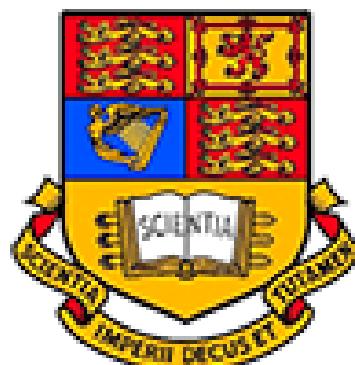


# MATH60131/MATH70131

## Consumer Credit Risk Modelling

Dr. Andrés Benchimol  
Room 535, Huxley Building  
[a.benchimol@imperial.ac.uk](mailto:a.benchimol@imperial.ac.uk)



**Imperial College  
London**

# Consumer Credit Risk Modelling

## Chapter 1: Introduction

## Overview



- **Credit scoring** is a systematic means by which banks and other financial institutions can gather information about borrowers or applicants for loans regarding their creditworthiness.
- Credit scores are used in automated decision processing and for financial risk management.
- Credit scoring makes use of many advanced and sophisticated methods from statistics and data mining.
- During this course, we will develop an understanding of the value of credit scoring and we will review the mathematical methods that underpin it.

## What will you learn? >

- The course content was designed with input from financial institutions.
- The emphasis is on ***practical*** application of statistical method, rather than statistical theory.
- Applied statistics requires an understanding of the business objective, and we will cover the consumer credit industry.
- Application of statistical method often requires ***subjective*** judgement in decision-making during statistical modelling:-
  - There may be more than one correct way to solve a problem;
  - But there are some solutions that are wrong.

## Course structure



- We will cover the following topic areas:
  - Building credit scoring models.
  - Evaluation of credit scoring models.
  - Data analysis and model refinement.
  - Lending goals: Uses of credit scores.
- Notes will be available on Blackboard.
- A discussion board will also be used on Blackboard for questions and discussion about the course.
- Problem sheets will be released during the course. These are voluntary but are an opportunity for you to test your understanding and receive feedback.
- Computer lab:
  - During the course, we will arrange one computer lab.
  - This is for you to get practical experience using credit scoring methods.
  - We will work in the R programming language.

## Course pre-requisites



Statistical modeling I (M2S2).

In particular, the following topics:

- Maximum likelihood estimation
- Hypothesis testing and use of p-values
- Likelihood ratio test
- Confidence intervals
- Linear regression

## Course assessment



Assessment	Weight
Project course work (due on early December)	25%
3 <sup>rd</sup> year's examination: 3 questions (1.5 hours)	75%
4 <sup>th</sup> year's examination: 3 questions + Mastery question (2 hours)	75%

### Project

- You will be provided with your own individual sample data set of credit data.
- You will build a credit scorecard, applying the methods you learn during the course.
- You will write a report describing your model and its performance.

## Recommended textbooks



- (1) Thomas, Lyn C. (2009).  
Consumer Credit Models: Pricing, Profit and Portfolios, OUP.
- (2) Anderson R (2007).  
The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation, OUP.
- (3) Thomas L.C., Edelman D.B. and Crook J.N. (2002).  
Credit scoring and its applications, SIAM, Philadelphia.

## Credit scoring: An overview



For the remainder of this chapter, we will cover the following introductory topics:

- What is credit scoring?
- The aim of credit scoring
- Formal models for credit scoring

## What is “scoring”?



Very simply:

*Scoring is a way to grade or rank the severity or risk of an event.*

Scores are used in many different application areas.

- **Glasgow coma score**

A score from 3 to 15 indicating the conscious state of an individual.  
(15=wide awake; 3=deep coma/dead).

- **Physiology score**

Provides a measure of the severity of illness of patient in critical care.

- **Marketing score**

Score on individuals or companies related to their chance of take-up of a marketing campaign.

So what you learn here will be applicable in many different areas.

You may be able to apply the methods you learn here for statistical decision making in your future career.

- **Consumer Credit**

For Consumer credit, the credit score is used to determine the creditworthiness of an individual applying for (or with) a loan, credit card or other financial product.

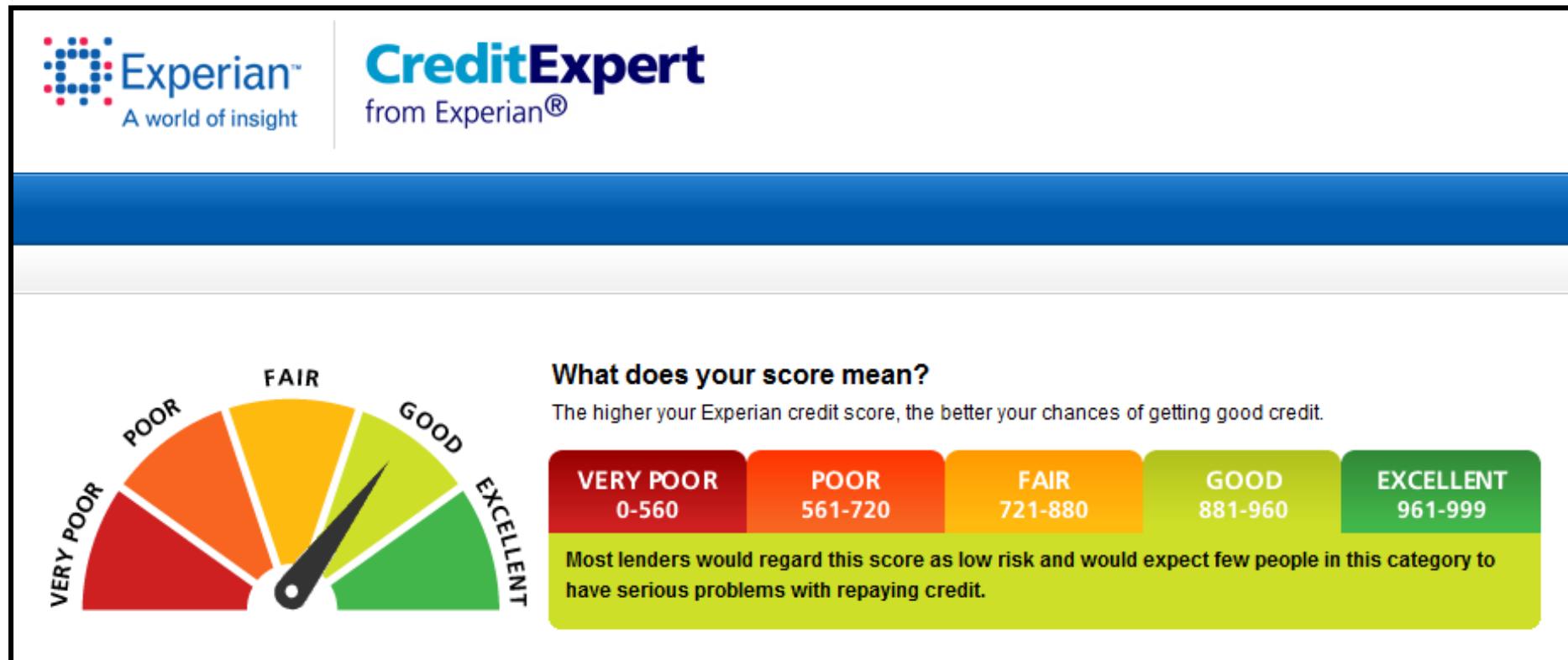
Credit scores may be calculated by banks but often generic scores are bought from credit bureaus such as Experian, Equifax or Callcredit.

### **Exercise 1.1**

Search the web and find *another* application of statistical scores.

*Examples.*

This is the Experian credit score (UK):



This is an Equifax credit score (UK):

**Your Score: 500**



**This reflects that:**

Your score is highly predictive that you would be a strong candidate for credit. You are well above the average for UK borrowers. Most Credit grantors would consider this score excellent. While many factors other than your score influence the decision to grant credit you will probably not be turned down for loans based on your score alone. Many Credit grantors may also offer you preferential terms such as higher credit limits or lower rates.

## **Exercise 1.2**

Have a look at the credit bureau websites.  
They contain a great deal of educational material.  
See what you can find out about credit scores there.

## Risk and Reward: for the Lender



Before describing the aims of credit scoring it is worth considering the goal of banks in lending.

They are looking to maximize reward with minimum risk.

What are the **rewards** of lending?

- Primarily, earning interest payments on loans and account fees.
- Also, establishing customer loyalty to sell products in the future.

What are the **risks** involved with lending?

- Primarily, the risk of the borrower not making repayments on the loan on time. This is called "**delinquency**".
- Ultimately, not paying back the whole of the loan with interest on time. This is called a "**default**" on the loan.
- Fraudulent use of credit that results in losses.

## Risk and Reward: for the Borrower



Of course, this works both ways.

There are risks and rewards for the borrower too.

- **Reward:** Get funds that are needed now to be paid back later.
  - Economists talk of the **time-value of money**: £100 is worth *more* now than in one year's time, even after taking inflation into account.
  - For instance, it would be almost impossible for many young people to buy a home if credit was not available to do so.
- **Risk:** Will need to pay back loan and interest payments over time – this is a financial burden.
  - There is a risk the borrower will not be able to pay back and will get into financial difficulties.

## The Five C's of Credit Scoring >

These are the five issues that are important to a lender when considering a borrower.

### **1. Capacity**

Is the borrower capable of repaying the loan? (eg income)

### **2. Capital**

What is at stake to the borrower if the loan goes bad?  
(eg mortgage deposit)

### **3. Collateral**

What assets can the borrower offer if the loan is not repaid?  
(eg mortgage)

### **4. Conditions**

What is the intended purpose of the loan?

### **5. Character**

What is the borrower like? Is he/she trustworthy?

For credit scoring, we are seeking more information about Capacity, Character and Collateral.

## Aims of Credit Scoring



Fundamentally, the aim of credit scoring is to provide banks with intelligence about the borrower (or applicant) that allows them to assess risk and potential reward.

Particular common aims can be categorized as either a part of a

- Decision process, or
- Probability estimation.

## Decision process →

- **Application scoring.**

Use scores to decide who to accept for a loan or other financial product and who to reject.

- **Behavioural scoring.**

Use scores to determine how well-behaved existing borrowers are and therefore to anticipate any problems in the future.

- **Fraud detection.**

Use scores to detect unusual credit use which may be the result of fraud.

- **Cross-selling.**

Decide who to target for additional financial products.

## Probability estimation →

- Predict probability of default.
- Allows for quantitative risk management and measurement of expected profitability/return.
- Valuable for capital requirement calculations and required by regulatory authorities.

## Why Formal Credit Models?



Up until the 1980's, most loan applications were assessed by a bank officer who would use his/her judgement to decide whether the loan was provided based on their personal view of the applicant. It would also be their responsibility to monitor the loan.

Now almost all loans are assessed and monitored using automated formal models.

*Why did this change?*

There are several reasons:

- **Performance.**

There is some evidence that, even with their expert judgement, bankers do not perform as well as formal models in choosing who is best to give a loan to.

- **Technology.**

Advances in computing technology enabled an explosion in consumer credit; it requires a fast automated system to process the huge number of applications this generated.

- **Objectivity.**

The judgement of banking staff is highly subjective, meaning there is a lack of consistency in their decisions. In contrast, formal systems provide a highly objective means of decision making.

- Two bank officers may make different decisions given the same individual.
- Or, even the same bank officer may make a different decision based on mood: whether they are tired, irritable, happy, etc.
- *Computers don't get moody or tired!*

- **Transparency.**

Formal systems provide a fully transparent system and equality between applicants (adjusting for genuine risk factors). This is a legal requirement and is demanded by regulatory authorities.

- **Cost.**

Using automated formal systems is cheaper.

## Exercise 1.3.

Suppose you are a bank manager.

You have six people come to you for a credit card.

Pick **three** of the applicants to receive a credit card and reject three.

Decide on the basis of who *you* think is least likely to default.

Employed ?	Monthly Income (£)	Home phone?	Residence type?	Months in residence	Months in current job	Accept or reject?
No	1,145	Yes	Home owner	48	12	
Yes	15,500	Yes	Renter	48	192	
Yes	900	Yes	Renter	96	12	
Yes	5,000	Yes	Renter	48	168	
No	400	Yes	Renter	12	0	
Yes	3,145	No	Home owner	96	36	

Later, we will return to this problem to see how good your judgement is compared to a formal model.

## Which Formal Models Are Available? >

The traditional and standard model for credit scoring is **logistic regression**.

However, around logistic regression, there are many different modelling techniques that we will cover.

For instance:

- Segmentation
- Variable selection
- Discretization of continuous variables
- Weights of evidence.

Additionally, there are several alternative models, which we will cover in this course:

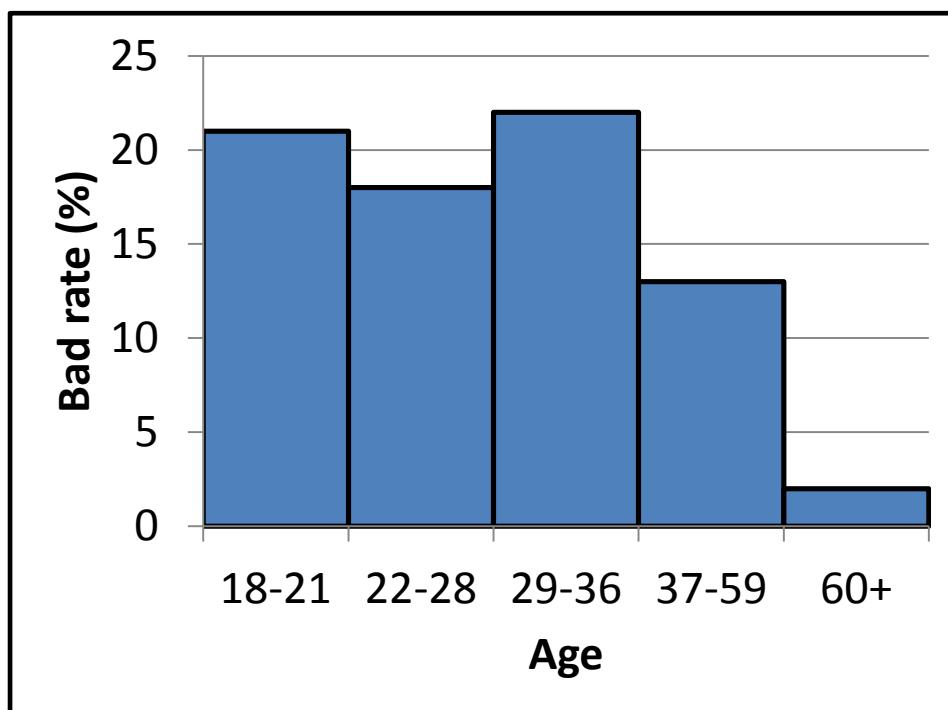
- Ordinary least squares linear regression
- Decision tree classifier
- Naive Bayes classifier

## Risk factors and credit score



There are many characteristics of borrowers which are risk factors in the sense that they are associated with default (or delinquency in loan repayment).

For instance, consider this graph linking age with default rates for a portfolio of loans.



There is a general reduction in risk with age.

*However, there is a spike for the 29-36 age group.*

Why?

Bad rates for different age groups for a sample of credit card holders (adapted from Thomas, Edelman and Crook, figure 8.2).

We want to build a **multivariate model** that combines all credit risk factors that are available to produce an overall assessment of risk.

The overall assessment of risk (or creditworthiness) is captured in the credit score.

Therefore, the credit score is *not* a quantity that is directly measured from an individual (eg it isn't tattooed on our arms!).

It is *computed* by the bank as an indicator of creditworthiness.

Therefore, the credit score is a **latent variable**.

It is *about* an individual, but is *not* directly measurable.

## Generic and custom credit scores >

There are two different types of credit score.

### 1. Generic credit score.

These are general scores about individuals regarding their

- *individual characteristics*  
(age, income, home ownership, etc) and
- *credit history*  
(number of times they have defaulted in the past, etc).

Generic scores are not linked to particular loan products.

Generic scores are often provided by credit bureaus.

## 2. Custom credit score.

These are built by banks and financial institutions to compute creditworthiness for specific loan products.

The credit scores will be computed from past data about borrowers with the same (or similar) loan product.

For a specific product, custom credit scores are likely to be more accurate than generic scores.

Generic scores are often used as a variable *within* a custom credit scoring model.

## Overview of Chapter 1



In this chapter we have:

1. Reviewed the course content and assessment.
2. Learned about the aims of credit scoring.
3. Had an overview of the formal models for credit scoring.

In the next chapter, we will look at how we can build a credit scoring model using ordinary least squares regression.

# Consumer Credit Risk Modelling

## Chapter 2: Credit Scoring Models

## Overview



In this chapter we shall:

1. Define a scorecard;
2. Review ordinary least squares (OLS) regression;
3. Look at how to develop a credit scoring model using OLS;
4. Show how to use the model to calculate credit scores.

## “Characteristics” and “variables”: some terminology >

In statistics, we sometimes refer to predictor variables as **covariates** or **independent variables**.

In credit scoring, the borrower variables are referred to as borrower **characteristics** (such as age, income, etc).

The word **attribute** is also used, but ambiguously: sometimes for a variable and sometimes for a level of a variable.

Additionally, in machine learning and data mining, a variable is also known as a **feature!**

In this course, we will generally refer to **predictor variables** ...

... But you should be familiar with all possible terminology.

## The Scorecard



We will link borrower characteristics to a credit score using a statistical model. This is the **scorecard**.

Traditionally, the credit score is a *linear* combination of weighted variable values. For  $m$  borrower characteristics we have the score

$$s(\mathbf{X}) = \beta_0 + \sum_{i=1}^m \beta_i X_i$$

where

- $\mathbf{X} = (X_1, \dots, X_m)$  is a random vector of borrower characteristics,
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)$  is a vector of weights on each characteristic, and
- $\beta_0$  is a constant term.

This formula is more easily expressed in vector notation:

$$s(\mathbf{X}) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{X}$$

The vector of weights  $\boldsymbol{\beta}$  form a **scorecard** which is used to score a given individual.

*Non-linear* scorecards are also possible and we will cover some types of these later in the course.

## Outcome of a loan



The credit score is typically linked to the risk of default.

Or, it could be linked to whichever event we are specifically modelling, such as repayment delinquency or fraudulent transactions.

We use a binary outcome variable  $Y \in \{0,1\}$  to represent the outcome with  $Y = 1$  indicating the outcome we are interested in. Examples are given below.

$Y = 1$	$Y = 0$
Default	Non-default
Delinquency	Non-delinquency
Fraudulent transaction	Legitimate transaction
Positive event	Negative event
Bad customer	Good customer

In general, we will refer to an outcome as a positive or negative outcome.

*Note: the terms "positive" and "negative" are here intended in a statistical sense, rather than moral sense.*

## Classification problem

&gt;

Lenders are interested to determine the outcome of a loan:  
Eg will it default or not?

For this reason, credit scoring is often viewed as a *classification problem*.

That is, default and non-default are treated as separate classes and we want to build a model that is able to *discriminate* between the two classes.

So, given loan details, we would like a model that gives an accurate estimate about whether it will default in the future, or not.

Many classification methods are available. We will cover some, but the *Methods for Data Science* course provides some more detail about the classification problem and available models.

## Multivariate regression model



Suppose we have a data set  $D$  of  $n$  observations  $[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)]$  with  $m$  predictor variables.

We can use this data to develop a scorecard.

Then,  $D$  is called the ***training set*** or ***design set***.

As we have seen, our scorecards are usually linear in the predictor variables, thus it is reasonable to try to construct a model that links the predictor variables linearly to outcome:

$$Y = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{X} + \varepsilon$$

where  $\varepsilon$  is a random variable error term.

This is a linear *multivariate regression model*:

- $\beta_0$  is a fixed constant called the **intercept**;
- $\boldsymbol{\beta}$  is a fixed  $m \times 1$  vector of **coefficients**; one for each predictor variable.

## Ordinary least squares regression



Ordinary least squares (OLS) regression is a commonly used method to estimate  $\beta_0$  and  $\beta$ .

It works by minimizing the sum of square error  $\varepsilon^2$  over the training set.

To demonstrate this, it is easiest to use matrix notation (we use a lower bar for matrices). So let

$$\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underline{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1 \\ 1 & \mathbf{x}_2 \\ \vdots & \\ 1 & \mathbf{x}_n \end{bmatrix}, \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \underline{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Therefore,

$$\underline{Y} = \underline{\mathbf{X}} \underline{\mathbf{b}} + \underline{\varepsilon}$$

with respect to an estimate  $\mathbf{b}$  of  $\beta_0$  and  $\beta$ .

Therefore, the sum of square error is given by loss function

$$\begin{aligned} L(\underline{\mathbf{b}}) &= \underline{\boldsymbol{\varepsilon}}' \underline{\boldsymbol{\varepsilon}} \\ &= (\underline{Y} - \underline{\mathbf{X}} \underline{\mathbf{b}})' (\underline{Y} - \underline{\mathbf{X}} \underline{\mathbf{b}}) \\ &= \underline{Y}' \underline{Y} - 2(\underline{Y}' \underline{\mathbf{X}}) \underline{\mathbf{b}} + \underline{\mathbf{b}}' (\underline{\mathbf{X}}' \underline{\mathbf{X}}) \underline{\mathbf{b}} \end{aligned}$$

Differentiate with respect to  $\mathbf{b}$  and set to zero to find the minima:

$$\frac{\partial L(\underline{\mathbf{b}})}{\partial \underline{\mathbf{b}}} = -2(\underline{Y}' \underline{\mathbf{X}}) + 2\underline{\mathbf{b}}' (\underline{\mathbf{X}}' \underline{\mathbf{X}}) = \mathbf{0}$$

Therefore,  $(\underline{\mathbf{X}}' \underline{\mathbf{X}}) \underline{\mathbf{b}} = \underline{\mathbf{X}}' \underline{Y}$  and so the OLS estimator gives

$$\underline{\mathbf{b}} = (\underline{\mathbf{X}}' \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}' \underline{Y}$$

Note that the first element of  $\underline{\mathbf{b}}$  is the estimate  $b_0$  of  $\beta_0$  and the remainder is the estimate  $\hat{\boldsymbol{\beta}} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$  of  $\boldsymbol{\beta}$ .

If we assume **strict exogeneity**, that is, the error term is independent of each observation, then  $E(\varepsilon | \mathbf{X} = \mathbf{x}) = 0$ .

It then follows that

$$E(Y | \mathbf{X} = \mathbf{x}) = E(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{X} + \varepsilon | \mathbf{X} = \mathbf{x}) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}$$

Since  $Y \in \{0,1\}$ , the regression formula could therefore be interpreted as the conditional probability of  $Y = 1$ , except that with OLS there is *no guarantee* that  $\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x} \in [0,1]$ .

## Credit scoring interpretation



The credit score is usually a model of *creditworthiness*, therefore if we take the outcome variable as non-default (ND):

$$Y_{\text{ND}} = \begin{cases} 0 & \text{if the borrower defaults} \\ 1 & \text{if the borrower does not default} \end{cases}$$

then the estimate  $\mathbf{b}$  forms a scorecard and the credit score for any borrower with observation  $\mathbf{x}$  is given by

$$s(\mathbf{x}) = b_0 + \hat{\boldsymbol{\beta}} \cdot \mathbf{x}$$

In particular, higher values of  $s(\mathbf{x})$  indicate better creditworthiness.

## Interpretation of association with default >

Each coefficient estimate  $\hat{\beta}$  gives the *relative effect* of each predictor variable in the scorecard.

In particular, the direction of each coefficient estimate on a predictor variable can be interpreted as follows:

Coefficient estimate	Association with creditworthiness	Association with default
Positive	Positive	Negative
Negative	Negative	Positive

The magnitude of a coefficient estimate is less easy to interpret since it depends on the unit of measurement of the predictor variable.

*Example 2.1.*

Suppose we have 12 observations with just two predictor variables:

- Emp months = consecutive months in current employment;
- Renter = whether the borrower rents his/her place of residence.

Also, for each observation, we have the outcome of a loan:

- 0=default, 1=non-default.

