Chapter 13
Model Validation

INTRODUCTION

Let us now assume that our probability of default (PD), loss given default (LGD), and exposure at default (EAD) models have been built. The next step is to then put them into production and start monitoring or validating them. Validation stems from the Latin word *validus* and refers to being effective, strong, or firm. It appears in many scientific disciplines, particularly in engineering, where it means a confirmation that a service or product meets the operator's requirements. Very similarly, in statistics, validation is a process of deciding whether the numerical results for quantification of hypotheses are proper descriptions of the data and is implemented, for example, via goodness-of-fit tests, residual analyses (as partly already discussed in earlier chapters), or backtesting of prediction power. In this chapter, we discuss various ways of validating credit risk models. The Basel Committee Validation Subgroup defines it in the following way:

> [T]he term validation encompasses a range of processes and activities that contribute to an assessment of whether ratings adequately differentiate risk, and whether estimates of risk components (such as PD, LGD, or EAD) appropriately characterize the relevant aspects of risk. (Basel Committee on Banking Supervision 2005b)

In what follows, you will first gain more insight into the regulatory aspects of validation, as well as learn key terms and principles. Next, you will learn about the characteristics of quantitative and qualitative validation.

REGULATORY PERSPECTIVE

We first start by discussing the regulatory perspective on validation. The Basel Committee on Banking Supervision stipulates various paragraphs on validation. We cannot state all of them, but rather we select several that are important for understanding the core concept of the regulatory perspective on model validation. Generally, a bank that uses models for the risk parameters is required

> to satisfy its supervisor that a model or procedure has good predictive power and that regulatory capital requirements will not be distorted as a result of its use. The variables that are input to the model must form a reasonable set of predictors. The model must be accurate on average across the range of borrowers or facilities to which the bank is exposed and there must be no known material biases. (Basel Committee on Banking Supervision 2006, §417)

In the same paragraph, it states:

> The bank must have a regular cycle of model validation that includes monitoring of model performance and stability; review of model relationships; and testing of model outputs against outcomes.

Banks therefore conduct validation on a regular basis. This usually occurs on a monthly basis and sometimes even more frequently. Obviously, the more frequently it's undertaken, the more quickly performance deviations or other critical issues can be detected. All this has to be accompanied by a proper documentation; see Basel Committee on Banking Supervision (2006, §418). The paragraphs §500 to §505 describe more details:

- §500: *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. 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*.
- §501: *Banks must regularly compare realised default rates with estimated PDs for each grade and be able to demonstrate that the realised default rates are within the expected range for that grade. Banks using the advanced IRB approach must complete such analysis for their estimates of LGDs and EADs. Such comparisons must make use of historical data that are over as long a period as possible. The methods and data used in such comparisons by the bank must be clearly documented by the bank. This analysis and documentation must be updated at least annually*.
- §502: *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. Banks' internal assessments of the performance of their own rating systems must be based on long data histories, covering a range of economic conditions, and ideally one or more complete business cycles*.
- §503: *Banks must demonstrate that quantitative testing methods and other validation methods do not vary systematically with the economic cycle. Changes in methods and data (both data sources and periods covered) must be clearly and thoroughly documented*.
- §504: *Banks must have well-articulated internal standards for situations where deviations in realised PDs, LGDs and EADs from expectations become significant enough to call the validity of the estimates into question. These standards must take account of business cycles and similar systematic variability in default experiences. Where realised values continue to be higher than expected values, banks must revise estimates upward to reflect their default and loss experience*.
- §505: *Where banks rely on supervisory, rather than internal, estimates of risk parameters, they are encouraged to compare realised LGDs and EADs to those set by the supervisors. The information on realised LGDs and EADs should form part of the bank's assessment of economic capital*.

§500 takes a very broad view on validation. A key point to remember here is that it is not only about the estimation of all relevant risk parameters, which constitutes a narrow view on validation, but also about the accuracy and consistency of the rating systems and processes as a whole, which is a much broader perspective on validation. In §501, comparing realized default rates with estimated PDs refers to the quantitative validation activity of *backtesting*. Also important here is that this must be done at least annually, although, as already indicated, banks will typically do this more often. §502 and §505 refer to the idea of *benchmarking*, another key quantitative validation activity. The purpose here is to compare internal models and/or estimates with an external reference model and/or estimates. The aim of benchmarking is to find model deficiencies and identify opportunities for improvement. §503 refers to *model stability*, which requires that the outputs of models and the validation results should not systematically change over time. §504 is basically stating that validation encompasses two activities. First, a validation framework provides a diagnosis; it should tell us if the rating system is robust or not. Second, once a diagnosis has been obtained, the validation framework should also foresee action plans. For example, if the diagnosis is that the rating system is not robust and that the PDs are systematically underestimated, then an action plan is needed to remedy this situation.

A general overview introduced by the BIS 14 Working Paper is given in Exhibit 13.1; see Basel Committee on Banking Supervision (2005a).

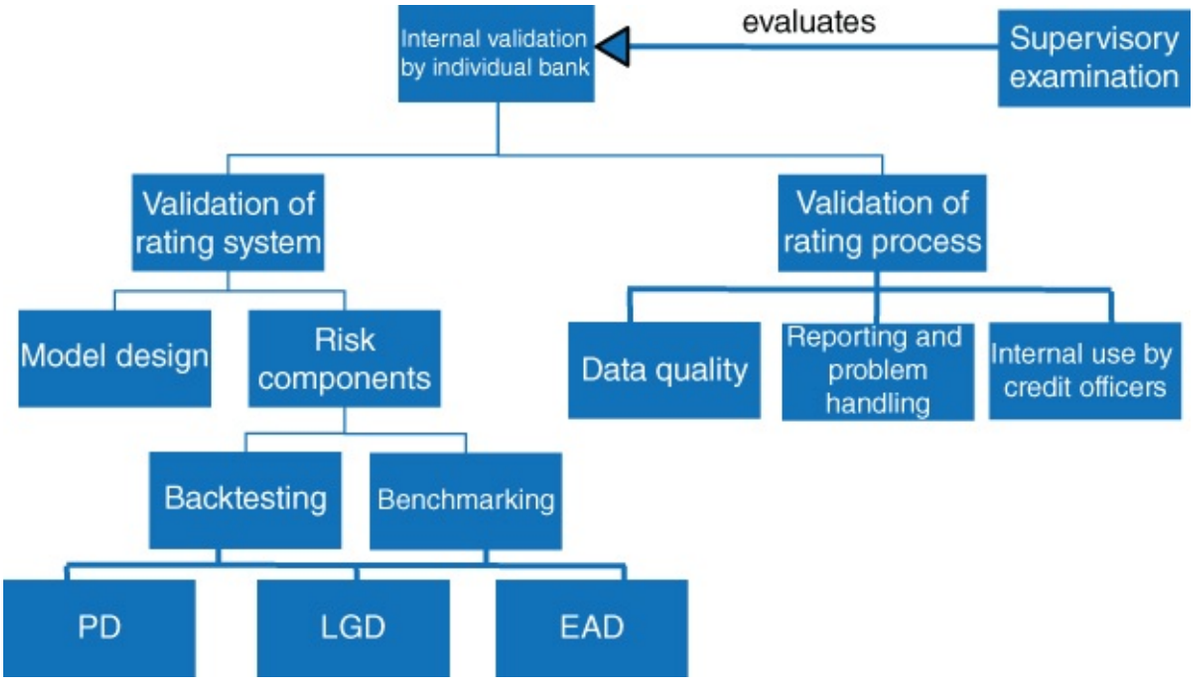


Exhibit 13.1 Validation Framework (Basel Committee on Banking Supervision 2005a)

The supervisor is the person who will evaluate the internal validation conducted by the bank. Usually, a bank will have both a modeling team and a validation or audit team. The purpose of the latter is to closely inspect and validate all PD, LGD, and EAD models, both quantitatively and qualitatively. Many banks adopt the principle of a Chinese wall separation between the modeling and validation teams to enforce an independent, unbiased, and fair evaluation. Validation encompasses both the validation of the rating system as well as the validation of the

rating process. In terms of the rating system, a first activity concerns the validation of the model design. This refers to the definition of the model, its perimeter and scope. Another key task is the validation of the risk components, which include both the backtesting and benchmarking of the PD, LGD, and EAD estimates. In terms of the rating process, a first validation activity concerns the issue of data quality. As said earlier, data is the key ingredient of an analytical PD, LGD, or EAD model, and data quality should thus be optimally safeguarded. A next activity concerns the reporting and problem handling. This basically relates to the reports and documentation available about the various steps and results of the rating process. Finally, it is of key importance that the analytical PD, LGD, and EAD models are used not only for Basel capital calculation purposes, but also for other activities and business purposes. This is the so-called use test, which we clarify in the following.

BASIC CONCEPTS OF VALIDATION

Defining Validation

Let us briefly refresh two key activities of quantitative validation: backtesting and benchmarking. Backtesting refers to comparing ex ante made estimates to ex post realized numbers. In Exhibit 13.2 you can see an example of this.

Rating Category	Estimated PD	Number of Observations	Number of Observed Defaults
A	2%	1000	17
B	3%	500	20
C	7%	400	35
D	20%	100	50

Exhibit 13.2 Example of Validation

Suppose we have four ratings with corresponding PD estimates: 2 percent, 3 percent, 7 percent, and 20 percent. Each rating has a number of observations assigned to it and a number of observed defaults, both of which allow us to calculate the default rate. Backtesting refers to comparing this default rate to the estimated PD. For example, for rating D, backtesting will conduct a statistical test whether the observed default rate of 50 percent is significantly different from the predicted PD of 20 percent. Benchmarking is another key quantitative validation activity, the primary concept here being to compare internal models and or estimates with a reference model and/or estimates. Note that validation is more than just backtesting and benchmarking. In this chapter, we will discuss both quantitative and qualitative validation. In terms of quantitative validation, we will take a closer look at backtesting and benchmarking. In terms of qualitative validation, we will discuss data quality, use test, model design, documentation, corporate governance, and management oversight.

Common Validation Issues

Before we continue the discussion, let us offer some initial observations. Banks employ a wide range of techniques to validate internal ratings, and the techniques used to assess corporate and retail ratings are substantially different. One of the reasons behind this is that, contrary to retail portfolios, in many corporate portfolios data availability is paramount. If there is a data shortage, credit risk model development will be altered, as will validation, an example of which is an expert-based qualitative credit risk model.

Ratings validation is not an exact science. Absolute performance measures are considered counterproductive by some institutions. It is challenging to derive minimum performance benchmarks that PD, LGD, or EAD models need to achieve in order to be considered satisfactory. This typically depends on the data characteristics, portfolio composition, and strategy of the financial institution. Hence, any performance metric reported should be interpreted in terms of its own specific context.

Expert judgment is critical. Data scarcity makes it almost impossible to develop statistically based internal ratings models in some asset classes. This refers to our earlier point concerning the lack of data, which can be observed in low default portfolios (LDPs): An insufficient number of defaulters ultimately prevents construction of a meaningful statistical model. Consult Chapter 8 for further information on LDPs. Thus, the role of credit experts and qualitative valuation becomes important.

Data issues center around both quantity and quality. Default data, in particular, is insufficient to produce robust statistical estimates for some asset classes. For PD modeling, the quality of the data is usually satisfactory, though sometimes the issue of quantity, in terms of number of defaulters, complicates the development of statistical models, as discussed previously. For LGDand EAD modeling, data quality is often a key concern. This is one of the major reasons why LGD and EAD models typically have a low predictive performance.

General Validation Principles

Here, you can see some General Validation Principles put forward by the Basel Committee Validation Subgroup; see Basel Committee on Banking Supervision (2005b).

- *Principle 1: Validation is fundamentally about assessing the predictive ability of a bank's risk estimates and the use of ratings in credit processes.*

  This refers to the ideas of backtesting and use testing.

- *Principle 2: The bank has the primary responsibility for validation;*

  The supervisor does not perform the validation; the bank has this responsibility. The supervisor reviews the validation only.

- *Principle 3: Validation is an iterative process.*

  Validation is not a single-shot, sequential activity. On the contrary, it is a continuous, iterative process, and sometimes quite ad hoc.

- *Principle 4: There is no single validation method.*

Validation is context dependent. This can refer to the type of portfolio, the strategy of the firm, the quality of the data, and so on.

- *Principle 5: Validation should encompass both quantitative and qualitative elements.*

  Quantitative validation refers to backtesting and benchmarking. Qualitative validation refers to data quality, use test, model design, documentation, and corporate governance and management oversight.

- *Principle 6: Validation processes and outcomes should be subject to independent review.*

  This refers to the supervisor reviewing the validation of the bank. Actually, validation is a very difficult activity to optimally organize from an organizational perspective. When adopting a strict split between the modeling and the validation team, where the latter is conceived as the watchdog of the former, then friction may arise between both teams. To be successful it's vital that validation is constructive and focuses on constructive feedback about the developed credit risk models. Validation does not provide a fixed decision but rather a suggestion for further action and study. Hence, both model diagnosticframeworks as well as action plans need to be developed. Finally, validation methods are not allowed to change with the economic cycle unless this is clearly and thoroughly documented.

Developing a Validation Framework

When implementing a validation framework, various things need to be considered. First, the validation needs must be unambiguously diagnosed. What credit risk models must be validated in which portfolios? Then the various validation activities need to be worked out in detail. All of this should be put into a timetable, specifying what validation activity should be conducted by when. Also the various statistical tests and analyses must be clearly defined. Obviously, this will highly depend upon the type of credit risk parameter, for example, PD, LGD, or EAD, to be validated. Finally, the actions must be defined in response to the potential findings. Suppose the validation exercise tells us that the LGD is systematically underestimated. To remedy this, an action plan must be defined. To summarize, the validation policy should clearly specify the why, what, who, how, and when of the whole validation exercise.

QUANTITATIVE VALIDATION

Introduction

In this section, we look closely at quantitative validation. The goal here is to verify how well the various ratings predict default (PD), loss (LGD), exposure, or credit conversion factor (CCF). As mentioned in the Basel II Accord, the burden is on the bank to satisfy its supervisor that a model or procedure has good predictive power and that regulatory capital requirements will not be distorted as a result of its use. Actually, the whole idea of quantitative validation boils down to comparing realized numbers to predicted numbers. It speaks for itself that those numbers will seldom be identical. Hence, various appropriate performance metrics and test statistics must be specified to assist in this comparison. When using these, appropriate cutoffs

such as significance levels must be set. The severity of each cutoff will then determine the severity of the whole validation exercise. A more severe cutoff will result in a more conservative validation which will more readily detect a performance difference. Also, when performing validation, you should think carefully about the split-up of the data. In other words, on what data are you going to calculate the variousvalidation performance metrics and statistics? Let us look closely at this next.

Data Set Split-Up

Exhibit 13.3 displays the various ways of splitting up your data for performing validation.

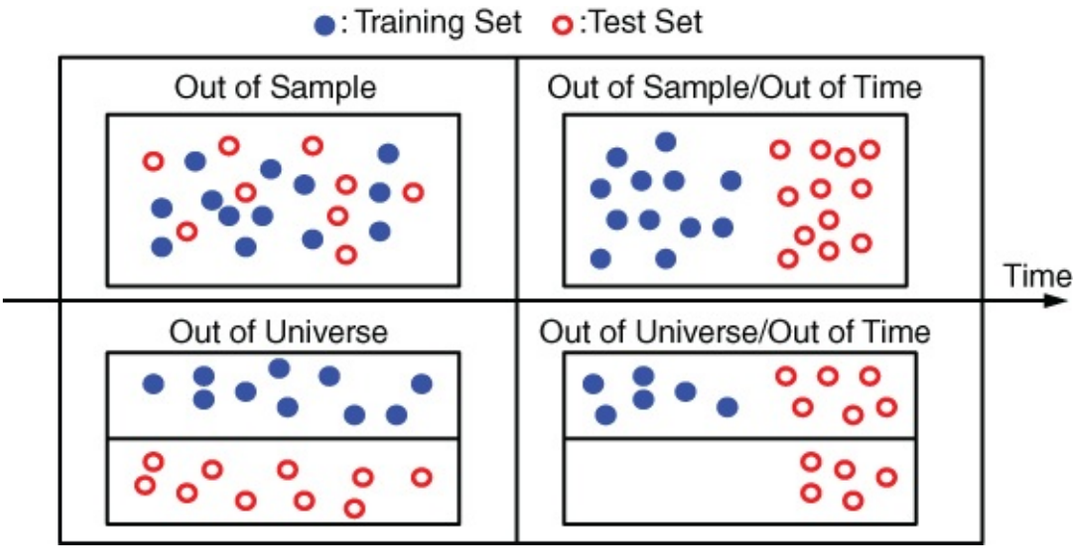


Exhibit 13.3 Data Set Split-Up

A first option to consider is out-of-sample validation. This works by splitting up a data set observed during a particular time frame into a training set and a test set. Remember, the training set is used to develop the model, whereas the test set is used to calculate its performance in an independent way. In this method, the training and test sets overlap in time. This is the most commonly used method of doing validation during the development of PD, LGD, and EAD models. In out-of-sample/out-of-time validation, there is a strict time difference between the test set and the training set. In other words, the test set comes from a subsequent time period. This is typically the type of validation that you will do during model usage, or after the models have been deployed and put into production. Out-of-universe validation entails the validation of a model in another population than that on which it was developed. As an example, think of a PD model developed on U.S. small and medium-sized enterprises (SMEs) being validated on a set of Canadian SMEs. Out-of-Universe validation directly relates to the model perimeter since it tells us how well a model generalizes beyond its original scope. This is very important given the many mergers and acquisitions seen in the financial industry lately, wherein banks suddenly have multiple credit risk models for similar types of portfolios. Finally, the most ambitious validation setup is out-of-universe and out-of-time validation, where a credit risk model is validated on data from another population and subsequent time frame.

Challenges

Various challenges arise when doing quantitative validation. One concerns the sources of variation you are being confronted with. For example, the difference between the predicted PDs and observed default rates can have at least three different causes. A first one concerns random sample variability. This is the variability due to the fact that the predicted PDs have been calculated using a limited sample of observations. A subsequent consideration is external effects, as macroeconomic up- or downturns will have an impact on default rates. Finally, there are internal or endogenous effects due to a change in portfolio composition, strategy shift, or a merger or acquisition, for example. Suppose now that you focus only on sample variation and you know that the PD for a rating grade with independent obligors is 100 basis points. Let us now say that you want to be 95 percent confident that the realized default rate is not more than 20 basis points off from the estimate. When modeling this using a binomial confidence interval, the number of obligors you would need equals or, in other words, Should be 9,500 obligors.

$$n = \left(\frac{1.96\sqrt{PD \cdot (1-PD)}}{0.002} \right)^2 \qquad \textbf{13.1}$$

For n independent obligors with identical PD, the default rate is approximately normally distributed with mean PD and variance $PD(1 - PD)/n$. Thus, a two-sided 95 percent confidence interval for the default rate is given by $PD \pm z(97.5) \cdot \sqrt{PD(1 - PD)/n}$, where $z(97.5)$ is the 97.5th percentile of the standard normal distribution, which is 1.96. Hence, if $PD = 0.01$ and the default rate should not be more than 0.002 off from the PD, it follows that $0.002 = 1.96 \cdot \sqrt{PD(1 - PD)/n}$. Rearranging yields equation (13.1). In retail portfolios, this is usually no problem. However, in corporate portfolios, the number of obligors may be substantially less, thereby increasing the confidence bounds and thus the uncertainty around the PD estimates.

