Validation of Internal Rating and Scoring Models

Dr. Leif Boegelein
Global Financial Services Risk Management
Leif.Boegelein@ch.ey.com

07.09.2005
Agenda

1. Motivation & Objectives
2. Measuring Discriminatory Power
3. Measuring the Quality of PD Calibration
4. Model Validation and “Rating Philosophies”
5. Relevance for Retail Credit Scoring
6. Summary
Motivation: Regulatory Compliance F-IRB / A-IRB

Regulatory Requirements on “Validation of internal estimates” (BII §§500-505)

“Banks must have a robust system in place to validate the accuracy and consistency of rating systems, processes, and the estimation of all relevant risk components…” [Basel II, § 500]

“…a bank must demonstrate to its supervisor that the internal validation process enables it to assess the performance of internal rating and risk estimation systems consistently and meaningfully.” [Basel II, § 500]

“Banks must regularly compare realised default rates with estimated PDs for each grade and be able to demonstrate that the realized default rates are within the expected range for that grade.” [Basel II, § 501]

“Banks must also use other quantitative validation tools and comparisons with relevant external data sources. The analysis must be based on data that are appropriate to the portfolio, are updated regularly, and cover a relevant observation period.” [Basel II, § 502]
Regulatory developments: Europe

- 2005:
  - Basel II Final Accord
  - Trading Book Review Consultative Paper April 2005
  - Final CRD 4th quarter 2005

- 2006:
  - FSA CP 05/03
  - CRD Legislated
  - FSA Prudential Source Book

- 2007:
  - Credit Risk Standardised
  - Operational Risk BI Standardised
  - Credit Risk Retail IRB FIRB
  - Operational Risk BI Standardised

- 2008:
  - Credit Risk Basel I
  - Operational Risk Advanced IRB
  - Credit Risk AMA
  - Operational Risk AMA

- QIS 4/5 Recalibration Of Accord
- FSA Application for Accreditation
- Parallel Running

Leif Boegelein, Credit Scoring & Credit Control IX. September 7th, 2005
Model Validation in the context of the UK Waiver application. CP05/03

Section A: Overview of Structure & Governance Framework

**CHALLENGES**
- Roll-out plan – THE KEY DECISION
- Single point of contact – the right person at the right level
- Effective governance framework

Section B: Self Assessment

**CHALLENGES**
- Basel II, CRD, CEBS, FSA
- Comprehensive assessment
- Sign-off

Section C: Summary of Firm’s Approach in Key Areas

**CHALLENGES**
- Section subject to most scrutiny
- Validation of Credit Risk Models
- Evidencing governance & use test

Section D: Capital Impact

**CHALLENGES**
- Quantitative impact study

Section E: Detail on Rating System

**CHALLENGES**
- Gathering of comprehensive information per model
- Evidencing compliance with requirements

Section F: CEO or Equivalent Sign-off

**CHALLENGES**
- Senior management held accountable
- How to involve them to the level required?
Motivation: Competitive Risk Management

- Credit Risk Management processes rely heavily on the quality of credit risk models.

- Model error influences all phases of the credit lifecycle from transaction origination to portfolio management.

- Systematically inaccurate estimates of risk will lead to inefficient pricing and portfolio management decisions.

Impact:

1. Deterioration of portfolio quality due to adverse selection effects.

2. Erosion of capital base due to unexpected losses, inadequate pricing and provisioning.
# The Model Validation Process - Components

## Methodology & Quantification
- **Developmental Evidence:** Model development and refinement stages
- Benchmarking: Comparison of results with other models
- **Backtesting:** Comparison of realized experience with that predicted by model
- Refinement: Review of key risk drivers

## Data Quality
- Completeness: Coverage of all transactions and portfolios; missing data protocols
- Appropriateness: Historical data is representative of current portfolio
- **Accuracy:** Input data validation; frequent information refresh – up to date data
- Consistency: Common identification and data definitions across organization

## Data Security & Controls
- Audit Trails for inputs and outputs
- Access & change controls
- **Business Continuity Procedures**

## Reporting
- Inclusion in reporting target audience of key decision-makers and reviewers
- Content matched to audience in depth, granularity, length, sufficiency
- Quality and accuracy appropriate for designated use
- Frequency appropriate for audience and designated use

## Governance
- Consistency of application by end users
- Use Test
- Transparency of model development, approval and validation processes
- Independence of model operation from transaction/position originator
- Board responsibilities and committee structure and senior management oversight
- Ownership of models, ownership of validation process
- Approval processes for new models
- Documentation standards for
  - Policies & procedures
  - Models and methodologies
  - Validation process

---

Leif Boegelein, Credit Scoring & Credit Control IX. September 7th, 2005
Agenda

1. Motivation & Objectives

2. Measuring Discriminatory Power

3. Measuring the Quality of PD Calibration

4. Model Validation and “Rating Philosophies”

5. Relevance for Retail Credit Scoring

6. Summary
Quantitative Validation: Discriminatory Power

- Frequency plots for validation sample of retail scorecard validation sample (N=1000) indicate that the system is moderately succeeding in separating defaulters from Non-Defaulters.