Observation	1	2	3	4	5	6	7	8	9	10	11	12
Emp months	2	48	6	36	3	22	0	10	28	3	58	24
Renter	1	0	1	0	0	1	1	0	0	1	1	0
Outcome <i>(non-default)</i>	0	1	1	1	0	0	1	1	1	0	1	1

From this training data, we compute the following OLS estimation of the scorecard:

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0.5756 \\ 0.0104 \\ -0.2330 \end{bmatrix}$$

### Interpretation:

- The intercept  $b_0$  is the same for all observation, so has no relative effect;
- The first coefficient  $b_1$  is for Emp months and is positive; it indicates a positive association between time in current employment and creditworthiness.
- The second coefficient  $b_2$  is for Renters and is negative; it indicates a negative association between renting one's place of residence and creditworthiness.
  - This is a relative association and is a contrast between renting and non-renting (eg this may mean home owner, or living in family home with no rent).

Given the scorecard, credit scores can be constructed for each of the observations using the model formula  $\hat{y} = b_0 + \hat{\beta} \cdot \mathbf{x}$  :

Observation	1	2	3	4	5	6	7	8	9	10	11	12
Emp months	2	48	6	36	3	22	0	10	28	3	58	24
Renter	1	0	1	0	0	1	1	0	0	1	1	0
Outcome <i>(non-default)</i>	0	1	1	1	0	0	1	1	1	0	1	1
Score	0.36	1.07	0.40	0.95	0.61	0.57	0.34	0.68	0.87	0.37	0.95	0.82

Notice:

- On average, the defaulters (outcome=0) have lower scores.
- Most scores are in the interval [0,1] and could be interpreted as probability of non-default, *but not always* (eg the score for observation 2 is 1.07).

## Credit score rescaling



Credit bureaus and banks usually adjust their credit scores so that they are positive integers within a given range.

For example, Experian scores are between 0 and 999.

This is a purely cosmetic step since non-mathematicians (the public, in general, or senior bank management) are not usually comfortable with negative real numbers.

The adjustment is usually a linear rescaling of the score given directly by the model.

However, for our purposes, we are interested in the “raw” model scores so this is what we will work with for the remainder of the course.

*Example 2.2.*

Suppose we have a raw model score  $s(\mathbf{x})$  which is typically in the range  $s_L$  to  $s_U$ , although we cannot guarantee this. We want to present a score between 0 and 999. We therefore truncate and rescale as follows:

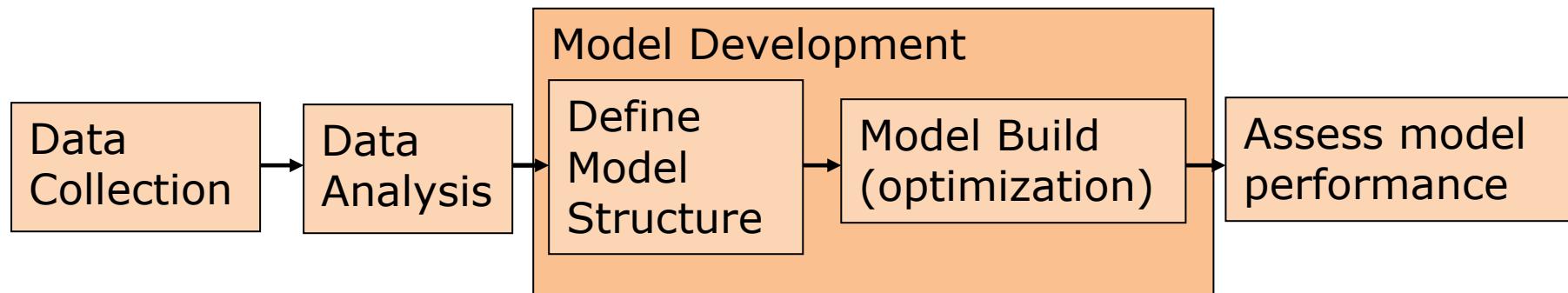
$$s_{\text{cosmetic}}(\mathbf{x}) = \begin{cases} 0, & \text{if } s(\mathbf{x}) < s_L \\ \left\lfloor \frac{999}{s_U - s_L} (s(\mathbf{x}) - s_L) \right\rfloor, & \text{if } s_L \leq s(\mathbf{x}) < s_U \\ 999, & \text{if } s(\mathbf{x}) \geq s_U \end{cases}$$

## Model Development



Building a scorecard using a statistical model, such as OLS regression, is only one step in model development.

Credit scorecard development is a much larger process involving several absolutely essential steps (this is true for *statistical modelling*, in general):-



We will be studying these steps in detail over the course of future lectures.

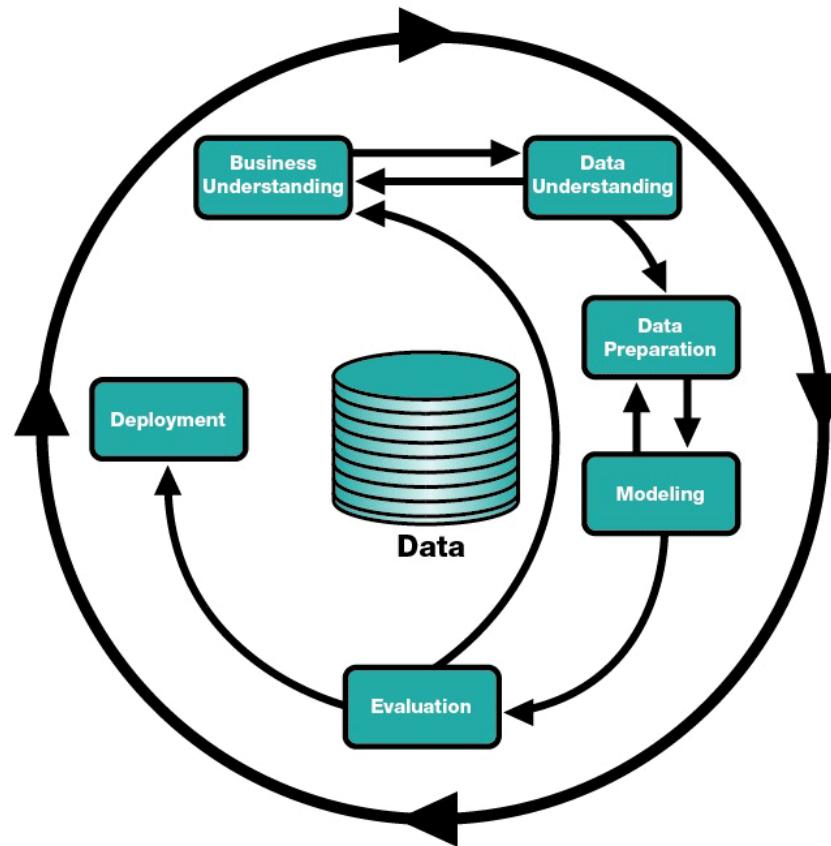
Based on this, course chapters can be *broadly* categorized into 5 key areas:-

<i>Credit scoring: Goals</i>	<i>Data collection &amp; analysis</i>	<i>Define Model Structure</i>	<i>Model Build</i>	<i>Assess model performance</i>
Introduction (1)	Credit data analysis (8)	Variable selection (10)	Credit scoring models (2)	Model interpretation (6)
Consumer credit (3)	Data preparation (9)	Interaction terms and segmentation (11)	Logistic regression (5)	Scorecard performance (7 & 15)
Lending process (4)		Segmentation and decision trees (12)	Selection bias and reject inference (14)	Sampling and testing (13)
				Probability calibration (16)
				Cost-based measures (17)

Chapter number in brackets.

**CRISP-DM**

Cross Industry Standard Process for Data Mining:



See <http://crisp-dm.eu/home/about-crisp-dm/>

## Overview of Chapter 2

&gt;

Topics covered in this chapter were:-

1. Scorecards;
2. Classification;
3. Ordinary least squares (OLS) regression;
4. Developing, interpreting and using a simple scorecard using OLS regression;
5. Data mining development cycle.

# Consumer Credit Risk Modelling

## Chapter 3: Consumer Credit

## Overview



In this chapter we will cover:

1. The consumer credit industry.
2. Types of credit available.
3. Bad debt

## Consumer credit institutions – the main players >

### Retail Banks

- Provide the majority of mortgages and consumer credit.
- Major retail banks are listed on stock exchanges.
- Usually part of multinationals that also include corporate and investment arms.

### Credit Unions and Building Societies

- These are not-for-profit cooperatives run by their members.
- Credit unions are popular in the USA where they have tax-exempt status.
- In the UK, building societies are mutually-owned financial institutions (eg Nationwide).
- Smaller local credit unions are becoming increasingly popular (eg the Islington and City Credit Union).

## **Credit card companies**

- Financial institutions specializing in credit cards.

## **Student Loans Company**

- Administers government-funded loans and grants in the UK.

## **Credit reference agencies, or credit bureaus**

- Collect and provide data about individuals to inform decisions on providing credit.
- Service to retail banks.

## Regulators

- Protect the consumer and user of financial services.
- Endeavour to provide stable financial markets:
  - eg ensure retail banks maintain sufficient capital reserves to back up loans.
- In UK, regulation is now divided between two bodies to cover these two functions:
  - Financial Conduct Authority (FCA);
  - Prudential Regulation Authority (PRA) – part of Bank of England;

## Types of credit >

Financial institutions generally have four different classes of loan.

1. **Loans to governments** (eg through sovereign bonds).
2. **Corporate loans** (to large corporations).
3. **Business loans** (to small and medium-size businesses).
4. **Consumer loans** (to individuals).

Since individual loans are so much higher in value for the first two classes and extensive government and corporate information is usually available, each potential loan is assessed in great detail and the methods used for lender decision are very different from that used for consumer credit.

In this course, our interest is consumer credit. However, scoring is also used for small businesses too.

## Types of consumer credit



Mortgage loan	Loan secured on a physical property.
Credit card *	Credit availability with upper credit limit and no fixed repayment plan.
Personal loan	Loan with fixed repayments over time.  Secured loans are based on physical property, car, etc.
Overdraft *	Credit limit available in current account for a fee and usually a limited duration.
Hire purchase	Hire item (eg car) with fixed payments, with final availability to own.

\* Examples of "revolving credit" (credit that is automatically renewed as debts are paid off).

## Bad Debt



Borrowers are expected to repay their loan plus interest and any fees.

- Usually repayment is made in regular instalments and typically each month.
- For mortgages and personal loans, the repayment amount will be fixed and agreed prior to acceptance by the borrower.
- For credit cards, there is usually a minimum repayment amount each month based on the balance outstanding on the card (eg 2% of balance).

As long as the borrower meets the repayments, there is no problem and they are considered a *good* borrower.

## Delinquency

When a borrower misses at least one payment they are contractually delinquent. The borrower may have accidentally missed the payment and they will make up the shortfall in the next month, so this is not necessarily a problem, but repeated delinquency is a problem.

## Default

If a borrower is delinquent multiple consecutive months or for an extended period then they are considered to be in default.

There is no single definition, but it is often defined as 90 days overdue on minimum repayment

- In particular, this definition was agreed internationally in the Basel Accords.

Defaults need to be followed up by the banks to try to recover the debt.

- In the first instance, this involves contacting the borrower.
- Ultimately it may lead to legal action;
- If the loan was a mortgage or secured in some way, the lender can take possession of the property/security.
- Or, the debt may be sold on to a debt collection agency.
- Following up defaults lead to additional costs for the bank, on top of the loss of loan amount and interest.

## Write-off / Charge-off

- At some point, the lender will write-off the debt.
- This means no further action will be taken for recovery.
- The bank will assess how much of the debt they have recovered and the overall loss they have made on the bad debt.
  - Quite often, bad debts are fully recovered and the banks have only incurred an administrative cost.

## Who will default and when? >

### Can't Pay

Basic requirement:

Income – Costs > Minimum repayment.

If Income declines then this may no longer hold.

Triggers: Loss of job, divorce, illness.

If Costs increase then it may also not hold.

Triggers: Inflation, family expansion, divorce, financial indiscipline.

## Won't Pay

Borrower has sufficient funds, but...

Feels no moral obligation to repay and feels it would be beneficial not to.

Fraud: Some individuals intentionally use credit (theirs or others) without the intention of repaying, and supply false information to do this.

## Hierarchy of repayments

When people cannot afford to repay debt, they usually prioritize repayments.

- Traditionally, the priority was on their mortgage, then transport, then credit card.
- However, this has changed and the preference now for many borrowers is to repay credit card debt first, since this is so important to pay for day-to-day expenses.

## How common is consumer default?



Default rates in USA:

Percentage of accounts in default (per annum).

<i>Month/Year</i>	<i>Oct 2010</i>	<i>Aug 2015</i>
First mortgage	2.91%	0.84%
Second mortgage	1.79%	0.57%
Bank cards	6.91%	2.71%
Auto loans	1.92%	0.90%

Source: S&P/Experian Consumer Credit Default Indices

Note: October 2010 are high due to the 2008 Credit Crunch.

## Overview of Chapter 3



In this chapter we have covered the following topics:

1. The consumer credit industry.
2. Types of credit.
3. Bad debt classifications.

# Consumer Credit Risk Modelling

## Chapter 4: The Lending Process

## Overview



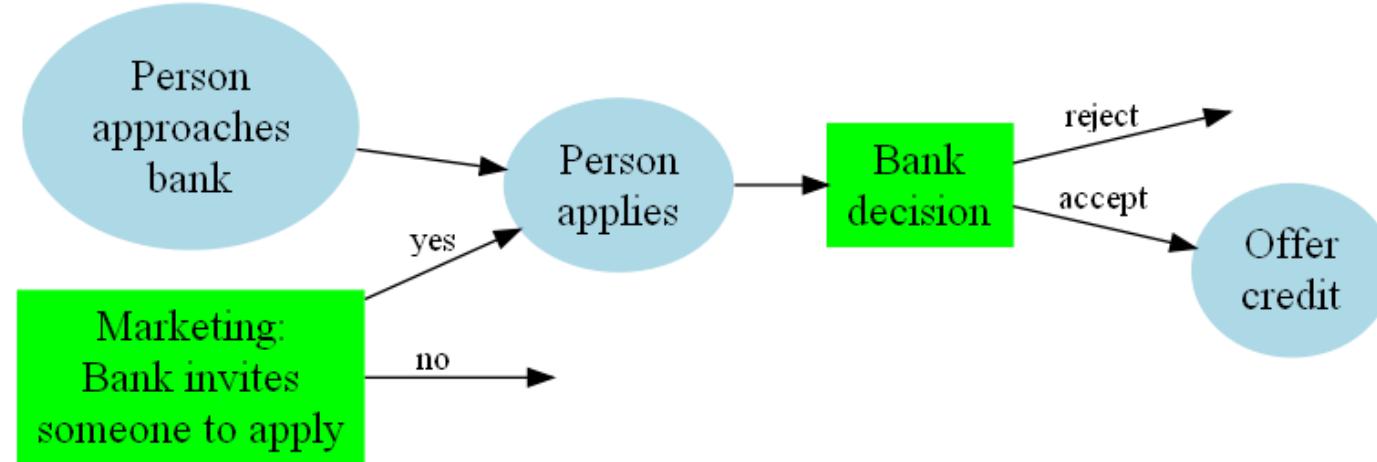
In this chapter we will cover:

1. The lending decision process.
2. Use of credit scores.
3. Risk grades.

## Consumer credit process

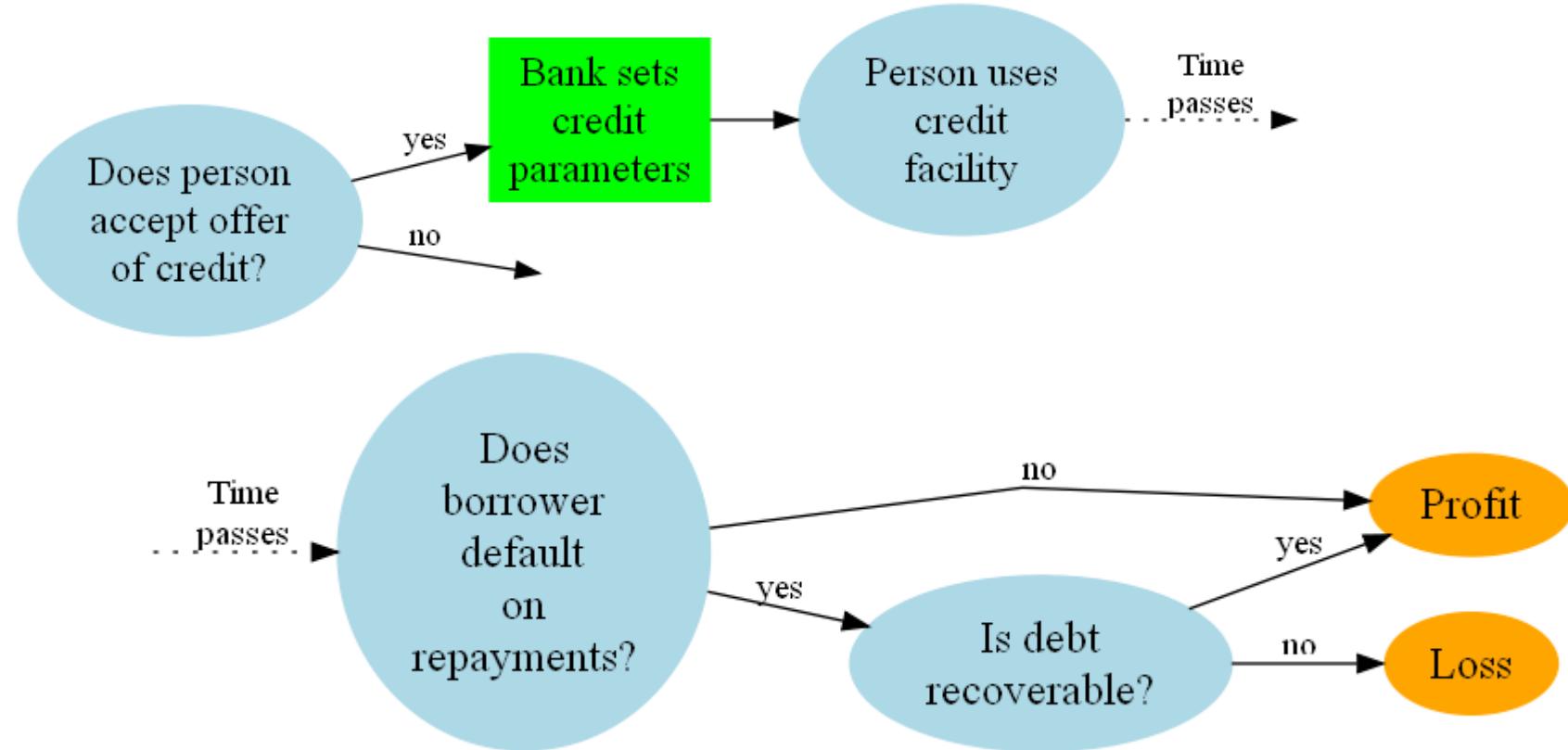


### 1. Application process



Lender decisions are in green.

## 2. Account management



Primary lender interests are in orange (profit and loss).

Note that possible credit parameters that can be set are:

- Term of loan (how long it is)
- Interest rate
- Fee
- Credit limit
- Repayment terms

Of course, some of these parameters are fixed by the offer  
(eg an *offer* of a 3-year loan fixes the loan term).

Lender decisions need to be informed sufficiently to maximize profit/minimize loss.

In particular, credit scoring informs the lender decision to accept or reject an application.

## Credit scores and the lending decision process >

Credit scores are developed to rate people's *propensity* to default. The *higher* the credit score, the *less risky* the individual is.

Some groups of people have a higher risk to default than others.

Credit scores are used to assist in the lending decision process and to predict probability of default.

In the lending process, there are three decision points for the lender:

1. Who to market a financial product to.
2. Which loan applicants to accept (application scoring).
3. What loan parameters to set (eg credit limit, repayment terms, interest rate).

Scoring models can be used for the first two decisions.

Application scoring is the most common use and we will look at this decision process in more detail.

Additionally, credit scoring can be used as part of the monitoring process for a portfolio of loans.

## Application Scoring >

In application scoring we want to accept those with a high credit score and reject those with a low credit score.

Hence, the lender specifies a **cut-off score** that will define who to accept and who to reject.

- If an applicant's score  $\leq$  cut-off, then reject the application;
- If an applicant's score  $>$  cut-off, then accept the application.

Therefore:

- A lower cut-off implies the lender is willing to accept a *higher risk* for *high volume*.
- A higher cut-off implies the lender is willing to accept *lower volume* for *low risk*.

There is always a trade-off between risk and volume:

- Too much risk will generate too many losses.
- Too little volume will not generate sufficient profit.

## Deciding the cut-off score >

The cut-off score is decided on the basis of a number of factors related to the lenders goals.

1. Specify a cut-off score which maximizes expected profits.
2. Specify a minimum volume or proportion of loans to be accepted.
  - A lender may do this if they want to maintain market share or promote a product.
  - They are willing to sacrifice immediate profit for long-term position in the market.

We will look at methods to optimize for these two goals in a later chapter.

Traditionally, retail banks have been conservative, only accepting loan applications they believe will be repaid.

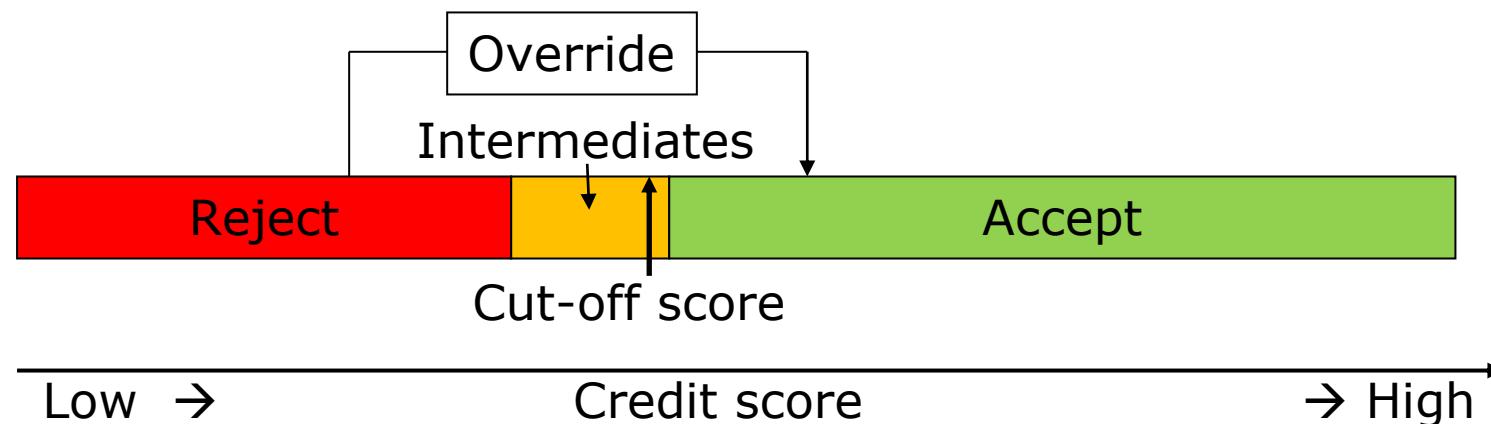
However, now banks are ready to accept some bad debt if it will maximize expected profit.

- The amount of bad debt is controlled using *risk management*.
- Sometimes this doesn't always work if the lenders are over-optimistic in their assessment of risk (eg the 2007 mortgage crisis in US).

## Intermediates and Overrides



Sometimes further analysis and manual intervention is required to follow up automated lending decisions. These are based on intermediate cases and decision overrides.



Sometimes lenders consider a third class of **intermediates** (or *indeterminates*).

- These are people that cannot be classed easily as good or bad.
- Generally intermediates are individuals with scores close to the cut-off score.
- To make a decision about intermediates requires further analysis, possibly manual. Also, the decision may legitimately be different at different times, depending on the demands of the financial product.
- Including intermediates makes this an intrinsically three class problem.

Occasionally, the automated reject/accept decision is changed by the lenders. This is called an **override**.

- This may be because further information on an individual is available;
- Or a customer is considered important and a reject is changed to an accept.

## Risk Grades



Sometimes it is useful to group borrowers by their general risk category.

