- Plotting all the points $(u, ROC(u))$, $u \in (0, 1)$; or
- Plotting all the points $(F_N(s), F_D(s))$, $s \in \mathbb{R}$.

Interpretations

- $100 \times ROC(u)\%$ is the percentage of defaulters that have been assigned a lower score than the highest score of the first $100 \times u\%$ non-defaulters.
- Points on the curve can also be seen as pairs of type I error and power that can arise when cut-off s is applied for testing H_0 non-default vs H_1 default.

Example on real data from two rating system

The same data as from Table 1 are used to construct the ROC curves in Figure 2.

Examples of ROC curves

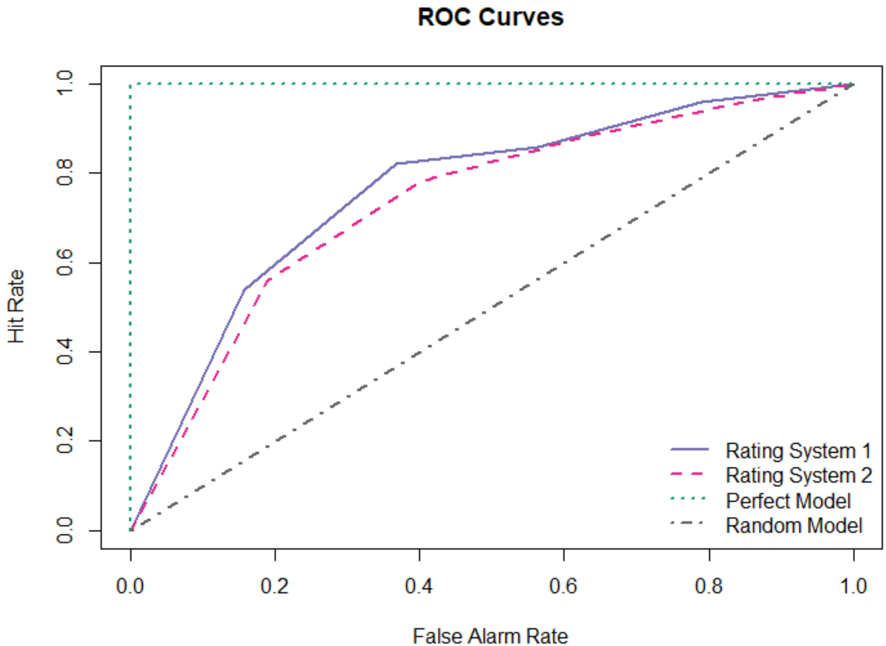


Figure 2 – ROC curves for two rating systems (cf. data in Table 1).

Area Under the Curve (AUC) : A Measure of Discriminatory Power

Definition (AUC)

The *Area Under the Curve* (AUC) is defined as the area under the ROC curve :

$$AUC = \int_0^1 ROC(u)\, du \in [0, 1].$$

- An AUC close to 1 indicates high discriminatory power.
- An AUC close to 0.5 indicates low discriminatory power.

Alternative Definition

The AUC can also be expressed as a probability :

$$AUC = P[S_D < S_N], \tag{1}$$

where $S_N \sim F_N$, $S_D \sim F_D$, and $S_N \perp\!\!\!\perp S_D$.

Relation Between AUC and AR

The AUC is linearly related to the AR, as it can be shown that :

$$AUC = \frac{AR + 1}{2}.$$

Area Under the Curve (AUC) : Connections to Other Metrics

Connection to the Kolmogorov-Smirnov Statistic

If the likelihood ratio f_D/f_N is monotonic, the maximum distance between the ROC curve and the diagonal can be expressed as an affine transformation of :

$$\max_s |F_D(s) - F_N(s)| = \frac{1}{2}\int_{-\infty}^{\infty} |f_D(s) - f_N(s)|\, ds \in [0, 1/2],$$

which represents the population equivalent of the *Kolmogorov-Smirnov (KS)* statistic, often used to test whether the distributions F_D and F_N are identical.

Connection to the Information Value (IV)

The *Information Value (IV)* is another measure of discriminatory power that quantifies the discrepancy between distributions using relative entropy. It is defined as :

$$IV = \int_{-\infty}^{\infty} (f_D(s) - f_N(s)) \ln \frac{f_D(s)}{f_N(s)}\, ds \in [0, \infty),$$

which can be interpreted as a weighted version of the (population) KS statistic, where the weights are given by the *weights of evidence*.

Testing for Discriminative Power

An important consequence of the representation of AUC as a probability in (1) is that the non-parametric *Mann-Whitney U test* (or *Wilcoxon rank sum test*) can be used to assess whether there is any discriminatory power in the model.