- Conditional Frequencies give us the likelihood of observing a certain Score given information about Default / Non – Default. This is not the sought after PD, i.e. the likelihood of observing Default for a given Score!

- Beyond eye-inspection, we need a metric for the degree of discrimination between Defaulters and Non-Defaulters.
Quantitative Validation: Discriminatory Power

**QQ plot of the distributions is the Cumulative Accuracy Profile (CAP).**

**The Accuracy Ratio (AR) is defined as the area between the model CAP and the random model, divided by the area between the perfect model and the random model.**

**AR is typically reported between 0 (random model) and 1 (perfect model). Possible values in [-1;1] with negatives indicating better discrimination than random but in reverse order.**
Quantitative Validation: Discriminatory Power (Example)

- CAP and Accuracy Ratio measure how well the model is able to separate defaulters from non-defaulters.

- The exemplary models produce values for AR between 0.26 and 0.63.

CAP curves for 3 models developed on a retail dataset. Validation sample 1000 cases. 2 out of the 3 models seem to have discriminatory power.

Validation of discriminatory power must be based on a representative sample that has not been used for model development!
Quantitative Validation: Discriminatory Power (Example)

- Default events are random. Even if we could predict all default probabilities without error, the measures could still indicate low discriminatory power for periods with “odd” defaults.

- For small samples, low default portfolios and models developed for pools of homogeneous obligors this issue becomes critical.

- AR is dependent on the sample and can therefore not provide an objective comparison between models of different samples!

- By simulating defaults for the portfolio and calculating AR in each scenario we gain an impression on the variability of AR itself.

*Bootstrapping: Logit CAP and 95% confidence bounds*

*Bootstrapping: Distribution of AR*
Quantitative Validation: Discriminatory Power (Example)

- For small (validation and development) sample sizes and low-default portfolios the influence of statistical noise on Accuracy ratios becomes more pronounced.

- AR analysis should include a test versus the random model with AR=0. In the example, the distribution of model AR is comfortably far to the right of the distribution of the Random model AR. Based on the simulation results (Validation sample size 200) we can estimate the confidence level of wrongly rejecting the hypothesis that the model is a random model (<<10exp-3 in the example).
Quantitative Validation: Discriminatory Power

- The Accuracy Ratio is a random variable itself. It is dependent on the portfolio structure, the number of defaulters, the number of rating grades etc. The metric is influenced by what it is measuring and should not be interpreted without knowledge of the underlying portfolio.

- Accuracy Ratios are not comparable across models developed from different samples.
Agenda

1. Motivation & Objectives

2. Measuring Discriminatory Power

3. Measuring the Quality of PD Calibration

4. Model Validation and “Rating Philosophies”

5. Relevance for Retail Credit Scoring

6. Summary
Quantitative Validation: Calibration

- The Calibration of Rating system influences the complete credit lifecycle, from pricing to portfolio management decisions.

- In practice, validation of PD accuracy is often considered a difficult exercise due to data constraints and lack of a proper definition of rating philosophy.

- Validation of PD accuracy covers all rating classes / retail pools. Different from the traditional focus on accuracy at the cut-off point in retail scoring.

- Simple statistical methods can be utilized to illustrate the “uncertainty” associated with PD estimates.
Quantitative Validation: Calibration

- The ultimate goal of rating models is to predict the probability of default events.
- PDs are either produced by the model or obtained via mapping of internal grades to external default experience.
- The accuracy of these estimates needs to be validated.
- Realized Default rates will deviate from estimated ones. Validation procedures need to examine whether the deviation is substantial and should lead to a review of the model or can be attributed to statistical noise. Focus:
  - Significance of Deviations
  - Monotony of PDs with regards to “risk”
- Binomial test / Normal test. Determine probability of observing the realized default rate under the hypothesis that the estimated value is correct for each rating class.
- Produce overall measure of calibration accuracy for the rating system.
Quantitative Validation: Binomial Test

Simple model to analyze calibration quality assuming independent default events:

Assume that our estimate of the PD for the reference period in a particular rating class $k$ is $\hat{p}_k$. We count $n_k$ obligors in this particular rating class. If default events happen independently, the number of total defaults $\hat{b}_k$ for this rating class during the reference period is binomially distributed with parameters:

$$\hat{b}_k \sim \text{bin}(n_k, \hat{p}_k)$$

Under this assumption, we can calculate the probability of observing a given range of the default rate. Alternatively we can fix a certain level of confidence and calculate the upper and lower bounds on the observed default rate under the hypothesis that $\hat{p}_k$ equals the true PD $p_k$.

