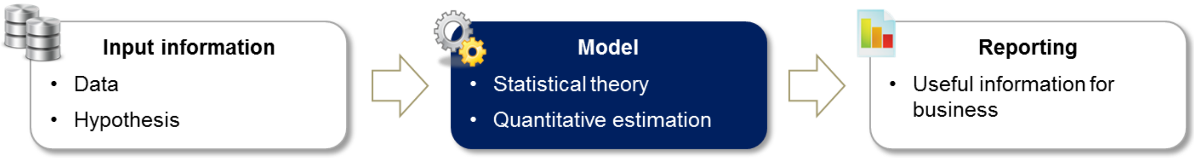
MODELLING WEEK 2014

IMPACT OF MODEL ESTIMATION ON MODEL RISK

March 2014

Introduction

Banks rely heavily on quantitative analysis and models in most aspects of financial decision making. They routinely use models for a broad range of activities, including underwriting credits; valuing exposures, instruments, and positions; measuring risk; managing and safeguarding client assets; determining capital and reserve adequacy; and many other activities. In recent years, banks have applied models to more complex products and with more ambitious scope, such as enterprise-wide risk measurement, while the markets in which they are used have also broadened and changed.



The expanding use of models in all aspects of banking reflects the extent to which models can improve business decisions but, what if models are misused or not well-defined? In 2012, large trading losses occurred at JPMorgan's Chief Investment Office, based on transactions booked through its London branch. The underlying story of the so called "London whale", talks about how one London-based quant was working on a new VaR (Value at Risk) model implemented in excel, and how decisions based on this model lead to enormous losses. After all this, an exhaustive investigation showed serious flaws in the model: volatility appeared lower than it should have because of issues with this model that can, in part, be traced back to how the model was used.

Similar cases have occurred in several financial institutions, causing severe losses due to poor models, misuse, wrong decisions based on them, lack of deep review and validation, etc., ending up in millions of euros in fines to those entities.

Within this framework, models can also be considered as a source of risk. The possible adverse consequences (including financial loss) of decisions based on models that are incorrect or misused is called **model risk**.

Model risk occurs primarily for two reasons:

- The model may have fundamental errors and may produce inaccurate outputs when viewed against the design objective and intended business uses. Model errors may include simplifications, wrong hypothesis, etc.
- The model may be used incorrectly or inappropriately. Even a fundamentally sound model producing accurate outputs consistent with the design objective of the model may exhibit high model risk if misapplied or misused.

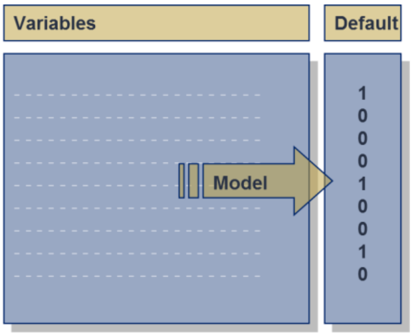
Banks should identify the sources of risk and assess the magnitude since every entity should have sufficient resources to absorb the losses of its activity. These resources are reflected in:

- **Expected losses**: reflect the average amount of losses that is really expected to be lost over a period of time. These are covered by provisions.

- **Unexpected losses**: represent the volatility of the credit losses around the level of expected losses. The resources which must legally be maintained by an Entity to cover the unexpected losses are the Regulatory Capital. As well, each entity determines a statistical confidence level indicating the desired level of capitalization.

Entities and regulators are becoming **increasingly interested in model risk**. Despite this fact and the existence of policies about model risk management, **its quantification and capital requirements are still an opened problem**.

As an example, one of the most extended models used daily in Banks is the scoring model. It is based on a formula that assigns points using known information to predict an unknown future outcome.



Credit scoring models are based on historical data from a bank's portfolio (retail banking, individuals and SMEs) that serves to predict the likelihood of a customer defaulting on a new account. They provide a score of every binomial customer-operation to conclude whether they are "good" or "bad".

This way, the process of admitting new operations can be automated and decisions are made based on objective data.

Variables to be included in a scoring model must cover a range of different profiles such as information of the operation, customer's income and outcome, negative credit data, etc., and they have to be statistically treated in order to ensure there are no errors on data (for instances, earlier ending date than start date), that the variable has not a big amount of missing values, etc.

1. Problem to be solved

The tasks to be completed are:

- **Statistical treatment of data** from a fictitious retail portfolio provided by Management Solutions:
  - Perform a univariate analysis for every variable so as to check whether variables are sufficiently informed (percentage of missing values), whether their values are within the expected range, etc.
  - At the sight of these analysis results, carry out a treatment for the missing and outlier values.
- Fit the all provided variables once they have been treated to a **logistic regression** to obtain the scoring model.
- Identify **potential sources of error** in the built model. As an example we could take into account:
  - The inherent error of the model
  - Confidence intervals
  - Lack of sample
  - Model use
- For every source of error, analyze possible **solutions** to avoid them or mitigate them.
- Estimate a **formula that quantifies the actual error** made.

2. Phases

2.1. Definition of the problem and clarification of doubts

In this first phase, Management Solutions will present the problem in greater detail, providing those ideas that have been developed so far, and will clarify any raising doubts about the understanding of the problem.

Management Solutions will also provide a polished and ready-to-use Excel file, so that the smallest amount of time will be spent on data processing. It will contain a bunch of selected variables that will be included in the model and the information of whether each operation has defaulted or not.

2.2. Phase 1: Estimation of scoring model parameters

Firstly, given the sample with the variables to use, students will perform a statistical analysis on the variables and then missing and outlier values will be treated.

Afterwards, the likelihood of default of each operation will be calculated by means of a logistic regression. Logistic or logit regression is used to predict a binary response (default or non-default) based on one or more predictor variables. Its expression is:

$$s = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \cdots - \beta_n x_n}}$$

where x_1 to x_n are the given variables, s is the scoring output and β_0 to β_n, the model estimators.

The binary response (default or non-default) is obtained after setting a cut point for the score.

2.3. Phase 2: Model error identification and quantification

List all possible errors made at the estimation process or when used. As an example, confidence intervals, variable misspecification, lack of sample, etc. could be analysed as a possible source of error and therefore model risk.

Once they are identified, try to quantify the error by providing a formula or a confidence interval of every individual error and discuss possible ways to avoid these errors or to mitigate them.

Finally, to complete this exercise, try to give a general formula that encloses all errors made in the model, as a measure of model risk.

2.4. Phase 3: Exposition and discussion of results

The case study will conclude with the students' presentation of the used methods and obtained results.