Mann-Whitney U Test

Assume that S_D has the same distribution as S_N shifted by an (unknown) amount θ, i.e. $F_D(s) = F_N(s-\theta)$. The hypothesis testing problem is then :

$$\begin{cases} H_0 : \quad \theta = 0 \iff F_D(s) = F_N(s) \quad \forall s \in \mathbb{R} \iff P[S_D < S_N] = \frac{1}{2}, \\ H_1 : \quad \theta \neq 0. \end{cases}$$

A test statistic for this problem is derived from the Mann-Whitney statistic U :

$$U' = \frac{U}{M(m-M)} = \hat{\mathrm{AUC}}, \quad U = M(m-M) + \frac{M(M+1)}{2} - W,$$

where m is the total number of borrowers, M is the number of borrowers who defaulted, and W is the rank sum for the defaulters. U' is an estimator for AUC.

Normal Approximation : When m and M are large, we can use a normal approximation to derive an asymptotic α-level test :

$$\phi_\alpha = \mathbb{1}\left(\left|\frac{U' - 1/2}{\sqrt{(m+1)/(12M(m-M))}}\right| > z_{\alpha/2}\right), \quad z_{\alpha/2} = \Phi^{-1}(1-\alpha/2).$$

Testing for the Difference of Two AUCs

When we have two rating systems providing two different AUCs, the question arises whether this difference is significant from a statistical point of view. This question can be answered using a test developed by DeLong.

DeLong Test

The hypothesis testing problem of the DeLong test on the difference of two AUCs can be formulated as follows :

$$\begin{cases} H_0 : & AUC_1 = AUC_2 \iff P[S_D^1 < S_N^1] = P[S_D^2 < S_N^2], \\ H_1 : & AUC_1 \neq AUC_2, \end{cases}$$

where S_y^1 and S_y^2 are the score variables generated by rating system 1 and 2, respectively, for $y \in \{D, N\}$. The test statistic for this problem derived by DeLong is defined as :

$$T = \frac{(U_1' - U_2')^2}{\hat{\sigma}^2_{U_1'} + \hat{\sigma}^2_{U_2'} - 2\hat{\sigma}_{U_1', U_2'}},$$

where $\hat{\sigma}^2_{U_1'}$, $\hat{\sigma}^2_{U_2'}$ and $\hat{\sigma}_{U_1', U_2'}$ are complicated functions of the observations. Under the null hypothesis, T is asymptotically χ^2_1-distributed such that an asymptotic α-level test is given by $\phi_\alpha = \mathbb{1}\left(T > q_\alpha\right)$, where q_α is $1 - \alpha$-quantile of the χ^2_1 distribution.

AUC Estimation and Comparison for two Rating Sytems

Applying the Mann-Whitney U test and the DeLong test to the data in Table 1 yields the following results.

Rating System	AUC	SE	95% CI LB	95% CI UB	p-value
1	0.7616	0.0336	0.6958	0.8274	1.84×10^{-10}
2	0.7354	0.0351	0.6665	0.8042	9.57×10^{-9}

Table 2 – Estimated AUC values from the two rating systems and the results of the Mann-Whitney U test, showing that the AUCs are significantly different from 0.5.

DeLong's Test for Two Correlated ROC Curves	
Statistic	**Value**
T	2.9377
p-value	0.0033
95% CI	(0.0087, 0.0438)
AUC of Rating System 1	0.7616
AUC of Rating System 2	0.7354

Table 3 – DeLong test results, demonstrating that the difference in AUCs between the two rating systems is statistically significant. As a result, Rating System 1 is preferred.

Outline

Measuring and Testing Calibration Accuracy

Definition : Calibration Accuracy

Calibration accuracy measures how closely our estimation of the conditional PD, i.e. our prediction, aligns with the actual default rate, i.e. the realization.

PIT vs. TTC and Independence Assumptions

The independence of default events in testing depends on whether conditional PDs given a score incorporate the current economic state :

- **Point-In-Time (PIT) :** The PIT (conditional) PD depends on the current state of the economy, e.g. $\pi(S) = \pi(S_t)$ where $S_t = f(X, Z_t)$, Z_t representing macroeconomic covariates like GDP growth. Given the observed macroeconomic covariates state z_t, defaults are treated as independent :

$$Y_i | [S_{i,t} = f(x_i, z_t)] \sim \mathrm{Bin}(1, \pi(f(x_i, z_t))), \quad i = 1, \ldots, m.$$

  Tests under this assumption are called *conditional tests*.
- **Through-The-Cycle (TTC) :** The TTC (conditional) PD is constant throughout the economic cycle and does not rely on the current economic state. Defaults are no longer assumed to be independent and tests accounting for this dependence are called *unconditional tests*.