Another complication is the statistical independence assumption that is typically assumed when building credit risk models, which is often untrue in reality. Think about the correlation between defaults and the correlation between PD, LGD, and EAD, for example. Hence, this further complicates the validation exercise. Finally, data availability can also be a general concern, especially in corporate portfolios.

Backtesting PD Models

In this section we discuss backtesting PD models. An overview on backtesting and benchmarking is given in Castermans et al. (2010). We adopt a multilevel perspective on a PD model. (See Exhibit 13.4.) At level 0, we start by checking the data stability. In other words, we measure to what extent the population that was used to construct the PD rating system is similar to the population that is currently being observed. At level 1, we measure how well the PD rating system provides an ordinal ranking of the risk measure considered. Finally, at level 2, the mapping of the rating to a quantitative risk measure (PD) is evaluated. A rating system is considered well-calibrated if the (ex ante) estimated risk measures deviate only marginally from what has been observed ex post.

Calibration	Mapping of a rating to a quantitative risk measure. A rating system is considered well-calibrated if the (ex ante) estimated risk measures deviate only marginally from what has been observed ex post.
Discrimination	Measures how well the rating system provides an ordinal ranking of the risk measure considered.
Stability	Measures to what extent the population that was used to construct the rating system is similar to the population that is currently being observed.

Exhibit 13.4 Backtesting

In the context of PD models, at level 0 the stability of the internal, external, and expert judgment data needs to be backtested. At level 1, the application and behavioral scorecard is evaluated. Both will typically have been constructed using logistic regression models. At level 2, the risk ratings and PD calibration are backtested.

When backtesting PD models, it is common to adopt a traffic light indicator approach to encode the outcomes of the various performance metrics or test statistics. (See Exhibit 13.5.) A green traffic light means that everything is okay and thus the model predicts well and no changes are needed. A yellow light indicates a decreasing performance and early warning that a potential problem may soon arise. An orange light is a more severe warning that a problem is very likely to occur and should be more closely monitored. A red light then indicates a severe problem that needs immediate attention and action. Depending upon the implementation, more or fewer traffic lights can be adopted. Note that within the context of PD modeling, dark green can also be used to refer to the fact that the risk measure is becoming too conservative. The traffic lights can be related to the p-values of a statistical test, for example, as follows:

PD	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av.
	0.26%	0.17%	0.42%_	0.53%_	0.54%	1.36%_	2.46%	5.76%_	8.76%_	20.89%	3.05
DR	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av.
1993	0.00%	0.00%	0.00%	*0.83%*	0.00%	0.76%	*3.24%*	5.04%	**11.29%**	**28.57%**	*3.24*
1994	0.00%	0.00%	0.00%	0.00%	0.00%	0.59%	1.88%	3.75%	7.95%	5.13%	1.88
1995	0.00%	0.00%	0.00%	0.00%	0.00%	*1.76%*	**4.35%**	*6.42%*	4.06%	11.57%	2.51
1996	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.17%	0.00%	3.28%	13.99%	0.78
1997	0.00%	0.00%	0.00%	0.00%	0.00%	0.47%	0.00%	1.54%	7.22%	14.67%	1.41
1998	0.00%	*0.31%*	0.00%	0.00%	*0.62%*	1.12%	2.11%	**7.55%**	5.52%	15.09%	2.83
1999	0.00%	0.00%	0.34%	0.47%	0.00%	*2.00%*	**3.28%**	**6.91%**	*9.63%*	20.44%	**3.35**
2000	*0.28%*	0.00%	**0.97%**	**0.94%**	*0.63%*	1.04%	**3.24%**	4.10%	**10.88%**	19.65%	3.01
2001	*0.27%*	*0.27%*	0.00%	0.51%	**1.38%**	**2.93%**	**3.19%**	**11.07%**	**16.38%**	**34.45%**	**5.48**
2002	**1.26%**	**0.72%**	**1.78%**	**1.58%**	**1.41%**_	*1.58%*	2.00%	**6.81%**	6.86%	**29.45%**	**3.70**
Av.	0.26%	0.17%	0.42%	0.53%	0.54%	1.36%	2.46%	5.76%	8.76%	20.90%	3.05

Exhibit 13.5 Traffic Lights Approach

- A p-value less than 0.01 corresponds to a red light.
- A p-value between 0.01 and 0.05 corresponds to an orange light.
- A p-value between 0.05 and 0.10 corresponds to a yellow light.
- A p-value higher than 0.10 corresponds to a green light.

In Exhibit 13.5 you can see an example of a traffic light indicator approach applied to backtesting PD models at the calibration level, where through years 1993 until 2002 a statistical test (see later) has been run for several rating grades. Note that green corresponds to normal type, yellow to italic, orange to bold, and red to underlined bold. It can be easily seen that from 2001 onwards the calibration is no longer satisfactory, because of the many red lights.

Backtesting PD at Level 0

When validating data stability at level 0, you should check whether internal or external environmental changes will impact the PD classification model. Examples of external environmental changes are new developments in the economic, political, or legal environment; changes in commercial law; or new bankruptcy procedures. Examples of internal environmental changes are alterations of business strategy, exploration of new market segments, or changes in organizational structure.

A two-step approach is suggested as follows:

Step 1: Check whether the population on which the model is currently being used is similar

to the population that was used to develop it.

Step 2: If differences occur in step 1, verify the stability of the individual variables.

For step 1, a population stability index (PSI) or system stability index (SSI) can be calculated. This is also called a deviation index in SAS. It is calculated by contrasting the expected or training e_k and observed or actual population percentages a_k across the various score ranges $k, k = 1, \ldots, K$. In other words, it is calculated as:

$$PSI = \sum_{k=1}^{K} (a_k - e_k) \cdot (\ln(a_k) - \ln(e_k))$$

The following example shows how the PSI can be computed via SAS/IML. After reading the data with score grades from 1 to 10 and expected/training as well as actual/observed percentages, IML computes the PSI columnwise and then sums up these values. Running the code and printing out the values shows that $PSI = 0.059$.

```
DATA data.psi1;
INPUT score expected actual;
DATALINES;
1 0.06 0.07
2 0.1 0.08
3 0.09 0.07
4 0.12 0.09
5 0.12 0.11
6 0.08 0.11
7 0.07 0.1
8 0.08 0.12
9 0.12 0.11
10 0.16 0.15
;
PROC IML;
USE data.psi1;
READ ALL VAR _NUM_ INTO DATA;
PSI_row = (data[,3]-data[,2])
# (LOG(data[,3])- LOG(data[,2]));
PSI = PSI_row[+];
PRINT PSI_row PSI;
QUIT;
```

Important to note is that the percentages reported in the data set are the percentages of the population and thus not default rates. In other words, they add up to 100 percent. Also observe that the PSI is defined in a similar way as the information value, which we discussed in the chapter on data preprocessing. A rule of thumb can then be defined as follows:

- PSI $<$ 0.10: no significant shift (green traffic light)
- 0.10 $\leq$ PSI $<$ 0.25: moderate shift (yellow traffic light)
- PSI $\geq$ 0.25: significant shift (red traffic light)

It is also recommended to monitor the system stability index through time as illustrated in the

next example where another column is added with actual/observed values at a later date, $t+1$. Then expected versus actual in the same period are compared as well as actual in t versus actual in $t+1$.

```
DATA data.psi2;
INPUT score expected actual_t actual_t1;
DATALINES;
1 0.06 0.07 0.06
2 0.1 0.08 0.07
3 0.09 0.07 0.1
4 0.12 0.09 0.11
5 0.12 0.11 0.1
6 0.08 0.11 0.09
7 0.07 0.1 0.11
8 0.08 0.12 0.11
9 0.12 0.11 0.1
10 0.16 0.15 0.15
;
PROC IML;
USE data.psi2;
READ ALL VAR _NUM_ INTO DATA;
PSI_row_e0 = (data[,3]-data[,2])
# (LOG(data[,3])- LOG(data[,2]));
PSI_row_e1 = (data[,4]-data[,2])
# (LOG(data[,4])- LOG(data[,2]));
PSI_row_t = (data[,4]-data[,3])
# (LOG(data[,4])- LOG(data[,3]));
PSI_e0 = PSI_row_e0[+];
PRINT PSI_row_e0 PSI_e0;
PSI_e1 = PSI_row_e1[+];
PRINT PSI_row_e1 PSI_e1;
PSI_t = PSI_row_t[+];
PRINT PSI_row_t PSI_t;
QUIT;
```

The first two values compare the observed or actual population with the expected or training population for two periods. The third one then compares the observed or actual population at time $t+1$ with the population at time t. This allows us to see the evolution of the PSI through time and detect when important changes occur. The same traffic light coding can be used as discussed previously. In the context of credit ratings and scores over time, the composition of the rating grades over time might change not only due to structural changes in the ratings but also simply due to the macroeconomy. In a point-in-time (PIT) rating system, obligors will ceteris paribus be upgraded in an economic upswing and therefore obligors will move into the upper rating grades, whereas the opposite might happen in a downswing. Thus, you have to be careful in diagnosing the reasons for instabilities.

When population instability has been diagnosed, you can then verify the stability of the individual variables. Again, a system stability index can be calculated at the variable level as illustrated in Exhibit 13.6 for the variables income and years client. The reader is encouraged to program this example by herself as an exercise. Note that also histograms and/or t-tests can be handy tools to diagnose variable instability. These are discussed in the chapter on

exploratory data analysis.

	Range	Expected (Training)%	Observed (Actual) at *t*	Observed (Actual) at *t* + 1
Income	0–1,000	16%	18%	10%
	1,001–2,000	23%	25%	12%
	2,001–3,000	22%	20%	20%
	3,001–4,000	19%	17%	25%
	4,001–4,000	15%	12%	20%
	5000+	5%	8%	13%
	S SI reference		0,029	0,208
	SSI *t* – 1			0,238
Years client	Unknown Client	15%	10%	5%
	0–2 years	20%	25%	15%
	2–5 years	25%	30%	40%
	5–10 years	30%	30%	20%
	10+ years	10%	5%	20%
	S SI reference		0,075	0,304
	SSI *t* – 1			0,362

Exhibit 13.6 PSI for Two Variables across Time

Another way of testing stability of a PD model is to include dummy variables for the in-sample and out-of-sample (or out-of-time) periods in a logistic regression model and test whether the dummies and time interactions are statistically significant. In the following example, we use the mortgage data set and define a dummy that is 1 if time < 60 and a dummy that is 2 if time is 60. We then estimate a probit regression using FICO scores, LTVs, and gross domestic product (GDP) as explanatory variables, as well as the time dummies and interactions between the dummies, and the explanatory variables LTV and FICO.

```
DATA tmp_pdstab;
SET data.mortgage;
time_dummy  = 0;
IF time < 60 THEN time_dummy =1;
IF time > 59 THEN time_dummy =2;
RUN;
PROC LOGISTIC DATA=tmp_pdstab DESCENDING ;
CLASS time_dummy (ref='1')/ PARAM = REF ;
MODEL default_time = FICO_orig_time
LTV_orig_time gdp_time time_dummy time_dummy*FICO_orig_time
    time_dummy*LTV_orig_time
/ LINK=Probit ;
STORE OUT=stab1;
RUN;
```

The output (Exhibit 13.7) shows that the time intercept as well as the interactions are not significant. This means that the model is stable across time periods 1–59 and 60. Moreover, the marginal coefficients for FICO and LTV can now be interpreted on a stand-alone basis and are highly significant.

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TMP_PDSTAB
Response Variable	default_time
Number of Response Levels	2
Model	Binary probit
Optimization Technique	Fisher's scoring

Number of Observations Read	622489
Number of Observations Used	622489

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	142575.79	137588.22
SC	142587.13	137667.61
-2 Log L	142573.79	137574.22

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	−1.0883	0.0455	571.5166	<.0001
FICO_orig_time		1	−0.00209	0.000049	1836.5919	<.0001
LTV_orig_time		1	0.00738	0.000352	439.1122	<.0001
gdp_time		1	−0.0805	0.00152	2795.0299	<.0001
time_dummy	2	1	−0.3047	0.6650	0.2100	0.6468
FICO_orig*time_dummy	2	1	0.000081	0.000728	0.0122	0.9119
LTV_orig_*time_dummy	2	1	−0.00250	0.00521	0.2304	0.6312

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	66.9	Somers' D	0.338
Percent Discordant	33.1	Gamma	0.338
Percent Tied	0.0	Tau-a	0.016
Pairs	9205923298	c	0.669

Exhibit 13.7 Stability Test with Interactions

Backtesting PD at Level 1

We now climb up one level in the credit risk model architecture and validate the discriminatory power of a scorecard or a PD model. It is recommended to first have a look at the scorecard and the model itself. For example, what was the logic behind the model used?

Were there any assumptions made such as independence or normality? Also important is to verify the sign of the regression coefficients. Are the signs as anticipated? Are there any unexpected signs? Suppose that your scorecard or your PD model tells you that a higher debt ratio corresponds to a better credit score or a lower PD. This is clearly counterintuitive and needs to be further investigated since no one will be prepared to use a scorecard with this pattern. It is important to inspect all the p-values and the model significance. Also, the input selection procedure adopted and any remaining multicollinearity issues need to be clarified. Finally, the various data preprocessing activities such as missing values, outlier handling, and coarse classification need to be verified.

Next, the *discrimination performance* should be considered at level 1. Two key performance metrics here are the *receiver operating characteristic (ROC)* curve and the *area underneath (area under ROC, or AUROC, or AUC)*, together with the *cumulative accuracy profile (CAP)* and the *accuracy ratio (AR)*. They can be derived in various ways. One way is to order the predicted scores from the riskiest score to the least risky. For the validation data set (in-sample, or better, out-of-sample / out-of-time), you then compute cumulative percentages of defaulters for the ordered scores. If this function is plotted in a diagram, it is called the CAP curve or Gini curve (as it is similar to a Gini coefficient in terms of computation).

As a metric measure, you can compute the AR. For this, you first compute the so-called ideal CAP curve, which is obtained if all defaulters are clustered in the riskiest scores. The AR then becomes the ratio of the area between the real CAP curve and the diagonal and the ideal CAP curve and the diagonal; that is,

$$AR = \frac{\text{Area between real CAP curve and diagonal}}{\text{Area between ideal CAP curve and diagonal}}$$

This gives a theoretical range of the AR between zero and one. The higher the value, the closer the real curve is to the ideal curve. The basic idea is that the better the discriminatory power of the score for the defaulters and nondefaulters, the more defaulters should cluster at the risky scores and the fewer should be in the less risky scores.

The ROC is computed in the following way. First, the scores are also ordered according to their riskiness. You then move along the scores and compute:

- The proportion of correctly classified defaulters among all defaulters (the so-called hit rate or sensitivity), and
- The proportion of incorrectly classified nondefaulters (the so-called false alarm rate or "1 − specificity").

A plot of these values yields the ROC. The area under the ROC gives the metric measure AUROC. It is important to know that AUROC is actually a transformation of AR and can be computed by the relation

$$AR = 2(AUROC - 0.5)$$

Both measures give the same information. However, AUROC has a minimum value of 0.5

(instead of 0 for AR) and therefore usually looks more optimistic. In SAS, the ROC curve, AUROC, and AR can be computed via PROC LOGISTIC using the option ROC as shown in the following example with the mortgage data set. We first divide the data set into an estimation data set (using time up to 59) and an (out-of-time) validation data set (using time 60). We then estimate three probit models: one with LTV only, one with LTV and FICO scores, and one with LTV, FICO scores, and GDP. Using the OUTMODEL option in the PROC LOGISTIC statement, the model output is stored. After estimating each model, this output is called into PROC LOGISTIC and the validation data set is scored using the SCORE command. Using the OUTROC option, the values required for the ROC analysis are stored. Next, the predicted PDs of the three models are merged and transformed into scores. Using these three sets of scores, three probit models are estimated for the validation data set, and the ROCs are compared. (See Exhibits 13.8 and 13.9.)

The LOGISTIC Procedure

ROC Model: model1

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	537.247	537.468
SC	544.235	551.444
-2 Log L	535.247	533.468

ROC Association Statistics							
ROC Model	Mann-Whitney				Somers' D	Gamma	Tau-a
	Area	Standard Error	95% Wald Confidence Limits				
model1	0.5718	0.0484	0.4770	0.6667	0.1437	0.1584	0.00154
model2	0.6362	0.0403	0.5572	0.7152	0.2724	0.2725	0.00291
model3	0.6366	0.0402	0.5579	0.7153	0.2732	0.2733	0.00292

ROC Contrast Coefficients		
ROC Model	Row1	Row2
model1	−1	−1
model2	1	0
model3	0	1

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = model1	2	6.5944	0.0370

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
model2 - model1	0.0644	0.0373	−0.00875	0.1375	2.9773	0.0844
model3 - model1	0.0648	0.0381	−0.00979	0.1394	2.8992	0.0886