Limitations:

- Default rates are typically influenced by common factors. Unconditional therefore appear to be "correlated". The closely related normal test incorporates correlation between default rates / default events.
- The chosen rating philosophy determines the volatility of PD estimates and should be considered when interpreting the results as we will discuss later on.
Quantitative Validation: Calibration

For 5 of the 7 rating classes we reject the hypothesis that the PD equals the long term average based on a two sided 95% interval and the assumption of independent defaults.
Quantitative Validation: Calibration

- The results from the Binomial test look at each rating class in isolation. The Hosmer Lemeshow statistic is an example of how the Goodness-of-Fit information can be condensed into one figure.

- The Hosmer-Lemeshow statistic is distributed Chi square with K degrees of freedom (if we assume that the \( p_k \) are estimated out of sample).

\[
HL = \sum_{k=1}^{K} n_k \left( \frac{p_k - \hat{p}_k}{\hat{p}_k(1 - \hat{p}_k)} \right)^2 \quad HL \sim \chi^2_k
\]

- Lower p-values for HL document decreasing Goodness-of-Fit, i.e. the hypothesis that the observed default rate equals the assumed value become increasingly unlikely.

- HL p-values allow us to compare rating models with different number of classes.

- For the example, the HL statistic is 44.7, which yields a p-value < 10exp-6 for 7 degrees of freedom. We would therefore reject \( H_0 \) that our estimated PDs are in line with realized defaults across rating classes.

- The test is based on the normal approximation and independence assumption of observed \( p_k \).
Agenda

1. Motivation & Objectives

2. Measuring Discriminatory Power

3. Measuring the Quality of PD Calibration

4. Model Validation and “Rating Philosophies”

5. Relevance for Retail Credit Scoring

6. Summary
Rating Philosophy

The volatility of observed PDs within rating grades and the migration frequency of obligors across rating grades are determined by the adopted “rating philosophy”.

When interpreting the previously discussed PD Accuracy measures, one has to take into account the philosophy underlying the estimations:

Point-in-time: Rating and PD estimate is based on current condition of the obligors’ risk characteristics, typically including the economic environment. This should result in high rating migration frequency. Default rates within rating classes should remain stable.

Through-the-cycle: Obligor ratings are not influenced by short term risk characteristics. Ratings do not change frequently, default rates vary.

Hybrid Approach: Mixture of both “pure” philosophies. The rating system incorporates current condition elements of obligor idiosyncratic and systematic risks. Data limitations force a “medium term” PD forecast.

<table>
<thead>
<tr>
<th></th>
<th>Default Rate Volatility</th>
<th>Rating Migration Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point in Time</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Through the Cycle</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

Point-in-time: Rating and PD estimate is based on current condition of the obligors’ risk characteristics, typically including the economic environment. This should result in high rating migration frequency. Default rates within rating classes should remain stable.
Agenda

1. Motivation & Objectives
2. Measuring Discriminatory Power
3. Measuring the Quality of PD Calibration
4. Model Validation and “Rating Philosophies”
5. Relevance for Retail Credit Scoring
6. Summary
Example of medium sized retail bank model infrastructure

As we have seen, uncertainty of measurements and constraints in application result mostly from data restrictions (low default, small sample, model assumptions etc.).

While on the retail banking book level, low-default and small may become important when portfolio splits are conducted to increase specificity of the model.
Agenda

1. Motivation & Objectives
2. Measuring Discriminatory Power
3. Measuring the Quality of PD Calibration
4. Model Validation and “Rating Philosophies”
5. Relevance for Retail Credit Scoring
6. Summary
Model Validation – Summary

- Validation is a rigorous process carried out by a bank to ensure that parameter estimates result in an accurate characterization of risk to support capital adequacy assessment.

- Validation is the responsibility of banks, not supervisors.

- It is not a purely statistical exercise. The appropriateness of metrics will depend on the particular institution's portfolios and data availability. However, quantitative measures are considered to play an integral part in validation process by institutions and regulators.

- In practice, data is often insufficient and validation will focus on developmental evidence in model design, data quality and benchmarking.

- In practice, model validation is often not implemented as an “actionable” and independent process. It often lacks a formal policy with statements of definitions of responsibilities, tests to be performed, metrics and benchmarks to use, thresholds for acceptable quality and actions to be taken if these are breached.
References


FSA, Strengthening capital standards, CP05/03, Jan. 2005