Regression Analysis and the Brier Score

Regression Framework and R^2

Let us recall that the conditional PD represents the expected value of the default indicator given the score, i.e. $\pi(S) = E[Y \mid S]$. This allows us to decompose the default indicator Y as :

$$Y = E[Y \mid S] + (Y - E[Y \mid S]) = \pi(S) + \epsilon,$$

where $\epsilon := Y - \pi(S)$ captures the unexplained variation.
In this regression setup, the *coefficient of determination* (R^2) measures how well the conditional PD explains the variance of the default state variable :

$$R^2 = \frac{Var[\pi(S)]}{Var[Y]} = \frac{Var[\pi(S)]}{p(1-p)} = 1 - \frac{E[(Y - \pi(S))^2]}{p(1-p)} \in [0, 1].$$

A high R^2 indicates that the default indicator is well explained by the conditional PD. Maximizing R^2 is equivalent to minimizing $E[(Y - \pi(S))^2]$.

Brier Score

A natural estimator of $E[(Y - \pi(S))^2]$ is the *Brier Score*, defined as :

$$\text{Brier Score} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - P[Y_i = D|S_i])^2. \tag{2}$$

Conditional Tests : The Spiegelhaler Test

Setup

Consider m borrowers with PD estimates $\hat{\pi}_i$ for the true $\pi_i := P[Y_i = D|S_i = s_i]$, for $i \in \{1, \ldots, m\}$. The Brier score in formula (2) is thus equivalent to the *Mean Squared Error* (MSE) of the PD estimates, i.e. $MSE = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{\pi}_i)^2$.

Spiegelhaler Test

We aim to test whether the estimated PDs match exactly the true PDs :

$$\begin{cases} H_0 : & \pi_i = \hat{\pi}_i \quad \forall i \in \{1, \ldots, m\} \\ H_1 : & \exists i \in \{1, \ldots, m\} \text{ such that } \pi_i \neq \hat{\pi}_i. \end{cases}$$

Under the null hypothesis, and assuming independence of default events, we have

$$E[MSE] = \frac{1}{m}\sum_{i=1}^{m}\hat{\pi}_i(1-\hat{\pi}_i) \quad \text{and} \quad Var[MSE] = \frac{1}{m^2}\sum_{i=1}^{m}\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i)^2.$$

When m is large, we can use the approximation $MSE \approx \mathcal{N}(E[MSE], Var[MSE])$. The α-level test under this approximation is :

$$\phi_\alpha = \mathbb{1}\left(\frac{MSE - E[MSE]}{\sqrt{Var[MSE]}} > z_\alpha\right), \quad z_\alpha = \Phi^{-1}(1-\alpha).$$

Conditional Tests : The Redelmeier Test

When we want to compare two rating systems calculated on the same data, e.g. with and without human expertise, we can use the *Redelmeier test* whose basic idea is to compare the MSEs of the two rating systems.

Setup

Consider m borrowers with PD estimates $\hat{\pi}_{i,j}$ for the true π_i, and associated $MSE_j = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{\pi}_{i,j})^2$ for $j \in \{1,2\}$, i.e. rating systems 1 and 2.

Redelmeier Test

We aim to test whether the two rating systems perform similarly in terms of expected MSE :

$$\begin{cases} H_0 : & E[MSE_1] = E[MSE_2] \\ H_1 : & E[MSE_1] \neq E[MSE_2]. \end{cases}$$

Under the null hypothesis, and assuming independence of defaults, the statistic

$$Z = \frac{\sum_{i=1}^{m}\left(\hat{\pi}_{i,1}^2 - \hat{\pi}_{i,2}^2 - 2(\hat{\pi}_{i,1} - \hat{\pi}_{i,2})Y_i\right)}{\sqrt{\sum_{i=1}^{m}(\hat{\pi}_{i,1} - \hat{\pi}_{i,2})^2(\hat{\pi}_{i,1} + \hat{\pi}_{i,2})(2 - \hat{\pi}_{i,1} - \hat{\pi}_{i,2})}}$$

is approximately standard normal when m is large, i.e., $Z \sim \mathcal{N}(0,1)$.
The corresponding α-level test is $\phi_\alpha = \mathbb{1}\left(Z > z_\alpha\right)$, where $z_\alpha = \Phi^{-1}(1-\alpha)$.

Conditional Tests : The Binomial Test for a Single Rating Grade

Consider the (conditional) PD for a rating grade defined by the score range $[s_L, s_U]$. Let m denote the number of borrowers assigned to this grade. If we assume independence of defaults, i.e. $S = S_t$, the number of defaults in the grade, M, is such that :

$$M = \sum_{i=1}^{m} Y_i \mid [s_U \geq S_{i,t} \geq s_L] \sim \text{Bin}(m, q), \quad q := P[Y = D \mid s_U \geq S_t \geq s_L]$$

Binomial Test

To test the accuracy of the forecasted PD $\hat{q}$ used by the bank to calculate regulatory capital, we formulate the following hypothesis test :

$$\begin{cases} H_0 : \quad q \leq \hat{q} \quad \text{(forecasted PD is conservative)} \rightarrow \text{Good} \\ H_1 : \quad q > \hat{q} \quad \text{(forecasted PD is too low)} \rightarrow \text{Bad.} \end{cases} \tag{3}$$