Exhibit 13.8 Out-of-Sample Association Statistics

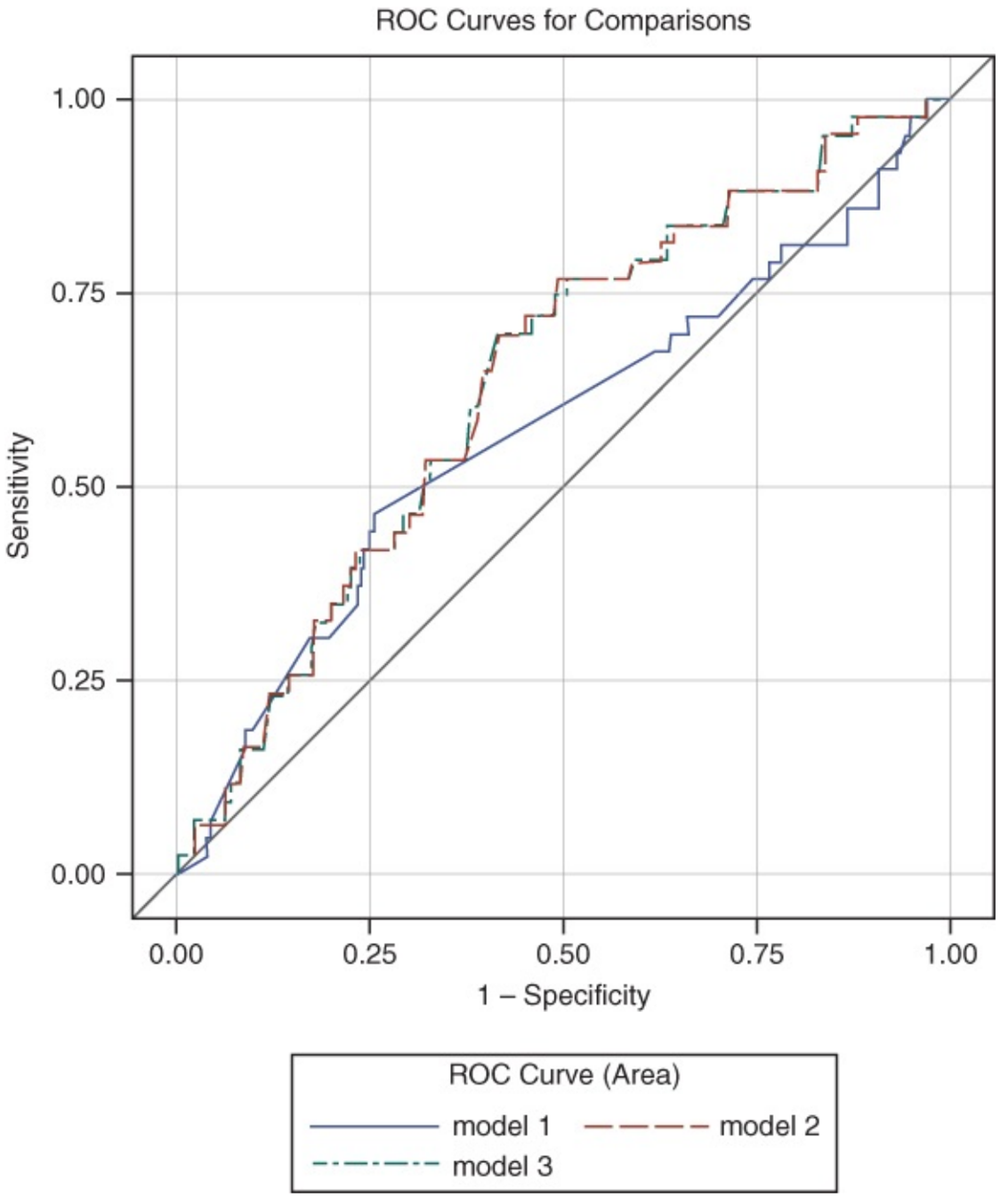


Exhibit 13.9 Out-of-Time ROC Curves

```
DATA tmp_pdvali1;
SET data.mortgage;
IF time < 60;
RUN;
DATA tmp_pdvali2;
SET data.mortgage;
IF time> 59;
RUN;
PROC SORT data = tmp_pdvali1;
BY id time;
RUN;
PROC SORT data = tmp_pdvali2;
BY id time;
```

```
RUN;
PROC LOGISTIC DATA=tmp_pdvali1 DESCENDING  OUTMODEL=model1;
MODEL default_time = /*FICO_orig_time */
LTV_orig_time  /*gdp_time*/
/ LINK=Probit ;
RUN;
PROC LOGISTIC /*DATA=model*/ DESCENDING  INMODEL=model1;
*MODEL default_time = FICO_orig_time
/*LTV_orig_time  gdp_time*/
/ LINK=Probit ;
SCORE DATA = tmp_pdvali2 OUT= pred1 OUTROC=outroc1;
RUN;
PROC LOGISTIC DATA=tmp_pdvali1 DESCENDING  OUTMODEL=model2;
MODEL default_time = FICO_orig_time
LTV_orig_time /* gdp_time*/
/ LINK=Probit ;
RUN;
PROC LOGISTIC /*DATA=model*/ DESCENDING  INMODEL=model2;
*MODEL default_time = FICO_orig_time
/*LTV_orig_time  gdp_time*/
/ LINK=Probit ;
SCORE DATA = tmp_pdvali2 OUT= pred2 OUTROC=outroc2;
RUN;
PROC LOGISTIC DATA=tmp_pdvali1 DESCENDING  OUTMODEL=model3;
MODEL default_time = FICO_orig_time
LTV_orig_time gdp_time
/ LINK=Probit ;
OUTPUT OUT=inpred3 predicted=Predicted3 Xbeta=xbeta3;
RUN;
PROC LOGISTIC /*DATA=model*/ DESCENDING  INMODEL=model3;
*MODEL default_time = FICO_orig_time
/*LTV_orig_time  gdp_time*/
/ LINK=Probit ;
SCORE DATA = tmp_pdvali2 OUT= pred3 OUTROC=outroc3;
RUN;
DATA pred2 ;
SET pred2 (RENAME = (P_1=P_2));
RUN;
DATA pred3 ;
SET pred3 (RENAME = (P_1=P_3));
RUN;
DATA pred;
MERGE pred1 pred2 pred3;
BY id time;
xbeta1 = PROBIT(P_1);
xbeta2 = PROBIT(P_2);
xbeta3 = PROBIT(P_3);
RUN;
ODS GRAPHICS ON;
PROC LOGISTIC DATA=PRED PLOTS=ROC(ID=id);
MODEL default_time(EVENT='1') = xbeta1 xbeta2 xbeta3 / NOFIT;
ROC 'model1' xbeta1;
ROC 'model2' xbeta2;
ROC 'model3' xbeta3;
ROCCONTRAST REFERENCE('model1') / ESTIMATE E;
```

```
STORE OUT=roc1;
RUN;
ODS GRAPHICS OFF;
```

SAS computes the AUROCs for the three models, including standard errors and confidence intervals (for the formulas we refer to the SAS manual; see SAS Institute Inc. (2015)). The AR measure is given as Somers' D, which has actually been a very popular measure used by statisticians for a long time; see Agresti (1984). You can see that the discriminatory performance increases with more included variables; in particular FICO adds much power. Moreover, the table computes statistical tests for a cross-wise AUROC comparison; see the SAS manual (SAS Institute Inc. 2015). While the AUROCs of model 2 and model 3 are significantly different from that of model 1, those of model 3 and 2 are not significantly different from each other. Similarly, the ROC curves show that models 2 and 3 are considerably better in terms of discrimination than model 1.

When exploring discriminatory power, you should keep a few caveats in mind. First, as the defaults are random events, the outcomes of the discriminatory power measures are random variables as well. Hence, comparisons should be done using statistical tests and confidence intervals, as shown previously. Second, it can be shown that the measures are *portfolio dependent*; see Hamerle, Rauhmeier, and Rösch (2003) and Blochwitz et al. (2005). This means that the outcomes depend on the PDs of the portfolio under consideration. When you compare portfolios with different PDs (and most portfolios *will* have different PDs unless they are exactly identical), the expectations of the measures will be different, and therefore also the outcomes, just because the portfolios are different. Similarly, when comparing AR values for a portfolio over time, the values *will* be different just because the PDs of the portfolio have changed. Therefore, it only makes sense to compare values for the same portfolio for the same time across different models. Moreover, as the values depend on the portfolio, requiring minimum levels, such as “AR should be at least 0.5,” or threshold levels at which the performance is good and alike, does not make sense. Third, the distributions of the measures are typically derived under the assumption of independence. For correlated defaults, it might be more difficult to conduct statistical tests. This, however, is not part of this book and is left for future research.

The importance of both the AR and the AUROC has been stressed by the Basel committee, as you can see in this quote from the BIS 14 working paper (see Basel Committee on Banking Supervision 2005a, 32):

> The group has found that the Accuracy Ratio (AR) and the ROC measure appear to be more meaningful than the other above-mentioned indices because of their statistical properties.

The BIS 14 paper then also provides a benchmark range as follows:

> Practical experience shows that the Accuracy Ratio has tendency to take values in the range 50% and 80%. However, such observations should be interpreted with care as they seem to strongly depend on the composition of the portfolio and the numbers of defaulters in the sample.

The paper therefore acknowledges the previous caveat of portfolio dependence. It is important to be aware that a benchmark is always relative and depends on the characteristics of the portfolio, application, and data quality.

Backtesting PD at Level 2

Let us now move to level 2 of our credit risk model architecture, which is the level of the calibration (see Exhibit 13.4). This is probably the most important level, as this gives us the PDs that we use to calculate the capital requirements. Key questions that should be answered here are:

- Do the ratings properly reflect the obligor's default risk?
- Are the credit characteristics of obligors in the same rating sufficiently homogeneous?
- Are there enough ratings to allow for an accurate and consistent estimation of default risk per rating?
- Are the assigned/estimated PDs in line with ex post observed default rates?

When backtesting PD at level 2, you should investigate whether the ratings provide a correct ordinal ranking of risk, and a correct cardinal measure of risk. In terms of the former, it should be verified whether the default rates (DRs) are properly ranked through the ratings. In other words, DR(A) < DR(B) < DR(C). In terms of cardinal measure of risk, the calibrated PD should be as close as possible to the realized default rates. Various test statistics can be used to compare the estimated PDs to the realized default rates. The most popular are the binomial test, the Hosmer-Lemeshow test, the Vasicek (ASRF) one-factor model, and the normal test. All these tests suffer from a couple of complications such as an insufficient number of defaults, the fact thatdefaults are typically correlated, and the issue of choosing an appropriate significance level. Since each statistical test has its shortcomings, they are typically used as early warning indicators. Note that the impact of the rating philosophy, point-in-time (PIT) or through-the-cycle (TTC), is also important to consider.

Brier Score

We start by defining a performance measure at level 2, the Brier score, which is defined as

$$BS = \frac{1}{n}\sum_{i}^{n}(\hat{\pi}_i - d_i)^2$$

where $\hat{\pi}_i$ is the estimated PD and d_i is the observed default ($d_i = 1$) or nondefault ($d_i = 0$) for obligor i. Obviously, the Brier score is always bounded between 0 and 1, and lower values are to be preferred. A higher discrimination in terms of higher granularity of rating grades and/or PD estimates might help to decrease the value if obligors are not homogeneous within a grade. Also, better calibration in terms of getting the right PDs will decrease the value. Note, however, that this score is only occasionally used in the industry. The following PROC IML code uses the out-of-time predicted PDs from the previous example and calculates the squared differences between the defaults and the predictions for each model, and then computes the

average of each. (See Exhibit 13.10.)

brier_1
0.0057251

brier_2
0.0057741

brier_3
0.0055243

Exhibit 13.10 Brier Scores

```
ODS GRAPHICS ON;
DATA brier;
SET pred (KEEP=default_time P_1 P_2 P_3);
RUN;
QUIT;
PROC IML;
USE brier;
READ ALL VAR _NUM_ INTO data;
brier_row_1 = (data[,2]-data[,1])##2;
brier_row_2 = (data[,3]-data[,1])##2;
brier_row_3 = (data[,4]-data[,1])##2;
brier_1 = brier_row_1[+]/NROW(data);
brier_2 = brier_row_2[+]/NROW(data);
brier_3 = brier_row_3[+]/NROW(data);
PRINT brier_1, brier_2, brier_3;
QUIT;
ODS GRAPHICS OFF;
```

As the values show, the best model fit is given by model 3 (which includes the GDP), which seems to be better calibrated than the other models using LTV and FICO only.

Binomial Test

Another (very popular) test for backtesting PD calibration is the binomial test. Three key assumptions of a binomial experiment are:

1. It should be an experiment with only two outcomes, success or failure.
2. It should be repeated multiple times with the same outcome probabilities.
3. There should be independence between the outcomes of the individual experiments.

Usually, when dealing with rating grades, two of these requirements are fulfilled, as we have only two outcomes, default or nondefault, and multiple obligors are considered with the assumption of identical PDs within a rating grade. Due to the correlation with the default behavior, the independence assumption is often not fulfilled. Hence, the binomial test will be used as a heuristic or early-warning indicator only and deliver results that are too

conservative. The null hypothesis, H_0, states that the PD of a rating grade, call it π_0, is correct. The alternative hypothesis can be two-sided or one-sided. From a regulatory perspective, it is important that capital is not underestimated, so let us make the alternative hypothesis, H_1, the PD of the rating is underestimated. As already explained, to use the binomial test, we are assuming that the default events are uncorrelated. Given a confidence level α (for example, 99 percent), the null hypothesis H_0 is rejected if the number of defaulters $d = \sum_{i=1}^{n} d_i$ in the rating is greater than or equal to k^*, which is obtained as follows: It is the minimum k such that the cumulative probability, as quantified using the binomial distribution, of observing between k and d defaulters is less than or equal to one minus α. The probability distribution function (PDF) of the binomial distribution is given by

$$P\left(\sum_{i=1}^{n} D_i = d|n, \pi_0\right) = P(d|n, \pi_0) = \binom{n}{d} \pi_0^d (1-\pi_0)^{n-d}, \quad d = 0, \ldots, n$$

Therefore,

$$k^* = \min\left\{ k \left| \sum_{j=k}^{n} \binom{n}{j} \pi_0^j (1-\pi_0)^{n-j} \leq 1-\alpha \right. \right\}$$

The central limit theorem can now be used for large n, and when $n\pi_0 > 5$ and $n(1-\pi_0) > 5\%$. The number of defaulters D can then be modeled as a normal distribution with expected value $n\pi_0$ and variance $n\pi_0(1-\pi_0)$ under H_0. We can now look for the k^* value such that the probability that D is less than or equal to k^* equals α. Hence, we have

$$P\left(Z \leq \frac{k^* - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}\right) = \alpha$$

with Z following a standard normal distribution. The critical value k^* can then be obtained as

$$k^* = \Phi^{-1}(\alpha)\sqrt{n\pi_0(1-\pi_0)} + n\pi_0$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function (CDF). In terms of a maximum observed default rate p^*, we simply divide k^* by n and have

$$p^* = \Phi^{-1}(\alpha)\sqrt{\frac{\pi_0(1-\pi_0)}{n}} + \pi_0$$

To summarize, we can reject H_0 at significance level α if the observed default rate is higher than p^*.

We apply the binomial test to a simple exercise using the mortgage data and the FICO score as a predictor. In the estimation and the validation data set, we create three rating grades with similar numbers of observations as follows:

- Grade 1: FICO $\geq$ 713

- Grade 2: 648 ≥ FICO < 713
- Grade 3: FICO < 648

using the following code:

```
DATA fico_class1;
SET tmp_pdvali1;
fico_class = 0 ;
IF FICO_orig_time>= 713        THEN  fico_class = 1;
IF 713> FICO_orig_time>= 648  THEN  fico_class = 2;
IF 648> FICO_orig_time         THEN  fico_class = 3;
RUN;
DATA fico_class2;
SET tmp_pdvali2;
fico_class = 0 ;
IF FICO_orig_time>= 713        THEN  fico_class = 1;
IF 713> FICO_orig_time>= 648  THEN  fico_class = 2;
IF 648> FICO_orig_time         THEN  fico_class = 3;
RUN;
```

Using PROC FREQ, we then compute for the estimation sample the in-sample default rates for each of the FICO grades and for each grade the respective out-of-sample default rate. Out-of-sample confidence intervals are also computed and a binomial test is requested. The confidence intervals are computed using the exact binomial test and the normal approximation.

```
ODS GRAPHICS ON;
PROC FREQ DATA = fico_class1;
TABLES fico_class * default_time / NOCOL NOPERCENT NOCUM;
RUN;
ODS GRAPHICS OFF;
```

The in-sample default rates are 1.37 percent, 2.54 percent, and 2.39 percent, respectively, as shown in Exhibit 13.11, computed from a total number of 614,485 observations and the three rating grades with roughly equal size. We set as H_0 for the out-of-sample test a value of 1 percent. The tables in Exhibit 13.12 show the output for each rating grade separately.

The FREQ Procedure

Table of fico_class by default_time			
fico_class	default_time		
	0	1	Total
1	193938 98.63	2694 1.37	196632
2	199243 97.46	5189 2.54	204432
3	206189 96.61	7232 3.39	213421
Total	599370	15115	614485

Exhibit 13.11 In-Sample Default Rates

The FREQ Procedure

fico_class=1

Binomial Proportion	
default_time = 1	
Proportion	0.0028
ASE	0.0011
95% Lower Conf Limit	0.0007
95% Upper Conf Limit	0.0049
Exact Conf Limits	
95% Lower Conf Limit	0.0011
95% Upper Conf Limit	0.0058

fico_class=1

Test of H_0: Proportion = 0.01	
ASE under H_0	0.0020
Z	−3.5936
One-sided Pr < Z	0.0002
Two-sided Pr > |Z|	0.0003

fico_class=2

Binomial Proportion	
default_time = 1	
Proportion	0.0039
ASE	0.0012
95% Lower Conf Limit	0.0015
95% Upper Conf Limit	0.0062
Exact Conf Limits	
95% Lower Conf Limit	0.0019
95% Upper Conf Limit	0.0071

fico_class=2

Test of H_0: Proportion = 0.01	
ASE under H_0	0.0020
Z	–3.1427
One-sided Pr < Z	0.0008
Two-sided Pr > |Z|	0.0017

fico_class=3

Binomial Proportion	
default_time = 1	
Proportion	0.0089
ASE	0.0017
95% Lower Conf Limit	0.0055
95% Upper Conf Limit	0.0123
Exact Conf Limits	
95% Lower Conf Limit	0.0058
95% Upper Conf Limit	0.0130

fico_class=3

Test of H_0: Proportion = 0.01	
ASE under H_0	0.0018
Z	–0.6145
One-sided Pr < Z	0.2695
Two-sided Pr > |Z|	0.5389

Exhibit 13.12 Out-of-Time Default Rates and Tests

```
PROC SORT DATA = fico_class2;
BY fico_class;
RUN;
ODS GRAPHICS ON;
PROC FREQ DATA = fico_class2;
BY fico_class;
TABLES default_time
/ BINOMIAL (LEVEL='1' P=0.01) ALPHA=.05;
RUN;
ODS GRAPHICS OFF;
```

The realized default rates are given as the proportions 0.0028, 0.0039, and 0.0089. These values are considerably lower than those for the estimation sample. This is due to the upswing of the economic cycle in the final observation period, as already shown in the chapter on PD models. Whether this difference is just random or systematic can be checked using the

confidence intervals or the binomial test. For grade 1 (where the in-sample default rate was about 1 percent), the 95 percent confidence limits are 0.0011 and 0.0058 (note also the slight difference between the exact and the asymptotic limits), which does not include 0.01 (or 1 percent). Thus, the test $H_0 = 0.01$ is rejected with one-sided and two-sided p-values < 0.001. Note, however, that the realized default rates are *lower* than the PD under H_0. That is, from a statistical perspective, H_0 is rejected. From a regulatory perspective, the value under H_0 might seem to be rather conservative and therefore the *upper* one-sided hypothesis would not be rejected. The results are quite similar for the second grade (for testing a different grade specific PD estimates in H_0 you should repeat the test with other values in the P = option in PROC FREQ). For grade 3, the null hypothesis is not rejected, although it seems that the PD estimates based on the FICO score alone are miscalibrated (compared to their historical PD estimate) and additional time-varying (macroeconomic) variables should be included, in order to provide better out-of-time forecasts.

Hosmer-Lemeshow Statistic

As the binomial backtests are usually done on a grade-by-grade basis, another test can be applied that tests all rating grades simultaneously. This is called the *Hosmer-Lemeshow statistic*; see Hosmer and Lemeshow (2000). It compares defaults and predictions for ordered groups. First, the observations are sorted in increasing order of their estimated default probability. The observations are then divided into approximately 10 groups according to a specific scheme. For technical details, we refer to the SAS manual; see SAS Institute Inc. (2015). Next, the following statistic is computed:

$$\chi^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - n_g \hat{\pi}_g)^2}{n_g \hat{\pi}_g (1 - \hat{\pi}_g)}$$

where G is the number of groups (or rating grades in our case), O_g is the observed total frequency of defaults in group g, n_g is the total number of observations in group g, and $\hat{\pi}_g$ is the average estimated predicted probability of default for the gth group. The statistic is χ^2 distributed with $g - 2$ degrees of freedom. Large values of χ^2_{HL} (and corresponding small p-values) indicate a lack of fit of the model. It can be computed using the LACKFIT option in PROC LOGISTIC as shown in the following code.

```
ODS GRAPHICS ON;
ODS OUTPUT LackFitPartition = LackFitPartition ;
PROC LOGISTIC DATA=pred ;
MODEL default_time(EVENT='1') = xbeta3 /LACKFIT RSQUARE;
EFFECTPLOT /NOOBS;
RUN;
ODS GRAPHICS OFF;
```

We compute the statistic for model 3 from the earlier ROC comparison example. The output (Exhibit 13.13) shows the 10-group partition for the test with the computed expected and observed frequencies from which the final statistic is evaluated. It is not significant at the 10 percent level, which shows that the model fit is okay. As you can see from the observed and

expected numbers in the table, the fractions of defaults divided by the total numbers increase only slightly through the groups.

