Heterogeneity Testing in IRB Models

When the P-Value $> 50\%$ is Informative

Andrija Djurovic

www.linkedin.com/in/andrija-djurovic

Regulatory Requirements

- Heterogeneity is a key aspect of analyzing the distribution and allocation of obligors and facilities in IRB models.
- It refers to adequately differentiating risk profiles across rating grades, pools, or buckets.
- Heterogeneity testing is required for all three risk parameters - Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD) - as part of both initial and ongoing model validation.
- The Capital Requirements Regulation (CRR), the ECB Guide to Internal Models, and the Commission Delegated Regulation (EU) 2022/439 set out key requirements for model structure, rating system validation, and risk differentiation.
- These regulations define minimum rating grade counts and require clear documentation of how risk is differentiated across grades or pools. They also set expectations for demonstrating homogeneity within grades and heterogeneity between them across all three risk parameters.
- Furthermore, regulations require institutions to ensure that the structure and distribution of exposures across grades or pools reflect genuinely distinct risk levels and avoid excessive concentration unless justified by the model framework or portfolio characteristics.

Heterogeneity Testing in Practise

Standard Practice

Standard practice recommends conducting heterogeneity testing based on the realized values of specific risk parameters, typically using one-sided hypothesis tests. The null hypothesis generally assumes that the average value of the risk parameter in the "better" grade, pool, or bucket is lower than in the "worse" one. Here, "better" and "worse" refer to adjacent categories, with "better" indicating a lower calibrated risk parameter value. A favorable outcome in heterogeneity testing is the rejection of this null hypothesis. Commonly used methods include the two-proportion test for PD models and the Welch t-test for LGD and EAD models. In addition, practitioners apply heterogeneity testing at different levels, such as the full development sample, the application sample, or an aggregated portfolio across multiple reference dates. The choice of testing level influences the statistical results and the overall conclusions regarding heterogeneity. Sometimes this leads to a relaxation of the testing criteria, with heterogeneity assessed solely based on the monotonicity of the realized values.

PD Models

The following formula shows the test statistic for the two-proportion test commonly used in PD models:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

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 $\frac{n_1\hat{p_1}+n_2\hat{p_2}}{n_1+n_2}$ and
- n_1 and n_2 denote the sizes of the "better" and "worse" rating.

The two-proportion test assumes that the test statistic follows a standard normal distribution. Consequently, the p-value is compared to a chosen significance level to draw conclusions from the test.

Heterogeneity Testing in Practise cont.

LGD and EAD Models

LGD and EAD are considered continuous variables, typically taking values within a specific range. A natural choice for testing heterogeneity is the 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 and
- 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:

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

Simulation Study

Simulation Design

The following simulation assumes heterogeneity testing for the PD model on the given rating scale. The rating scale below consists of seven grades, accompanied by data (`no` - number of obligors, `nb` - number of defaults, and `odr` - realized default rate) for a specific application portfolio at a single reference date. The simulation aims to perform heterogeneity testing on adjacent ratings using the two-proportion test and to interpret the resulting p-values in the context of model heterogeneity.

Rating Scale

```
##   rating  no  nb    odr
## 1    R01 170   3  1.76%
## 2    R02 118  10  8.47%
## 3    R03 274  50 18.25%
## 4    R04 105  17 16.19%
## 5    R05  91  43 47.25%
## 6    R06 196 122 62.24%
## 7    R07  51  44 86.27%
```

Simulation Study cont.

Heterogeneity Testing

	rating	no	nb	odr	p-value
1	R01	170	3	1.76%	
2	R02	118	10	8.47%	0.35%
3	R03	274	50	18.25%	0.68%
4	R04	105	17	16.19%	68.08%
5	R05	91	43	47.25%	0.00%
6	R06	196	122	62.24%	0.84%
7	R07	51	44	86.27%	0.06%

Simulation Results

Comparing the p-values from the one-sided two-proportion tests to the commonly used 5% significance level, it can be concluded that there is sufficient evidence of model heterogeneity for all adjacent rating pairs except R03 - R04. For the R03 - R04 pair, the p-value exceeds 50%, **indicating a disruption of monotonicity**. Given the testing level, practitioners sometimes relax the heterogeneity criteria, accepting monotonicity alone as sufficient for a limited number of rating pairs. In addition to allowing p-values below 50% for specific pairs, the standard framework can be extended to identify how many rating pairs are expected to fail even under monotonicity conditions, thereby enhancing the robustness of the testing process.