Normal Approximation : When m is large, we can use the approximation $M \approx \mathcal{N}(mq, mq(1-q))$. The α-level test under this approximation is :

$$\phi_\alpha = \mathbb{1}\left(\frac{M - mq}{\sqrt{mq(1-q)}} > z_\alpha\right), \quad z_\alpha = \Phi^{-1}(1-\alpha).$$

Conditional Tests : Hosmer-Lemeshow Test for Multiple Ratings

Testing multiple PDs across rating grades using the binomial test individually increases the likelihood of at least one hypothesis being erroneously rejected.

Setup

Let $q_1, \ldots, q_k$ be the (conditional) PDs for k rating grades, and m_j the number of borrowers assigned to grade $j \in \{1, \ldots, k\}$. Denote by M_j the number of defaults in grade j. The *Hosmer-Lemeshow* statistic defined as

$$H = \sum_{j=1}^{k} \frac{(m_j q_j - M_j)^2}{m_j q_j (1 - q_j)}$$

is asymptotically χ^2_k-distributed under the independence of defaults assumption.

Hosmer-Lemeshow Test

The test evaluates whether the forecasts match the true PDs across all grades :

$$\begin{cases} H_0 : & q_j = \hat{q}_j, \quad \forall j \in \{1, \ldots, k\} \\ H_1 : & \exists j \in \{1, \ldots, k\} \text{ such that } q_j \neq \hat{q}_j. \end{cases}$$

The approximate α-level test is given by $\phi_\alpha = \mathbb{1}\left(H > q_\alpha\right)$, where q_α is the $1 - \alpha$ quantile of the χ^2_k distribution.

Unconditional Tests : Normal Test

In a TTC framework, where the PD is constant across the economic cycle, defaults can no longer be regarded as independent. However, if a time series of default rates is available, assuming independence *over time* might be reasonable.

Setup

Consider a fixed rating grade with m_t borrowers at the beginning of year $t \in \{1, \dots, T\}$, of whom M_t default during the year. Assuming that defaults in different years are independent, the annual default rates M_t/m_t form an independent sequence of random variables.

Normal Test

To assess the accuracy of the forecasted (TTC) PD $\hat{q}$ as formulated in (3), we use the normal approximation $\bar{X} \approx \mathcal{N}(q, \hat{\sigma}^2/T)$, for large T, where

$$\bar{X} = \frac{1}{T}\sum_{t=1}^{T}\frac{M_t}{m_t} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{M_t}{m_t} - \frac{1}{T}\sum_{t=1}^{T}\frac{M_t}{m_t}\right)^2 .$$

The α-level test under this approximation is :

$$\phi_\alpha = \mathbb{1}\left(\frac{\bar{X} - q}{\hat{\sigma}/\sqrt{T}} > z_\alpha\right), \quad z_\alpha = \Phi^{-1}(1-\alpha).$$

Framework for the Vasicek Test

Vasicek One-Factor Model

As before, let m denote the number of borrowers at the beginning of the period, and M the number of defaults during the period. Under the *Vasicek one-factor model*, M is given by :

$$M = \sum_{i=1}^{m} \mathbb{1}(X_i \leq d), \quad X_i = -\sqrt{\rho} Z + \sqrt{1-\rho}\epsilon_i, \quad i \in \{1, \ldots, m\}.$$

For a given i, X_i is called the *asset value* and $Z, \epsilon_1, \ldots, \epsilon_m \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

The (unobserved) *systemic* risk factor Z represents the state of the economy while $\epsilon_1, \ldots, \epsilon_m$ are *idiosyncratic* risk components. ρ is the asset correlation, i.e. $\text{Corr}(X_i, X_j) = \rho$ for all $i \neq j$, and is specified by the regulator under Basel II.

Conditional PD given the Systemic Risk Factor

Since $X \sim \mathcal{N}(0,1)$, we define the default threshold as $d = \Phi^{-1}(q)$, ensuring that the true (TTC) PD satisfies $q = \Phi(\Phi^{-1}(q)) = P(X \leq \Phi^{-1}(q)) = P(X \leq d)$.