- It helps in communicating credit score data to the general public and senior bank management.
- It also allows the riskiness of a whole portfolio of loans to be assessed in terms of summaries of cases in each group.
- Probabilities of outcome within each group can be used to compute overall risk of default.
- Risk groups are defined by ranges of scores.

*Example 4.1*

The risk grades for the Experian generic credit score:

<i>Score</i>	<i>Creditworthiness assessment</i>
0 - 560	Very poor
561 - 720	Poor
721 - 880	Fair
881 - 960	Good
961 - 999	Excellent

## Overview of this chapter



In this chapter we have covered the following topics:-

1. The lending decision process.
2. Use of credit scores.
3. Risk grades.

# Consumer Credit Risk Modelling

## Chapter 5: Naïve Bayes classifier and Logistic regression

## Overview

In this chapter we shall cover:

1. Probability of default.
2. Log-odds function.
3. Naive Bayes classifier.
4. Logistic regression.
5. Maximum likelihood estimation (MLE) of logistic regression.
6. Look at how to develop a credit scoring model using logistic regression.
7. Show how to use the model to calculate credit scores.

## Credit score and Probability of outcome

It is difficult to get accurate exact predictions of default, especially in retail credit, therefore we focus on getting estimates of probability of outcome.

- Let  $Y \in \{0,1\}$  be an outcome variable with 1 representing “bad” outcome (eg default or delinquency).
- Let  $s(\mathbf{X})$  represent a score on a random vector of characteristics  $\mathbf{X} = (X_1, \dots, X_m)$ .

We quantify the risk by assigning a probability of outcome to each credit score  $s$  using a **link function**  $f_L$  as follows:

$$P(Y = 0 | \mathbf{X} = \mathbf{x}) = f_L(s(\mathbf{x})).$$

Typically, we want increasing scores to reflect increasing creditworthiness, therefore the link function should increase with score:

$$\text{For all } s_1 < s_2, \quad f_L(s_1) < f_L(s_2).$$

Also, of course,  $0 \leq f_L(s) \leq 1$  for all  $s$ .

Remember in Chapter 2, OLS regression did not fulfil the criteria  $0 \leq s(\mathbf{X}) \leq 1$  and there is not a natural link function which will guarantee this, whilst ensuring the probabilities of outcome are accurate.

A truncation function is possible but this is awkward and arbitrary.

Therefore, we will introduce other models which are better suited to model probability of outcome:

- The Naive Bayes classifier
- Logistic regression

## Probability of default

&gt;

When the outcome of interest is default, then this gives the conditional **probability of default** (PD) as

$$P_D(\mathbf{x}) = P(Y = 1 | \mathbf{X} = \mathbf{x}) = 1 - f_L(s(\mathbf{x})).$$

The PD is much used in the credit industry and by regulators.

## Log-odds link function

&gt;

It is natural to use a log-odds function for scores:

$$s(\mathbf{x}) = \log\left(\frac{P(Y = 0|\mathbf{X} = \mathbf{x})}{P(Y = 1|\mathbf{X} = \mathbf{x})}\right)$$

[note: this is log to base  $e$ ]

This gives the inverse log-odds link function

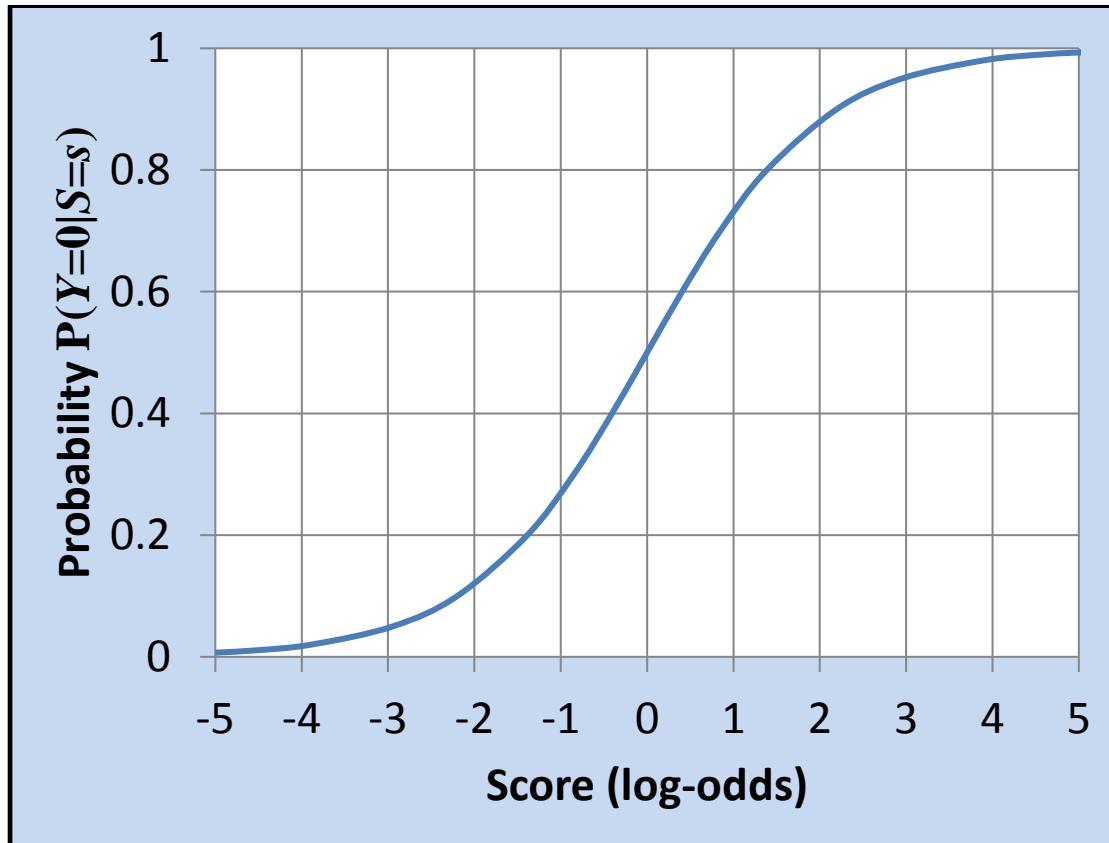
$$P(Y = 0|\mathbf{X} = \mathbf{x}) = f_{LO}(s(\mathbf{x})) = \frac{1}{1+e^{-s(\mathbf{x})}}.$$

The inverse log-odds link function has two main advantages:

1. The range is  $(0,1)$ , hence can be used to represent probabilities.
2. Probability is monotonically increasing with score.
3. Commonly used in the industry.
4. The logistic distribution is close to normal, but is easier to manipulate and implement.

Note, a normal link function could be used: this is called a *probit* link function.

The log-odds link function:



Because the “bad” outcomes are rare, typically  $P(Y = 0|\mathbf{X} = \mathbf{x}) > 0.5$ .  
This implies that most scores will be greater than 0  
(although this is not guaranteed!).

## Naive Bayes classifier



The Naive Bayes method follows naturally from the log-odds formulation of a scorecard *if we assume independence between the covariates* in the model, and using Bayes Rule.

Let  $\mathbf{X}$  be a vector of *categorical* variables.

By Bayes Rule,

$$\frac{P(Y = 0|\mathbf{X} = \mathbf{x})}{P(Y = 1|\mathbf{X} = \mathbf{x})} = \frac{P(\mathbf{X} = \mathbf{x}|Y = 0)}{P(\mathbf{X} = \mathbf{x}|Y = 1)} \frac{(1 - p_1)}{p_1}$$

where  $p_1 = P(Y = 1)$ .

If the covariates  $X_j$  in  $\mathbf{X}$  are conditionally independent of each other then

$$\frac{P(Y = 0|\mathbf{X} = \mathbf{x})}{P(Y = 1|\mathbf{X} = \mathbf{x})} = \frac{1 - p_1}{p_1} \prod_{j=1}^m \frac{P(X_j = x_j|Y = 0)}{P(X_j = x_j|Y = 1)}$$

and therefore, taking logs of both sides, and using log-odds scores,

$$s(\mathbf{x}) = w_0 + \sum_{j=1}^m w(x_j)$$

where  $\mathbf{x} = (x_1, \dots, x_m)$ ,  $w_0 = \log[(1 - p_1)/p_1]$  is the log-odds of the negative event and

$$w(x_j) = \log \left[ \frac{P(X_j = x_j | Y = 0)}{P(X_j = x_j | Y = 1)} \right]$$

is the **weights of evidence** (WOE) of a particular value of the  $j$ th predictor variable.

The prior probabilities and conditional probabilities in the WOE can be estimated from a training data set.

*Note: WOE will turn up several times through the module and we will cover it in more detail later.*

## Naive Bayes classifier: comments >

The Naive Bayes classifier is a very simple method to produce a linear scorecard based only on the WOE of each covariate.

However, the assumption of independence is almost never met.

For example, both *employment status* and *income* are common covariates to include in a scorecard. We would expect an association with different incomes for employed, unemployed, retired and self-employed.

Nevertheless, if we are careful about which variables we include, Naive Bayes may be sufficiently robust to be a good choice of classifier.

## Logistic regression



If we use the log-odds interpretation of the credit score, we arrive at the most commonly used credit scoring model based on logistic regression:

$$\log\left(\frac{P(Y = 0|\mathbf{X} = \mathbf{x}, \beta_0, \boldsymbol{\beta})}{P(Y = 1|\mathbf{X} = \mathbf{x}, \beta_0, \boldsymbol{\beta})}\right) = s(\mathbf{x}) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}$$

where

- $\beta_0$  is an intercept;
- and  $\boldsymbol{\beta}$  is a vector of coefficients, one for each predictor variable.

Note:

This is actually known as **binary** logistic regression, because it models a binary outcome  $Y \in \{0,1\}$ . Other versions of logistic regression are available for non-binary classification (eg ordinal and multinomial), but we do not need to cover these in this course.

## Training a logistic regression model >

How do we determine  $\beta_0$  and the vector of coefficients  $\beta$ ?

This is done by **training** on an existing data set of borrowers for which we already know their outcome.

Formally a **training set** is a sequence of predictor variable/outcome pairs for some  $n$  observations:

$$[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \dots, (\mathbf{x}_n, y_n)]$$

In the case of logistic regression, model fit to training data is achieved using maximum likelihood estimation (MLE).

## Maximum Likelihood Estimation (MLE) >

MLE is a general purpose method for parametric model estimation. We will make use of it to estimate the logistic regression.

If we have a model with parametric structure  $\theta$ , we can compute the **likelihood** that the model will generate a sequence of  $n$  observations  $D_1, \dots, D_n$ .

$$L(\theta; D_1, \dots, D_n) = P(D_1, \dots, D_n | \theta)$$

The model which best fits the data is selected as the one which maximizes this likelihood.

$$\hat{\theta} = \arg \max_{\theta} L(\theta; D_1, \dots, D_n)$$

If we *assume independence between the observations*, this then gives

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n P(D_i | \theta)$$

This MLE can be expressed more conveniently in terms of log-likelihoods (since log is monotonically increasing):

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^n \log P(D_i | \boldsymbol{\theta})$$

Remember:

- We do not know the true value of the parameter  $\boldsymbol{\theta}$ , but we want to estimate it.
- To distinguish the estimate from the true value, in our notation, we put a “hat” on the estimate:  $\hat{\boldsymbol{\theta}}$ .

MLE has several nice asymptotic properties:

- Consistency
- Asymptotic normality
- Efficiency.

## MLE for logistic regression

&gt;

Consider a training data set  $D_n$  with  $n$  observations (borrowers).

Remember

- $\mathbf{X}_i$  denotes values for predictor variables for observation  $i$ .
- $Y_i$  denotes the outcome for observation  $i$ , either 0 or 1.

Then the likelihood of the outcome for each observation  $i$  is given by

$$\begin{aligned} P(Y_i = 0 | \mathbf{X}_i = \mathbf{x}_i, \beta_0, \boldsymbol{\beta}) &\quad \text{if } Y_i = 0, \\ 1 - P(Y_i = 0 | \mathbf{X}_i = \mathbf{x}_i, \beta_0, \boldsymbol{\beta}) &\quad \text{if } Y_i = 1 \end{aligned}$$

which is

$$P(Y_i = y_i | \mathbf{X}_i = \mathbf{x}_i, \beta_0, \boldsymbol{\beta}) = P(Y_i = 0 | \mathbf{X}_i = \mathbf{x}_i, \beta_0, \boldsymbol{\beta})^{1-y_i} (1 - P(Y_i = 0 | \mathbf{X}_i = \mathbf{x}_i, \beta_0, \boldsymbol{\beta}))^{y_i}$$

giving log-likelihood for each observation:

$$(1 - y_i) \log P(Y_i = 0 | \mathbf{X}_i = \mathbf{x}_i, \beta_0, \boldsymbol{\beta}) + y_i \log(1 - P(Y_i = 0 | \mathbf{X}_i = \mathbf{x}_i, \beta_0, \boldsymbol{\beta}))$$

Assuming independence between observations, and using the log-odds link function, this gives the log-likelihood function for  $\beta_0$  and  $\beta$ :

$$\log L(\beta_0, \beta; D_1 \dots D_n) = \sum_{i=1}^n (1 - y_i) \log\left(\frac{1}{1 + e^{-(\beta_0 + \beta \cdot \mathbf{x}_i)}}\right) + y_i \log\left(\frac{1}{1 + e^{\beta_0 + \beta \cdot \mathbf{x}_i}}\right)$$

Differentiating by each coefficient in  $\beta$  and setting the derivative equal to zero to find the maxima gives

$$\sum_{i=1}^n \left( 1 - y_i - \left( \frac{1}{1 + e^{-(\beta_0 + \beta \cdot \mathbf{x}_i)}} \right) \right) = 0$$

and

$$\sum_{i=1}^n x_{ij} \left( 1 - y_i - \left( \frac{1}{1 + e^{-(\beta_0 + \beta \cdot \mathbf{x}_i)}} \right) \right) = 0$$

for each attribute  $j=1$  to  $m$ , where  $\mathbf{x}_i = (x_{i1} \dots x_{im})$ .

These are non-linear equations that can be solved by computer intensive processes such as Newton-Raphson methods.

## Standard errors on the MLE



Since  $\hat{\theta}$  is only an estimate of the best model to explain the data, it is possible to derive standard errors  $\hat{s}$  on the estimates.

[ Note: We do not cover the details of this in this course ]

Asymptotic normality for MLE is such that

$$\frac{(\hat{\theta}_j - \theta_j)}{\hat{s}_j} \rightarrow N(0,1) \text{ as } n \rightarrow \infty$$

where  $\hat{\theta}_j$ ,  $\theta_j$  and  $\hat{s}_j$  are the  $j$ th components of  $\hat{\theta}$ ,  $\theta$  and  $\hat{s}$  respectively and  $N(0,1)$  is the standard normal distribution.

This property then allows us to:

- Conduct hypothesis testing using the Wald test.
- Construct confidence intervals for the actual parameters.

## Hypothesis Test

We test the hypothesis that an estimated coefficient is not zero against the null hypothesis that it is zero.

That is, we testing if a parameter has a *genuine* measurable effect within the model.

- Null hypothesis:  $H_0: \theta_j = 0$
- Alternative hypothesis:  $H_1: \theta_j \neq 0$

The Wald test says reject  $H_0$  if  $\frac{|\hat{\theta}_j|}{\hat{s}_j} > z_{\alpha/2}$  for some significance level  $\alpha$ , where  $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$  and  $\Phi$  is the CDF for the standard normal distribution.

## Confidence intervals

The asymptotic normality property also allows us to compute confidence intervals (CIs):

$$P(\hat{\theta}_j - z_{\alpha/2} \hat{s}_j < \theta_j < \hat{\theta}_j + z_{\alpha/2} \hat{s}_j) \rightarrow 1 - \alpha$$

as  $n \rightarrow \infty$ .

This is a range of possible values of the parameter within a given confidence level  $1 - \alpha$ .

Note: the larger the confidence level, the broader the confidence interval.

*Example 5.1*

For large enough sample size, if we have an estimate of  $\theta=2.3$  with  $\hat{s}=0.2$  then calculate the 95% CI for the estimate.

- $\alpha = 1 - 0.95 = 0.05$ .
- Then  $z_{\alpha/2} \approx 1.96$ .
- So the CI is approximately  $(2.3 - 1.96 \times 0.2, 2.3 + 1.96 \times 0.2) = (1.908, 2.692)$ .

## Likelihood Ratio Test of Goodness-of-Fit

The maximized likelihood gives a measure of how well the model fits the data (1=perfect fit, 0=no fit). The ratio of likelihoods between two models, B “nested” in A, can be used to test whether the fit of A improves on B.

### Definitions

Suppose we have two models A and B with the same parametric structure except A has more parameters than B:

$$\boldsymbol{\theta}_A = (\theta_1, \dots, \theta_{m+r}) \text{ and } \boldsymbol{\theta}_B = (\theta_1, \dots, \theta_m)$$

Then B **is nested in** A.

The **null model** is a model with no parameters  $\boldsymbol{\theta}_\emptyset = ()$ .

Therefore the null model is nested in all other models.

The **Log-likelihood ratio (LR) statistic** is

$$\lambda = 2 \log \left( \frac{L(\hat{\theta}_A; D)}{L(\hat{\theta}_B; D)} \right)$$

given a training data set  $D$ .

We can use this as the basis of a hypothesis test to compare the likelihoods that the data was drawn from each of the models.

- Null hypothesis  $H_0$ : Model B represents the data:  $\theta = \hat{\theta}_B$ .
- Alternative hypothesis  $H_1$ : Model A represents the data:  $\theta = \hat{\theta}_A$ .

In particular, model B can be set to the null model to get a basic test of model fit for any logistic regression model.

**Theorem**

Under  $H_0$ :  $\lambda$  approximates a chi-square distribution with  $r$  degrees of freedom.

[ *Proof not given here* ]

This result can be used to compute a chi-square statistical significance test of model fit:

If  $1 - \chi_r^2(\lambda) < \alpha$  then reject  $H_0$ .

where  $\chi_r^2(\lambda) = F_\chi(\chi_r^2 \leq \lambda)$  is the CDF of the chi-square distribution with degree  $r$ , and  $\alpha$  is the test significance level.

### Example 5.2

The following logistic regression output was produced on a data set of 40,000 credit cards.

Log-likelihood Ratio = 1081 (p-value < 0.001)

Variable	Coeffi- cient	Estimate	Standard error	$z$	$P(> z )$
Intercept	$\beta_0$	0.116	0.0703	1.65	0.099
$X_1$ : Employed	$\beta_1$	+0.245	0.0285	8.62	<0.001
$X_2$ : Income (log)	$\beta_2$	+0.0774	0.0132	5.87	<0.001
$X_3$ : Home phone?	$\beta_3$	+0.637	0.0298	21.4	<0.001
Home owner		0			
$X_4$ : Renter	$\beta_4$	-0.121	0.0391	-3.10	0.002
$X_5$ : Lives with parents	$\beta_5$	-0.0531	0.0440	-1.21	0.227
$X_6$ : Other residence	$\beta_6$	+0.103	0.0671	1.53	0.126
$X_7$ : Months in residence	$\beta_7$	+0.000479	0.000102	4.69	<0.001
$X_8$ : Months in current job	$\beta_8$	+0.00363	0.000231	15.7	<0.001

\* Notice that the Home owner category is set as base residency category and so has no coefficient estimate. We will discuss this in a later lecture.

We have used logistic regression to model the “negative” outcome (ie  $Y = 0$ ).

- This may seem strange given that the outcome of interest is the positive one (eg default).
- However, this model ensures the log-odds scores are the right way round: ie increasing scores imply increasing creditworthiness.
- There is no material difference. If we had modelled  $Y = 1$ , the signs on the coefficient estimates would be reversed, but everything else would be the same.

We will look at how to interpret a logistic regression model in the next chapter.

## Sample results

Remember in the exercise in Chapter 1 we gave details of six borrowers. You were asked to select three to accept and three to reject.

Here the scores assigned by the model in Example 5.2 are shown. The observations with the three lowest scores are rejected by the model. The actual outcome in each case is also shown.

*How does your performance compare with the model?*

Emp-loyed ?	Monthly Income (£)	Home phone?	Residence type?	Months in residence	Months in current job	Score	Model accept or reject?	Actual outcome
No	1,145	Yes	Home owner	48	12	1.36	Reject	Good
Yes	15,500	Yes	Renter	48	192	2.34	Accept	Good
Yes	900	Yes	Renter	96	12	1.49	Accept	Good
Yes	5,000	Yes	Renter	48	168	2.17	Accept	Bad
No	400	Yes	Renter	12	0	1.10	Reject	Bad
Yes	3,145	No	Home owner	96	36	1.16	Reject	Bad

*Example 5.3*

Take the first borrower and apply the scorecard.

Variable	Value	Coefficient	Estimate	Value × Estimate
Intercept	n/a	$\beta_0$	0.116	0.116
$X_1$ : Employed	0	$\beta_1$	+0.245	0
$X_2$ : Income (log)	$\log(1145) = 7.04$	$\beta_2$	+0.0774	+0.545
$X_3$ : Home phone?	1	$\beta_3$	+0.637	+0.637
Home owner	1		0	0
$X_4$ : Renter	0	$\beta_4$	-0.121	0
$X_5$ : Lives with parents	0	$\beta_5$	-0.531	0
$X_6$ : Other residence	0		+0.103	0
$X_7$ : Months in residence	48	$\beta_6$	+0.000479	+0.023
$X_8$ : Months in current job	12	$\beta_7$	+0.00363	+0.044
<b>Score (sum)</b>				<b>+1.365</b>

*Example 5.3 continued*

Compute the PD of the borrower.

Score = 1.365

Remember, score

$$s(\mathbf{x}) = \log\left(\frac{P(Y = 0|\mathbf{X} = \mathbf{x})}{P(Y = 1|\mathbf{X} = \mathbf{x})}\right)$$

Therefore

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1+e^{s(\mathbf{x})}} \approx 0.20.$$

## Logistic regression in R >

Logistic regression is available in most statistics packages (eg SAS, Stata, R).

In particular, in R, logistic regression can be run using `glm`.

*Example 2.5*

```
train <- read.delim("train.txt")

glm.out <- with(train,
  glm(good ~ emp + log(income2+1) + res_phone
    + res_type_A + res_type_C
    + months_in_res + months_in_the_job,
  family = binomial("logit")) )

summary(glm.out)
```

## Overview of Chapter 5



Topics covered in this chapter were:

1. Log-odds credit scores.
2. Naive Bayes classifier.
3. Logistic regression.
4. Developing and using a simple scorecard using logistic regression.

Advanced reading about logistic regression and GLMs:

Dobson AJ and Barnett AG (2008), An introduction to Generalized Linear Models (3<sup>rd</sup> ed.), CRC Press

# Consumer Credit Risk Modelling

## Chapter 6: Model interpretation

## Overview

Once we have a logistic regression model, we want to understand what it means.

In particular, we want to know about:

- (1) *Model Fit*: Does the model adequately explain the training data?
- (2) *Association*: How are predictor variables associated with outcome (if at all)?

In this chapter, we will look at ways to interpret the model to answer these questions.

## Model Fit

The first question is addressed by the likelihood ratio (LR).

The higher this is, the better the fit to the data.

LR can be used to determine if the model is a statistically significantly better fit than the null model (ie one without any predictor variables).

The null hypothesis is that including the predictor variable does not give better fit than the null model.

We choose a significance level  $\alpha$  and reject the null hypothesis if the p-value for LR is less than  $\alpha$ .

## Association between variables

We talk about an association between predictor variables and an outcome variable.

Notice that association does **not** necessarily imply a *causal* link.

A **confounding** variable may explain the association.

*Example 6.1.*

- Suppose we find that drinking coffee is weakly associated with lung cancer?
- We do not conclude that drinking coffee causes lung cancer.
- A plausible explanation is that cigarette smokers also drink more coffee than non-smokers, therefore explaining the association (here cigarette smoking is the confounding variable).

*Example 6.2.*

- Often there is an association between housing type and default. In particular, tenants tend to be more likely to default than home owners.
- However, this is not a direct causal link.
- It is plausible to postulate common socio-economic conditions for an individual being both a property renter and being at risk of default.