The LOGISTIC Procedure

Partition for the Hosmer and Lemeshow Test					
Group	Total	default_time = 1		default_time = 0	
		Observed	Expected	Observed	Expected
1	800	1	1.87	799	798.13
2	804	4	2.39	800	801.61
3	801	2	2.78	799	798.22
4	809	2	3.21	807	805.79
5	798	2	3.59	796	794.41
6	800	4	4.07	796	795.93
7	800	8	4.63	792	795.37
8	801	6	5.39	795	795.61
9	801	7	6.47	794	794.53
10	790	7	8.61	783	781.39

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.7581	8	0.6743

Exhibit 13.13 Hosmer-Lemeshow Statistics

You can also check calibration graphically, as is done in the following code. (See Exhibit 13.14.) The ODS OUTPUT statement in PROC LOGISTIC in the preceding code creates the Partition Table with Expected and Observed Frequencies as a SAS data set. We then compute the relative frequencies and plot them against each other in order to see how well the 10 classes are calibrated. Ideally, they should all lie close to the diagonal. Here again the fit is rather moderate. The graphical analysis also helps to discover those PD regions with better and poorer fit.

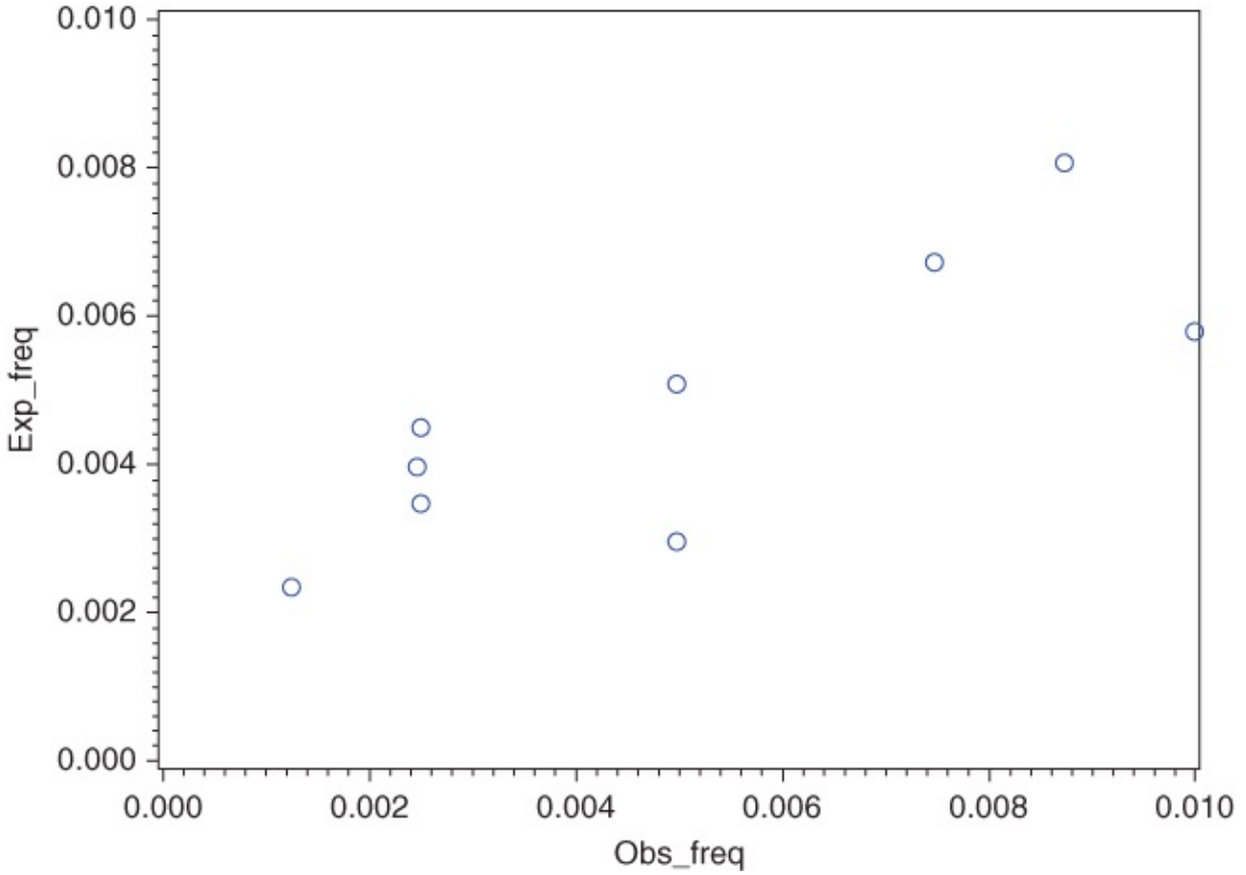


Exhibit 13.14 Calibration Diagram

```
DATA Calibration;
SET LackFitPartition;
Obs_freq = EventsObserved/ Total;
Exp_freq = EventsExpected/ Total;
RUN;
ODS GRAPHICS ON;
SYMBOL1 INTERPOL=NONE
VALUE=CIRCLE
CV=BLUE
WIDTH=4 HEIGHT=4;
PROC GPLOT DATA = Calibration;
PLOT Exp_freq * Obs_freq
/HAXIS=0 TO 0.01 BY 0.002 VAXIS = 0 TO 0.01 BY 0.002;
RUN;
ODS GRAPHICS OFF;
```

Similarly, if we plot the predicted PDs, produced by the LACKFIT option in the MODEL statement of PROC LOGISTIC, against the linear predictor (the score, named xbeta3) as shown in Exhibit 13.15, we see that there is only a moderate increase when moving up the linear predictor. Additional variables might be needed to develop a better model that might deliver better forecasting results.

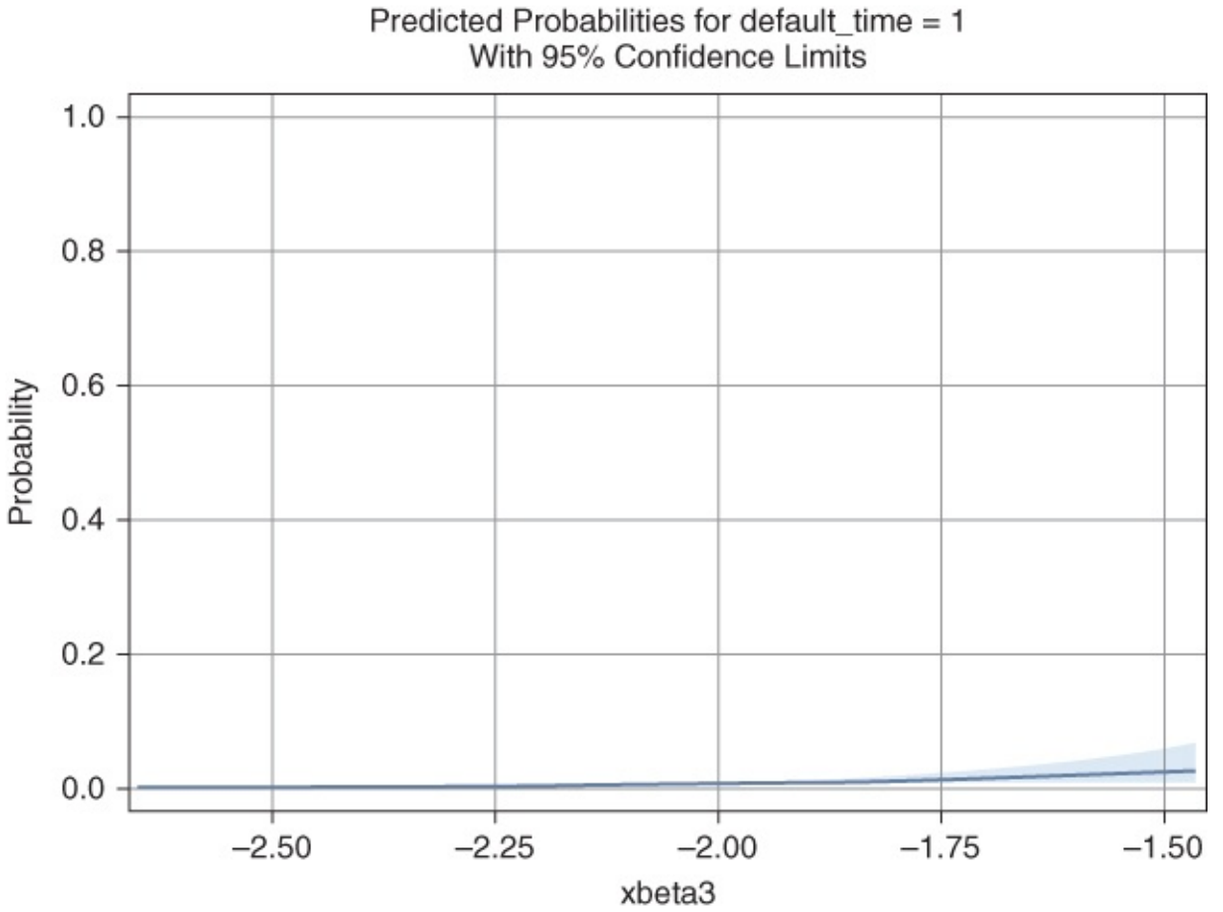


Exhibit 13.15 Out-of-Sample PD Predictions

Binomial Test with Correlation

The previous tests assume uncorrelated defaults. As we know from the chapter about default correlations, the distribution of defaults becomes wider when defaults are correlated. As a correlation model, it is convenient to use the Basel one-factor model from the earlier chapter. The unconditional probability distribution for the number of defaults is given as:

$$P(D = d) = \int_{-\infty}^{\infty} \binom{n}{d} CPD(x)^d (1 - CPD(x))^{n-d} \phi(x) dx, d = 0, \ldots, n$$

where:

$$CPD(X) = \Phi\left(\frac{c - \sqrt{\rho}X}{\sqrt{1-\rho}}\right)$$

and $c = \Phi^{-1}(PD)$. The critical value for $PD = \pi_0$ is then given by the respective quantile against which the observed default rate is compared. If the observed default rate is higher, H_0 is rejected (one-sided). The following code computes and plots the critical values for $\alpha = 0.99$ and 100 obligors as a function of the PD and the correlation ρ. As can be seen, the critical values sharply increase with the correlation. (See Exhibit 13.16.)

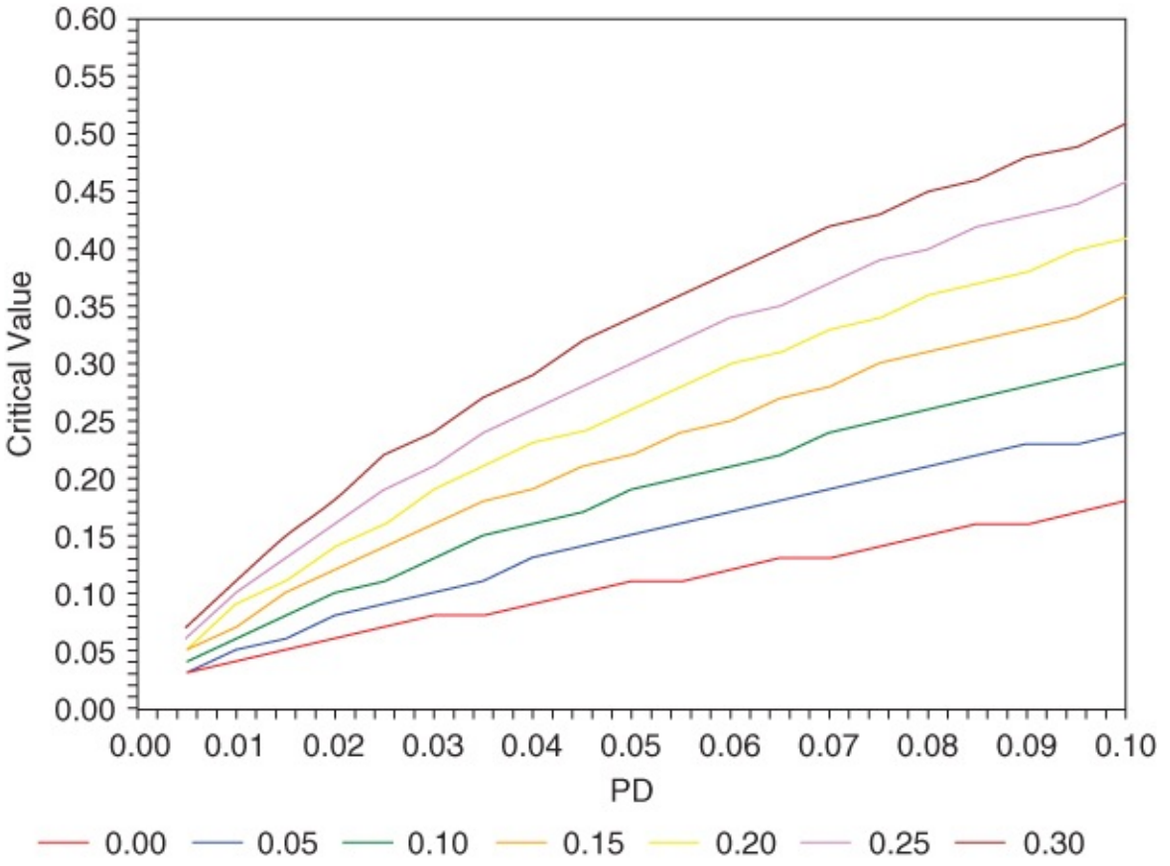


Exhibit 13.16 Critical Values under Extended Binomial Model with Various Correlations

```
ODS GRAPHICS ON;
PROC IML SYMSIZE=10000000 WORKSIZE = 10000000;
/*Number of Obligors*/
N           = 100;
/*Vary Asset Correlation*/
DO rho = 0 TO 0.3 BY 0.05;
/*Vary PD*/
DO  p = 0.005 TO 0.1 BY 0.005;
START fun(x) GLOBAL(k,p,N,rho);
pi = CONSTANT('PI');
/*Compute the Integral*/
CPD = PROBNORM((1/SQRT(1-rho)) * (PROBIT(p)-SQRT(rho)*x));
v   = PROBBNML(CPD,n,k) * (exp(-(x*x)/2))/(SQRT(2*pi));
RETURN(v);
FINISH fun;
k=0;
/*Increase k up to desired Quantile*/
DO UNTIL(z>0.99);
/*Call QUADRATURE */
a   = {.M .P};
eps = 1.34E-15;
CALL QUAD(z,“fun”,a,eps);
quantile=k;
default_rate= quantile/N;
k=k+1;
prob=z;
END;
```

```
out = out//(N||rho||p||quantile||default_rate||z||prob);
END;
END;
CREATE  crit_reg FROM out;
APPEND FROM out;
QUIT;
GOPTIONS RESET=GLOBAL GUNIT=PCT NOBORDER CBACK=WHITE
COLORS=(BLACK BLUE GREEN RED)
FTITLE=SWISSB FTEXT=SWISS HTITLE=3 HTEXT=3;
SYMBOL1 COLOR=RED   INTERPOL=JOIN
WIDTH=3 VALUE=NONE  HEIGHT=0;
SYMBOL2 FONT=MARKER VALUE=NONE
COLOR=BLUE  INTERPOL=JOIN
WIDTH=3 HEIGHT=0;
SYMBOL3 COLOR=GREEN INTERPOL=JOIN
VALUE=NONE WIDTH=3  HEIGHT=0;
SYMBOL4 COLOR=ORANGE INTERPOL=JOIN
WIDTH=3 VALUE=NONE  HEIGHT=0;
SYMBOL5 COLOR=YELLOW INTERPOL=JOIN
WIDTH=3 VALUE=NONE  HEIGHT=0;
SYMBOL6 COLOR=VIOLET INTERPOL=JOIN
WIDTH=3 VALUE=NONE  HEIGHT=0;
SYMBOL7 COLOR=BROWN INTERPOL=JOIN
WIDTH=3 VALUE=NONE  HEIGHT=0;
AXIS1 ORDER=(0 TO 0.1 BY 0.01) OFFSET=(0,0)
LABEL=('PD')
MAJOR=(height=1) MINOR=(height=1)
WIDTH=3;
AXIS2 ORDER=(0 TO 0.6 BY 0.05) OFFSET=(0,0)
LABEL=('critical value')
MAJOR=(HEIGHT=1) MINOR=(HEIGHT=1)
WIDTH=3;
LEGEND1 LABEL=NONE;
ODS GRAPHICS ON;
PROC GPLOT DATA=crit_reg;
PLOT COL5*COL3 = COL2/ OVERLAY LEGEND=LEGEND1
HAXIS=AXIS1
VAXIS=AXIS2;
RUN;
QUIT;
ODS GRAPHICS OFF;
```

The binomial test under correlation can also be approximated by the ASRF formula (Vasicek model). (See Exhibit 13.17.) The critical value p^* is then given as the α-quantile of the ASRF model:

$$p^* = \Phi\left(\frac{\Phi^{-1}(\pi_0) + \sqrt{\rho}\Phi^{-1}(\alpha)}{\sqrt{1-\rho}}\right)$$

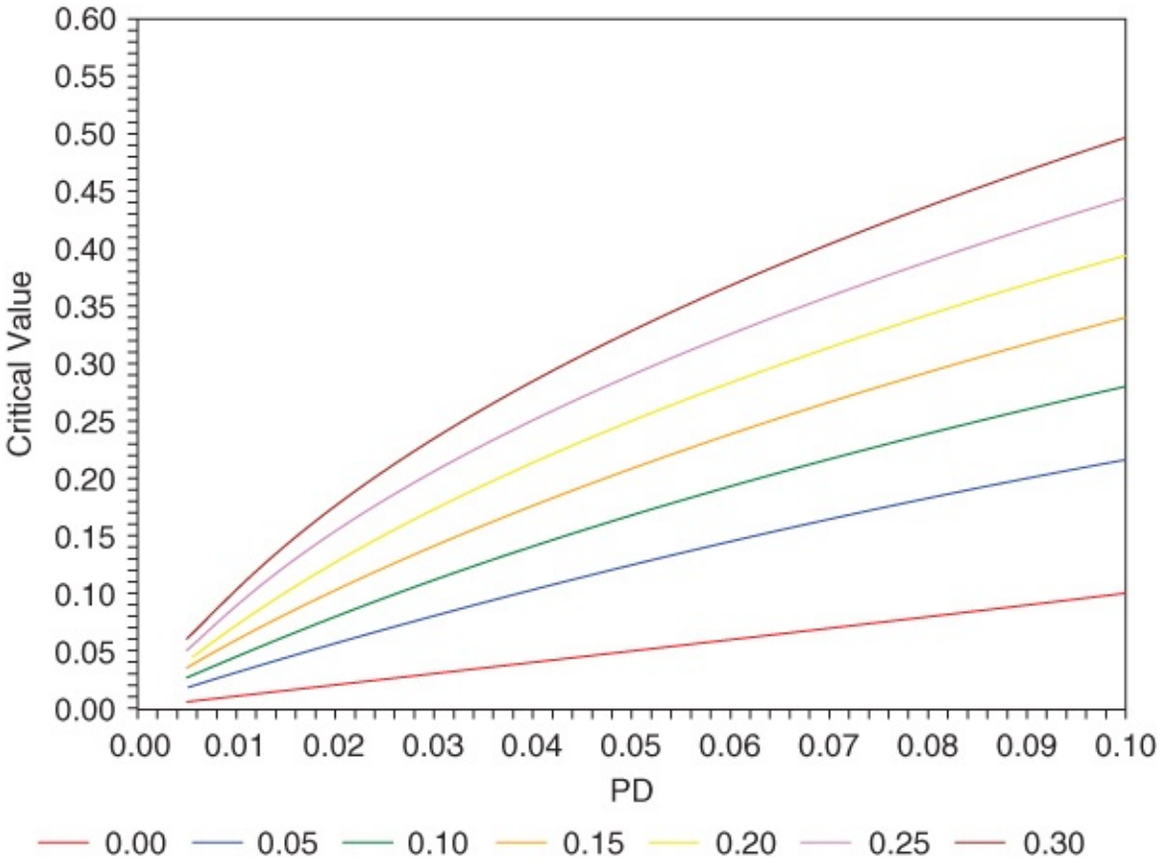


Exhibit 13.17 Critical Values under ASRF Model with Various Correlations

The following code computes the respective critical values for $\alpha = 0.99$.

```
ODS GRAPHICS ON;
/*ASRF model*/
DATA ASRF_crit;
DO PD = 0.005 TO 0.1 BY 0.005;
DO rho = 0 TO 0.3 BY 0.05;
p_crit = PROBNORM((PROBIT(PD)+SQRT(rho)*PROBIT(0.99))/SQRT(1-rho));
OUTPUT;
END;
END;
RUN;
ODS GRAPHICS ON;
PROC GPLOT DATA=ASRF_crit;
PLOT p_crit*PD = rho/ OVERLAY LEGEND=LEGEND1
HAXIS=AXIS1
VAXIS=AXIS2;
RUN;
QUIT;
ODS GRAPHICS OFF;
```

Other Important Issues

Confidence Level, α- versus β-Error

Whenever you conduct a statistical test, the question about the error probability of the test

decision arises. Given a null hypothesis H_0 and an alternative hypothesis H_1, you can differentiate between the true but unknown states of the world (either H_0 or H_1 is true), and the decision based on the statistical test (either rejection of H_0 or no rejection of H_0). Thus, four scenarios can arise:

1. H_0 is true and the test decision is *not* to reject H_0 → This is a correct decision.
2. H_0 is true but the test decision is to (erroneously) reject H_0 → This is a wrong decision.
3. H_0 is not true (and H_1 is true instead) and the test decision is to reject H_0 → This is a correct decision.
4. H_0 is not true (and H_1 is true instead) but the test decision is *not* to reject H_0 → This is a wrong decision.