Under the one-factor model, the conditional PD given the systemic factor Z is :

$$P(X \leq d \mid Z) = P\left(-\sqrt{\rho}Z + \sqrt{1-\rho}\epsilon \leq \Phi^{-1}(q) \mid Z\right) = \Phi\left(\frac{\Phi^{-1}(q) + \sqrt{\rho}Z}{\sqrt{1-\rho}}\right).$$

Unconditional Tests : Vasicek Test

Quantile of the empirical default rate distribution

Assume the Vasicek one-factor model where, given a systemic factor Z, the default state variables $Y_1, \ldots, Y_m$ are independent and identically distributed as :

$$Y_1, \ldots, Y_m \mid [Z = z] \overset{\text{iid}}{\sim} \text{Bin}\left(1, \Phi\left(\frac{\Phi^{-1}(\hat{q}) + \sqrt{\rho}z}{\sqrt{1-\rho}}\right)\right),$$

where $\hat{q}$ is our estimate of q, the true TTC PD.
Denote by $Q_m(x)$ the quantile function of the distribution of the empirical default rate M/m, for $x \in (0,1)$. The limiting $1-\alpha$-quantile is then given by :

$$Q_V(1-\alpha) := \lim_{m\to\infty} Q_m(1-\alpha) = \Phi\left(\frac{\Phi^{-1}(\hat{q}) + \sqrt{\rho}\Phi^{-1}(1-\alpha)}{\sqrt{1-\rho}}\right). \tag{4}$$

Vasicek Test

To assess the accuracy of the forecasted PD $\hat{q}$ as formulated in (3), we use the limiting result (4) to construct an approximate α-level test :

$$\phi_\alpha = \mathbb{1}\left(M/m > Q_V(1-\alpha)\right).$$

Unconditional Tests : Moment Matching Approximation

An alternative approach for constructing a test for a fixed sample size m is to approximate the distribution of M/m using a *Beta distribution*.

Beta Distribution

If $Z \sim B(\alpha, \beta)$, the parameters α and β can be expressed in terms of the moments of the distribution :

$$\alpha = \frac{E[Z]}{Var[Z]}\left(E[Z](1-E[Z]) - Var[Z]\right), \quad \beta = \frac{1-E[Z]}{Var[Z]}\left(E[Z](1-E[Z]) - Var[Z]\right). \tag{5}$$

Moment Matching Approximation

Under the Vasicek one-factor model, the expectation and variance of M/m are :

$$E[M/m] = q, \quad \text{and} \quad Var[M/m] = \frac{m-1}{m}\Phi_2(\Phi^{-1}(q), \Phi^{-1}(q), \rho) + \frac{q}{m} - q^2. \tag{6}$$

To assess the accuracy of the forecasted PD $\hat{q}$, as formulated in (3), we can thus use the approximate α-level test :

$$\phi_\alpha = \mathbb{1}\left(M/m > Q_B(1-\alpha)\right),$$

where $Q_B(1-\alpha)$ is the $1-\alpha$-quantile of the Beta distribution with parameters obtained by substituting (6), with q replaced by $\hat{q}$, into (5).

Simulation Study

In this simulation study, we want to compare the performance of three tests — the exact binomial test, the Vasicek test, and the moment-matching approximation test — for the following hypothesis testing problem :

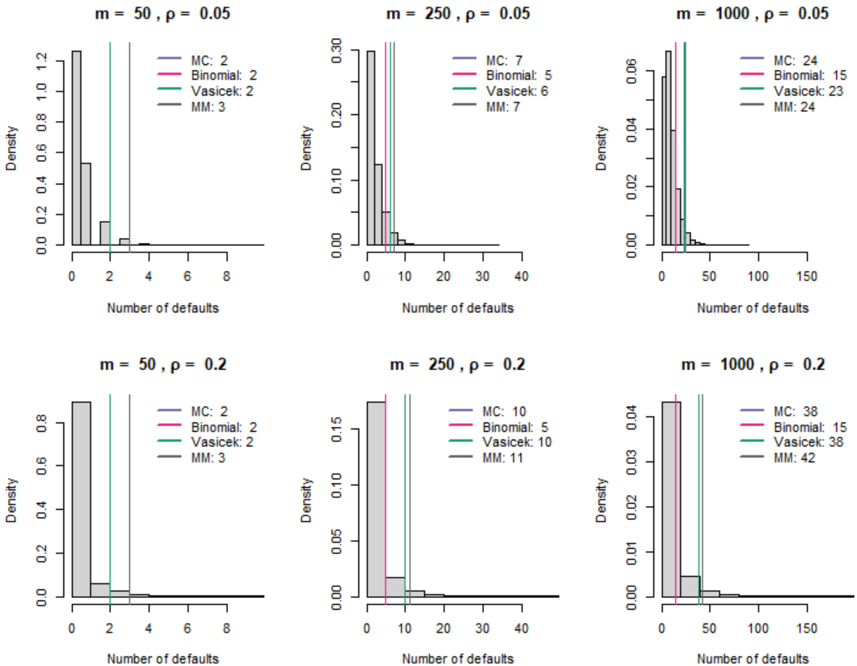
$$\begin{cases} H_0 : & q \leq 1\% \\ H_1 : & q > 1\%. \end{cases}$$

The simulation study consists of two main parts where defaults are simulated using the Vasicek one-factor model under different values of the portfolio size m and asset correlation ρ.