**Causal links or associations?**

- However, note that for Credit Scoring, knowing causal links is not particularly important.
- It is the associations that are important to model, since these will lead to improved predictions of default and losses.

## Evidence of association

The second question requires we look at each predictor variable separately.

Then, there are several points of interpretation:

- (2a) Does the predictor variable genuinely have an association with outcome?
- (2b) What is the direction of the association?
- (2c) What is the magnitude of the association?

The coefficient tells us the association between the predictor variable and outcome.

- Notice, however, that the association of a predictor variable has to be interpreted in light of all the other predictor variables in the model, which are called *controlled variables*. If the model had included different predictor variables, then the evident association *may have been different*.
- If a coefficient is not zero, then there is an association between the predictor variable and outcome since a change in the value of the predictor variable implies a change in (the probability of) outcome.

## Testing for a non-zero coefficient

The model gives only a coefficient *estimate*. So if the estimate is sufficiently close to zero, it may be that the true value of the coefficient **is** zero.

We use the Wald statistic to test the null hypothesis that the coefficient is zero.

- Set a significance level  $\alpha$ .
- Reject the null hypothesis if the p-value is less than  $\alpha$ .
- If the p-value is greater than  $\alpha$ , then there is insufficient evidence to suppose an association between the predictor variable and outcome.

## Direction of association

The sign on the coefficient gives the direction of association in answer to (2b):

- A positive sign means a positive association between the predictor variable and outcome.
- A negative sign means a negative association between the predictor variable and outcome.

## Magnitude of association

The magnitude of the association (2c) is not so easily interpreted.

Very simply, the size of coefficient  $|\beta_j|$  tells us the magnitude of change for each unit change of the predictor variable.

However, there are two difficulties with this interpretation:-

- Firstly, it clearly means that the sizes of the coefficients are not directly comparable, since the magnitude of their association is also dependent on the unit of measurement, scale and general distribution of the predictor variable.
- Secondly, the coefficient size is linearly related only to the log-odds score and not to the outcome itself or probability of outcome. This is because they are non-linearly related to the predictor variables, through the log-odds link function.

## Marginal Effect

To determine the association further we need to consider the marginal effect.

### Definition

Suppose  $\mathbf{X}$  is a vector of random variables  $\mathbf{X} = (X_1, \dots, X_m)$

The marginal effect of a continuous variable  $X_j$  on a function  $f(\mathbf{x})$  is the slope of  $f(\mathbf{x})$  with respect to  $x_j$ , holding all other variables constant:

$$m_j = \frac{\partial f(\mathbf{x})}{\partial x_j}$$

The marginal effect on the *log-odds score* is then given by

$$\frac{\partial [\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}]}{\partial x_j} = \beta_j$$

and the marginal effect on the *probability of outcome* is given by

$$\begin{aligned}\frac{\partial P(Y = 0 | \mathbf{X} = \mathbf{x})}{\partial x_j} &= \frac{\partial \left[ \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x})}} \right]}{\partial x_j} \\ &= \frac{-1}{[1 + e^{-(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x})}]^2} \cdot e^{-(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x})} \cdot -\beta_j \\ &= \frac{e^{\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}}}{[1 + e^{\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}}]^2} \beta_j\end{aligned}$$

- Therefore, the marginal effect on probability of outcome depends on the values given for the other predictor variables and so is itself variable.
- Hence, we will normally interpret an association using the marginal effect on the log-odds score, which is just the coefficient value.
- In Chapter 12 we will see the value of marginal effects for more complex model structures.

*Example 6.3.*

Let us go back to interpret the model given in Example 5.2.

Log-likelihood Ratio = 1081 (p-value < 0.001)

Variable	Coefficient	Estimate	Standard error	z	P(> z )
Intercept	$\beta_0$	0.116	0.0703	1.65	0.099
Employed	$\beta_1$	+0.245	0.0285	8.62	<0.001
Income (log)	$\beta_2$	+0.0774	0.0132	5.87	<0.001
Home phone?	$\beta_3$	+0.637	0.0298	21.4	<0.001
Home owner		0			
Renter	$\beta_4$	-0.121	0.0391	-3.10	0.002
Lives with parents	$\beta_5$	-0.0531	0.0440	-1.21	0.227
Other residence	$\beta_6$	+0.103	0.0671	1.53	0.126
Months in residence	$\beta_7$	+0.000479	0.000102	4.69	<0.001
Months in current job	$\beta_8$	+0.00363	0.000231	15.7	<0.001

## Interpretation:

1. The model is a better fit than the null model since  $LR=1081$  with  $p<0.001$ .
2. There is sufficient evidence, at 1% significance level, that there is an association with default for several variables: being employed, having a residential phone, being a renter (relative to home owners), months in residence and months in current job and income.
3. Living with parents and other residence type do not show evidence of association with default, at a 1% significance level, relative to home owners.

4. The direction of effects is such that

- Being employed, having higher income, having a residential phone, months in residence and months in current job are all positively association with creditworthiness;
- Being a renter has a negative association with creditworthiness, relative to home owners.

5. Marginal effects on log-odds credit score are given directly by the coefficient estimates.

## Odds ratios

Odds Ratios (ORs) can also be used to interpret predictor variable effect size in logistic regression. They are commonly used in medical statistics.

If we have two exclusive events that have probabilities  $p_1$  and  $p_2$  of occurring then they have odds  $p_1/(1 - p_1)$  and  $p_2/(1 - p_2)$  respectively.

Their odds ratio is then defined as

$$OR = \frac{p_2/(1 - p_2)}{p_1/(1 - p_1)}$$

Now, suppose we want to contrast the probabilities of an outcome ( $Y = 0$ ) given two different instances  $\mathbf{x}_1$  and  $\mathbf{x}_2$ .

Then, for the log-odds score, this gives

$$\begin{aligned} OR(\mathbf{x}_1, \mathbf{x}_2) &= \frac{P(Y = 0 | \mathbf{X} = \mathbf{x}_2) / (1 - P(Y = 0 | \mathbf{X} = \mathbf{x}_2))}{P(Y = 0 | \mathbf{X} = \mathbf{x}_1) / (1 - P(Y = 0 | \mathbf{X} = \mathbf{x}_1))} \\ &= \frac{\exp(s(\mathbf{x}_2))}{\exp(s(\mathbf{x}_1))} = \exp(s(\mathbf{x}_2) - s(\mathbf{x}_1)) \\ &= \exp(\boldsymbol{\beta} \cdot (\mathbf{x}_2 - \mathbf{x}_1)) \end{aligned}$$

In particular, suppose we want to see the effect of a  $k$ -unit change on the  $j$ th covariate, *ceteris paribus* (ie keeping all other terms constant).

Let  $\mathbf{x}_1 = (x_1, \dots, x_j, \dots, x_m)$  and  $\mathbf{x}_2 = (x_1, \dots, x_j + k, \dots, x_m)$

Then

$$OR_j = OR(\mathbf{x}_1, \mathbf{x}_2) = \exp(\boldsymbol{\beta} \cdot (\mathbf{x}_2 - \mathbf{x}_1)) = \exp(k\beta_j)$$

Therefore, the effect of each variable on the odds of outcome is easily given by the coefficient itself.

*Example 6.4.*

The odds ratios of some of the covariates given in Example 6.3 are given below along with an interpretation:

- The OR of being employed  $\approx \exp(1 \times 0.245) \approx 1.28$ .

Therefore being employed means the odds of not defaulting increases by a multiple of 1.28, relative to being unemployed.

- The OR of being a renter  $\approx \exp(1 \times -0.121) \approx 0.886$ .

Therefore being a renter means the odds of not defaulting decreases by a multiple of 0.886, relative to being a home owner.

- The OR of each 12 months in current job  $\approx \exp(12 \times +0.00363) \approx 1.045$ .

Therefore 12 more months in current job means the odds of not defaulting increases by a multiple of 1.045.

## Overview of Chapter 6

Topics covered in this chapter were:

- Model interpretation
- Model fit
- Association between variables
- Marginal effect
- Odds Ratios

# Consumer Credit Risk Modelling

## Chapter 7: Scorecard performance, Part 1

## Overview

In this chapter we will cover:

- Types of model performance measures
- Cut-off decisions and classification
- Receiver Operating Characteristic (ROC) curve
- Area under the ROC curve (AUC)
- Divergence and Information Gain

## Introduction

We will look at several credit score modelling methods. But first we need to introduce some techniques to determine what makes a “good” credit scoring model.

The model build may give measures of goodness of fit and statistical significance.

For example, for logistic regression:

The Wald statistics on covariates and the overall likelihood ratio.

But...

*What we are interested in is how well the model performs in predicting new cases.*

For application scoring, this usually means how well does the model perform in predicting default.

We use a **validation data set** to assess the performance of a model, comparing model *prediction* against *actual outcome*.

## Types of assessment

There are broadly three types of performance measure that we consider:

1. How good is the model at classifying borrowers?
2. How good are the models at estimating probabilities?
3. How well do the models estimate the profit/loss of an individual borrower?

As we will see, the three measures are linked but it is convenient to distinguish them into the three groups.

In this chapter, we will look at the first set of performance measures.

The next two types will be presented in future chapters.

## The cut-off decision and classification errors

Remember that a cut-off score is set to determine when to accept or reject an application.

We want to accept as many applications as possible, but with as few defaults as possible.

Therefore, if we have a fixed cut-off score  $c$  and a borrower with credit score  $S$ , then the *predicted* outcome  $\hat{y}$  for the borrower is

$$\begin{aligned}\hat{y} &= 1 \text{ if } S \leq c, \\ \hat{y} &= 0 \text{ if } S > c.\end{aligned}$$

Remember that we take an outcome 1 as a “bad” outcome (eg default) and outcome 0 as a “good” outcome (eg non-default).

Let  $y \in \{0,1\}$  be the *actual* outcome.

Then, there are two types of classification error:

Type I	False positives	$\hat{y} = 1$ and $y = 0$	Predicted positive, but actually negative.
Type II	False negatives	$\hat{y} = 0$ and $y = 1$	Predicted negative, but actually positive.

## Definitions

Define

- $F_0(c) \triangleq P(S \leq c | Y = 0)$  as the cumulative distribution function (CDF) of scores that are predicted positive amongst those that are negative (corresponding to Type I error);
- $F_1(c) \triangleq P(S \leq c | Y = 1)$  as the CDF of scores that are predicted positive amongst those that are positive.

Therefore,

- $1 - F_0(c)$  is the complementary CDF of scores that are predicted negative amongst those that are negative;
- $1 - F_1(c)$  is the complementary CDF of scores that are predicted negative amongst those that are positive (corresponding to Type II error).

Let  $p_1 \triangleq P(Y = 1)$ . Therefore,  $1 - p_1 = P(Y = 0)$ .

## Confusion matrix

A confusion matrix shows predictive outcome as a matrix with predictions in the rows and actual outcomes in the columns.

		Actual outcome	
		Positive	Negative
Prediction	Positive	$nF_1(c)p_1$	$nF_0(c)(1 - p_1)$
	Negative	$n(1 - F_1(c))p_1$	$n(1 - F_0(c))(1 - p_1)$

Shaded area indicates errors.

Based on a validation data set with  $n$  observations.

Unfortunately, the confusion matrix gives a summary of accuracy in four measures.

To compare models, we would prefer just one.

## Empirical frequencies

In practice, these distributions need to be given as empirical frequencies based on score data. These can be expressed as follows.

Suppose we have  $n$  observations of scores  $s_1, s_2, \dots, s_n$  with corresponding outcomes  $y_1, y_2, \dots, y_n$ . Then,

- $n_y = \sum_{i=1}^n I(y_i = y)$  is the count of observations with outcome  $y$ ;
- $\tilde{p}_1 = n_1/n$  is the empirical estimate of  $P(Y = 1)$ ;
- $\tilde{F}_y(c) = \frac{\sum_{i=1}^n I(s_i \leq c \text{ & } y_i = y)}{n_y}$  is the empirical distribution for  $F_y(c)$

where  $I(\cdot)$  is the indicator function, returning 1 if its argument is true and 0 otherwise.

Notice that when empirical frequencies are used, the cells of the confusion matrix become simple counts over the score data.

## Error rate

A naive measure of classification error is the proportion of cases that are either type of error. This is the error rate given by

$$\text{err}(c) \triangleq F_0(c)(1 - p_1) + (1 - F_1(c))p_1.$$

**However, there are problems with error rate and it is generally not used in credit scoring.**

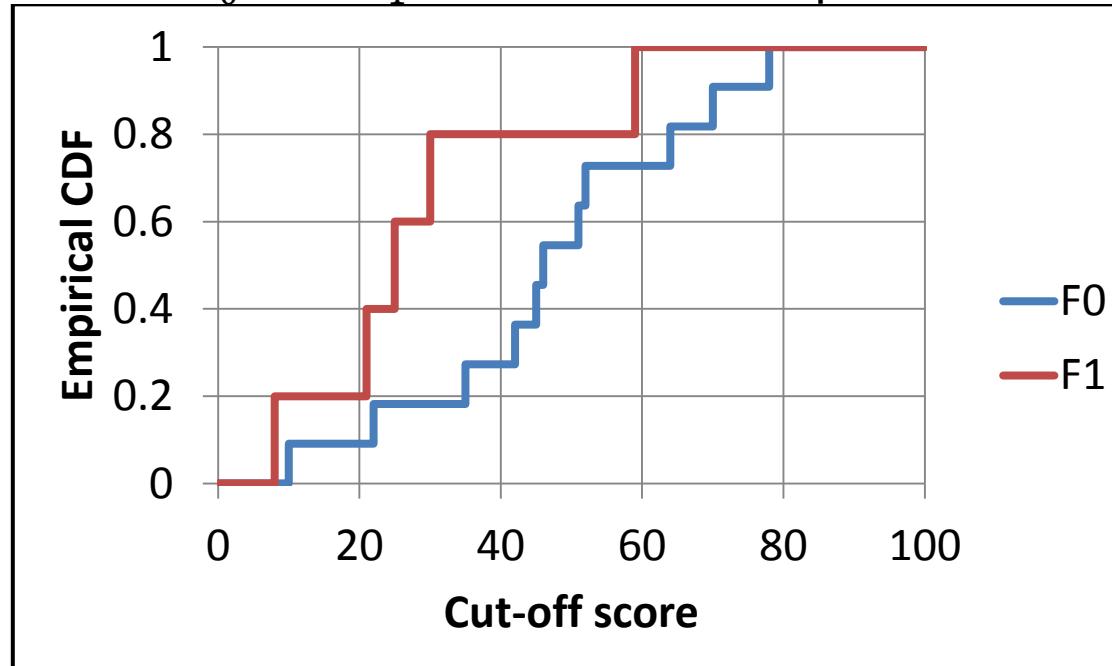
1. For credit scoring, usually the proportion of “bad events” (defaults) is much less than “good” events and so Type I errors dominate the error rate.
2. This will encourage models that do not reject sufficient bad loans.
3. Ultimately we would expect the loss from a single default to *outweigh* the benefit of many rejected “good” borrowers.  
That is, the two types of errors should not be treated equally.

*Example 7.1*

Consider 16 applicants with different scores and outcomes.

Score	8	10	21	22	25	30	35	42	45	46	51	52	59	64	70	78
Actual outcome	1	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0

Empirical distributions  $\tilde{F}_0$  and  $\tilde{F}_1$  for the 16 example borrowers.



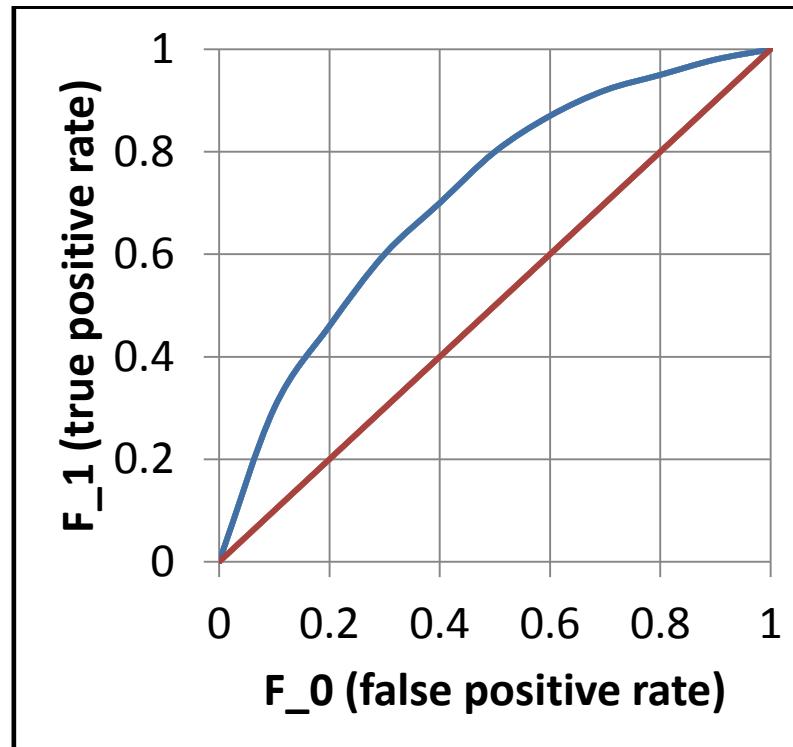
Notice that  $\tilde{F}_0$  and  $\tilde{F}_1$  diverge.

## ROC curve

A widely used tool to assess class discrimination accuracy is the **Receiver Operating Characteristic** (ROC) curve.

- Developed originally for signal detection theory, hence the name.
- Has the merit of being independent of any specific cut-off score or class distribution.
- Plots  $F_1$  on vertical axis against  $F_0$  on the horizontal axis: that is, false positive rate against true positive rate.

Typically, a ROC curve looks like this:



The blue line is the ROC curve.

The red line is a reference line (it represents an uninformative model).

## Characteristics of the ROC curve

- True positive rate ( $F_1(c)$ ) is also known as **sensitivity**.  
True negative rate ( $1 - F_0(c)$ ) is also known as **specificity**.
- The ROC curve shows the trade-off between true positive rate and true negative rate. In general, as one is increased, so the other decreases.
- Must pass through point (0,0) since this is the extreme case when cut-off is so low, no scores are less (eg all applications are accepted).
- Must pass through point (1,1) since this is the extreme case when cut-off is so high, all scores are less (eg all applications are rejected).
- The best model has ROC curve that passes through (0,1) since this is the case when there are no errors of either type  
(ie  $F_0(c)=0$  and  $1-F_1(c)=0$ ).

- A model that has no discriminatory power is such that  $F_0(c) = F_1(c)$  for all  $c$ . This is represented by a straight line from  $(0,0)$  to  $(1,1)$ : the red line in the example above.

Proof. If a model has no discriminatory power, then score is independent of the outcome: ie  $F_0(c) = P(S \leq c | Y = 0) = P(S \leq c)$  and  $F_1(c) = P(S \leq c | Y = 1) = P(S \leq c)$ , hence  $F_0(c) = F_1(c)$ .

- Where  $F_0$  and  $F_1$  are both differentiable, the slope on the ROC curve is  $F_1'(c)/F_0'(c)$ .

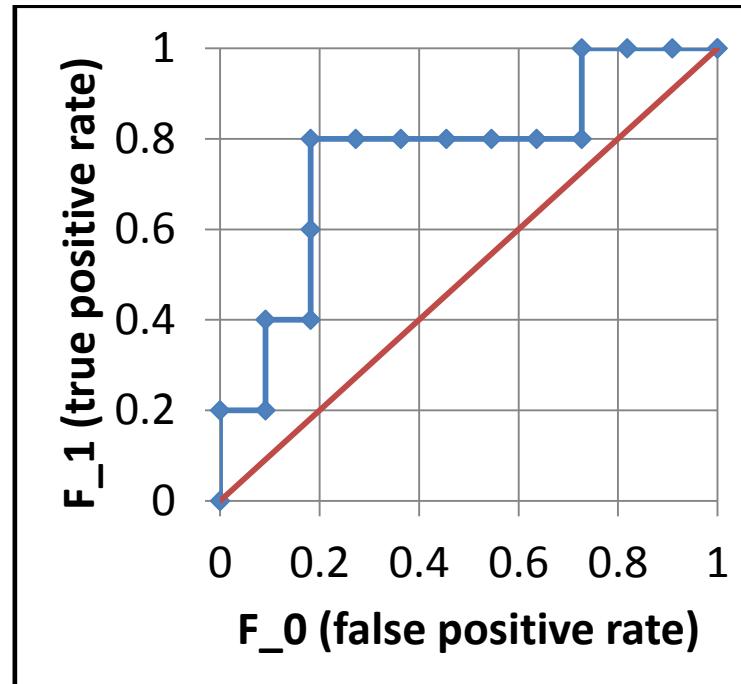
Proof:  $x = F_0(c), y = F_1(c)$ , therefore  $\frac{dy}{dx} = \frac{dF_1(c)/dc}{dF_0(c)/dc} = \frac{F_1'(c)}{F_0'(c)}$ .

*Example 7.2*

Again, consider the 16 applicants from example 7.1, with different scores and outcomes.

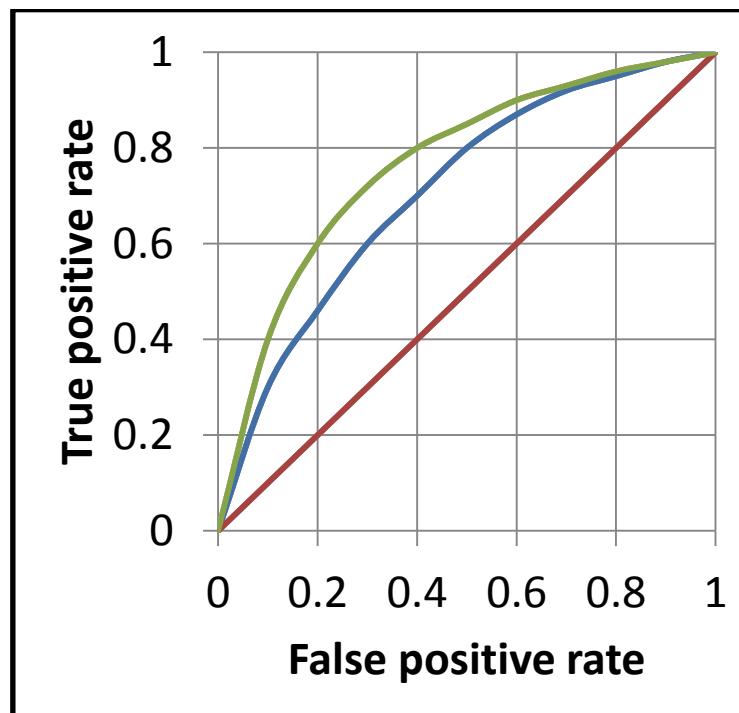
Score	8	10	21	22	25	30	35	42	45	46	51	52	59	64	70	78
Outcome	1	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0

The ROC curve is given as follows.



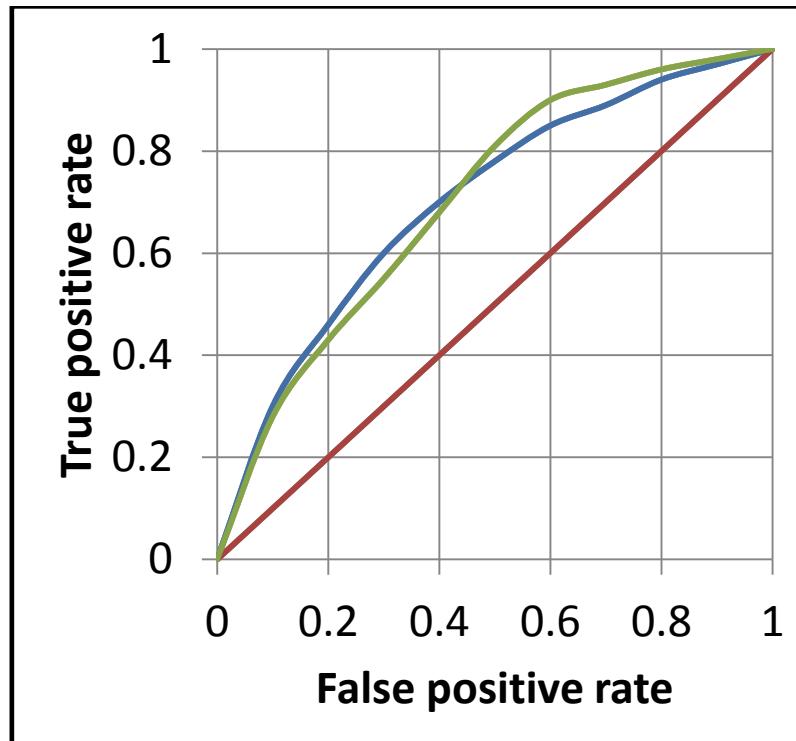
## Comparing models using the ROC curve

Consider two models A and B that produce two different ROC curves on the same validation data set.