Situations 1 and 3 are not problematic; situations 2 and 4 might hold, as you make a wrong decision. Situation 2, however, is under control, because the probability for a wrong decision when H_0 is true is given by the confidence level α. Therefore, a general question concerns the confidence level to be chosen, as it can be set by the decision maker who controls the α-error (or type 1 error) probability per test construction. The Hong Kong Monetary Authority (HKMA) has given some further recommendations about confidence levels to be adopted for the binomial test. More specifically, it states:

> For example, if a Binomial test is used, authorized institutions (AIs) can set tolerance limits at confidence levels of 95% and 99.9%. Deviations of the forecast PD from the realized default rates below a confidence level of 95% should not be regarded as significant and remedial actions may not be needed. Deviations at a confidence level higher than 99.9% should be regarded as significant and the PD must be revised upward immediately. Deviations which are significant at confidence levels between 95% and 99.9% should be put on a watch list, and upward revisions to the PD should be made if the deviations persist. (Hong Kong Monetary Authority (HKMA) 2006)

Situation 4, however, is somewhat problematic. Here, the test does not signal that H_0 is untrue, although H_1 holds in reality. The issue is that the probability for this type of error (the β - or type 2 error) depends on the parameter under H_1, which is unfortunately unknown (otherwise you would not need a statistical test). The way out here is to conduct a "what-if" analysis and check how the error would be under specific scenarios for the parameters (PD).

Let P be the random default rate (i.e., $P = D/n$), π_0 be the PD under H_0, and π be the true PD where $\pi_0 \neq \pi$. The β-error in the normal approximation of the simple binomial test (without correlation) can then be computed as

$$\begin{aligned} Prob_\beta &= P(P < p^* | \pi, \pi_0 \neq \pi) = P(P \leq \pi_0 + \Phi^{-1}(\alpha)\sqrt{\pi_0(1-\pi_0/n)} | \pi, \pi_0 \neq \pi) \\ &= \Phi\left(\frac{\pi_0 - \pi + \Phi^{-1}(\alpha)\sqrt{\pi_0(1-\pi_0/n)}}{\sqrt{\pi(1-\pi/n)}}\right) \end{aligned}$$

and returns the probability that the default rate P does not exceed the critical value p^*, as a

function of the true parameter π. As can be seen from the formula, the closer the true value π is to π_0, the higher the type 2 error becomes, ceteris paribus. The following code computes the β-error for various values of π_0 and π, for $\alpha = 0.99$ and $n = 1{,}000$ obligors. The four curves are computed such that π is $\pi_0 + h$ where $h \in \{0.005, 0.01, 0.025, 0.05\}$.

```
/*Beta error under simple Binomial Test*/
DATA Binomial_simple_beta_error;
ALPHA = 0.99;
N=1000;
DO PD0 = 0.005 TO 0.1 BY 0.005;
PD1 = PD0 + 0.005;
PD2 = PD0 + 0.01;
PD3 = PD0 + 0.025;
PD4 = PD0 + 0.05;
P_beta_005 = PROBNORM( (PD0-PD1 + PROBIT(alpha)*SQRT(PD0*(1-PD0)/n))
/ (SQRT(PD1*(1-PD1)/n)) );
P_beta_01 = PROBNORM( (PD0-PD2 + PROBIT(alpha)*SQRT(PD0*(1-PD0)/n))
/ (SQRT(PD2*(1-PD2)/n)) );
P_beta_025 = PROBNORM( (PD0-PD3 + PROBIT(alpha)*SQRT(PD0*(1-PD0)/n))
/ (SQRT(PD3*(1-PD3)/n)) );
P_beta_05 = PROBNORM( (PD0-PD4 + PROBIT(alpha)*SQRT(PD0*(1-PD0)/n))
/ (SQRT(PD4*(1-PD4)/n)) );
OUTPUT;
END;
RUN;
QUIT;
ODS GRAPHICS ON;
PROC GPLOT DATA=Binomial_simple_beta_error;
PLOT P_beta_005*PD0  P_beta_01*PD0 P_beta_025*PD0 P_beta_05*PD0
/ OVERLAY LEGEND=LEGEND1
HAXIS=0 TO 0.1 BY 0.02
VAXIS=0 TO 1 BY 0.2;
RUN;
QUIT;
ODS GRAPHICS OFF;
```

As can be seen in Exhibit 13.18, the type 2 error decreases with the difference between π_0 and π rising. It should also be noted that the α- and β-error are related. The higher the confidence level α (that is, the lower the type 1 error probability), the higher the type 2 error probability will be, ceteris paribus. Therefore, there is a trade-off between the two error probabilities. We leave it as an exercise to the reader to check this for our example.

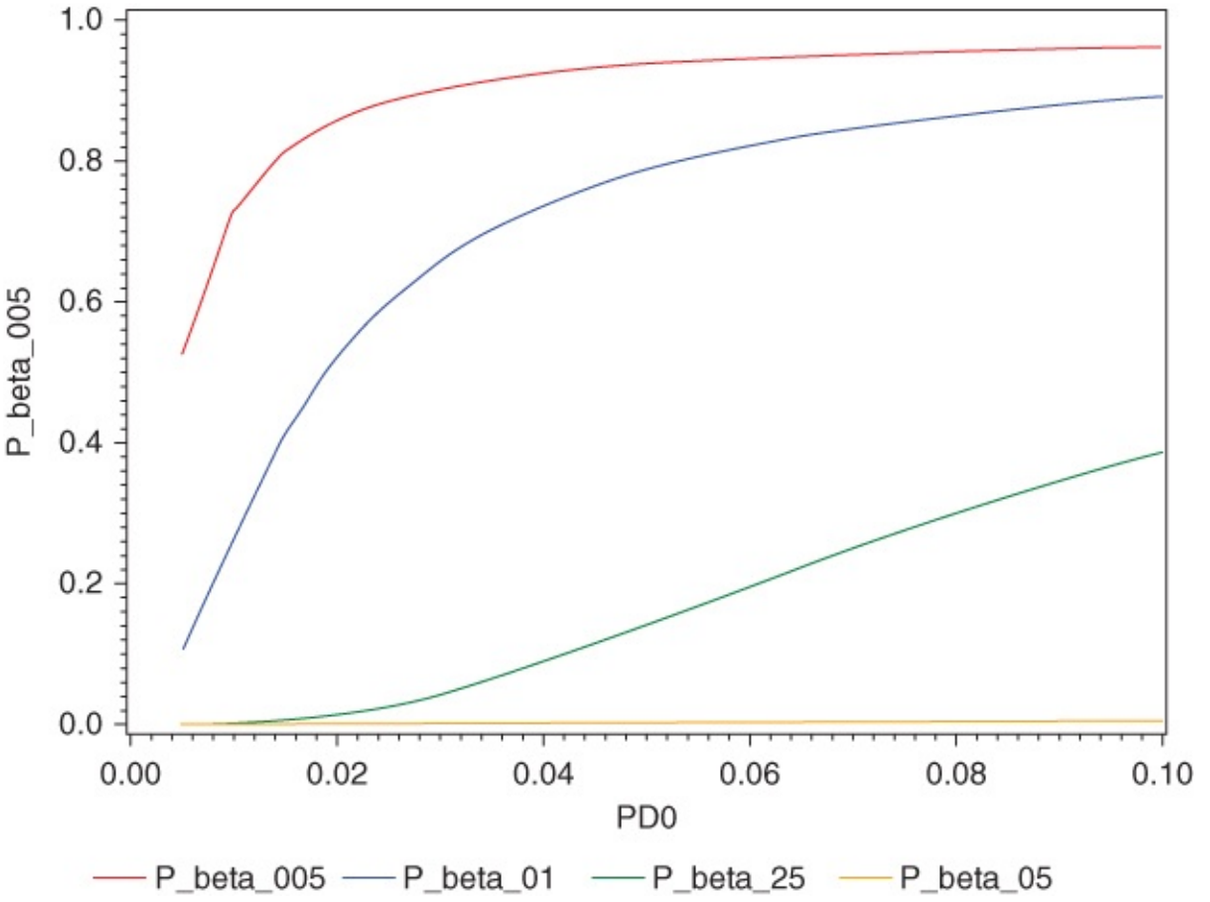


Exhibit 13.18 Beta Error for Simple Binomial Test

Unfortunately, the type 2 error probability becomes larger (all other parameters fixed) when defaults are correlated, that is, when we apply the extended binomial test under correlation. To see this, we look at the ASRF model with the critical value p^*. The type 2 error probability is then given as the probability that $P < p^*$ given that π is the true value rather than π_0. In the correlation chapter, we have seen that this probability is the CDF of the ASRF model, evaluated at π, and thus,

$$\begin{aligned} Prob_\beta &= P(P < p^* | \pi, \pi_0 \neq \pi) \\ &= \Phi\left(\frac{\sqrt{1-\rho}\cdot\Phi^{-1}(p^*) - \Phi^{-1}(\pi)}{\sqrt{\rho}}\right) \end{aligned}$$

for a given correlation ρ. The following code computes and plots these critical values under the assumption that $\pi = \pi_0 + 0.05$ for various values of $\rho \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$.

```
/*Beta error under correlated Binomial Test*/
DATA Binomial_corr_beta_error;
ALPHA = 0.99;
DO PD0 = 0.005 TO 0.1 BY 0.005;
PD1     = PD0 + 0.05;
rho1    = 0.01;
rho2    = 0.05;
rho3    = 0.1;
```

```
rho4    = 0.2;
rho5    = 0.3;
p_crit_rho01
= PROBNORM( (PROBIT(PD0)+SQRT(rho1)*PROBIT(alpha))/ SQRT(1-rho1));
p_beta_rho01
= PROBNORM( (SQRT(1-rho1)*PROBIT(p_crit_rho01) -PROBIT(PD1) )
/ SQRT(rho1) );
p_crit_rho05
= PROBNORM( (PROBIT(PD0)+SQRT(rho2)*PROBIT(alpha))/ SQRT(1-rho2));
p_beta_rho05
= PROBNORM( (SQRT(1-rho2)*PROBIT(p_crit_rho05) -PROBIT(PD1) )
/ SQRT(rho2) );
p_crit_rho10
= PROBNORM( (PROBIT(PD0)+SQRT(rho3)*PROBIT(alpha))/ SQRT(1-rho3));
p_beta_rho10
= PROBNORM( (SQRT(1-rho3)*PROBIT(p_crit_rho10) -PROBIT(PD1) )
/ SQRT(rho3) );
p_crit_rho20
= PROBNORM( (PROBIT(PD0)+SQRT(rho4)*PROBIT(alpha))/ SQRT(1-rho4));
p_beta_rho20
= PROBNORM( (SQRT(1-rho4)*PROBIT(p_crit_rho20) -PROBIT(PD1) )
/ SQRT(rho4) );
p_crit_rho30
= PROBNORM( (PROBIT(PD0)+SQRT(rho5)*PROBIT(alpha))/ SQRT(1-rho5));
p_beta_rho30
= PROBNORM( (SQRT(1-rho5)*PROBIT(p_crit_rho30) -PROBIT(PD1) )
/ SQRT(rho5) );
OUTPUT;
END;
RUN;
QUIT;
ODS GRAPHICS ON;
PROC GPLOT DATA=Binomial_corr_beta_error;
PLOT P_beta_rho01*PD0 P_beta_rho05*PD0  P_beta_rho10*PD0
P_beta_rho20*PD0  P_beta_rho30*PD0  / OVERLAY LEGEND=LEGEND1
HAXIS=0 TO 0.1 BY 0.02
VAXIS=0 TO 1 BY 0.2;
RUN;
QUIT;
ODS GRAPHICS OFF;
```

As can be seen in Exhibit 13.19, the β error probability rapidly increases toward 1 if the correlation is increased. Note that this is even the case for small values of π_0 where the difference between π_0 and π is huge (in relative terms). Such an analysis of type 2 errors should accompany any kind of PD backtesting procedure.

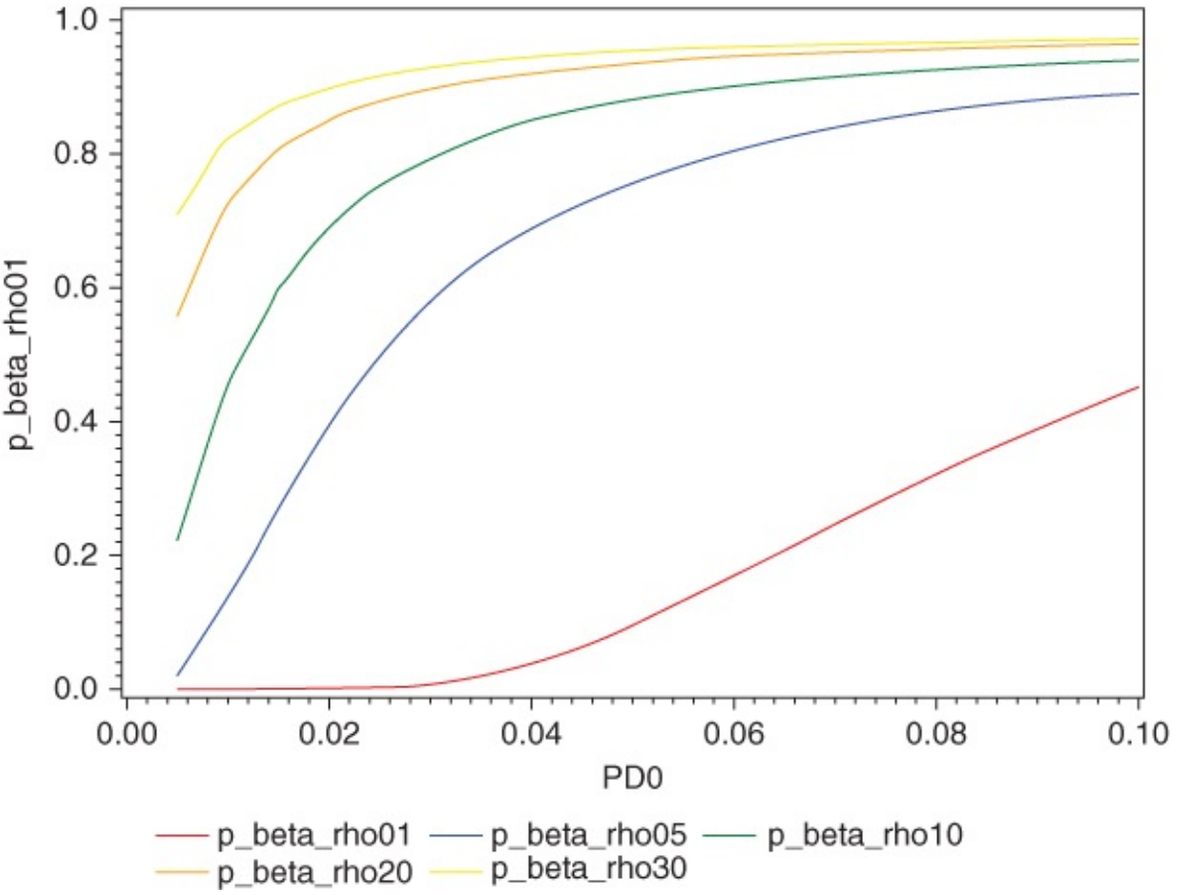


Exhibit 13.19 Beta Error for Extended Binomial Test under Correlations

Data Aggregation

Data aggregation can be interesting to consider during backtesting. Let's assume we have a portfolio with n obligors and G ratings. This implies that there are approximately n/G observations per rating. Hence, more ratings makes backtesting more difficult, since there will be fewer observations per rating, which will increase the standard errors and consequently the critical values. To improve the significance of the backtesting, data aggregation can be considered. One example would be to merge ratings with a low number of observations, for example, AA+, AA, and AA− into one overall rating AA. The aggregation can then also be considered for important segments or even at the overall portfolio level.

Risk Philosophy

The risk philosophy should also be taken into account during backtesting. In our chapter on time-discrete hazard models, we introduced the PIT and TTC rating philosophies. PIT ratings take into account both cyclical and noncyclical information. Hence, the backtesting should find that the realized default rates are close to the forecast PD, or, in other words, the PIT PDs should be validated against the 12-month default rates. TTC ratings consider only noncyclical information and are thus more stable. Hence, backtesting should find that the realized default rates vary around the forecast PD, rising in downturns and falling in upturns, or, the TTC PDs should be validated against cycle average default rates; see also Rösch (2005) for an analysis.

Setting up a Traffic Light Indicator Dashboard for PD Backtesting

Up until now, we have discussed various backtesting performance metrics and statistics to backtest PD models at levels 0, 1, and 2 of the credit risk model architecture. It is important that all these backtesting statistics are now combined in a traffic light indicator dashboard. In Exhibit 13.20, you can see an example of this starting at level 2.