- **Part 1 : Critical Values Comparison**
  We construct the distribution of the number of defaults M through Monte Carlo (MC) simulation. The 95% quantile of the simulated distribution is then compared with the critical values from the three tests (results are displayed in Figure 3).
- **Part 2 : Test Power Comparison**
  Using again simulation, we calculate the rejection rate for each test under different true PD values. When the true PD exceeds 1%, the rejection rate is equal to the power of the tests. (results are displayed in Table 4).

Simulation Study : Comparison of Critical Values

Figure 3 – Comparison of 95%-critical values for PD tests of $H_0 : q \leq 1\%$.



Simulation Study : Comparison of Test Power

ρ	True PD	$m = 50$ Binomial	$m = 50$ Vasicek	$m = 50$ MM	$m = 250$ Binomial	$m = 250$ Vasicek	$m = 250$ MM	$m = 1000$ Binomial	$m = 1000$ Vasicek	$m = 1000$ MM
5%	0.5	0.48	0.48	0.07	1.53	0.71	0.33	2.58	0.36	0.28
	1.0	2.40	2.40	0.55	9.61	5.77	3.44	17.78	5.08	4.35
	1.5	5.72	5.72	1.68	22.36	15.36	10.45	38.80	16.21	14.45
	2.0	10.06	10.06	3.58	36.36	27.42	20.36	57.61	30.95	28.40
	2.5	15.05	15.05	6.08	49.29	39.56	31.26	71.83	45.85	42.98
20%	0.5	1.63	1.63	0.65	5.00	1.30	1.04	7.70	1.36	1.07
	1.0	4.79	4.79	2.29	13.15	4.57	3.79	19.15	4.97	4.08
	1.5	8.54	8.54	4.52	21.56	8.84	7.55	29.84	9.66	8.18
	2.0	12.54	12.54	7.15	29.34	13.57	11.81	39.05	14.78	12.78
	2.5	16.51	16.51	9.92	36.54	18.50	16.33	47.13	20.08	17.63

Table 4 – Simulation study : Rejection rate (%) of the binomial, Vasicek, and moment matching (MM) tests across PD values, portfolio sizes (m), and asset correlations (ρ).

- For small portfolios ($m = 50$), the binomial and Vasicek tests offer similar—though limited—power and outperform the moment matching method. Correlation (ρ) has little influence in this regime.
- For larger portfolios ($m = 250$ and $m = 1000$), the binomial test loses its size control, while the Vasicek test emerges as the most powerful α-level test. The moment matching method performs comparably, though overall power remains modest.

Outline

Outline

Basic Setting

Definition (Loss Given Default)

The *Loss Given Default* (LGD) represents the proportion of a credit exposure that is lost when a borrower defaults, after accounting for recoveries. It is typically expressed as a percentage of total exposure at the time of default.

Mathematical Framework

Let $Y \in [0, 1]$ denote the realized loss rate (as a fraction of total exposure at the time of default). Given a set of covariates X associated with the borrower's creditworthiness, the LGD is defined as the conditional expectation $E[Y \mid X]$, such that :

$$Y = \text{LGD} + \epsilon, \tag{7}$$

where the residual term $\epsilon := Y - \mathbb{E}[Y \mid X]$ captures the unexplained variation.

Beta Regresion

An example for the conditional distribution of Y given X is the beta distribution :

$$Y \mid X \sim B(\text{LGD}\phi, (1 - \text{LGD})\phi),$$

where we have $E[Y \mid X] = \text{LGD}$ and $Var[Y \mid X] = Var[\epsilon \mid X] = \frac{\text{LGD}(1-\text{LGD})}{1+\phi}$.

Outline

How Can Discriminatory Power Be Measured and Tested ?

In contrast to the PD setup where the response variable is binary, the realized loss rate Y is a continuous random variable taking value between 0 and 1.

- Standard tools such as the CAP and ROC curves used in PD validation are not directly applicable ;
- Instead, LGD-specific validation concepts must be used to assess discriminatory power.

Definition (Discriminatory Power for LGD)

In the context of LGD, discriminatory power refers to the ability of the LGD to distinguish borrowers based on their realized loss rates. In other words :

- It reflects how well the LGDs preserve the ordering of the actual losses Y ;
- Equivalently, it is a measure of how closely the LGD values align with the quantity Y they aim to predict.

Commonly Used Tools

In the following, we present two widely used tools to measure the discriminatory power of LGD models :

- **Concentration Curves (CC) :** A generalization of the CAP curve adapted to continuous outcomes.
- **Rank Correlation :** With particular focus on Spearman's r_S coefficient.

Concentration Curve (CC) : Concept and Definition

Recall the definition of the CAP curve in the PD validation setup :

$$CAP(u) = F_D(F^{-1}(u)) = \frac{E[Y\mathbb{1}\left(S \leq F^{-1}(u)\right)]}{E[Y]}, \quad u \in (0,1).$$

Extending this concept to LGD validation, where Y is the (continuous) realized loss rate, and where S is replace by LGD, yields the concentration curve.