ROC curve for model A is blue and model B is green.

Model B outperforms model A over the whole range of the curve, since its curve is always higher, so B seems to be the better model.  
However, not all comparisons between ROC curves are so clear-cut.  
Consider:



ROC curve for model A is blue and model B is green.

Model A is good for low false positive rates, whereas model B is good for high false positive rates. Therefore, it is difficult to determine a “best” model.

For credit scoring, it is low cut-offs (eg rejecting few applications) that are usually considered, so it is the lower left of the ROC curve which is usually most important. However, where do we draw the line for such comparisons?

The ROC curve is useful to view behaviour of a model across different cut-off scores.

However, it does not give a single measure of discrimination, which is what we really want for model comparison.

## Area under the ROC curve

A popular measure of discrimination is the *area under the ROC curve* (AUC) given by

$$A \triangleq \int_C F_1(c)F_0'(c)dc$$

In particular,

- AUC=0.5 corresponds to a model with no classification power.
- AUC=1 corresponds to a model with maximal classification power.
- Models can be directly compared using their AUC.  
If model A has a higher AUC than model B then it is considered the better model in terms of discriminatory power.

## Estimate of AUC

Suppose we have a validation data set with  $n$  observations and instances of scores indexed in rank order:

$$s_1 \leq s_2 \leq \dots \leq s_n$$

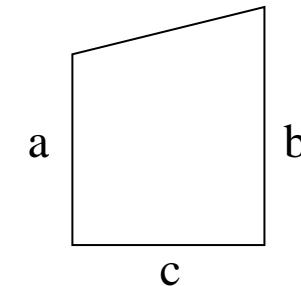
with empirical estimates  $\tilde{F}_0$  and  $\tilde{F}_1$  for  $F_0$  and  $F_1$  respectively.

Since  $A \triangleq \int_c F_1(c)F_0'(c)dc$ , we estimate AUC as

$$\hat{A}_n = \sum_{i=1}^n \frac{1}{2}(\tilde{F}_1(s_{i-1}) + \tilde{F}_1(s_i))[\tilde{F}_0(s_i) - \tilde{F}_0(s_{i-1})]$$

and using  $\tilde{F}_k(s_0) = 0$ .

This uses the trapezoid rule to estimate the area of segments of the ROC curve where multiple observations exist with the same score but different outcome.



$$Area = \frac{1}{2}(a+b)c$$

*Example 7.3*

Again, consider the 16 applicants from Example 7.1.

Score	8	10	21	22	25	30	35	42	45	46	51	52	59	64	70	78
Outcome	1	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0

Compute empirical CDFs as follows:

Cut-off $c$	$\tilde{F}_0(s_i)$	$\tilde{F}_1(s_i)$	$\hat{a}_i$	Cut-off $c$	$\tilde{F}_0(s_i)$	$\tilde{F}_1(s_i)$	$\hat{a}_i$
8	0	0.2	0	45	0.455	0.8	0.0727
10	0.091	0.2	0.0182	46	0.545	0.8	0.0727
21	0.091	0.4	0	51	0.636	0.8	0.0727
22	0.182	0.4	0.0364	52	0.727	0.8	0.0727
25	0.182	0.6	0	59	0.727	1	0
30	0.182	0.8	0	64	0.818	1	0.0909
35	0.273	0.8	0.0727	70	0.909	1	0.0909
42	0.364	0.8	0.0727	78	1	1	0.0909

where  $\hat{a}_i = \frac{1}{2}(\tilde{F}_1(s_{i-1}) + \tilde{F}_1(s_i))[\tilde{F}_0(s_i) - \tilde{F}_0(s_{i-1})]$ .

Therefore AUC estimate is  $\hat{A}_n = \sum_{i=1}^n \hat{a}_i = 0.764$ .

## Information Gain

Another useful measure is Information Gain (IG).

- It is an information measure and can be interpreted as the amount of information there is about the classes within the conditional distributions over a categorical variable.
- This measure is different to AUC.

Suppose we divide the borrowers into  $K$  bins by some categorical variable  $V$  taking distinct values  $v_1, v_2, \dots, v_K$ .

Then, IG is calculated as:

$$I_G(V) = \sum_{j=1}^K \left( P(V = v_j | Y = 0) - P(V = v_j | Y = 1) \right) w(v_j)$$

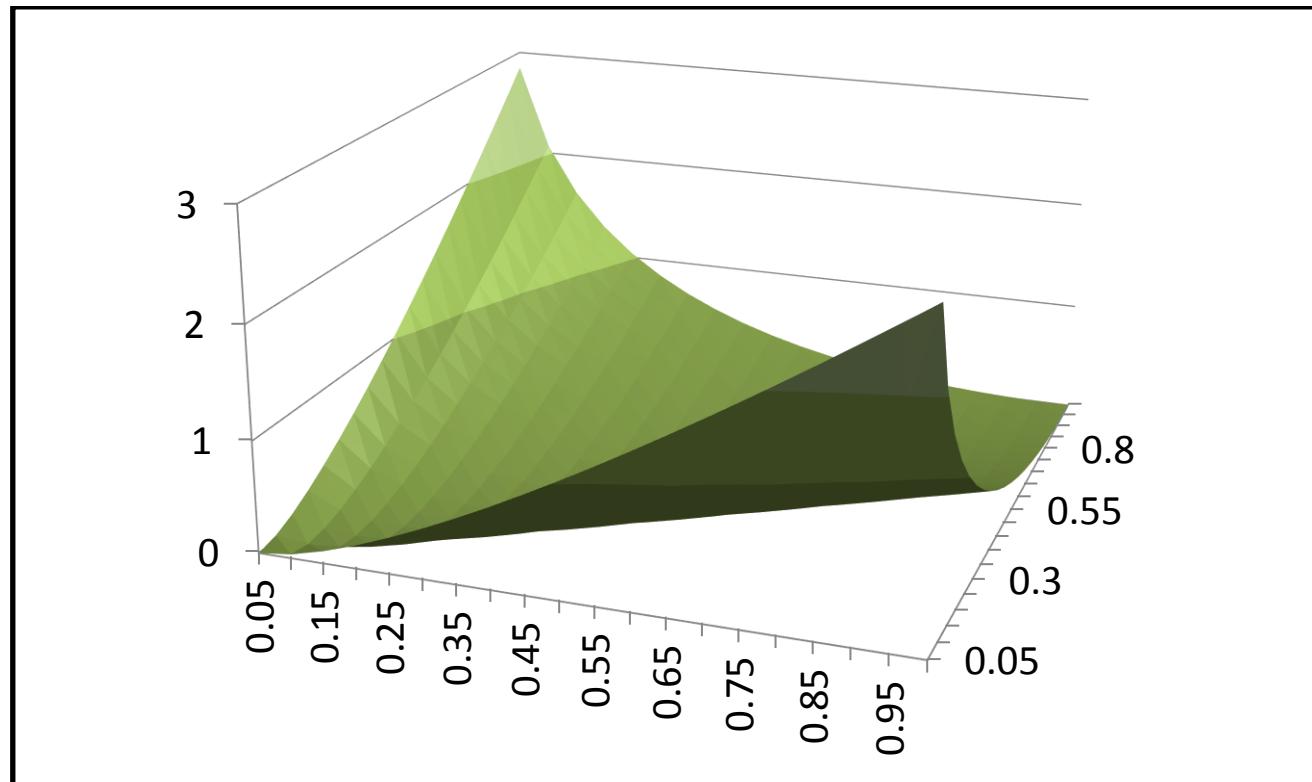
where  $P(V = v | Y = 1)$  is probability of value  $v$  given a positive outcome,

$P(V = v | Y = 0)$  is probability of value  $v$  given a negative outcome,

and  $w(v) = \log \left( \frac{P(V=v|Y=0)}{P(V=v|Y=1)} \right)$  is the weights of evidence (WOE) of value  $v$ .

## Shape of Information Gain

The shape of the IG function for one category value  $v$  and over the range of values for the  $P(V = v|Y = 0)$  and  $P(V = v|Y = 1)$  is shown in this graph.



## Comments

- Intuitively, IG tells us how much information  $Y$  give us about  $V$ .
- IG is independent of any specific cut-off score.
- IG is highly dependent on how we divide scores into grades.
- Strictly speaking, information is measured in base 2, hence the logarithm should be taken in base 2. However, it is usual to use the natural log, so long as we are consistent in its use. We will use natural log in this course.
- There is an information theoretical foundation to this measure which we will return to later in the course (Chapter 15).

## Information Gain over Risk grades

In particular, to calculate IG by score  $S$ , we divide the borrowers into  $K$  grade bins so that the  $j$ th grade of scores is in the interval  $(g_j, g_{j+1}]$ .

That is construct a categorical variable with  $K$  levels such that

$$V = j \text{ iff } g_j < S \leq g_{j+1}.$$

### Example 7.4

Again, consider 16 applicants from Example 5.2.

Score	8	10	21	22	25	30	35	42	45	46	51	52	59	64	70	78
Outcome	1	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0

Consider the following score grades:

Grade A. Scores 50 to 100;

Grade B. Scores 30 to 49;

Grade C. Scores 1 to 29.

Then we can calculate the following:

	Grade A	Grade B	Grade C
#Positive ( $Y = 1$ )	1	1	3
#Negative ( $Y = 0$ )	5	4	2
$P(g_j < S \leq g_{j+1}   Y = 1)$	0.2	0.2	0.6
$P(g_j < S \leq g_{j+1}   Y = 0)$	0.46	0.36	0.18
Weights of evidence	0.82	0.60	-1.19
Information Value	0.21	0.10	0.50

$$\text{Total Information Gain} = 0.21 + 0.10 + 0.50 = 0.81.$$

But suppose we consider alternative score grades:

Grade A'. Scores 55 to 100;

Grade B'. Scores 25 to 54;

Grade C'. Scores 1 to 24.

Then we can calculate the following:

	Grade A'	Grade B'	Grade C'
#Positive ( $Y = 1$ )	1	2	2
#Negative ( $Y = 0$ )	3	6	2
$P(g_j < S \leq g_{j+1}   Y = 1)$	0.2	0.4	0.4
$P(g_j < S \leq g_{j+1}   Y = 0)$	0.27	0.55	0.18
Weights of evidence	0.31	0.31	-0.79
Information Value	0.023	0.045	0.172

$$\text{Total Information Gain} = 0.023 + 0.045 + 0.172 = 0.24.$$

This is much less than the first grading structure.

*Which is the best to use?*

## Overview of Chapter 7

In this chapter we have covered:-

- Types of model performance measures
- Cut-off decisions and classification
- Receiver Operating Characteristic (ROC) curve
- Area under the ROC curve (AUC)
- Information Gain

We have only looked at measuring classification power of the scorecard.

We have not finished with scorecard performance measures. We will return to this subject later in the course.

# Consumer Credit Risk Modelling

## Chapter 8: Credit Data Analysis

## Overview

In this chapter we consider the data and variables available to build credit scoring models.

In particular, we will look at

- Types of variables
- Data available for application scorecards, and
- Data validation issues.

## Types of variables

There are several different types of variables that we need to consider.

Type	Description
Continuous	Real number and possibly negative valued.
Ordinal, discrete	Ordered and discrete values.
Integer	Special case of ordinal and discrete.
Categorical	Variable whose values are separate categories that do not follow any special order. Each possible value is called a <i>level</i> .
Binary	Special case of Categorical with only two levels. For example: <i>true</i> and <i>false</i> .
Text	Descriptive data with no particular structure.

## Application variables

For application scoring, the lender has four sources of information available.

1. **Application form data:** Data provided by the applicant on his/her application form.
2. **Credit bureau data:** Credit bureaus can provide lenders with further generic personal information.
3. **Past credit history:** May be immediately available to the lender or may be provided by a credit bureau.
4. **Customer details:** If the applicant is an existing customer (eg a current account holder) then further account behaviour data is also available.

## 1. Application form

Typical information requested on loan or credit card application forms are listed below.

<b>Variable</b>	<b>Type</b>
Name	Text
Address	Text
Telephone number	Text
Email address	Text
Age	Integer
Assets (property, savings)	Multiple continuous variables
Profession	Categorical
Employment status	Categorical
Employer	Text
Income	Continuous
Outgoings	Continuous
Residential status (owner, renter etc)	Categorical
Length of time in current residence	Integer
Number of dependents	Integer

This is not an exhaustive list, nor are all variables collected for all products.

Note that it is illegal for lenders to use Ethnicity or Sex as part of their lending decision process (in most countries) even if these details are available.

## 2. Credit reference agency data

More general information about individuals is collected by credit bureaus.

<b>Variable</b>	<b>Type</b>
Generic credit score	Integer
Other credit card products	Integer, or text
Past credit history and repayment records	Various
Past county court judgments (CCJs), in UK	Text and integer
Demographics (where the applicant lives and his/her social group)	Categorical

### 3. Past credit history

<b>Variable</b>	<b>Type</b>
Past account repayments	Various
Severity of delinquency (how long past due)	Continuous
Number of delinquent items	Integer
Amounts owed	Continuous
Number of credit lines	Integer
Number of recent applications for credit	Integer
Length of credit history	Continuous

Again, this is not an exhaustive list and different data may be collected and used by different organizations.

## Data Validation

When data is provided, it needs to be checked to ensure the data is valid.

We look for several possible data problems:

- Data inconsistency
- Missing values
- Unusual distributions
- Outliers

## Data inconsistency >

Application data is primarily given by the applicant. This may be in error for two reasons.

- 1.The applicant may have made a typing mistake.
- 2.The applicant may be deliberately falsifying information in order to gain an advantage (fraud).

Either way the lender will need to **validate** some or all of the data entered by the applicant. For example:

- Identity can be validated from a passport or driving licence.
- Address can be validated from a recent utility bill or checking the electoral register.
- Phone numbers can be checked using public directories.
- Evidence of income and employment can also be requested in the form of pay slips or contracts of employment.

Clearly it is important that lenders can verify the veracity of their data.

It is also important to analyse the data to try to find logical inconsistencies.

For example, if we have a field for Age and we have the value "-23" this is clearly wrong.

Or we may have inconsistencies amongst groups of variables.

For example, if we have a record that a borrower has made her last payment, but a default flag is set to "True", then this is inconsistent.

*Remedy.*

- False or inconsistent data is an indication of data entry error. If possible, data should be sent back to the provider for re-entry.
- If this is not possible, then it is safest to discard the record.

## Missing values

Some variables may have missing values. This is likely an indication that the value was genuinely unknown, although it could be a data entry problem.

We use a variable  $M_X \in \{0,1\}$  to denote a missing value on variable  $X$ ; ie it takes a value 1 when  $X$  has a missing value and 0 when a value of  $X$  is not missing.

*Example*

$X$	2.3	-4.2		5.0		6.3	7.6	-0.2
$M_X$	0	0	1	0	1	0	0	0

There are several types of missing values that are considered and listed in the following slides.

- **Missing by definition.**

This means the values are missing by design of the data set. Also, known as structural missing.

*For example, if an individual has already indicated she is single then clearly a field for spouse's income must be missing.*

- **Missing completely at random (MCAR).**

This means that  $M_X$  is independent of the variables in the data.

That is, there is no connection between the occurrence of missing values and the observations.

- **Missing at random (MAR).**

MCAR is sometimes too strong an assumption. MAR assumes that  $M_X$  is independent of  $X$ , but may possibly be dependent on other variables in the data.

*Example: Suppose income has missing values. We may assume that the occurrence of missing values is dependent on age, but perhaps not income itself. This would be the MAR assumption.*

MAR is a central assumption in many automated approaches to missing values.

- **Nonignorable missing (NI).**

This means the missing values are not MAR. That is,  $M_X$  is dependent on  $X$ , and possibly on other variables too.

Data with NI missing values requires special attention to determine the structure of missing values and therefore select the best option to deal with the problem.

*Possible Remedies.*

There are several different ways to deal with the problem of missing data.

**1. Listwise deletion.**

A very common solution is to simply discard the records with missing values. This is plausible if only a small proportion of the data is removed in this way and if the original data set was sufficiently large.

However, if the missing values are not MCAR, this could lead to biased analysis.

## 2. Imputation.

It is also possible to *impute* a value where it is missing.

- Consider a data set of  $n$  observations where  $X_i$  is the  $i$ th observation of the variable and  $M_{X,i}$  is the missingness indicator for that observation.

### ***Imputing the mean***

A simple way to do this is to impute the sample mean of values where they are not missing. That is, for each observation where  $M_{X,i} = 1$ , set the value of  $X_i$  to

$$\bar{X} = \frac{1}{n - n_M} \sum_{i=1}^n (1 - M_{X,i}) X_i$$

where

$$n_M = \sum_{i=1}^n M_{X,i}$$

However, this is not always the best solution since this will under-estimate the variance of the variable, hence altering the data distribution. Also, it is somewhat ad hoc.

*For example, does it makes sense to impute an average age of 40 for everyone where it is missing?*

### **Demonstration of reduced variance**

Let  $\sigma^2$  be sample variance of  $X$  were values are not missing.

Let  $\bar{X}_I$  and  $\sigma_I^2$  be sample mean and variance across all observations *after mean imputation*, respectively.

Then  $\bar{X} = \bar{X}_I$  and

$$\sigma_I^2 = \left(1 - \frac{n_M}{n-1}\right) \sigma^2 < \sigma^2 \text{ if } n_M > 0$$

*Proof is the subject of exercise 1 in problem sheet 2.*

## ***Regression imputation***

It may be better to impute from other variables.

This can be done by regressing the known values from other predictor variables in the data set.

So using only observations where  $M_{X,i} = 0$ , we estimate a regression model  $\hat{f}$  from

$$X_i = f(\mathbf{X}_i^R) + \varepsilon_i$$

where

- $\mathbf{X}_i^R$  denotes the vector of *remaining* variables for the  $i$ th observation (ie with variable  $X$  removed);
- $\varepsilon_i$  is an appropriate error term (eg assumed normal).

Then values are imputed for each observation where  $M_{X,i} = 1$  as  $\hat{f}(\mathbf{X}_i^R)$ .

- If  $X_i$  is a numeric term then OLS regression can be used.
- If  $X_i$  is a categorical variable then logistic regression can be used.

However, regression imputation needs to be used with caution and only if the association between the variables is strong.

*Example: an association between age and income might allow us to sensibly impute income from age.*

Imputation is a promising approach to dealing with missing data, especially the regression approach.

- However, a lot depends on which regression model is used. If the relationship between variables is not linear, eg, then problems occur with inconsistent estimates.
- Also, if the data is not MCAR then potentially there are problems with the regression approach because of the dependency of the occurrence of missing on the regression variables. This can lead to biased estimates.

## Example: OLS regression imputation

- Use model structure

$$X_i = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{X}_i^R + \varepsilon_i$$

and estimate  $\hat{\beta}_0, \hat{\boldsymbol{\beta}}$  using OLS regression (ie minimize sum of square errors; see Chapter 2) using all observations where  $M_{X,i} = 0$  as training data.

- Apply model to estimate for observations when  $M_{X,i} = 1$ ,

$$X_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}} \cdot \mathbf{X}_i^R$$

### 3. Model-based procedures

There are computer-intensive advances on imputation that allow us to formalize the imputation process as part of the model specification.

For instance, the EM algorithm can be used to formalize the regression imputation method. Usually requires some distributional assumption on the data (typically, multinomial normal).

*Note: we will not cover this method in this course.*

In Chapter 11 we consider a method of segmentation which can be used to deal with structurally missing values as part of the model specification.

## 4. Structural indicator

If there is some structural reason for the missing value, then include  $M_X$  as an indicator variable in the model (and use an arbitrary value to replace the missing values in  $X$ , eg impute the mean).

*For example, if an applicant does not complete a field indicating an employer, then this may indicate unemployment and may be a useful variable.*

In SAS software, this is called “informative missing”.

## 5. Variable deletion

If there are too many missing values for a variable, then it may be necessary to simply discard the variable.

Further reading on the problem of missing data:

- Little RJA and Rubin DB (1987), *Statistical analysis with missing data*, Wiley (available in the library)

*Example 8.2*

Consider this data set of values for Home owner (1=true, 0=false) and Months in residence, for which there are some missing values.

Home owner	1	0	1	1	0	0	1	0	1	0	1	0	1	0	0	0	1	1
Months in res	68	12	24	3	8	1	120	38	29	6	12	15						

- Impute the mean for each missing months in residence.
- Impute missing months in residence using OLS regression on Home owner.

*Solution*

- Mean =  $\frac{68+12+24+3+8+1+120+38+29+6+12+15}{12} = 28$
- Regress months in residence on home owner  $\approx 13.3 + 29.3 \times [\text{Home owner}]$

Therefore,

Home owner	1	0	1	1	0	0	1	0	1	0	1	0	0	0	1	1
Months in res	68	12	24	3	8	1	120	38	29	6	12	15				
$M_{\text{Res}}$	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
IM.													28	28	28	28
RI.													13.3	13.3	42.6	42.6

IM = imputed mean, RI=regression imputation

## Unusual distributions

The distribution of values for a variable could alert us to problems. For example, if 90% of borrowers are under 30, this would be suspicious (however, not implausible if the product had been targeted to a young market).

Or, if the distribution is bimodal, this may indicate that different units of measurements are being used.

For example, a bimodal distribution of income may indicate that it is measured in two currencies (say, £ and euros), or in different scales.

*Remedy.*

- This is potentially a serious problem and the data provider would need to be consulted. Hopefully, the distributions could be explained or resolved.
- If the unusual distribution cannot be understood, then it may be necessary to discard the entire variable.

## Outliers

There may be some values that are very extreme and therefore highly unlikely.

For example, an age of 121.

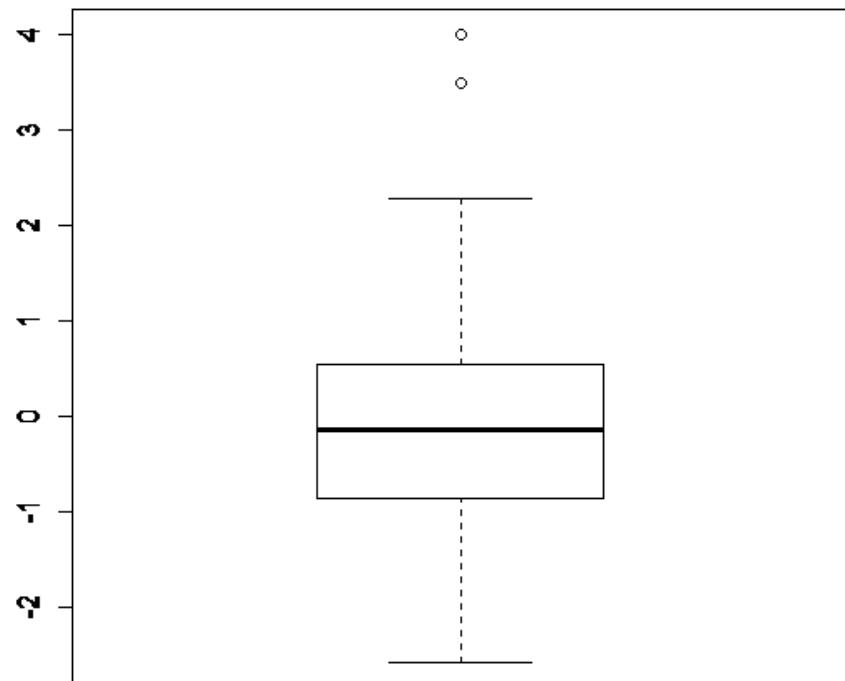
Also, in some fields special codes may be used. For example, age=999 may be an indicator for missing age. It may also be detected as an outlier (or, an inconsistent value).