Level 2: Calibration	Quantitative	Binomial	Not significant at 95% level	Significant at 95% but not at 99% level	Significant at 99% level
		Hosmer-Lemeshow	Not significant at 95% level	Significant at 95% but not at 99% level	Significant at 99% level
		Vasicek	Not significant at 95% level	Significant at 95% but not at 99% level	Significant at 99% level
		Normal	Not significant at 95% level	Significant at 95% but not at 99% level	Significant at 99% level
	Qualitative	Portfolio distribution	Minor shift	Moderate shift	Major shift
		Difference	Correct	Over-estimation	Under-estimation
		Portfolio stability	Minor migrations	Moderate migrations	Major migrations

Exhibit 13.20 Backtesting PD at Level 2

You can see that both quantitative tests as well as qualitative tests are included. The binomial, Hosmer-Lemeshow, Vasicek, and normal tests are the quantitative tests, possibly accompanied by an analysis of α- and β-errors. Qualitative tests are more subjective and are based on expert evaluation. Here, we include an inspection of the portfolio distribution, the overall difference between the estimated and realized default rates, and an evaluation of the portfolio stability. The result of each of these tests has been encoded using three traffic lights.

This can then be continued at level 1. In Exhibit 13.21, you can see quantitative tests based on the accuracy ratio, the area under the ROC curve (both are related), and the overall model significance. Qualitative checks inspect the data preprocessing activities conducted, the coefficient signs of the scorecard, the number of overrides, and the model documentation available. Again, the outcome of each of these tests is represented as a traffic light.

Level 1: Discrimination	Quantitative				
		AR difference with reference model	<5%	Between 5% and 10%	>10%
		AUC difference with reference model	<2,5%	Between 2.5% and 5%	>5%
		Model significance	*p*-value < 0.01	*p*-value between 0.01 and 0.10	*p*-value > 0.10
	Qualitative	Preprocessing (missing values, outliers)	Considered	Partially considered	Ignored
		Coefficient signs	All as expected	Minor exceptions	Major exceptions
		Number of overrides	Minor	Moderate	Major
		Documentation	Sufficient	Minor issues	Major issues

Exhibit 13.21 Backtesting PD at Level 1

Finally, as depicted in Exhibit 13.22 for level 0, the quantitative tests include the system stability index at the population level and the attribute level, and a t-test at the attribute level (not discussed in this chapter). It can be accompanied by a test of model stability. Qualitative checks include characteristics analysis and histogram inspection.

Level 0: Data	Quantitative				
		SSI (current versus training sample)	SSI < 0.10	0.10 < SSk 0.25	SSI > 0.25
		SSI attribute level	SSI < 0.10	0.10 < SSk 0.25	SSI > 0.25
		t-test attribute level	*p*-value > 0.10	*p*-value between 0.10 and 0.01	*p*-value < 0.01
	Qualitative	Characteristic analysis	No change	Moderate change	Major change
		Attribute histogram	No shift	Moderate shift	Major shift

Exhibit 13.22 Backtesting PD at Level 0

Action Plan

Based on all these tests, a decision needs to be made as to whether the PD model is performing well. Remember, if there are issues with the PD model, a backtesting action plan should be available to remedy the situation. This plan will specify what to do in response to the findings of the PD backtesting exercise.

In Exhibit 13.23, you can see an example of an action scheme for PD backtesting. If the model calibration is okay, you can continue to use the PD model, since the capital is appropriately calculated and there is no problem. If the model calibration is not okay, you need to verify the model discrimination or ranking at level 1. If this is OK, then the solution might be to simply recalibrate the probabilities upward or downward using a scaling factor. If not, the next step is to check the data and model stability at level 0. If the stability is still okay, you may consider tweaking the model to see if you can remedy the situation. This is not always straightforward and will often involve completely reestimating the PD model, as is the case when the stability is not okay.

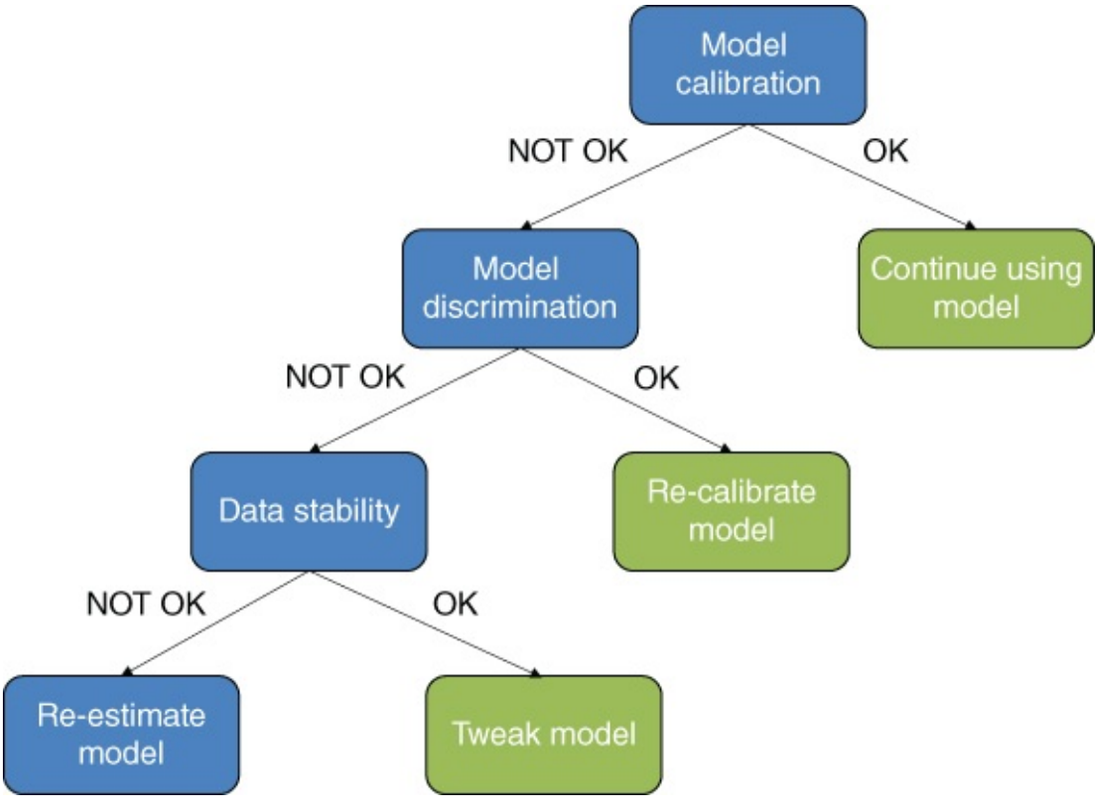


Exhibit 13.23 Action Plan for PD Backtesting

Backtesting LGD and EAD

In what follows, we discuss backtesting LGD and EAD models. An empirical study on LGD model comparison is given by Loterman et al. (2012). Remember, both models are typically developed using similar methodologies. Hence, the backtesting procedures will also be similar. We illustrate the techniques using LGDs only, but application to EADs is straightforward using the same codes. As with PD backtesting, we also make use of the multilevel model architecture. At level 0, we backtest the stability of the data. At level 1, we verify the discrimination and at level 2 the calibration (see Exhibit 13.24).

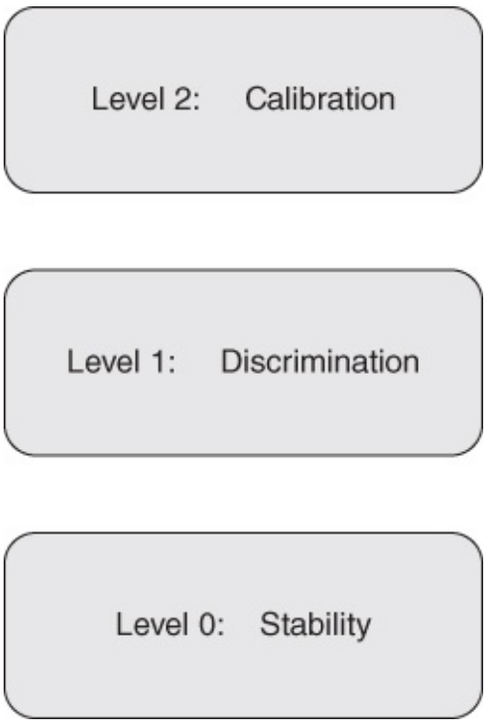


Exhibit 13.24 Backtesting LGD and EAD

For backtesting the stability of the data, a system stability index can be used, just as with PD. This SSI will then contrast the distribution of the actual population with that of the training population across the various LGD ranges. Similarly, a test on model stability can be conducted using time dummies.

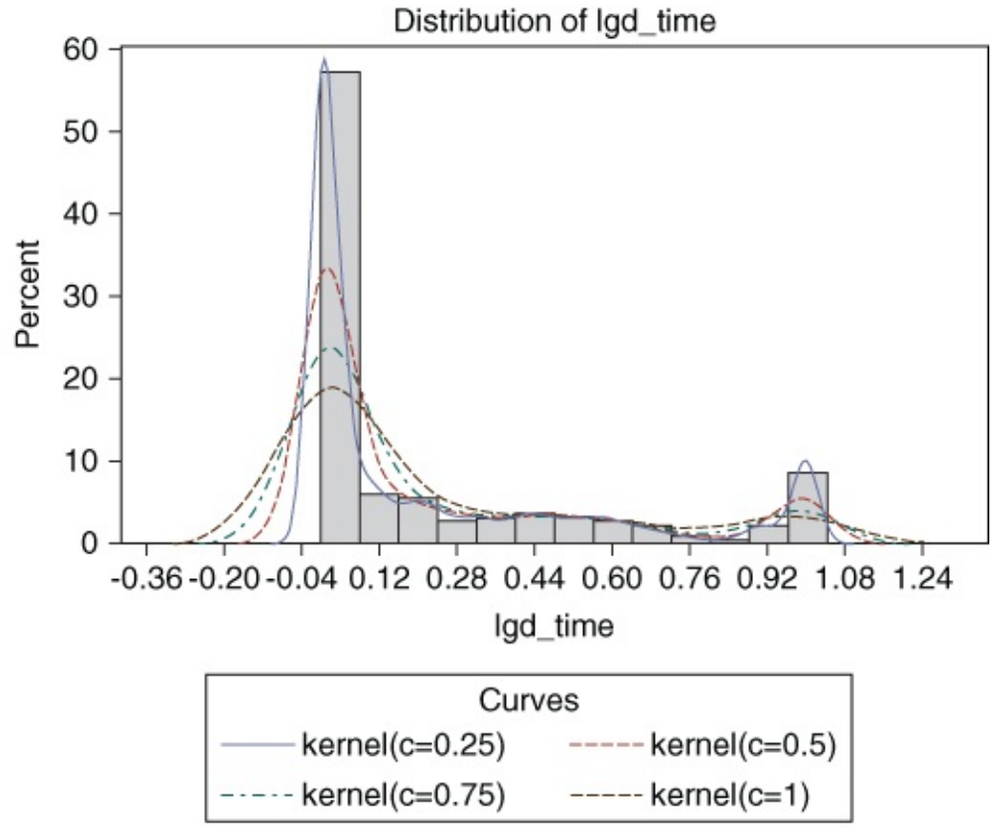
Backtesting at Level 1

At level 1, the discrimination can be verified using ROC plots and corresponding AUROCs and ARs. However, defaults are binary variables whereas LGDs and EADs are metric variables. Therefore, some adjustments have to be made. To provide insight into the implementation in SAS, we divide the data that we used in the chapter on LGDs into a training and an out-of-sample validation data set (named lgd_is and lgd_os). For ease of exposition, we show our examples with the linear models. The following code estimates the LGD for the in-sample training data with PROC REG and scores the validation sample using PROC PLM; that is, it computes the out-of-sample predicted values.

```
ODS GRAPHICS ON;
PROC REG DATA=data.lgd_is OUTEST=Reg1Out;
MODEL lgd_time= LTV purpose1/  DETAILS=ALL;
STORE OUT=model1;
RUN;
QUIT;
PROC PLM SOURCE=model1;
SCORE DATA = data.lgd_os OUT = lgd_os1;
RUN;
ODS GRAPHICS OFF;
```

We then compare the predictions in the validation sample with the actual LGDs. We first use

PROC UNIVARIATE and PROC BOXPLOT in order to compare the distributions of actual and predicted values (see Exhibits 13.25 and 13.26).



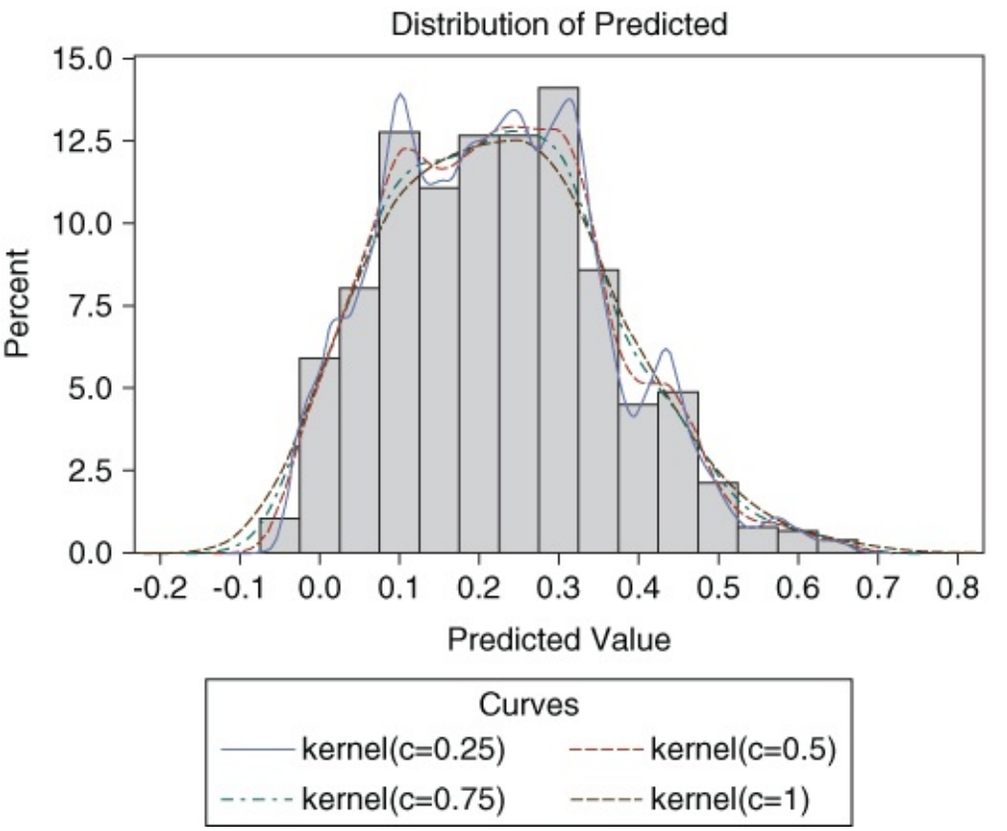
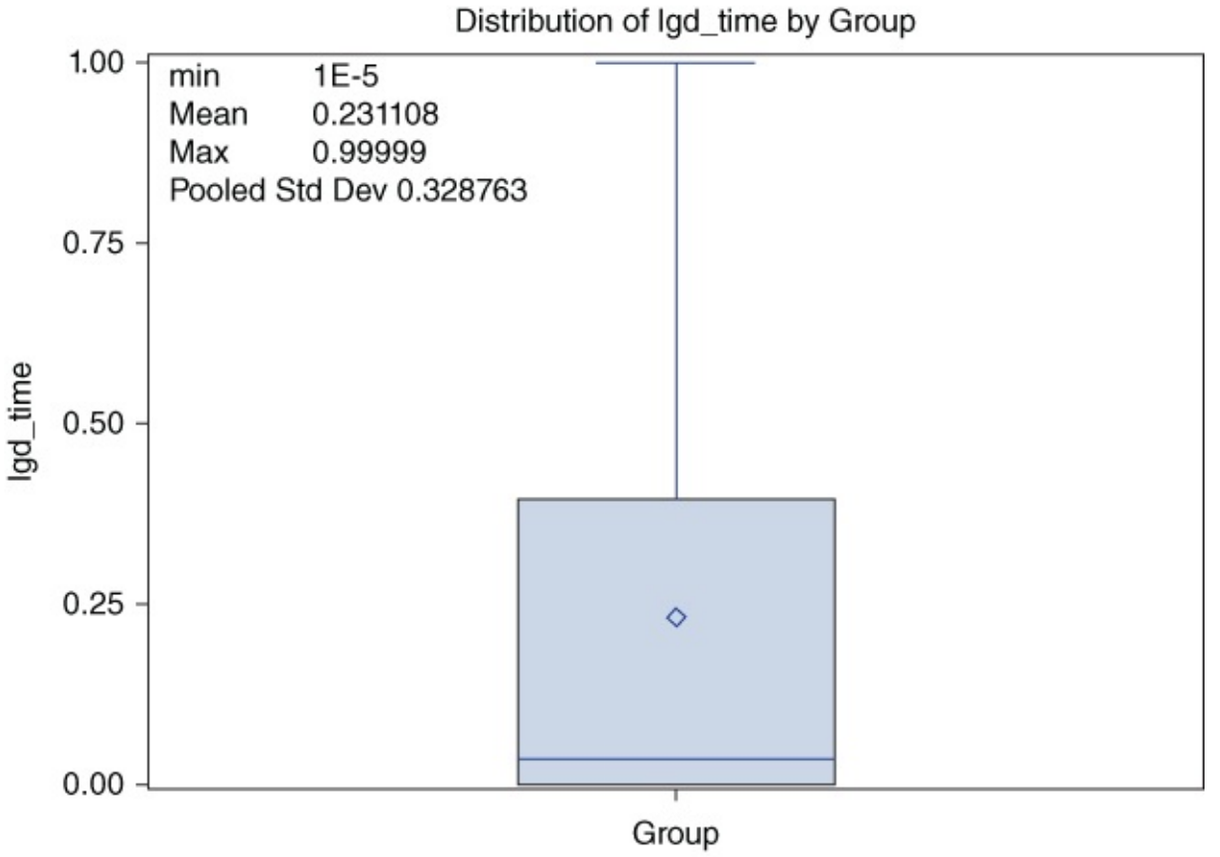


Exhibit 13.25 Histograms of Actual and Predicted LGDs



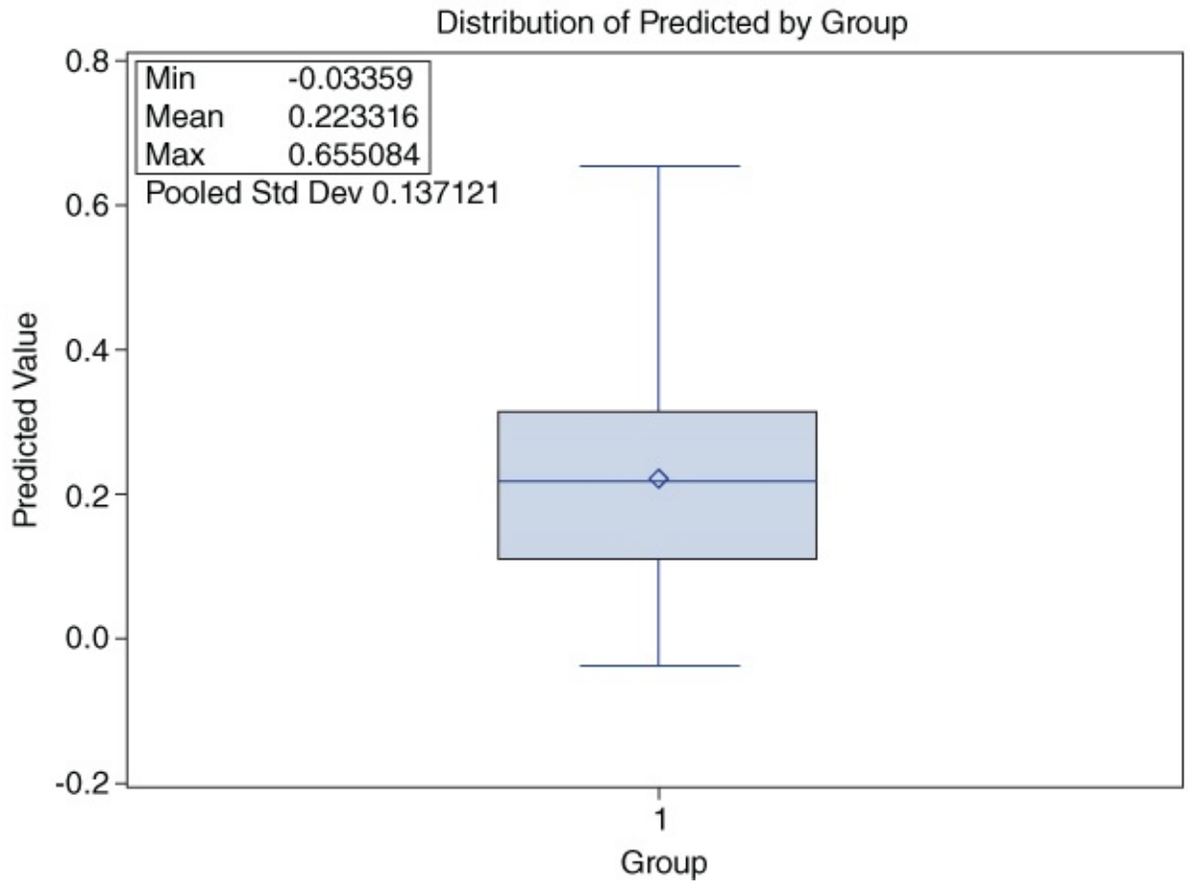


Exhibit 13.26 Box Plots of Actual and Predicted LGDs

```
ODS GRAPHICS ON;
PROC UNIVARIATE data = lgd_os1;
HISTOGRAM lgd_time predicted / KERNEL(c = 0.25 0.50 0.75 1.00
l = 1 20 2 34
NOPRINT);
RUN;
ODS GRAPHICS OFF;

ODS GRAPHICS ON;
PROC BOXPLOT data = lgd_os1;
PLOT lgd_time*group; INSET MIN MEAN MAX STDDEV ;
PLOT predicted*group; INSET MIN MEAN MAX STDDEV ;
RUN;
ODS GRAPHICS OFF;
```

The histograms (Overlaid by kernel densities) show serious deviations between actual and predicted values. This is particularly attributable to the nonnormal shape of the LGD distribution, which cannot be properly fitted by the OLS regression model with only two explanatory variables. Similarly, the different distributions can be seen in the box plots. The means (the diamond in the plot) are quite similar but the medians (the line in the box), the 25 percent and the 75 percent percentiles (the frames of the boxes), and the maximum and minimum values are quite different.

