Heterogeneity Testing in IRB Credit Risk Models

Regulatory Requirements, Methods, Challenges, and Considerations

Andrija Djurovic

www.linkedin.com/in/andrija-djurovic

Model Heterogeneity

- One key aspect of analyzing the distribution and allocation of obligors and facilities in IRB credit risk models is heterogeneity.
- Heterogeneity refers to adequate differentiation in risk profiles across ratings, pools, or buckets.
- It is commonly tested in Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD) models, often using tests like the two-proportion test and Welch t-test.
- Practitioners closely examine the heterogeneity of the rating system as a regulatory requirement. However, challenges such as defining specific thresholds for the testing procedure and drawing overall conclusions about the level of heterogeneity remain insufficiently explored in practice.

Regulatory Requirements for Model Heterogeneity

- In the regulatory context, heterogeneity is typically assessed as part of the overall model structure. In this regard, heterogeneity testing is usually considered alongside discriminatory power and homogeneity testing.
- Different regulatory documents set varying requirements and expectations for model structure and heterogeneity.
- The Capital Requirements Regulation (CRR) establishes the foundation for rating system requirements, ensuring that models effectively differentiate risk levels and maintain adequate granularity.
- The Commission Delegated Regulation (EU) 2022/439 provides additional clarifications and assessment criteria for competent authorities when verifying compliance with CRR requirements.
- The ECB Guide to Internal Models, in the credit risk section, sets more detailed expectations for risk differentiation across grades or pools.
- Although still in draft form, the Draft Guidelines on Credit Conversion Factor Estimation under Article 182(5) of Regulation (EU) No 575/2013 also indicates that institutions should avoid significant overlaps in the distribution of conversion risk between grades and pools.

Common Practices in Heterogeneity Testing

- Heterogeneity testing typically relies on statistical hypothesis tests comparing adjacent grades or pools.
- For PD models, the two-proportion test is most commonly applied, while for LGD and EAD, practitioners use the Welch t-test for mean differences.
- Both tests are one-sided, formulated such that the parameter of the "better" grade should be smaller than that of the "worse" one.
- Usually practitioners run these tests across all adjacent pairs and reference dates, and the overall heterogeneity conclusion is drawn by assessing the proportion of pairs that pass the test at chosen significance levels typically 1%, 5%, or 10%.

Common Practices in Heterogeneity Testing cont.

Two-Proportion Test

The following formula presents the test statistic for the two-proportion test:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where:

- p_1 is the observed proportion of the "better" rating grade;
- p_2 is the observed proportion of the "worse" rating grade;
- p is the pooled proportion calculated as $\frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$;
- n_1 and n_2 denote the sizes of the "better" and "worse" rating.

The test statistic Z is assumed to follow a standard normal distribution, and the p-value is derived accordingly.

Common Practices in Heterogeneity Testing cont.

Welch T-Test

The following formula represents the test statistic for the Welch t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{x}_1$ is the “better” pool/bucket observed mean;
- $\bar{x}_2$ is the “worse” pool/bucket observed mean;
- s_1^2 and s_2^2 are the sample variances;
- n_1 and n_2 are the sample sizes of the “better” and “worse” pool/bucket.

The test statistic above is assumed to follow a t-distribution, with degrees of freedom calculated using the Satterthwaite approximation, as given by the following formula:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

where:

- s_1^2 and $s_2^2$are the variances of the pools or buckets;
- n_1 and n_2 are the sizes of the respective pools or buckets.

Given the assumed t-distribution of the test statistic and the corresponding degrees of freedom, the p-value can be derived.

Practical Challenges in Heterogeneity Testing

- Despite its routine use, practitioners face several challenges when testing model heterogeneity.
- Test results are often interpreted solely based on observed averages, overlooking variability driven by differing sample sizes and data dispersion across testing levels.
- Unified thresholds for the acceptable number of failed pairs are applied across portfolios, even when their underlying sample characteristics differ.
- Moreover, heterogeneity testing is rarely connected to model calibration or to the expected monotonic pattern of risk parameters, which can lead to inconsistent or misleading conclusions.
- Standard practices can be improved by using statistical power analysis and monotonicity analysis to address some of the most common challenges practitioners face in heterogeneity testing.

Statistical Power Analysis

- Statistical power analysis can provide an additional perspective to support current practices in heterogeneity testing.
- Power represents the probability that a test correctly rejects a false null hypothesis - the likelihood of detecting true differences between consecutive grades or pools.
- It can support both model development and validation.
- During model development, it helps determine the number of rating grades or pools needed to detect differences in risk parameters reliably.
- During validation, it enables estimating how many pairs are expected to fail the heterogeneity test purely by chance, given the sample sizes and true effect sizes, under the assumption of a well-calibrated model.
- Finally, statistical power analysis enables practitioners to set more defensible thresholds for determining whether the model exhibits adequate heterogeneity overall.

Monotonicity Disruption

- Because heterogeneity testing is performed using a one-sided test - examining whether the average risk parameter of a “better” rating, pool, or bucket is smaller than that of a “worse” one - monotonicity is a minimally necessary condition.
- Depending on whether the test is applied to the whole modelling dataset (all reference dates) or to one or a few reference dates, heterogeneity test results are strongly influenced by the number of observations per rating grade, pool, or bucket.
- Therefore, when testing heterogeneity as part of periodic model validation, practitioners can combine the results of statistical tests with observed monotonicity, using monotonicity as a softer criterion for assessing this model aspect.
- A recurring challenge in this context is determining the threshold for the expected number of grades at which monotonicity may be disrupted.
- Designing monotonicity analysis to investigate its disruption under the assumption of a well-calibrated model can help practitioners probabilistically assess whether an observed disruption should be interpreted as a genuine problem or simply as noise in the data.

Conclusions

- Regulatory frameworks require meaningful differentiation of risk across rating grades, pools, or buckets in PD, LGD, and EAD models.
- Standard practice (two-proportion test for PD, Welch t-test for LGD/EAD) often faces challenges with threshold setting, proper consideration of calibration, and accounting for monotonicity disruptions.
- Statistical power analysis quantifies the ability to detect heterogeneity and can support testing during both model development and validation.
- Monotonicity disruption analysis assesses breaks in expected risk ordering, complementing formal hypothesis testing.
- By combining statistical power and monotonicity disruption analyses, practitioners can improve interpretation, distinguish genuine model issues from random variation, and support robust model development and validation.