Some caution is required, since determining an outlier depends on the underlying distribution.

For example, when a distribution has a long right tail then relatively high values will be expected.

There are automated means to check for outliers. However, it is reasonable to use a manual approach by looking at the data. For example, a histogram of the data or “box and whiskers” plot will identify potential outliers.

A modified “box and whiskers” plot is shown below for a sample taken from a normally distributed variable.



The box is defined as the first and third quartile of the data.

Therefore, 25% of data is above the box and 25% below.

The thick line running through the box is the median.

The whiskers represent end points of the data, but they are no more than 1.5 times the distance outside the box than the interquartile range (IQR).

Observations beyond the whiskers are shown as circles. They are extreme values that need to be investigated as outliers.

### *Example 8.4*

In the figure shown above there are two potential outliers.

This is actually not too surprising given the way the data was generated.

The following R code shows how the data was generated and how the box plot was produced.

```
> x <- c(rnorm(100), 3.5, 4)  
> boxplot(x, range=1.5)
```

### *Remedies.*

- See if the problem can be resolved by the data provider (either enter a corrected value or confirm the outlier is indeed true).
- Otherwise, discard the records with outliers.

## Overview of Chapter 8

We have looked at credit application data and data validation issues, including:

- Types of variables
- Application credit data
- Data inconsistency
- Missing values
- Unusual distributions
- Outliers

# Consumer Credit Risk Modelling

## Chapter 9: Data Preparation

## Overview

The raw data is rarely in a condition for direct use in a model.

We need to ensure that variables enter the model in a form that gives optimal model fit. They may need to be transformed prior to model build.

This is known as ***data wrangling***.

We consider four key methods of data preparation.

- Including categorical variables in linear models
- Distribution change
- Discretization
- Weights of evidence (WOE)
- Dealing with descriptive variables

## Including categorical variables in the model

Categorical variables are usually included in a model as a series of binary **indicator variables** for each possible value (also called **dummy** variables).

Formally, if we have a categorical variable  $X_{\text{CAT}}$  taking  $K$  possible values  $x_1, x_2, \dots, x_K$  then we create  $K$  new indicator variables,

$$X_j^I \triangleq I(X_{\text{CAT}} = x_j) \text{ for all } j=1 \text{ to } K.$$

Then the indicator variables  $X_j^I$  are included in the model instead of  $X_{\text{CAT}}$ .

However, usually one of them is excluded. This is called the *base category* or the **excluded category**.

## Why exclude a category value?

At least one category value must be excluded since the combination of all categories is colinear (ie one of them must be true).

That is, the sets of events expressed by  $X_j^I$  are mutually exclusive and exhaustive. In particular:  $\sum_{j=1}^K X_j^I = 1$ .

Including all the indicator variables would lead to a non-unique coefficient estimate. Therefore, the model cannot be estimated with a unique solution.

### Proof.

Suppose a model provides coefficient estimates  $\hat{\beta}_j^I$  for each of a colinear set of indicator variables  $X_j^I$  for  $j \in \{1, \dots, K\}$ .

Let  $\hat{\gamma}_j^I = \hat{\beta}_j^I - \alpha$ , for any constant  $\alpha$ . Then

$$\sum_{j=1}^K \hat{\beta}_j^I X_j^I = \sum_{j=1}^K (\hat{\beta}_j^I - \alpha) X_j^I + \alpha \sum_{j=1}^K X_j^I = \alpha + \sum_{j=1}^K \hat{\gamma}_j^I X_j^I$$

Let  $\mathbf{x}$  be the vector of all predictor variables. Hence it is constructed from  $(X_1^I, \dots, X_K^I)$  and a remainder of other predictors  $\mathbf{x}^R$ .

Suppose the log-likelihood function is constructed as

$$L = \sum_{i=1}^n f(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}_i)$$

for  $n$  observations, as is the case with OLS and logistic regression, in particular.

Then, reducing the vector of predictor variables,

$$L = \sum_{i=1}^n f\left(\beta_0 + \boldsymbol{\beta}^R \cdot \mathbf{x}_i^R + \sum_{j=1}^K \beta_j^I X_{ij}^I\right)$$

where  $\boldsymbol{\beta}^R$  are coefficients on the remainder of predictors.

Let coefficient estimates  $\beta_0 = \hat{\beta}_0$ ,  $\boldsymbol{\beta}^R = \hat{\boldsymbol{\beta}}^R$  and  $\beta_j^I = \hat{\beta}_j^I$ , for  $j \in \{1, \dots, K\}$  maximize  $L$  for fixed  $\mathbf{x}$ .

Then the maxima is

$$L_{\max} = \sum_{i=1}^n f \left( \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^R \cdot \mathbf{x}_i^R + \alpha + \sum_{j=1}^K \hat{\gamma}_j^I X_j^I \right)$$

which implies that the likelihood function can also be maximized with the estimates  $\beta_0 = \hat{\beta}_0 + \alpha$ ,  $\boldsymbol{\beta}^R = \hat{\boldsymbol{\beta}}^R$  and  $\beta_j^I = \hat{\beta}_j^I - \alpha$ .

Hence there is no unique solution when optimizing the model.

What happens in practice?

- The `glm` function in R will still run by *forcing* an excluded category value if you try to include all of them.
- Other software packages behave in a similar way.
- So be careful to look out for this!

## Interpreting category variables in models

Because at least one category value is excluded, this means coefficient estimates for all other values are *relative* to the base category.

We can interpret the base category as having a coefficient of 0.

However, for logistic regression, the choice of base category does not materially effect the coefficient estimates, only which category the coefficients are relative to.

That is, suppose  $\hat{\beta}_i^I$  are coefficient estimates for a categorical variable with base category on index  $b$  and  $\hat{\gamma}_i^I$  are coefficient estimates for the same variable on the same data but with base category on index  $c$ . Then,

$$\hat{\gamma}_i^I = \hat{\beta}_i^I - \hat{\beta}_c^I \text{ for } i \neq b \text{ and } \hat{\gamma}_b^I = -\hat{\beta}_c^I.$$

Therefore  $\hat{\gamma}_c^I = 0$ , and  $\hat{\beta}_i^I = \hat{\gamma}_i^I - \hat{\gamma}_b^I$ .

The proof of this claim will be an exercise in problem sheet 2.

*Example 9.4*

Create indicator variables for age ranges 18-29, 30-49 and 50+ for the following records.

Age	18	20	20	22	26	30	32	34	38	40	47	48	52	55	59	62
-----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

*Solution*

Age	18	20	20	22	26	30	32	34	38	40	47	48	52	55	59	62
<i>Categories:</i>																
Age 18-29	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Age 30-49	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0
Age 50+	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

The following table shows coefficient estimates for each category level for two models using two different base categories for a model of non-default.

	Model 1		Model 2	
Category level	Excluded category?	Coefficient estimate	Excluded category?	Coefficient estimate
Age 18-29	No	-0.69	Yes	0
Age 30-49	No	-0.37	No	+0.32
Age 50+	Yes	0	No	+0.69

Notice the relationship between coefficient estimates for the two models: the relative difference between any two levels is always the same.

*Interpretation:*

Since negative estimates indicate less association with non-default, we conclude that:

Younger people (less than 30) have a higher risk of default, relative to older people and those older (50 years or older) are the least risky group.

## Distribution change

Some continuous variables may be naturally skewed with long tails of extreme values.

In order to produce a robust model it is a good idea to transform these variables to avoid extreme values entering the model.

*Why?*

1. Because extreme values in predictor variables will lead to extremes in estimates of the dependent variable.
2. Risk may be non-linearly related to raw values of variable, with a specific pattern.

If values of a continuous variable are greater or equal to zero and distributed with a right-skew, then a **log transformation** is typically used.

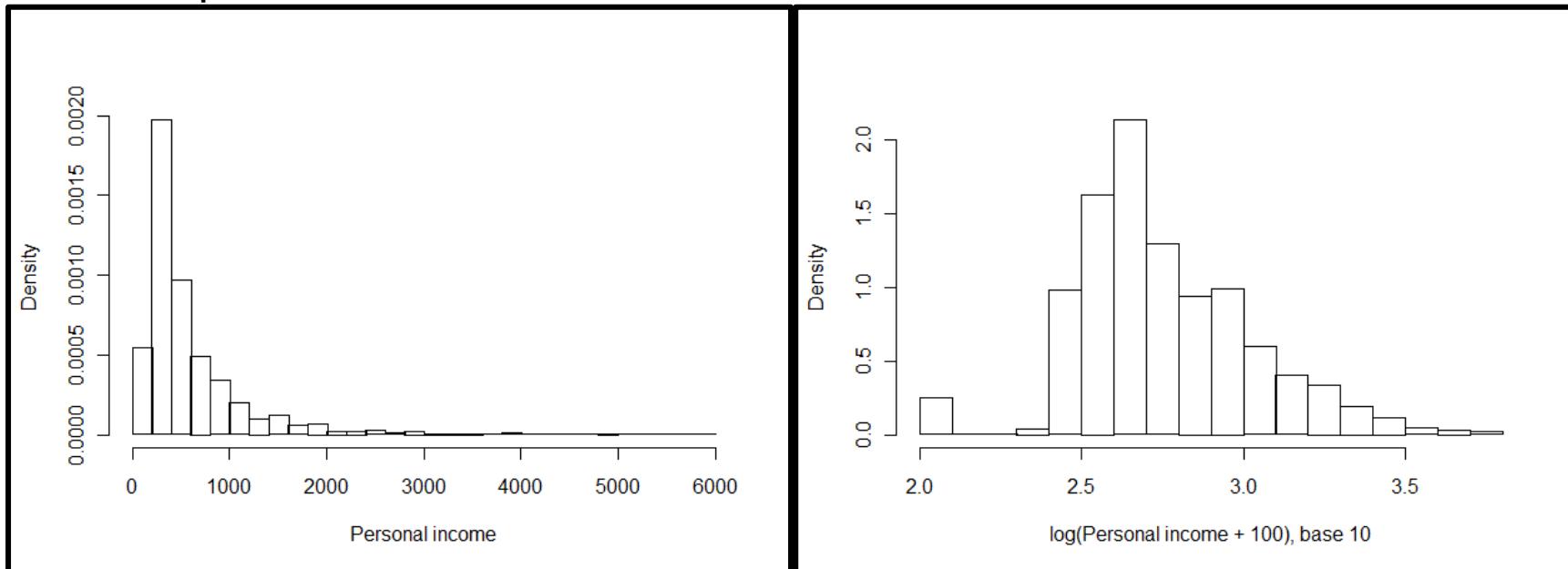
- This is often the case with monetary values and sometimes with elapsed time variables (eg age).

### Example 9.1

Income is naturally a positive value with right-skew.

It is typical to use  $\log(\text{Income})$  in the model instead of Income, since this has a distribution closer to normal.

An example from a credit card data set:



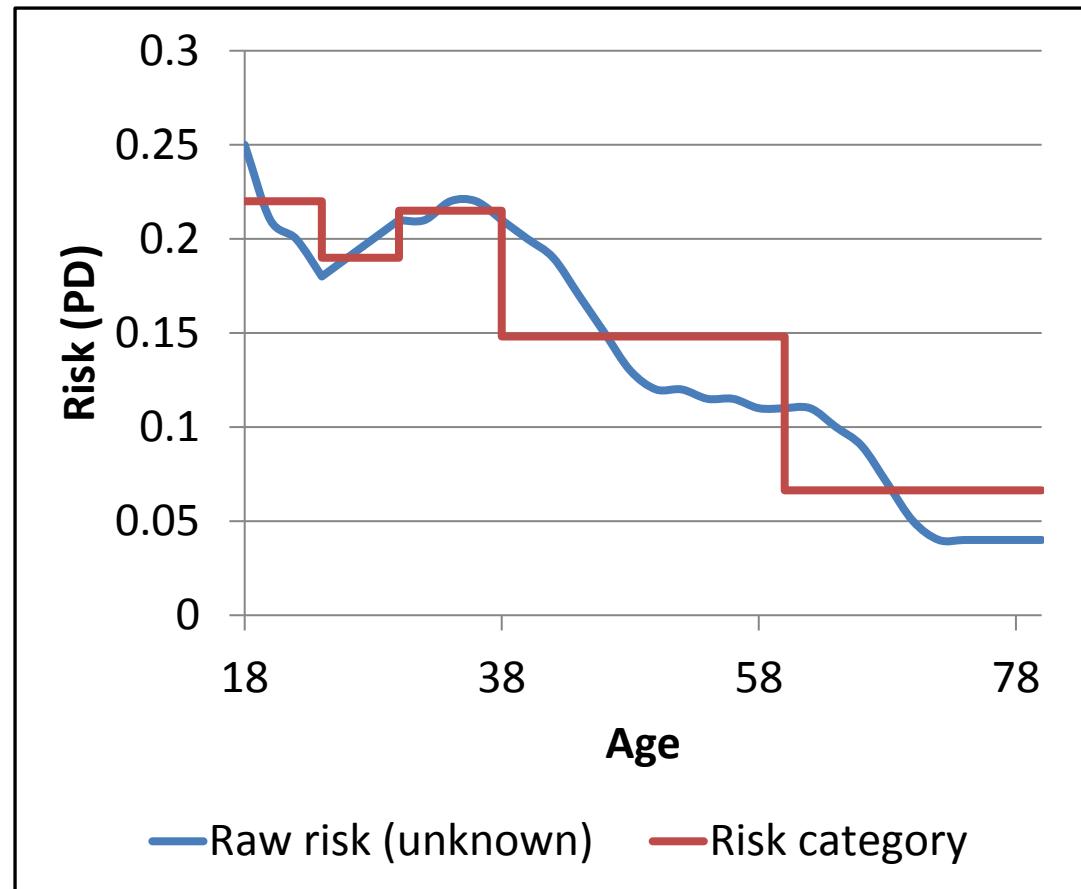
Same data, different transformation.

## Discretization

Continuous variables may need to be transformed into categories. This is because:

1. the relationship between a continuous variable and the classification event is not always linear,
2. using categories promotes a robust model, in the sense that it is expected to be resilient to changes in population distribution or risk factors over time in the future (ie population drift; see Chapter 13).

Discretization is also known as ***coarse classification*** in the credit industry. It may also be referred to as ***categorization***.

*Example 9.2 (fictitious data)*

The blue line shows a risk association between age and default that is non-linear.  
Generally this risk is unknown.  
It is modelled using categories over age. In this case: 18-24, 24-30, 30-38, 38-60 and 60+.  
The red line shows estimated risk within each category. This can be interpreted as expected risk within each age range.

There are three general approaches to dividing a continuous variable into categories.

1. Use ***quantiles***:

Divide into several groups corresponding to approximately the same number of observations in each.

2. Use some meaningful division of the range.

For example, age ranges above and below retirement would make sense (eg 65 plus, to distinguish retired people).

3. ***Automated Discretization***:

Divide the continuous variable into ranges which maximize overall discrimination across the categories.

The important points with discretization are to

1. ensure sufficient levels to capture risk profile of variable and
2. that each category is representative of a good proportion of the population.

## Automated Discretization

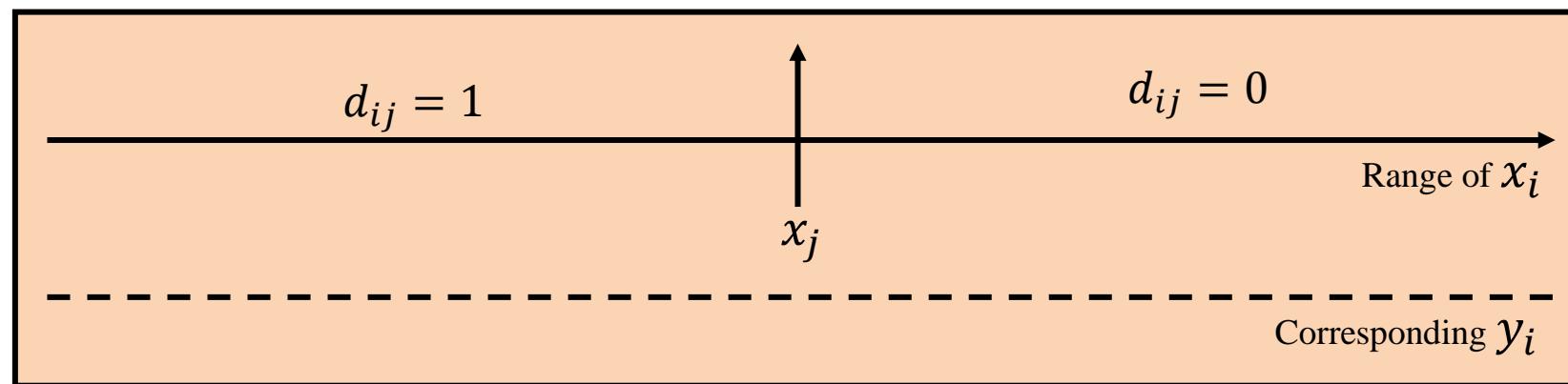
A search of the range of a continuous variable is performed to find a split point that maximizes total discrimination using a given discrimination measure.

- Consider a training sample of  $n$  observations having values of a continuous variable  $X$  as  $x_1, \dots, x_n$  with corresponding (binary) outcomes  $y_1, \dots, y_n$ .
- Suppose we have a discrimination measure  $M$  which takes a sequence of value/outcome pairs as arguments and calculates a real number measure of discriminatory power.
  - $M: (\mathbb{R} \times \{0,1\})^n \rightarrow \mathbb{R}$
  - For example,  $M$  could be AUC or Information Gain.

- The optimal point to divide  $X$  into categories to achieve the best discrimination is estimated from the data as  $x_{j_{\text{OPT}}}$  where

$$j_{\text{OPT}} \triangleq \arg \max_j [M((d_{1j}, y_1), \dots, (d_{nj}, y_n))]$$

and  $d_{ij} = I(x_i \leq x_j)$  are indicator variables on the range of values for each  $j$ .



- Typically, a computationally intensive search through the range of values is necessary to find this maximum.

- Thus, a new categorical variable is created as follows:

$$X_{\text{CAT}} = \begin{cases} 0 & \text{if } X \leq x_{j_{\text{OPT}}} \\ 1 & \text{if } X > x_{j_{\text{OPT}}} \end{cases}$$

- This procedure can be repeated recursively on each sub-range until we have the number of splits we require.
- The variable  $X_{\text{CAT}}$  is then used in the model instead of  $X$ .
- A typical measure  $M$  for discretization is the Information Gain (IG) measure we have already introduced in Chapter 7.
  - This gives an entropy-based measure, although cost-based measures can also be used (and may be superior).
  - Notice that when  $V$  is a binary variable,  $V \in \{0,1\}$ , as is the case for discretization,

$$I_G(V) = [P(V = 0|Y = 0) - P(V = 0|Y = 1)][w(0) - w(1)]$$

➤ Note: empirical frequencies are sometimes 0, in which case WOE is not well-defined. Use method on Slide 23 to deal with this problem.

### Example 9.3

Suppose we want to categorize age into two groups. We have data for age and outcome of 17 borrowers.

Age	18	19	22	25	26	32	34	35	36	38	40	42	45	51	53	60	64
Outcome	1	0	1	1	0	1	0	1	0	1	0	0	0	1	0	0	0

Use IG to compute performance with split point  $x_j=35$ .

Then, discriminator variable has following values (note  $j=8$ ):

Age	18	19	22	25	26	32	34	35	36	38	40	42	45	51	53	60	64
$y$	1	0	1	1	0	1	0	1	0	1	0	0	0	1	0	0	0
$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$d_{ij}$	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0

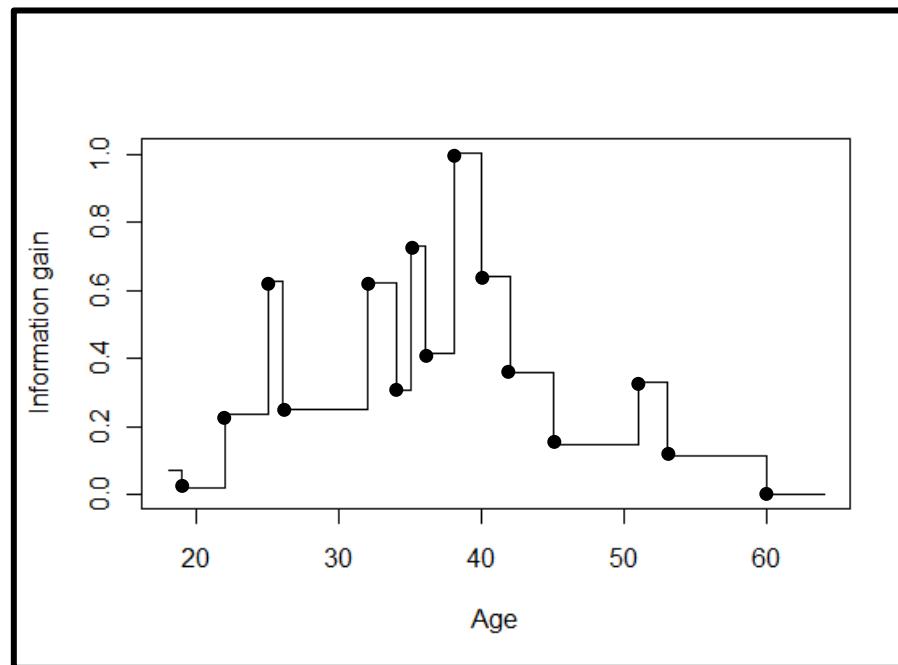
From this we can compute empirical probabilities for IG:

$P(d_{ij} = 1 Y = 1) = 5/7$	$P(d_{ij} = 0 Y = 1) = 2/7$
$P(d_{ij} = 1 Y = 0) = 3/10$	$P(d_{ij} = 0 Y = 0) = 7/10$

Using the formula for IG we get

$$\text{IG} = \left(\frac{7}{10} - \frac{2}{7}\right) \left(\log\left(\frac{7/10}{2/7}\right) - \log\left(\frac{3/10}{5/7}\right)\right) \approx 0.731.$$

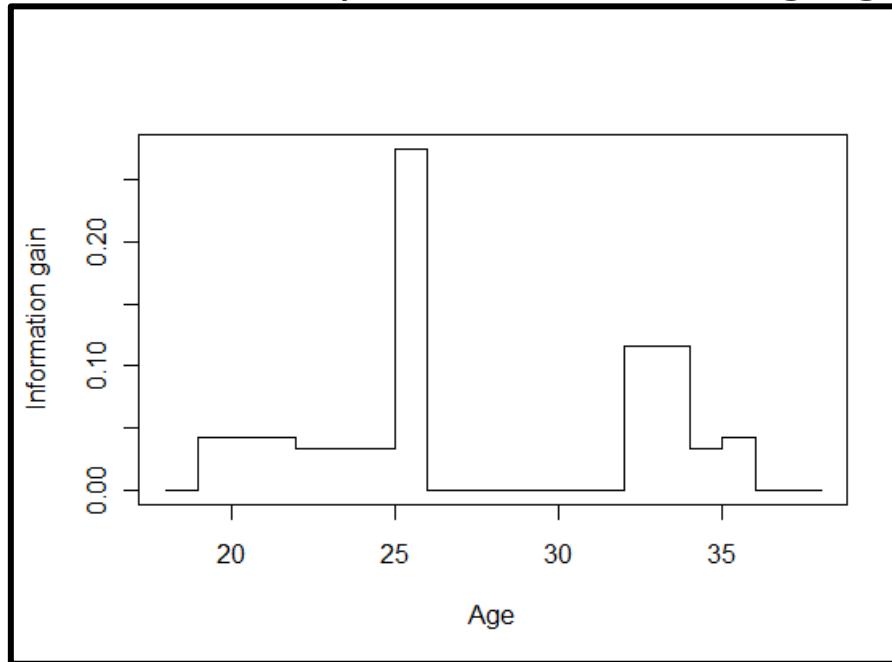
If we compute IG across all ages, based on this data set, we get this graph.



Therefore the optimal IG is to split with age groups 18-39 and 40+.

This example splits age into a two category variable.  
However, we can apply the procedure recursively to get finer discretization.

For example, what is the best split of the 18-39 age group?



So do a further split at age 26.