Next, we analyze the ROC curve. Here you must keep in mind that LGD is a continuous variable whereas for PD we had a binary outcome. Therefore, we follow the suggestion in Gupton and Stein (2005) and transform the observed LGDs into a binary variable using

$$d(\mathrm{LGD}_i) = \begin{cases} 1 & \mathrm{LGD}_i \geq \overline{LGD} \\ 0 & \mathrm{LGD}_i < \overline{LGD} \end{cases}$$

where $\overline{LGD}$ is the mean observed LGD. As can be seen from the box plot, the mean of the LGDs is about 0.2311. The transformation is given by the following code.

```
DATA lgd_os2;
SET lgd_os1;
D_LGD = 0;
IF lgd_time> 0.2312 THEN D_LGD = 1;
RUN;
```

Next, PROC LOGISTIC is used to plot the ROC curve and compute association statistics of the model which regresses the dummy against the predicted values. (See Exhibit 13.27.)

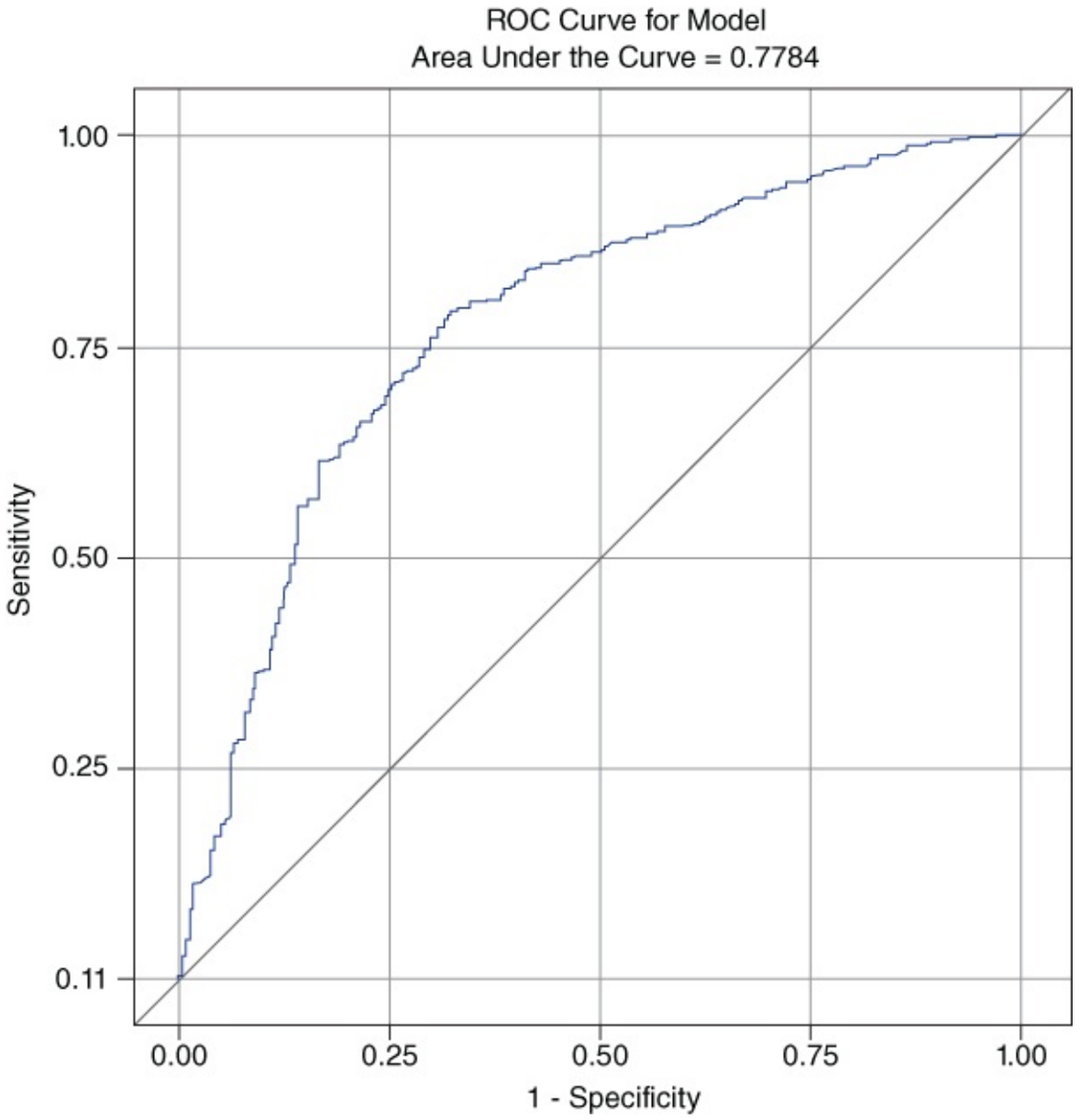


Exhibit 13.27 ROC for Predicted LGDs

```
ODS GRAPHICS ON;
PROC LOGISTIC data = lgd_os2 PLOTS(ONLY)=ROC;
CLASS D_LGD;
MODEL D_LGD = predicted;
RUN;
QUIT;
ODS GRAPHICS OFF;
```

The discriminatory power of high (above-mean) versus low (below-mean) LGDs can be seen from the plot and from the AUROC measure, which is 0.7784. As for PDs, the measure and the curve depend on the portfolio composition, and in order to make an assessment about the quality of the discrimination, the ROC should ideally be compared relative to a benchmark model.

Finally, we can compute correlation coefficients for measuring the association between the predicted and the observed LGDs. The linear *Bravais-Pearson* correlation should be obvious

from the chapter on exploratory data analysis and therefore a formula is not provided. The nonparametric *Spearman* rank correlation is given by:

$$\theta = \frac{\sum_i ((R_i - \overline{R})(S_i - \overline{S}))}{\sqrt{\sum_i (R_i - \overline{R})^2 \ \sum (S_i - \overline{S})^2}}$$

where R_i is the rank of $\text{LG } D_i^{obs}$, S_i is the rank of LGD_i^{pred}, $\overline{R}$ is the mean of the R_i values, and $\overline{S}$ is the mean of the S_i values.

Another popular nonparametric measure for association is *Kendall's* τ-b, which is given by:

$$\tau = \frac{\sum_{i<j} (sgn(LGD_i^{obs} - LGD_j^{obs}) sgn(LGD_i^{pred} - LGD_j^{pred}))}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}$$

where $T_0 = n(n-1)/2$, $T_1 = \sum_k t_k(t_k - 1)/2$, and $T_2 = \sum_l u_l(u_l - 1)/2$. t_k is the number of tied LGD^{obs} values in the k th group of tied LGD^{obs} values, u_l is the number of tied LGD^{pred} values in the l th group of tied LGD^{pred} values, n is the number of observations, and $sgn(z)$ is defined as

$$sgn(z) = \begin{cases} 1 & if\ z > 0 \\ 0 & if\ z = 0 \\ -1 & if\ z < 0 \end{cases}$$

All measures (and some more) can be computed via PROC CORR as shown in the following code (see Exhibit 13.28).

The CORR Procedure

Pearson Correlation Coefficients, *N* = 760 Prob > |r| under *H0*: Rho = 0					
		lgd_time	Predicted	Plgd_time	PPredicted
lgd_time		1.00000	0.43601		<.0001
Predicted	Predicted Value	0.43601	1.00000	<.0001	

Spearman Correlation Coefficients, *N* = 760 Prob > |r| under *H0*: Rho = 0					
		lgd_time	Predicted	Plgd_time	PPredicted
lgd_time		1.00000	0.47656		<.0001
Predicted	Predicted Value	0.47656	1.00000	<.0001	

Kendall's Tau b Correlation Coefficients, *N* = 760 Prob > |tau| under *H0*: Tau = 0					
		lgd_time	Predicted	Plgd_time	PPredicted
lgd_time		1.00000	0.33440		<.0001
Predicted	Predicted Value	0.33440	1.00000	<.0001	

Exhibit 13.28 Measures for Correlation and Association

```
ODS GRAPHICS ON;
PROC CORR DATA = lgd_os2 Pearson Spearman Kendall;
VAR lgd_time predicted ;
RUN;
QUIT;
ODS GRAPHICS OFF;
```

The linear correlation is about 0.43, which shows a medium size association (remember that the correlation ranges between -1 and $+1$). Similarly, the Spearman measure, which is 0.48, and Kendall's τ, which is 0.33, show moderate positive association.

Backtesting at Level 2

Next, we check the calibration at level 2. We analyze several measures that can be obtained by running a regression of actual versus predicted values as illustrated in the following code (see Exhibits 13.29 and 13.30).

The REG Procedure
Model: MODEL1
Dependent Variable: lgd_time

Root MSE	0.29606	***R*-Squared**	0.1901
Dependent Mean	0.23111	**Adj *R*-Sq**	0.1890
Coeff Var	128.10582		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|
Intercept	Intercept	1	−0.00234	0.02053	−0.11	0.9092
Predicted	Predicted Value	1	1.04538	0.07837	13.34	<.0001

Exhibit 13.29 Real-Fit Diagnostics

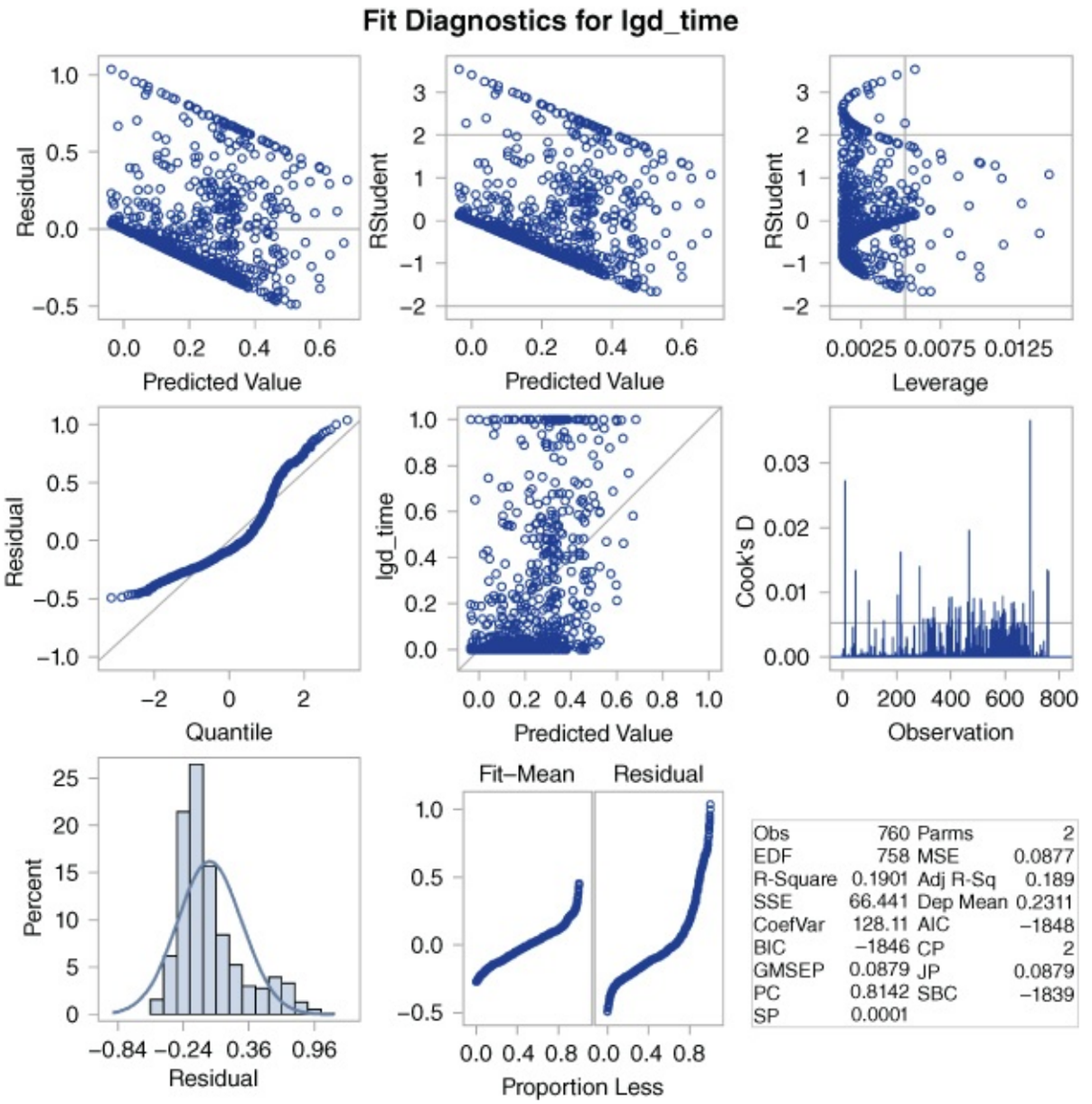


Exhibit 13.30 Linear Regression

```
ODS GRAPHICS ON;
PROC REG DATA=lgd_os1
    PLOTS(MAXPOINTS= 10000 STATS= ALL)= (CRITERIA QQ);
MODEL lgd_time= predicted  ;
RUN;
QUIT;
ODS GRAPHICS OFF;
```

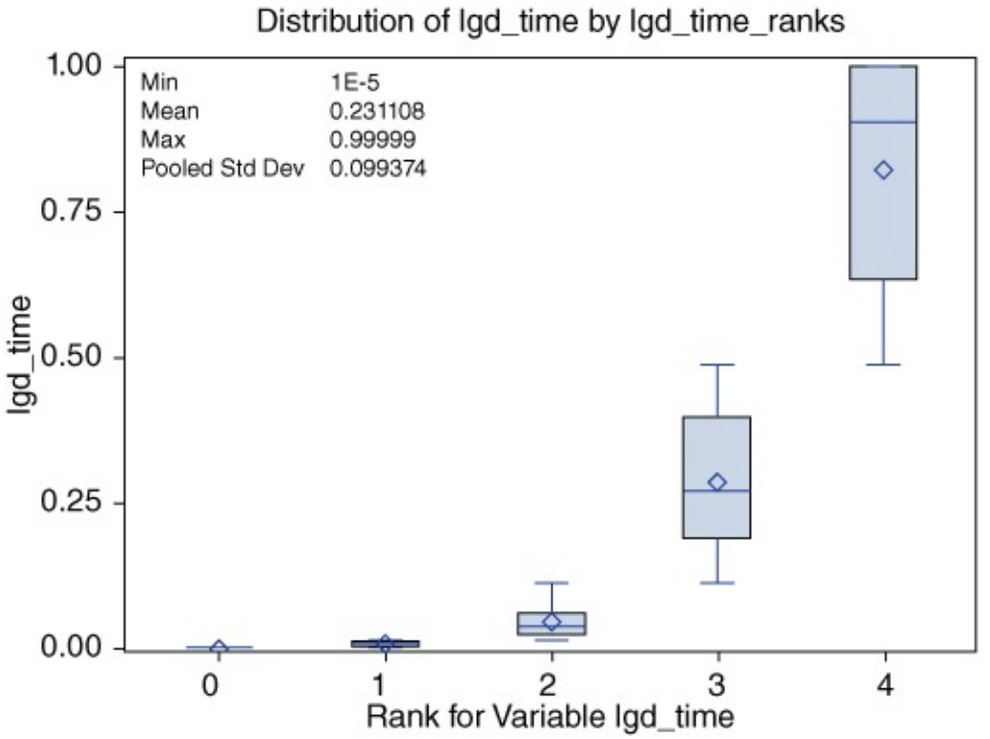
The intercept is close to zero and the slope close to one, which shows good calibration for the mean. However, the model returns an adjusted $R^2 = 0.19$ only and a root MSE of 0.30. In other words, the average deviation of the LGDs is 0.3. As can also be seen from the plots, the different shape of the distributions translates into a very moderate fit. Again, such figures would call for better covariates to be included.

We extend the preceding analysis from an individual level to a bucket-wise level. The following code uses PROC RANK (which you already know from the PD chapter) to create

five groups of LGD buckets according to the observed LGDs. In a next step, the data are sorted according to their ranks, and mean LGDs are computed for the predictions and the observations for each bucket.

```
PROC RANK DATA=lgd_os1 OUT=LGD_ranks GROUPS=5;
VAR lgd_time;
RANKS lgd_time_ranks;
RUN;
PROC SORT DATA = lgd_ranks;
BY lgd_time_ranks;
RUN;
PROC MEANS DATA = lgd_ranks;
VAR lgd_time predicted;
BY lgd_time_ranks;
OUTPUT OUT = means1 MEAN(lgd_time predicted)=lgd_time predicted;
RUN;
```

Next, we compute box plots for each bucket in order to learn about the groupwise distribution (see Exhibit 13.31).