Definition (Concentration Curve (CC))

Let F denote the distribution function of the LGD, the *concentration curve* is defined as :

$$CC(u) = \frac{E[Y\mathbb{1}\left(\text{LGD} \leq F^{-1}(u)\right)]}{E[Y]}, \quad u \in (0,1).$$

Two Extreme Situations

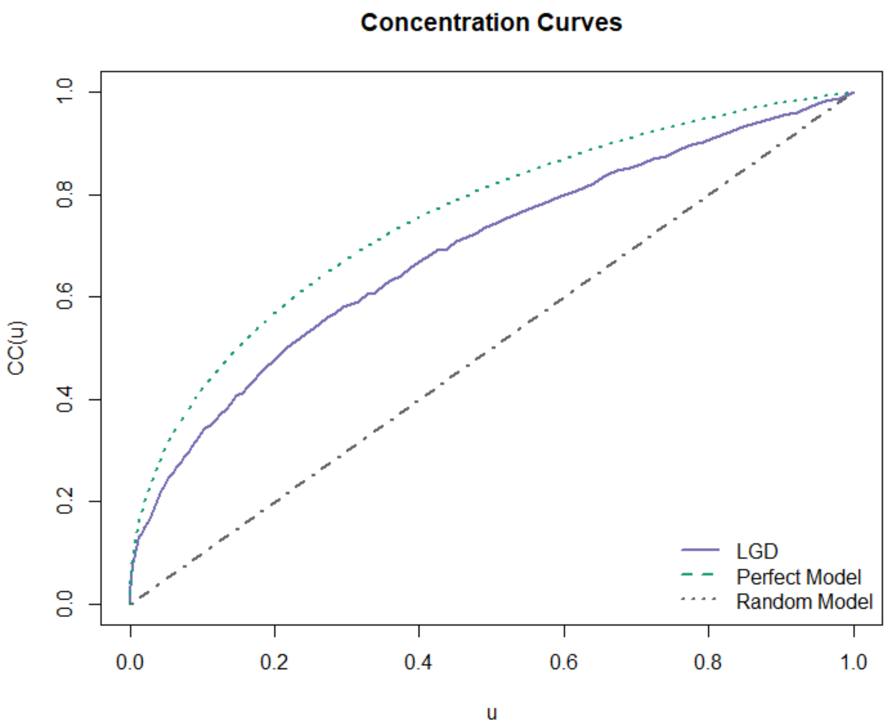
- **Random LGD :** The LGD carries no information on the loss rate :

$$\text{LGD} \perp\!\!\!\perp Y \quad \Longleftrightarrow \quad \text{CC}(u) = \frac{E[Y]P[\text{LGD} \leq F^{-1}(u)]}{E[Y]} = u.$$

- **Perfect LGD :** The LGD predicts with no error the realized loss rates :

$$CC(u) = \frac{E[Y\mathbb{1}\left(\text{Y} \leq F^{-1}(u)\right)]}{E[Y]}.$$

Examples of Concentrations Curves



Rank Correlation as a Measure of Discriminatory Power

A statistical test of discriminatory power can be obtain using rank correlation measuring the association between two random variables based on their ranks. Two commonly used rank correlations are Spearman's r_S and Kendall's τ. Here, we focus on Spearman's r_S due to its similarity to the classical Person correlation.

Spearman's r_S

Let $R(Y_i)$ denote the rank of Y_i among $Y_1, \cdots, Y_M$ and $R(\text{LGD}_i)$ denote the rank of LGD_i among $\text{LGD}_1, \dots, \text{LGD}_M$. Spearman's r_S is then given by :

$$r_S = \frac{\sum_{i=1}^{M} R(Y_i)R(\text{LGD}_i) - \frac{1}{M}\left(\sum_{i=1}^{M} R(Y_i)\right)\left(\sum_{i=1}^{M} R(\text{LGD}_i)\right)}{\sqrt{\sum_{i=1}^{M} R(Y_i)^2 - \frac{1}{M}\left(\sum_{i=1}^{M} R(Y_i)\right)^2}\sqrt{\sum_{i=1}^{M} R(\text{LGD}_i)^2 - \frac{1}{M}\left(\sum_{i=1}^{M} R(\text{LGD}_i)\right)^2}}$$

$$= 1 - \frac{6\sum_{i=1}^{M}(R(Y_i) - \text{LGD}_i)^2}{n(n^2-1)} \quad \text{if there are no ties.}$$

This statistic can be used to test the hypothesis of no association between the realized loss rate Y and the LGD, i.e. LGD has no discriminatory power.