Therefore we have a final categorical variable for Age as:

$$X_{\text{CAT}} = \begin{cases} 0 & \text{if } X < 26, \\ 1 & \text{if } 26 \leq X < 40, \\ 2 & \text{if } X \geq 40 \end{cases}$$

## Reference

One of the first discussions of automated discretization, in the context of decision tree learning (we'll cover that in chapter 12) is:

Fayyad, Usama M.; Irani, Keki B. (1993)  
"Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning", *Proceedings of the International Joint Conference on Uncertainty in AI* (Q334 .I571 1993), pp. 1022-1027

## Weights of Evidence (WOE)

If too many values exist for a categorical variable then it is not feasible to enter them in the model as a series of indicator variables.

There will be too many of them and consequently coefficients will be poorly estimated.

A good solution is to substitute the categorical variable with the continuous WOE for each value.

Recall:

$$w(x) = \log \left[ \frac{P(X = x|Y = 0)}{P(X = x|Y = 1)} \right]$$

is the WOE for value  $x$  of variable  $X$ .

- WOE greater than 0 indicates a greater association with the negative event ( $Y = 0$ ).

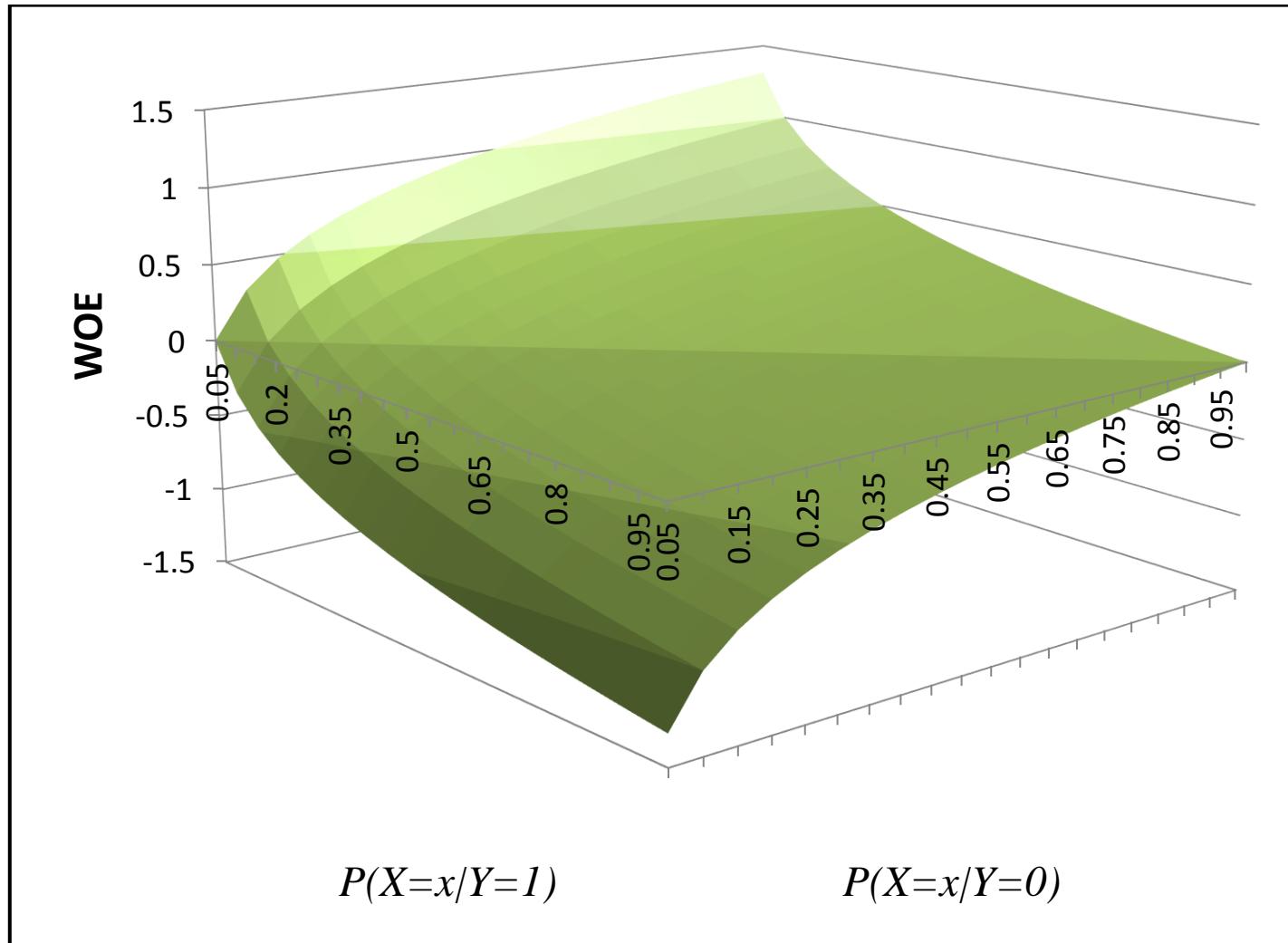
- WOE less than 0 indicates a greater association with the positive event.
- So substitute categorical variables  $x_i$  in data with continuous variable  $x_i' = w(x_i)$ .
- This has the additional nice feature that the non-linearity of risk of  $X$  is captured in the WOE transformation.
- Sometimes WOE is given in terms of WOE *of the positive event*. That is,

$$w^+(x) = \log \left[ \frac{P(X = x | Y = 1)}{P(X = x | Y = 0)} \right]$$

However, it is clear that the two versions of WOE are only different in sign:  $w^+(x) = -w(x)$ .

So, either versions can be used in the model, without any substantial differences to model estimation.

## Shape of WOE function



## Computing WOE : practical issues

WOE is not defined when either conditional frequency is zero. Hence, for practical purposes, we need to adjust WOE to allow for that.

- Let WOE be computed from a data set of  $n$  observations with values of  $X$  given as  $x_1, \dots, x_n$  and outcome  $y_1, \dots, y_n$  as usual;
- Let  $n_y = |\{i: i \in \{1, \dots, n\}, y_i = y\}|$  be count of number of occurrences of  $y$ ;
- Let  $n_{xy} = |\{i: i \in \{1, \dots, n\}, (x_i, y_i) = (x, y)\}|$  be count of number of occurrences of observations  $(x, y)$ ;
- Then empirical probabilities are given by  $\hat{P}(X = x|Y = y) = n_{xy}/n_y$ .

Two possible solutions to deal with zero values of frequency:

1. Use a Bayesian estimate; eg  $\hat{P}(X = x|Y = y) = \frac{(n_{xy}+1)}{(n_y+K)}$ . This prevents zero frequencies being generated.
2. Adjust the formula for WOE to allow for a minimum value of occurrences.

In this course we use option 2.

$$\hat{w}(x) = \log \left[ \frac{\hat{P}(X = x | Y = 0)}{\hat{P}(X = x | Y = 1)} \right] = \log \left[ \frac{n_{x0}/n_0}{n_{x1}/n_1} \right] = \log \left[ \frac{n_{x0}n_1}{n_{x1}n_0} \right].$$

So, instead, to ensure WOE is well-defined, use

$$\hat{w}_2(x) = \log \left[ \frac{\max(n_{x0}, 1) n_1}{\max(n_{x1}, 1) n_0} \right].$$

### Example 9.5

Profession may be a useful indicator of creditworthiness, but usually there are many possible values, so this is a case where WOE would be useful.

Here is a simple example data set summary.

Profession	$n_{x0}$	$n_{x1}$	$P(X = x Y = 1)$	$P(X = x Y = 0)$	$w(x)$
Lecturer	12	4	0.125	0.070	-0.583
Solicitor	8	1	0.0313	0.0465	0.398
Student	120	20	0.625	0.698	0.11
Postman	32	7	0.219	0.186	-0.162

### ***Alternative to using WOE***

If there is some natural grouping of the categories then these broader groupings could be used instead of the original categorical variable or WOE. This may lead to a loss of information. However, it may be more robust.

### *Example 9.6*

Standard Industrial Classification (SIC) codes are standard hierarchical codes for classifying industries. They are sometimes used in credit scoring to:

- Categorize the employer of an individual borrower;
- Categorize the industry of a business for business credit scoring.

There are 10 top level SIC code categories (see table in next slide), but hundreds of subcategories exist.

So we have a choice:

1. WOE on many subcategories, or
2. Only use top level categories and include as indicator variables.

The choice depends very much on sample size.

- If we have a large enough data set that WOE can be computed with precision then this is possibly the better choice.
- With smaller sample sizes using the second option may be better.

Highest level SIC codes (source: NAICS)

<b>Div</b>	<b>Industry Title</b>
<b>01-09</b>	<b>Agriculture, Forestry, And Fishing</b>
<b>10-14</b>	<b>Mining</b>
<b>15-17</b>	<b>Construction</b>
<b>20-39</b>	<b>Manufacturing</b>
<b>40-49</b>	<b>Transportation, Communications, Electric, Gas, And Sanitary Services</b>
<b>50-51</b>	<b>Wholesale Trade</b>
<b>52-59</b>	<b>Retail Trade</b>
<b>60-67</b>	<b>Finance, Insurance, And Real Estate</b>
<b>70-89</b>	<b>Services</b>
<b>91-99</b>	<b>Public Administration</b>

## Descriptive variables

There are some variables that are unique for each individual. These may be text fields or phone numbers, for example.

- One way to deal with them is to search and categorize by key word.
- Another way is to simply acknowledge their presence with a binary (true/false) variable.

- **Text mining** is being used increasingly in predictive analytics at banks.
- That is, we want to associate a piece of text (eg a customer review) to an outcome of interest.
- Methods now exist to allow us to do this; eg using the “bag-of-words” method to represent documents and text.
- However, these text mining algorithms are outside the scope of this course.

## Overview of Chapter 9



We have considered four key methods of data preparation:

- Distribution change
- Discretization
- Weights of evidence (WOE)
- Dealing with descriptive variables

# Consumer Credit Risk Modelling

## Chapter 10: Variable selection

## Overview

Before building the model, we need to decide which characteristics to include as predictor variables.

We will cover the following topics:

1. Variable selection
2. Univariate variable selection
3. Multivariate variable selection:
  - Stepwise variable selection
  - LASSO penalty

## Variable selection



Our sample data sets may provide many candidate predictor variables (eg 100s for application or behavioural data).

We usually want to reduce the number of variables in a scorecard (say, 10-20 as a rule of thumb).

There are several reasons for reducing the number of predictor variables:

1. Models with fewer variables are more robust: they may be less likely to overfit the training data and make for better predictors.  
*(We will look at overfitting in more detail later in the course).*
2. Credit scoring models need to be explanatory models, so that lending decisions can be explained, and reducing the number of variables makes this easier.
3. Reduce **multicollinearity**. Many variables are likely to be highly correlated. Removing some highly correlated variables should not affect predictive performance but should give more precise coefficient estimates with lower standard errors.

The first step is *manual variable selection*. Initially credit scoring experts will determine which variables are appropriate to use.

For instance, the colour of the borrower's hair will not be used, even if it is predictive of default! More seriously, the sex of the borrower will not be included on legal grounds.

After manual variable selection, we can then use statistical methods to select candidate predictor variables.

There are two main approaches: univariate and multivariate variable selection.

Note: in machine learning jargon variable selection is also called ***feature selection***.

A simpler model based on a restricted subset of predictor variables is often called a ***parsimonious model***.

## Univariate Variable selection

We remove those variables that do not have significant discriminatory power when considered on their own.

We can use any of the discrimination measures we have already considered (eg AUC or Information Gain).

Suppose we have a validation data set with  $n$  observations and  $m$  characteristics where  $x_{ij}$  denotes the value of the  $j$ th predictor variable of the  $i$ th observation and  $y_i \in \{0,1\}$  denotes the outcome of the  $i$ th observation.

Then we use a performance measure on each characteristic

$$M_j = M((x_{1j}, y_1), \dots, (x_{nj}, y_n))$$

Then, either set a threshold  $t_N$  on the number of variables to be selected or the magnitude of performance  $t_M$  and select the set of predictor variables as

$$V_{\text{select1}} = \{j: |\{k: M_k \geq M_j\}| \leq t_N\}$$

or

$$V_{\text{select2}} = \{j: M_j \geq t_M\}$$

### *Example 10.1*

We could build a log-odds model on each variable  $j$  alone to generate scores, then calculate AUC.

*Example 10.2*

We could compute information gain for each variable, as a measure of discrimination

$$M_j = I_G(V_j)$$

- See chapter 7 for the formula for information gain.
- Note, continuous variables need to be discretized before information gain can be used.

Order by information gain and select those variables with the largest information gain for inclusion in the model.

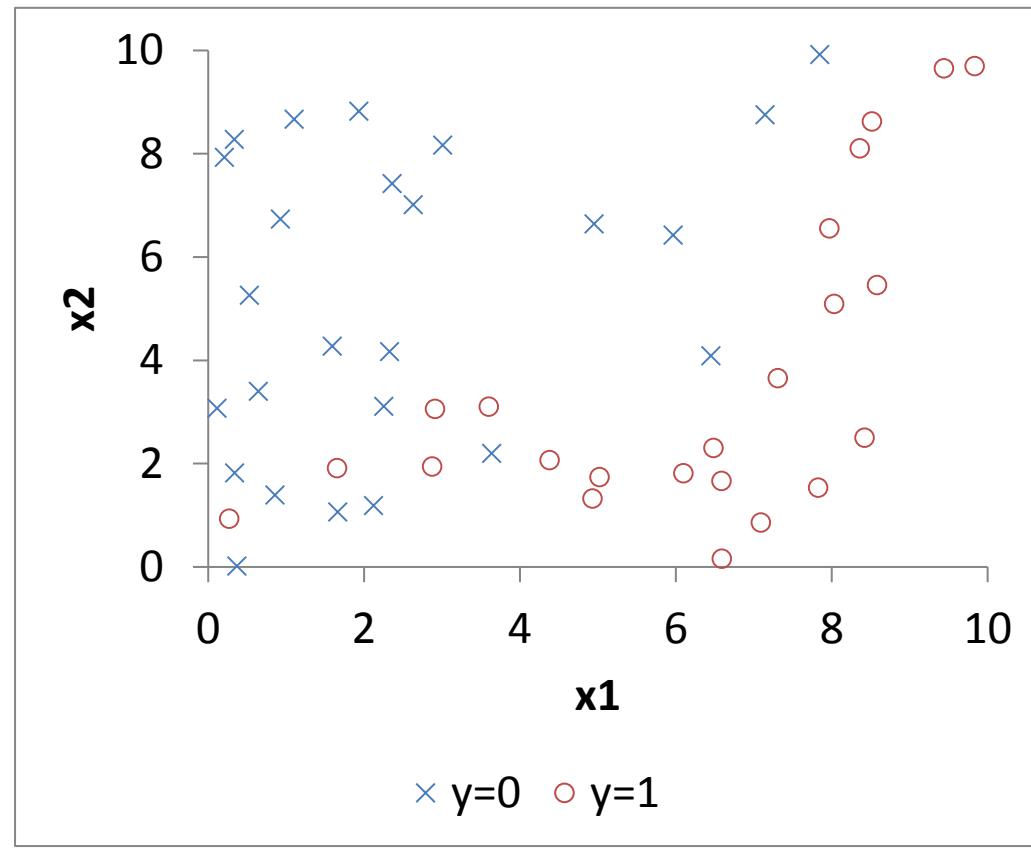
## Multivariate Variable selection

- The problem with univariate feature selection is that it does not consider how variables behave in the model in combination with each other.
- For instance, individually, two variables may have low discriminatory power, but together they have good discriminatory power.
- For this reason we use computationally intensive variable selection procedures that involve all variables.

*Example 10.3*

Consider two variables  $X_1$  and  $X_2$  with an outcome  $Y \in \{0,1\}$ .

A sample data set is shown in the graph below.



There is clearly a relationship between the variables and outcome, but neither  $X_1$  nor  $X_2$  on its own can fully represent that relationship.

### *Example 10.3 continued*

If we run three logistic regression models (for  $x_1$  only,  $x_2$  only and both together), we get the following output.

Variable	Est.	z	P(> z )
Intercept	-2.3	-3.33	0.0009
$X_1$	0.53	3.73	0.0002

Variable	Est.	z	P(> z )
Intercept	0.71	1.33	0.18
$X_2$	-0.18	-1.76	0.079

Variable	Est.	z	P(> z )
Intercept	-1.2	-1.50	0.13
$X_1$	1.08	3.22	0.0013
$X_2$	-0.83	-2.68	0.0073

- If  $X_2$  were considered in isolation, it may be rejected as a predictor variable (since its p-value is high: ie it does not even meet a 5% significance level).
- When considered together, both variables are statistically significant (at 1% level).
- The coefficient estimate for each variable is different when they are included in the model together.

## Stepwise variable selection

The basic tool we need is a comparative performance measure to compare the model fitness of two sets of variables. Let us call this a function

$$m_{\text{fit}}(S, T) \rightarrow \mathbb{R}$$

where  $S$  and  $T$  are some sets of variables.

- It measures how much better  $S$  is than  $T$ .
- In particular, if  $m_{\text{fit}}(S, T) > 0$ , we say  $S$  is better than  $T$ .
- We will return to specify  $m_{\text{fit}}$  later.

## Stepwise forward selection



Start with the null model (ie no variables). Add one variable at a time, in order of how well they improve the model. Keep adding variables until the model fit performance no longer improves.

*Formally:*

Suppose  $C$  is a set of all candidate variables.

Let the variable selection set start with  $S_o = \emptyset$ .

Let step  $i = 0$ .

Repeat:

    Let  $k = \arg \max_{j \in C \setminus S_i} m_{\text{fit}}(S_i \cup \{j\}, S_i)$ .

    If  $m_{\text{fit}}(S_i \cup \{k\}, S_i) > 0$  then

$S_{i+1} = S_i \cup \{k\}$ .

        Let  $i = i + 1$ .

        If  $S_i = C$  then Stop

    Else

        Stop

Return variable selection set  $S_i$ .

## Stepwise backward selection

This is the opposite of forward selection. Start with a model with all available variables. Remove variables, in order of how poorly they improve the model. Keep removing variables until the model fit is not improved.

*Formally:*

Suppose  $C$  is a set of all candidate variables.

Let the variable selection set start with  $S_o = C$ .

Let step  $i = 0$ .

Repeat:

    Let  $k = \arg \max_{j \in S_i} m_{\text{fit}}(S_i \setminus \{j\}, S_i)$ .

    If  $m_{\text{fit}}(S_i \setminus \{k\}, S_i) > 0$  then

$S_{i+1} = S_i \setminus \{k\}$ .

        Let  $i = i + 1$ .

        If  $S_i = \emptyset$  then Stop

    Else

        Stop

Return variable selection set  $S_i$ .

Another method is ***stepwise general selection*** which is like stepwise forward selection but allows for backward selection during the process. We do not cover this method in detail.

Which is the best stepwise procedure?

All three methods are available in statistical packages and are used. There is no general consensus as to which is “best”.

## Variable inclusion criterion

All of these methods require a model fitness measure  $m_{\text{fit}}$ .

There are several alternatives: eg p-values on coefficient estimates or F-test on the model.

However, typically a model fitness measure is used, to avoid multiple hypothesis testing. In particular:

- Akaike's information criterion (AIC)

## Akaike's Information Criterion (AIC)

What we would like to do is to choose the model that gives the best model fit.

- For example, for MLE, we could choose the model that maximizes the likelihood.

The problem is that the addition of a variable to a model always improves model fit.

- The likelihood will always increase (even if only by a small amount).
- This is because each additional variable gives an extra degree of freedom to fit the data.
- Therefore, with a model fit criteria, *all* variables would be selected.

The solution to this problem is to penalize the model fitness based on the number of predictor variables entered into the model. This is what the AIC does.

- AIC is a popular variable inclusion measure for predictive models where model fit can be expressed as a likelihood measure.
- AIC includes a penalty term for number of variables.

Therefore, there will be a point at which AIC increases with the inclusion of further variables that do not give sufficient increase in model fit to compensate for the penalty.

Formally, AIC is given for a model  $M_S$  with a set  $S$  of predictor variables, in general, as

$$\text{AIC}_S = 2|S| - 2 \log L_S$$

where  $L_S$  is the likelihood of  $M_S$  (remember  $0 < L_S < 1$ ).

Smaller AIC indicates an improved model.

Therefore, we use the fitness measure  $m_{\text{fit}}^{\text{AIC}}(S, T) = \text{AIC}_T - \text{AIC}_S$ .

*Example 10.4*

Let good-payer (0 or 1) be a dependent variable.

Consider the following predictor variables:

Application method, Employment status, Income (log), Residential phone, Months in current job and Location (binary).

Use stepwise backward selection with the AIC criterion for the inclusion of predictor variables.

Iteration steps of the backward variable selection process are as follows.

<b>Step (i)</b>	<b>Variable removed (j)</b>	<b>AIC</b>	$S_i$
0	-	38833	{Application method, Employment status, Income (log), Residential phone, Months in current job, Location}
1	Application method	38831	{ Employment status, Income (log), Residential phone, Months in current job, Location}
2	Location	38829	{ Employment status, Income (log), Residential phone, Months in current job }
3	Removing any more variables does not decrease AIC.		Stop

The following model was generated from this variable selection procedure.

Likelihood Ratio = 1043 (p-value < 0.001)

Variable	Estimate	Standard error	z	P(> z )
Intercept	+0.168	0.0678	2.49	0.013
Employed?	+0.250	0.0284	8.79	<0.001
Income (log)	+0.0750	0.0131	5.71	<.0001
Residential phone	+0.644	0.0298	21.6	<.0001
Months in current job	+0.00374	0.000229	16.3	<.0001

## Shrinkage method: the LASSO penalty



- The stepwise method is heuristic, since it adds or removes variables one at a time using an iterative process, dependent on variables selected or so far. Critically, it does not search all possible variable combinations.
- Adding a penalty term on the *magnitude* of coefficient estimates to the objective function is an alternative approach that allows for variable selection as part of the model fit optimization process.
- The penalty will tend to **shrink** some coefficients to zero, hence essentially de-selecting them.
- Therefore, the final selection of variables gives optimal model fit.
- This is called the LASSO penalty: **Least Absolute Shrinkage and Selection Operator**.

Note: the *objective function* is the function to be minimized to achieve best fit. For OLS regression, this is square-error, for MLE, it is the negative log-likelihood function.

## Including LASSO penalty

&gt;

The LASSO is including in this way:

$$\min_{\beta} [f(\beta; D) + \lambda \|\beta\|_1]$$

where

- $D$  is a training data set,
- $\beta = (\beta_1, \dots, \beta_m)^T$  a vector of coefficients in the model,
- $f$  is the objective function,
- $\|\cdot\|_1$  is the  $L^1$ -norm:

$$\|\beta\|_1 = \sum_{i=1}^m |\beta_i|$$

- $\lambda > 0$  is a user-defined parameter, controlling the level of shrinkage.

## Why does LASSO work?

&gt;

- Firstly, the penalty term will tend to shrink the sizes of coefficient estimates, since this will tend to minimize the Lasso-objective.
- Secondly, the additional penalty term will ensure that setting a coefficient  $\beta_i = 0$  minimizes the Lasso-augmented objective, unless a non-zero  $\beta_i$  contributes to minimizing the objective  $f$ .
- The LASSO is also useful when the number of parameters is large, compared to number of training examples, since the optimization problem is still well-defined.

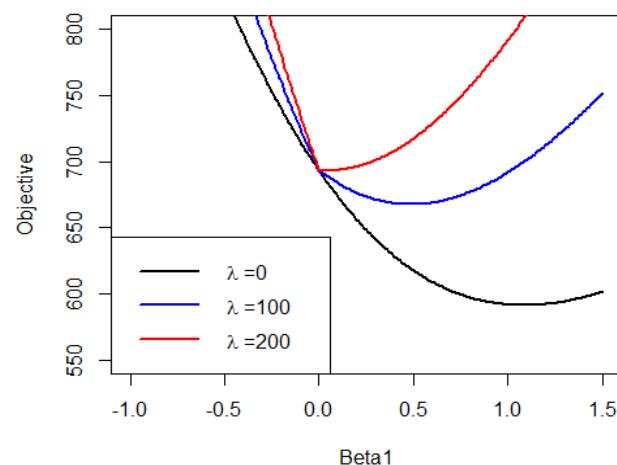
### Example 10.5

Suppose we use logistic regression with one continuous predictor variable  $x_i$  and binary outcome variable  $y_i$ , with a training data set  $n = 1000$  examples.