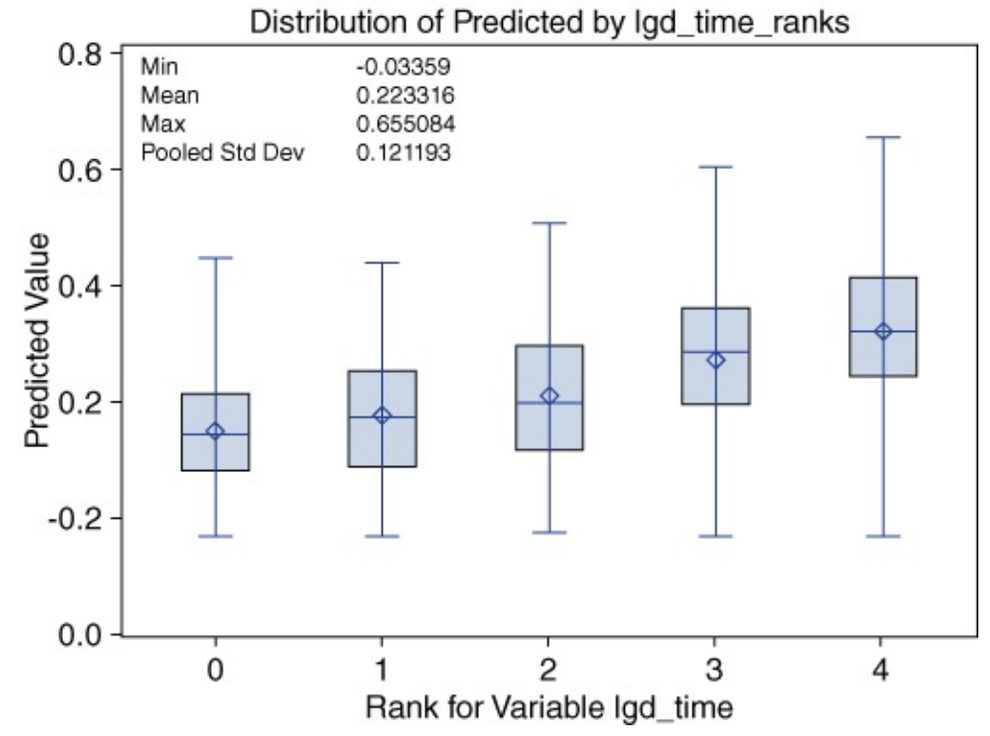


Exhibit 13.31 Box Plots of Actual and Predicted LGDs

```
ODS GRAPHICS ON;
PROC BOXPLOT data = lgd_ranks;
PLOT lgd_time*lgd_time_ranks; inset min mean max stddev ;
PLOT predicted*lgd_time_ranks; inset min mean max stddev ;
RUN;
ODS GRAPHICS OFF;
```

As the box plots reveal, we find increasing predicted LGD distributions for each bucket, which is desired. However, the distributions of the observed LGDs steeply increase from low to high buckets. Similarly, in the next step, you can plot the mean actual LGDs versus the mean predicted LGDs.

```
ODS GRAPHICS ON;
SYMBOL1 INTERPOL=NONE
VALUE=CIRCLE
CV=BLUE
WIDTH=4 HEIGHT=4;
PROC GPLOT DATA = means1;
PLOT lgd_time * predicted / HAXIS=0 TO 1 BY 0.2
VAXIS=0 TO 1 BY 0.2;
RUN;
ODS GRAPHICS OFF;
```

The scatter plot (Exhibit 13.32) supports the former analyses and shows that for the lower LGD buckets the predicted values are too high whereas they are too low for the higher LGD buckets. The next steps would now be to apply some of the other LGD modeling methods discussed in the LGD chapter and/or include better explanatory variables.

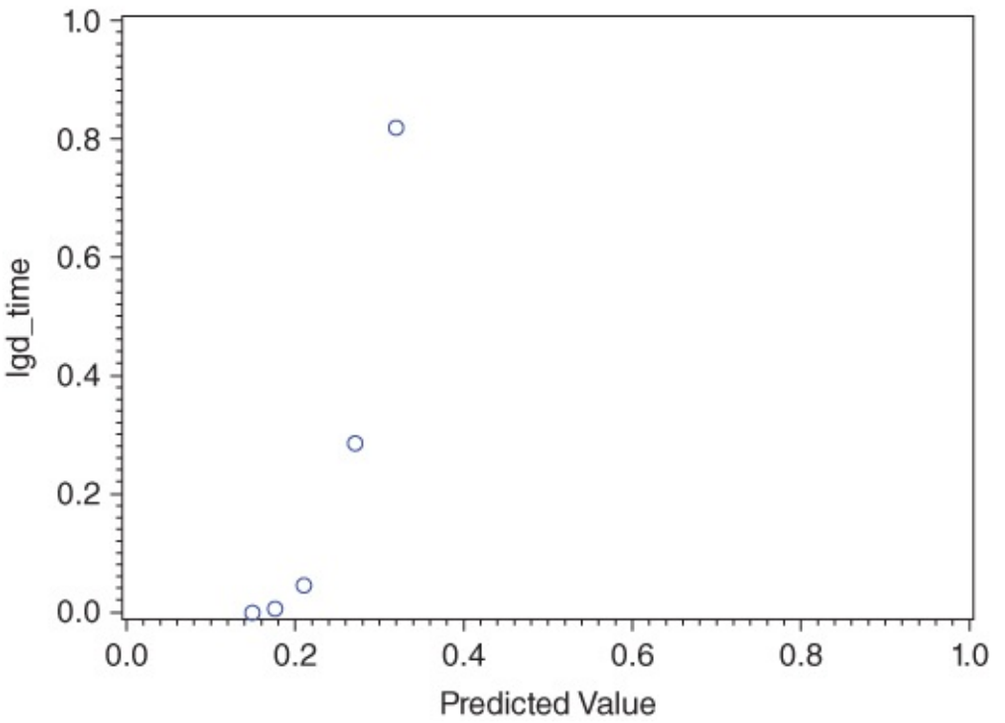


Exhibit 13.32 Scatter Plot of Actual and Predicted LGDs

We conclude with references to more backtesting techniques. First, more discrimination measures can be computed, as shown in Fischer and Pfeuffer (2014). Second, statistical tests can be applied for the variances of the LGD distributions, or a bootstrap procedure can be used for conducting a test for R^2 or MSE (see Loterman et al. 2012).

Benchmarking

Benchmarking is another important quantitative validation activity. The idea here is to compare the output and performance of the analytical PD, LGD, or EAD model with a reference model or benchmark. This is recommended as an extra validity check to make sure that the current credit risk model is the optimal one to be used. Various credit risk measurements can be benchmarked, for example credit scores, ratings, calibrated risk measurements such as PDs, LGDs, CCFs, or even migration matrices. Various benchmarking partners can be considered. Examples are credit bureaus, rating agencies, data poolers, and even internal experts. As an example of a simple benchmarking exercise, consider benchmarking an application score against a FICO score.

The benchmark can be externally or internally developed. Various problems arise when performing external benchmarking. Firstly, there is no guarantee that external ratings are necessarily of good quality. Think about what happened during the credit crisis, when many rating agencies were criticized because their ratings turned out to be overly optimistic. Next, the external partner might also have a different portfolio composition and adopt different model development methodologies and/or processes, making a comparison less straightforward. Also, different rating philosophies might be used, where the benchmark rating system is either more point-in-time or through-the-cycle. The default and loss definitions might differ, and different LGD weighting schemes can be adopted, along with different discount factors or collection policies. External benchmarking might also be complicated because of legal constraints where information cannot be exchanged due to banking secrecy regulation. Credit risk is typically also an endogenous phenomenon, which is highly dependent upon the internal credit culture and/or process. There is also a risk of cherry-picking where a close-match external benchmark is selected without further critical evaluation.

Given these complications with external benchmarking, the idea of internal benchmarking has been advocated. It was first introduced by the Hong Kong Monetary Authority (HKMA), as illustrated by this quote:

> Where a relevant external benchmark is not available (e.g., PD for SME and retail exposures, LGD, and EAD), an authorized institution (AI) should develop an internal benchmark. For example, to benchmark against a model-based rating system, an AI might employ internal rating reviewers to re-rate a sample of credit on an expert-judgment basis. (Hong Kong Monetary Authority (HKMA) 2006)

The internal benchmark can be a statistical or an expert-based benchmark. Consider, for example, a PD model built using a plain-vanilla logistic regression model. You can then consider building a neural network benchmark where the performance of both the logistic regression and the neural network can then be contrasted. Although the neural network is clearly a black box model, and thus cannot be used as the final credit risk model, the result of this benchmarking exercise will tell whether there are any nonlinear effects in the data. If it turns out that the neural network performs better than the logistic regression, you can then start looking for nonlinear effects or interactions and try to add them to your logistic regression model to further boost its performance. The benchmark can also be expert based. Remember, an expert-based benchmark is a qualitative model based on expert experience and/or common

sense. An example of this could be an expert committee ranking a set of small and medium-sized enterprises (SMEs) in terms of default risk by merely inspecting their balance sheet and financial statement information in an expert-based, subjective way. The ranking obtained by an expert-based rating system can then be compared to the ranking obtained by the logistic regression.

A champion challenger approach can be used when benchmarking. The current model is the champion that is challenged by the benchmark. If the benchmark beats the champion in performance, then it can become the new champion. This way, models are continuously challenged and further improved.

As tools for comparison, the previously discussed techniques that were used for backtesting can be applied, (e.g., correlation and measures of association).

QUALITATIVE VALIDATION

To conclude this chapter, we touch on qualitative validation and consider these topics:

- Use testing
- Data quality
- Model design
- Documentation
- Corporate governance and management oversight

Use Testing

The idea of use testing is to use the IRB models and estimates not only for Basel capital calculation, but also for other business activities such as credit pricing, credit approval, and economic capital calculations. This can be illustrated with the following regulatory articles:

- *"Internal ratings and default and loss estimates must play an essential role in the credit approval, risk management, internal capital allocations, and corporate governance functions of banks using the IRB approach"* (§444, Basel Committee on Banking Supervision 2006).
- *"The systems and processes used by a bank for risk-based capital purposes must be consistent with the bank's internal risk management processes and management information reporting systems"* (Federal Register 2007).
- *"When institutions use different estimates for the calculation of risk weights and for internal purposes, it shall be documented and be reasonable"* (Art. 179, European Union 2013).

So to summarize, the IRB estimates must play an essential role in other business activities. This, however, does not mean an exclusive role. Earlier, the Financial Services Authority

(FSA) put forward three conditions that should be satisfied in order to meet the use test requirement.

1. Consistency: The information the IRB estimates, PD, LGD, and EAD, are based on should be consistent with internal lending standards and policies.
2. Use of all relevant information: Any relevant information used in internal lending standards and policies must also be used in calculating the IRB estimates.
3. Disclosure: If differences exist between the calculation of the IRB estimates and those used for internal purposes, then they must be documented and the reasonableness demonstrated.

Some examples of use test issues are as follows. For application scoring, many firms use a time window of 18 months. Remember, however, that for PD, Basel requires a time window of only 12 months. Also, the Basel default definition is 90 days, whereas some financial institutions like to use their own definitions. Most regulators will tolerate these differences, as long as they are properly documented and the reasonableness thereof is demonstrated.

Another issue concerns the use of downturn LGD for other business activities. This is often perceived to be too conservative for other applications. FSA has indicated:

> Firms can use different LGDs for business purposes to those used for regulation and not fail the use test, provided that the rationale for their use and differences/transformation to capital numbers is understood.

Hence, average LGD values can be used for economic capital calculation, IFRS provisions, and other accounting applications.

Data Quality

Data is the key ingredient to any credit risk model, be it a PD, LGD, or EAD model. It speaks for itself that to have good models, data should be of high quality. About this, the regulators said:

- *"The institution shall have in place a process for vetting data inputs into the model, which includes an assessment of the accuracy, completeness and appropriateness of the data"* (Art. 173, European Union 2013).
- *"The data used to build the model shall be representative of the population of the institution's actual obligors or exposures"* (Art. 173, European Union 2013).
- *"The PRA expects a firm to set standards for data quality, aim to improve them over time and measure its performance against those standards"* (Section 10, Prudential Regulation Authority 2013).

Data quality can be measured in various ways. A first important dimension is accuracy. The aim here is to verify if the inputs measure what they are supposed to measure. The FSA introduced the data accuracy scorecard to measure this. Bad data accuracy can be caused by data entry errors, measurement errors, and outliers.

Another important dimension is data completeness. Observations with missing values can be removed only if sound justifications can be given. About this, the CEBS mentioned:

> While missing data for some fields or records may be inevitable, institutions should attempt to minimize their occurrence and aim to reduce them over time. (Committee of European Banking Supervisors (CEBS) 2005)

Data timeliness refers to the recency of the data. Data should be updated at least annually, although higher updating frequencies are recommended for the riskier obligors. Data should also be appropriate in the sense that there should be no biases or unjustified data truncation. Furthermore, data should be unambiguously defined. As an example, consider the definition of a ratio variable, commonly used in corporate credit risk models: Ratios are defined as a numerator divided by a denominator, and both should be clearly defined. It should also be mentioned what happens to the ratio when the denominator equals zero; missing values should also be defined in an unambiguous way and not coded as zero, as is often the case.

To summarize, it is very important that financial institutions set up master data management and data governance initiatives. This applies not only to internally collected but also to externally obtained data. Moges et al. (2013) conducted a worldwide survey with more than 50 banks on the topic of data quality. Note that the focus of the survey was on credit risk analytics. The main findings were:

- Most banks indicated that between 10 to 20 percent of their data suffers from data quality problems.
- Manual data entry is one of the key problems, together with the diversity of data sources and the consistency of corporate-wide data representation.
- Regulatory compliance is the key motive to improve data quality, rather than strategic or competitive advantage, for example.

Model Design

A next qualitative validation activity relates to model design. Some example questions that need to be answered here are:

- When was the model designed and by whom?
- What is the perimeter of the model in terms of counterparty types, geographical region, industry sectors? For example, was the model developed using Belgian SMEs active in the agricultural sector?
- What are the strengths and weaknesses of the model?
- What data was used to build the model? How was the sample constructed? What is the time horizon of the sample? Which default definition was adopted?
- Is human judgment used, and if so, how?

Being able to adequately answer all these questions is very important for correctly using the

model and for facilitating model maintenance. It is also essential that all this is properly documented.

Documentation

All steps of the credit risk model development and monitoring process should be adequately documented. This can be illustrated by means of the following regulatory quotes:

- *"All material elements of the internal models and the modeling process and validation shall be documented"* (Art. 188, European Union 2013).
- *"Documentation should be transparent and comprehensive"* (Federal Register 2007).
- *"Documentation should encompass, but is not limited to, the internal risk rating and segmentation systems, risk parameter quantification processes, data collection and maintenance processes, and model design, assumptions, and validation results"* (Federal Register 2007).

Documentation is needed both for internally developed as well as for externally purchased models. It is advisable to use document management systems with appropriate versioning facilities to keep track of all document versions. An ambitious goal here is to aim for a documentation test that verifies whether a newly hired analytical team could use the existing documentation to continue development or production of the existing analytical PD, LGD, and EAD models.

Corporate Governance and Management Oversight

A final qualitative validation activity concerns corporate governance and management oversight. The idea here is to pursue active involvement of the board of directors and senior management in the implementation and validation process of the various credit risk models. About this, the EU regulation mentioned:

> All material aspects of the rating and estimation processes shall be approved by the institution's management body or a designated committee thereof and senior management. These parties shall possess a general understanding of the rating systems of the institution and detailed comprehension of its associated management reports. (Art. 189, European Union 2013)

Senior management should demonstrate active involvement on an ongoing basis, assign clear responsibilities, and put into place organizational procedures and policies that will allow the proper and sound implementation and validation of the IRB systems. The outcome of the validation exercise must also be communicated to senior management and, if needed, accompanied by an appropriate response.

PRACTICE QUESTIONS

1. Discuss the most important validation issues and validation principles.

2. Compute the PSI/SSI for the example in Exhibit 13.6 with the variables income and years client.
3. Estimate logistic regression PD models with LTV only and FICO only, and compare the ROC curves. Use data set mortgage.
4. Why are discrimination measures portfolio dependent? Research the literature and provide examples.
5. Conduct the binomial test for the example using the PD estimates of the other two rating grades in the Quantitative Valuation section, and interpret the results.
6. Show via the examples of type 1 and type 2 errors that there is a trade-off between both error probabilities.

References

Agresti, A. 1984. *Analysis of Ordinal Categorical Data*. New York: John Wiley & Sons.

Basel Committee on Banking Supervision. 2005a. "Studies on the Validation of Internal Rating Systems (Revised)."

Basel Committee on Banking Supervision. 2005b. "Update on Work of the Accord Implementation Group Related to Validation under the Basel II Framework."

Basel Committee on Banking Supervision. 2006. "International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Comprehensive Version."

Blochwitz, S., A. Hamerle, S. Hohl, R. Rauhmeier, and D. Rösch. 2005. "Myth and Reality of Discriminatory Power for Rating Systems." *Wilmott Magazine*, 2–6.

Castermans, G., D. Martens, T. Van Gestel, B. Hamers, and B. Baesens. 2010. "An Overview and Framework for PD Backtesting and Benchmarking." *Journal of the Operational Research Society, Special Issue on Consumer Credit Risk Modeling* 61: 359–373.

Committee of European Banking Supervisors (CEBS). 2005. "Guidelines on the Implementation, Validation and Assessment of Advanced Measurement (AMA) and Internal Ratings Based (IRB) Approaches." Technical report, CP10 consultation paper.

European Union. 2013. "Regulation (EU) no 575/2013 of the European Parliament and of the Council of 26 June 2013."

Federal Register. 2007. "Proposed Supervisory Guidance for Internal Ratings-Based Systems for Credit Risk, Advanced Measurement Approaches for Operational Risk, and the Supervisory Review Process (Pillar 2) Related to Basel II Implementation."

Fischer, M., and M. Pfeuffer. 2014. "A Statistical Repertoire for Quantitative Loss Given Default Validation: Overview, Illustration, Pitfalls and Extensions." *Journal of Risk Model*

Validation 8: 3–29.

Gupton, G., and R. Stein. 2005. “Losscalc v2: Dynamic Prediction of LGD, Moody's KMV Company, Modeling Methodology.”

Hamerle, A., R. Rauhmeier, and D. Rösch. 2003. “Uses and Misuses of Measures for Credit Rating Accuracy.” Available at SSRN 2354877.

Hong Kong Monetary Authority (HKMA). 2006. “Validating Risk Rating Systems under the IRB Approaches.”

Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression*. 2nd ed. New York: John Wiley & Sons.

Loterman, G., T. Brown, D. Martens, C. Mues, and B. Baesens. 2012. “Benchmarking Regression Algorithms for Loss Given Default Modeling.” *International Journal of Forecasting* 28: 161–170.

Moges, H., K. Dejaeger, W. Lemahieu, and B. Baesens. 2013. “A Multidimensional Analysis of Data Quality for Credit Risk Management: New Insights and Challenges.” *Information and Management* 50: 43–58.

Prudential Regulation Authority. 2013. “Internal Ratings Based Approaches.” Working paper.

Rösch, D. 2005. “An Empirical Comparison of Default Risk Forecasts from Alternative Credit Rating Philosophies.” *International Journal of Forecasting* 25(1), 37–51.

SAS Institute Inc. 2015. *SAS/STAT 14.1 User's Guide: Technical Report*. Cary, NC: SAS Institute.