Under H_0, the ranks of the LGD predictions are randomly assigned, allowing us to compute the distribution of r_S under H_0 by evaluating all possible permutations. An exact α-level test is then $\phi_\alpha = \mathbb{1}(|r_S| > r_0)$ where r_0 is such that $P(|r_S| > r_0) = \alpha$ under the null hypothesis.

Outline

Measuring and Testing Calibration Accuracy

Definition : Calibration Accuracy for LGD

Similar to the PD case, a well-calibrated LGD model produces estimates that are close to the true LGD values.

PIT vs. TTC and Independence Assumptions

Here, we adopt a Point-in-Time (PIT) approach by treating realized loss rates as independent realizations of random variables.

In this framework, LGD is allowed to vary with the economic environment :

$$\text{LGD} = f(X, Z_t)$$

where :

- X : borrower-specific characteristics ;
- Z_t : macroeconomic covariates (e.g., GDP growth).

Commonly Used Tools

The following tests are widely used to assess LGD calibration accuracy :

- F-**test** : evaluates model fit via regression-based specification testing.
- t-**test** : compares average estimated LGDs and realized loss rates.

Calibration Tools : F-test

Let $\hat{\text{LGD}}_1, \ldots, \hat{\text{LGD}}_M$ be the estimated LGDs for the M borrowers who defaulted during the period (typically one year) and let $Y_1, \ldots, Y_M$ be the corresponding (independent) realized loss rates.

Consider the linear model :

$$Y_i = \alpha + \beta \hat{\text{LGD}}_i + \epsilon_i, \quad i = 1, \cdots, M. \tag{8}$$

F-test

Based on the definition of LGD in (7), we can now formulate the following hypothesis testing problem in terms of the α and β parameters in (8) :

$$\begin{cases} H_0 : & \alpha = 0 \quad \text{and} \quad \beta = 1 \\ H_1 : & \alpha \neq 0 \quad \text{and/or} \quad \beta \neq 1. \end{cases}$$

Under the assumption that $\epsilon_1, \cdots, \epsilon_M \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, an exact α-level test is :

$$\phi_\alpha = \mathbb{1}\left(F > F_{2,M-2,\alpha}\right),$$

where $F_{2,M-2,\alpha}$ is the $(1-\alpha)$ quantile of the $F_{2,M-2}$-distribution.

When the sample size M is large, asymptotic tests such as the Wald test, the likelihood ratio test or the Lagrange mulitplier test may be used instead.

Calibration Tools : *t*-test

t-test

We test whether the estimated LGDs match the true LGDs :

$$\begin{cases} H_0 : & \mathrm{LGD}_i = \hat{\mathrm{LGD}}_i \quad \forall i \in \{1, \dots, M\} \\ H_1 : & \exists i \in \{1, \dots, M\} \text{ such that } \mathrm{LGD}_i \neq \hat{\mathrm{LGD}}_i. \end{cases}$$

Under H_0, and assuming $\epsilon_1, \cdots, \epsilon_M \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ in (7), we have :

$$E[\bar{Y} \mid \mathcal{X}] = \frac{1}{M} \sum_{i=1}^{M} \hat{\mathrm{LGD}}_i = \bar{\mathrm{LGD}} \quad \text{and} \quad \mathrm{Var}[\bar{Y} \mid \mathcal{X}] = \frac{\sigma^2}{M}.$$

An exact α-level test is thus given by :

$$\phi_\alpha = \mathbb{1}\left(\left| \frac{\bar{Y} - \bar{\mathrm{LGD}}}{\hat{\sigma}/\sqrt{M}} \right| > t_{M,\alpha/2} \right)$$

where $t_{M,\alpha/2}$ is the $(1 - \alpha/2)$-quantile of t_M and $\hat{\sigma} = \frac{1}{M} \sum_{i=1}^{M} (Y_i - \hat{\mathrm{LGD}}_i)^2$.

When M is large, a normal approximation can be used. Non-parametric alternatives, e.g. the sign test or the Wilcoxon signed-rank test, are also available.

Outline

References

- Engelmann, B., & Rauhmeier, R. (2011). *The Basel II Risk Parameters : Estimation, Validation, Stress Testing with Applications to Loan Risk Management* (2nd ed.). Springer.
- Fischer, M., & Pfeuffer, M. (2014). A statistical repertoire for quantitative loss given default validation : overview, illustration, pitfalls and extensions. *Journal of Risk Model Validation*.
- McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative Risk Management : Concepts, Techniques and Tools*. (2nd ed.) Princeton Series in Finance. Princeton University Press.
- Scheule, H., Rösch, D., & Baesens, B. (2017). *Credit Risk Analytics : The R Companion*. Wiley.
- Tasche, D. (2006). Validation of internal rating systems and PD estimates.
- Tasche, D. (2003). A traffic lights approach to PD validation.
- Wackerly, D., Mendenhall, W., & Schaeffer, R. (2008). *Mathematical Statistics with Applications* (7th ed.). Cengage.