The objective function to minimize is then,

$$-\sum_{i=1}^n (1 - y_i) \log\left(\frac{1}{1 + e^{-(\beta_0 + \beta x_i)}}\right) + y_i \log\left(\frac{1}{1 + e^{\beta_0 + \beta x_i}}\right) + \lambda|\beta|$$

(refer to chapter 5). Further, simplify with  $\beta_0 = 0$ . This can then be plotted to observe the minima on a graph:



This demonstrates the shrinkage of the coefficient estimate with larger values of  $\lambda$ .

In particular,  $\lambda = 200$ , leads to estimate  $\hat{\beta} = 0$ .

## Estimation with the LASSO penalty



- There are many different ways to estimate a model that includes a Lasso penalty. Here we will focus on the method of **Pathwise Coordinate Optimization** which is an efficient method to do this.
- Pathwise Coordinate Optimization is a gradient descent method, but each iteration is restricted to one dimension of the coefficient space.
- We will consider specifically within the context of OLS linear regression (reviewed in Chapter 2).
- Then the least-squares solution with LASSO penalty is to minimize

$$f(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^n (y_i - \boldsymbol{\beta} \cdot \mathbf{x}_i)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

for predictor vector  $\mathbf{x}_i = (x_{i1}, \dots, x_{im})$ .

- Without loss of generality, suppose the training data is standardized; ie  $\sum_{i=1}^n x_{ij} = 0$  and  $\sum_{i=1}^n x_{ij}^2 = 1$  for all  $j$ .

## Estimate for univariate model

&gt;

- Firstly, consider the simple case when  $m = 1$ , ie only one predictor.
- Then, the standard OLS solution (ie without the LASSO penalty) is

$$\underline{\mathbf{b}} = (\underline{\mathbf{X}}' \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}' \underline{\mathbf{Y}}$$

from Chapter 2.

- Hence for one predictor variable,

$$\hat{\beta}_{\text{OLS}} = \frac{1}{\sum_{i=1}^n x_{i1}^2} \sum_{i=1}^n x_{i1} y_i = \sum_{i=1}^n x_{i1} y_i$$

- The function  $f(\beta)$  is piecewise continuous, hence in univariate case,

$$\begin{aligned} f(\beta) &= \frac{1}{2} \sum_{i=1}^n (y_i^2 - 2y_i \beta x_{i1} + (\beta x_{i1})^2) + \lambda |\beta| \\ &= \left( \frac{1}{2} \sum_{i=1}^n y_i^2 \right) - \beta \hat{\beta}_{\text{OLS}} + \frac{1}{2} \beta^2 + \lambda |\beta| \end{aligned}$$

- Since this is piecewise continuous, for  $\beta \neq 0$ ,

$$\begin{aligned}\frac{df(\beta)}{d\beta} &= -\hat{\beta}_{OLS} + \beta + \lambda \text{sign}(\beta) = 0 \\ \Rightarrow \beta &= \hat{\beta}_{OLS} - \lambda \text{sign}(\beta)\end{aligned}$$

Then consider each of these combinations of  $(\hat{\beta}_{OLS}, \beta)$ :

$\hat{\beta}_{OLS}$	$\beta$	$\Rightarrow$	
$> \lambda$	$> 0$	$\beta = \hat{\beta}_{OLS} - \lambda > 0$	
$> \lambda$	$< 0$	Not possible	
$-\lambda \leq \hat{\beta}_{OLS} \leq \lambda$	$> 0$	$\beta < 0$	Contradiction
$-\lambda \leq \hat{\beta}_{OLS} \leq \lambda$	$< 0$	$\beta > 0$	Contradiction
$< -\lambda$	$> 0$	Not possible	
$< -\lambda$	$< 0$	$\beta = \hat{\beta}_{OLS} + \lambda < 0$	

Therefore, there is a single solution for the case  $|\hat{\beta}_{OLS}| > \lambda$  but no solution for  $-\lambda \leq \hat{\beta}_{OLS} \leq \lambda$ , hence by *reductio ad absurdum*, it must be that  $\beta = 0$  in this case.

Indeed, when  $-\lambda \leq \hat{\beta}_{OLS} \leq \lambda$ ,  $\beta < 0 \Rightarrow \frac{df(\beta)}{d\beta} < 0$  and  $\beta > 0 \Rightarrow \frac{df(\beta)}{d\beta} > 0$   
 which shows that  $f(\beta)$  is monotonically decreasing and increasing each side of  $\beta = 0$ , which makes  $\beta = 0$  the minima.

Therefore

$$\beta = S(\hat{\beta}_{OLS}, \lambda)$$

where

$$S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda & \text{if } \hat{\beta} > \lambda \\ \hat{\beta} + \lambda & \text{if } \hat{\beta} < -\lambda \\ 0 & \text{if } |\hat{\beta}| \leq \lambda \end{cases}$$

- This result demonstrates how the LASSO shrinks the coefficient estimate.

## Estimate for multivariate model



Use coordinate descent.

Find minima by iteratively moving in the direction of the minima in one dimension/coordinate at a time.

### Algorithm

1. Start with an initial set of proposed estimates  $\beta^{[0]}$ , eg  $\beta^{[0]} = \hat{\beta}_{OLS}$ .
2. Set iteration  $t = 1$ .
3. Repeat until convergence:

For  $j = 1$  to  $m$ :

$$\beta_j^{[t]} = \arg \min_{\beta_j} f(\beta_j ; \beta^{[t-1]})$$

Set  $t \leftarrow t + 1$ .

Under mild conditions, this method will converge to the minima for  $f(\beta)$ .

- Following a similar argument to the univariate case, for OLS linear regression,

$$\beta_j^{[t]} = S\left(\beta_j^{[t-1]} + \sum_{i=1}^n x_{ij}(y_i - \boldsymbol{\beta}^{[t-1]} \cdot \mathbf{x}_i), \lambda\right)$$

at each update stage of the algorithm.

Proof is an exercise in problem sheet 2.

- Reference:

Friedman J, Hastie T, Hofling H, Tibshirani R (2007), Pathwise Coordinate Optimization, *Annals of Applied Statistics*, Vol. 1, No. 2, pp.302-332.

## LASSO for logistic regression

- We have shown how the LASSO can be implemented for OLS linear regression, but how about logistic regression?
- The log-likelihood function for logistic regression can be expressed as an iteratively reweighted least squares (IRLS) problem. With this approach, a similar derivation using coordinate descent can be made.
- Details will not be given here. For further reading,  
Friedman J, Hastie T, Tibshirani R (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent*, Journal of Statistical Software, Jan 2010, Vol 33, Issue 1.

### Example 10.6

Consider a data set of mortgages (source: Freddie Mac, USA).

Models of non-default within 2 year from loan origination:

Predictor variable	Data type	LR coefficient estimate	LR: $\text{Pr}(> z )$	LR with LASSO coefficient estimate
(Intercept)		-2.45	0.881	4.23
FICO score	continuous	0.0127	<0.0001	0.0117
Debt-to-Income (DTI)	continuous	-0.0392	<0.0001	-0.0314
Loan-to-Value	continuous	-0.0545	<0.0001	-0.0396
Log(loan value)	continuous	-0.341	<0.0001	-0.146
Interest rate	continuous	-1.16	<0.0001	-1.05
First-time-homebuyer	0/1	0.134	0.0837	0
First-time-homebuyer unknown	0/1	0.287	<0.0001	0.0065
Number of units	0,1,2	-0.0882	0.199	0
Occupancy status				
= Investment	0/1	0.068	0.436	0
= Second home	0/1	-0.242	0.0882	0
= Owner occupier	0/1	excluded		
DTI>65%	0/1	0.282	0.0673	0

Channel				
= Broker	0/1	0.183	0.0125	0.0118
= Correspondent	0/1	-0.178	0.00612	0
= Unspecified	0/1	-0.533	<0.0001	-0.488
= Retail	0/1	excluded		
Property type				
= Condo	0/1	0.242	0.00382	0
= Planned unit	0/1	0.242	0.00344	0
= Others	0/1	1.31	0.0511	0
= Leasehold	0/1	excluded		
Loan purpose				
= Cash-out refinance	0/1	-0.0927	0.169	-0.00571
= Purchase	0/1	0.722	<0.0001	0.41
= No cash-out	0/1	excluded		
Original loan term	20 to 30	0.0281	0.537	0
Number of borrowers	1 or 2	0.871	<0.0001	0.651

0 indicates coefficients shrunk to zero and hence de-selected.

Note, both models give the same predictive performance, measured by AUC, on an independent test set.

## Review of Chapter 10



We need to decide which characteristics to include as predictor variables.

We have not covered all variable selection techniques – many are available.

We have covered the following topics:

1. Variable selection
2. Univariate variable selection
3. Multivariate variable selection
  - a) Stepwise variable selection
  - b) Akaike's Information Criterion
  - c) LASSO penalty

# Consumer Credit Risk Modelling

## Chapter 11: Interaction Terms and Segmentation

## Overview

Before building the model, we need to consider the model structure.

We will consider methods in model structure development that are important for credit scoring:

1. Including interaction terms.
2. Segmentation of the population.

## Interaction terms

It may be that some variables work together to explain the riskiness of an individual.

In particular, the value of one variable will effect the riskiness of another.

In this case, these variables need to be included together in the model.

This is done by adding a *new variable* called the ***interaction term***:

If  $X_1$  and  $X_2$  are two variables, then

$$X_{1,2} = X_1 X_2$$

is their interaction term.

The model coefficient on the interaction term will explain their interacting effect.

It is important that if  $X_{1,2}$  is included in the model then  $X_1$  and  $X_2$  should also be included, otherwise it is difficult to interpret the effect of the interaction.

### Example 11.1

In general, home owners will be considered low risk. However, this may not be true for young families who will be loaded with extra costs as well as a mortgage. Young families usually have income from people in the 26-39 age group.

Therefore we might have the following situation.

<i>Age</i>	<i>Home owner</i>	<i>Not home owner</i>
18-25	Low risk	Medium risk
26-39	High risk	Low risk
40+	Low risk	Medium risk

We may then consider including the interaction term  $Age \times Home\ owner$ .

## Finding interaction terms

How do we discover variable interactions?

1. Expert (human) judgement can be used in the first instance to consider which variables are likely to interact and then test those candidate interactions statistically. Financial risk managers and their teams will have this expertise.
2. By manual inspection of the variables and their association with the outcome.
  - This can be done by plotting a graph of data points and analyzing the patterns (by eye) or building small models to test whether interaction terms are significant.

3. An automated search of interaction terms using variable selection is also possible.

- In particular, multiple interaction terms can be included as candidate variables in the full model structure and one of the variable selection methods described in Chapter 10 can be used to remove those interaction terms that do not contribute to model fit.
- However, some care is needed since there will be many possible combinations of possible interactions and some control is needed to avoid generating too many false discoveries.

[ An analogy... if I throw 2 dice enough times I will get a 12 by chance... if I search over a large space of interaction terms, one of them will come out as significant, just by chance].

## Interpreting interaction terms

When interaction terms are included, it is not straightforward to interpret the effect of individual variables in the model.

If we include an interaction term between two variables, then we end up with three coefficient estimates: two for the original variables and one for the interaction term:

$$s(\mathbf{x}) = \cdots + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 + \cdots$$

We can use the p-value on the estimate of  $\beta_{1,2}$  to determine the statistical significance of the interaction.

That is, if the null hypothesis that  $\beta_{1,2} = 0$  cannot be rejected then there may be no interaction effect.

## Plot effect of variables on score

The easiest way to see the changing effect of one variable on another is to plot a graph of score with changes to the two variables, *ceteris paribus* (ie keeping all other terms constant).

This will be a 3D graph if both variables are non-categorical,

Or, can be plotted as multiple 2D graphs if one of the variables is categorical.

- Y-axis plots score against the x-axis plot of the non-categorical value for multiple values of the categorical variable.
- See Example 11.2 below.

Remember, for logistic regression, the score is calculated as  $s(\mathbf{x}) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}$ .

## Marginal effect on log-odds score

Remember from Chapter 6 the **marginal effect** of variable  $X_1$  is given by

$$\frac{\partial f(\mathbf{x})}{\partial x_1}.$$

For a variable which is included in a single interaction term  $x_1x_2$  this is

$$\beta_1 + \beta_{1,2}x_2.$$

- Clearly then the effect of  $x_1$  is dependent linearly on the value of  $x_2$ .
- And, there is a similar marginal effect of  $x_2$ , of course.
- This is the marginal effect on the credit score; the relationship to probability of default is indirect (not linear).

The marginal effect is linear with  $x_2$  and it is useful to plot the effect on a graph with changes in  $x_2$ .

*Example 11.2.*

Two predictor variables are included in a scorecard: whether the borrower is a renter and months in residence. The dependent variable is non-default.

This logistic regression shows the effects of each variable.

Variable	Estimate	Standard Error	z	P(>  z )
Intercept	1.10	0.05	22.8	<0.001
Renter	-0.077	0.039	2.0	0.047
Months in residence (log)	0.0688	0.0101	6.79	<0.001

Interpretation:

- Generally, renters have lower credit scores.
- Credit score improves with number of months in current residence.

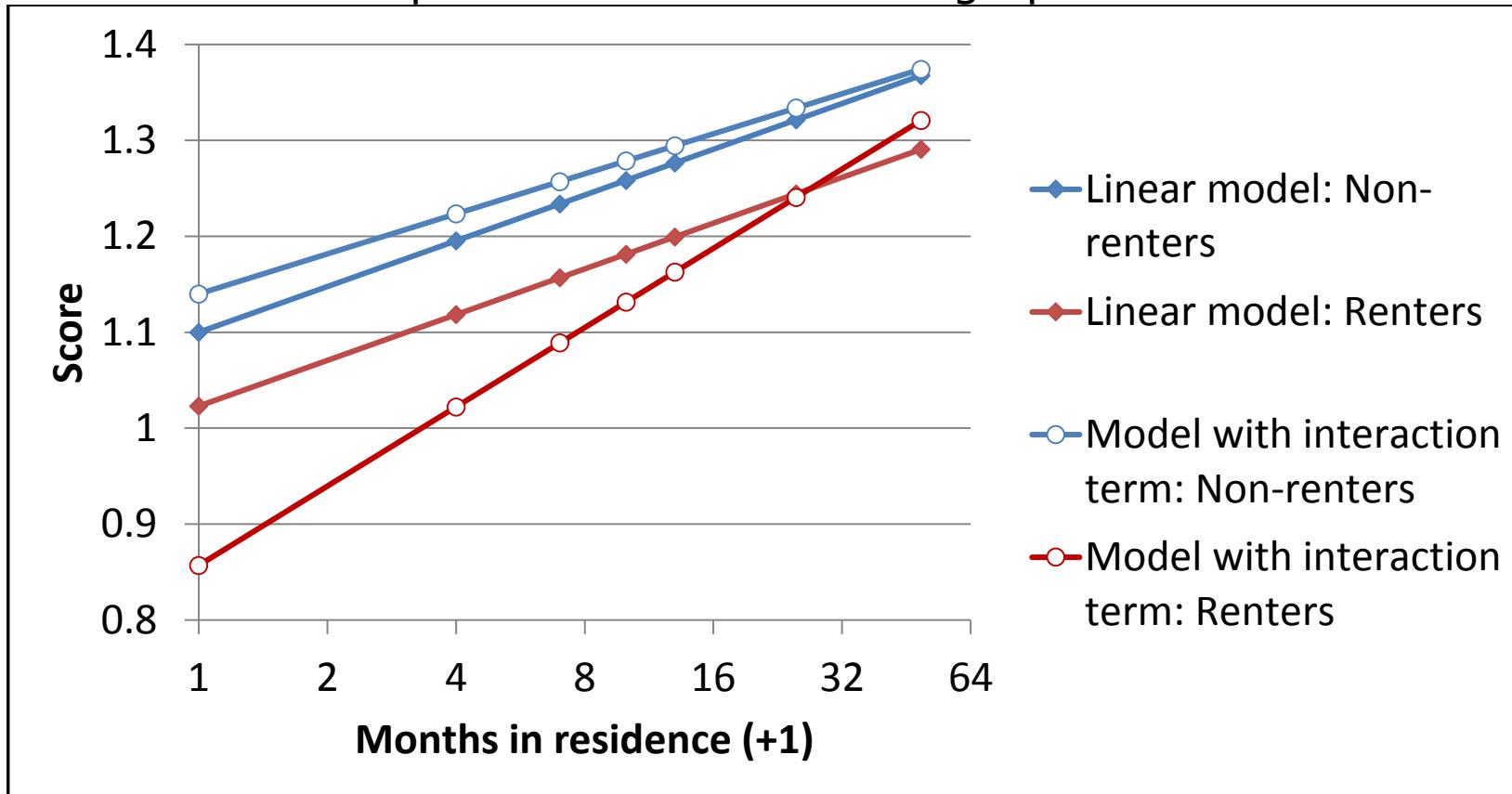
*Example 11.2 continued.*

However, it is reasonable to hold the prior assumption that the effect of months in residence may be different for renters and non-renters. Therefore the interaction term *Renter\*Months in residence* is included in the model. The updated model is shown below.

Variable	Estimate	Standard Error	z	P(>  z )
Intercept	1.14	0.05	21.8	<0.001
Renter	-0.283	0.107	2.65	0.008
Months in residence (log)	0.0602	0.011	5.47	<0.001
Renter * Months in residence (log)	0.059	0.029	2.05	0.040

The interaction term is statistically significant at 5% level.

The effect is best interpreted with the aid of a graph.



$$\text{Score is } s(x) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}.$$

This shows that including the interaction term,

- differentiates the effects for renters and non-renters further;
- indicates a greater months in residence effect for renters.

Marginal effects are calculated as follows.

- Marginal effect of being a Renter is

$$\beta_1 + \beta_{1,2}x_2 = -0.283 + 0.059 \text{ Months in residence (log)},$$

relative to being a non-Renter.

- Marginal effect of Months in residence (log) is

$$\beta_2 + \beta_{1,2}x_1 = 0.0602 + 0.059 \text{ Renter};$$

ie 0.1192 if Renter and 0.0602 if not a Renter.

## Extending interaction terms to more than 2 variables

It is possible that more than two variables have an interaction effect. If this is the case, then the principles outlined for the 2-variable case can be generalized to more than 2 variables.

For example, if three variables,  $X_1$ ,  $X_2$  and  $X_3$ , interact then the following 4 interaction terms will need to be included in the model:

$$X_{1,2} = X_1 X_2, X_{1,3} = X_1 X_3, X_{2,3} = X_2 X_3, X_{1,2,3} = X_1 X_2 X_3$$

The marginal effect of  $X_1$  is then given as

$$\frac{\partial [\beta_1 x_1 + \beta_{1,2} x_1 x_2 + \beta_{1,3} x_1 x_3 + \beta_{1,2,3} x_1 x_2 x_3]}{\partial x_1} = \beta_1 + \beta_{1,2} x_2 + \beta_{1,3} x_3 + \beta_{1,2,3} x_2 x_3$$

with similar expressions for  $X_2$  and  $X_3$ .

## Segmentation

Sometimes the values of a variable  $X$  represent such a major change in the population of risk factor characteristics that it would mean including far too many interaction terms.

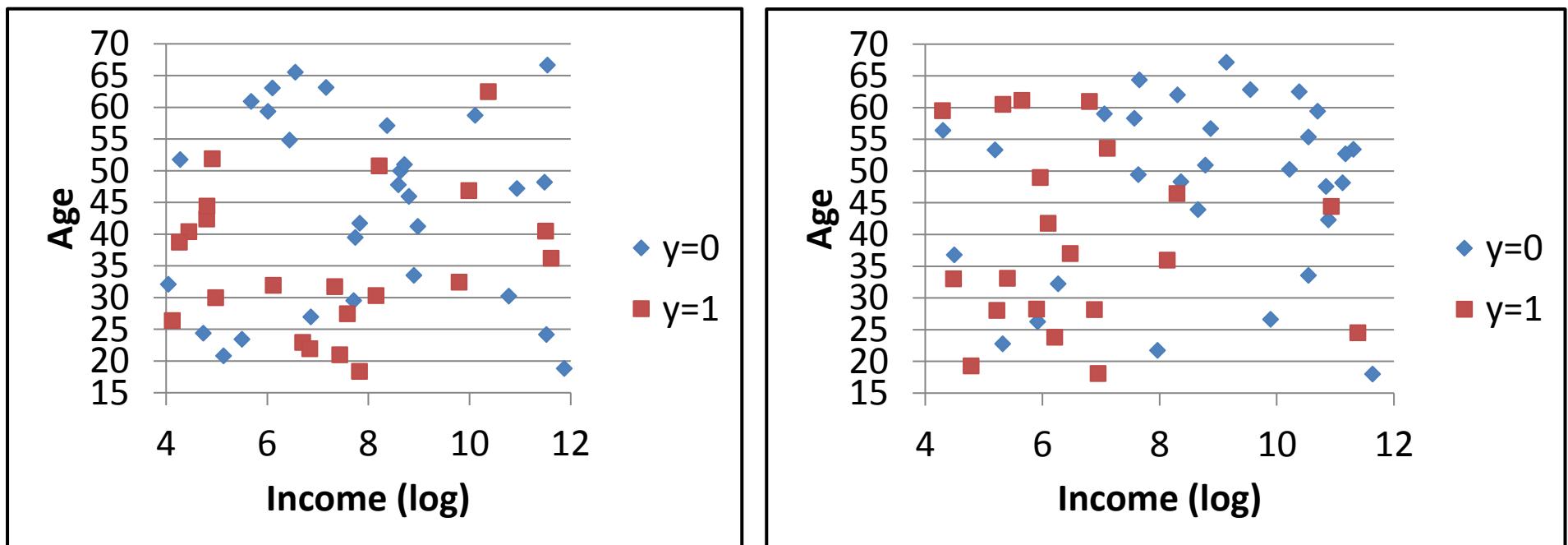
This would be a problem for two reasons:

1. It is unlikely to result in a robust model.
2. It will be difficult to interpret the model  
(something which is desirable from the point of view of risk management in banks and regulation).

In this case, the data can be segmented by values in  $X$  and then separate scorecard models built on each separate data segment.

*Example 11.3*

Consider the country of Itovia. For a different data set of personal loans, the following samples of age/income association with default ( $Y = 1$ ) emerge for those living in the South (left) and North (right).



- These graphs show a distinct association between age, income and outcome for the two geographical areas.
- Hence the North/South divide is a good candidate for segmentation.

## Business reasons for segmentation

Apart from the statistical reasons, there are two other business reasons why we may want to segment the data.

1. There may be a difference in information between two or more groups.  
For example, since there will be little credit history available for young people, they may want to be treated differently to older people.
2. The lender may want to treat different groups in different ways.  
For example, offering higher earners preferential rates.

A particular instance of point (1) is to use of segmentation to offer a solution to the ***missing values problem*** if values are structurally missing (eg missing by definition – review Chapter 8).

- That is, observations where data are missing on a variable are taken as a separate segment and a sub-model built without that variable, whilst a model with no missing data is built on the segment where values are available for that variable.

*Example 11.4*

Suppose we have a field indicating home owner (Y/N) and another for home value (£).

Then clearly home value is structurally missing depending on whether home owner=Y.

Therefore segment on home owner and

1. Build a model with home value included if home owner=Y;
2. Build a second model with home value excluded if home owner=N.

## Segmented Model

Formally, let  $X_C$  be a categorical variable with distinct values  $v_1, v_2, \dots, v_K$  (if it isn't, we know how to discretize a continuous variable).

We segment on  $X_C$  as follows.

For each  $j \in \{1, \dots, K\}$ , build a scorecard  $m_j(\mathbf{X})$  using only observations where  $X_C = v_j$ , where  $\mathbf{X}$  is a vector of other predictor variables.

- Note that, although all variables are available for all segments generally, for a specific segment model build we can choose to exclude them simply by fixing their coefficient to zero.

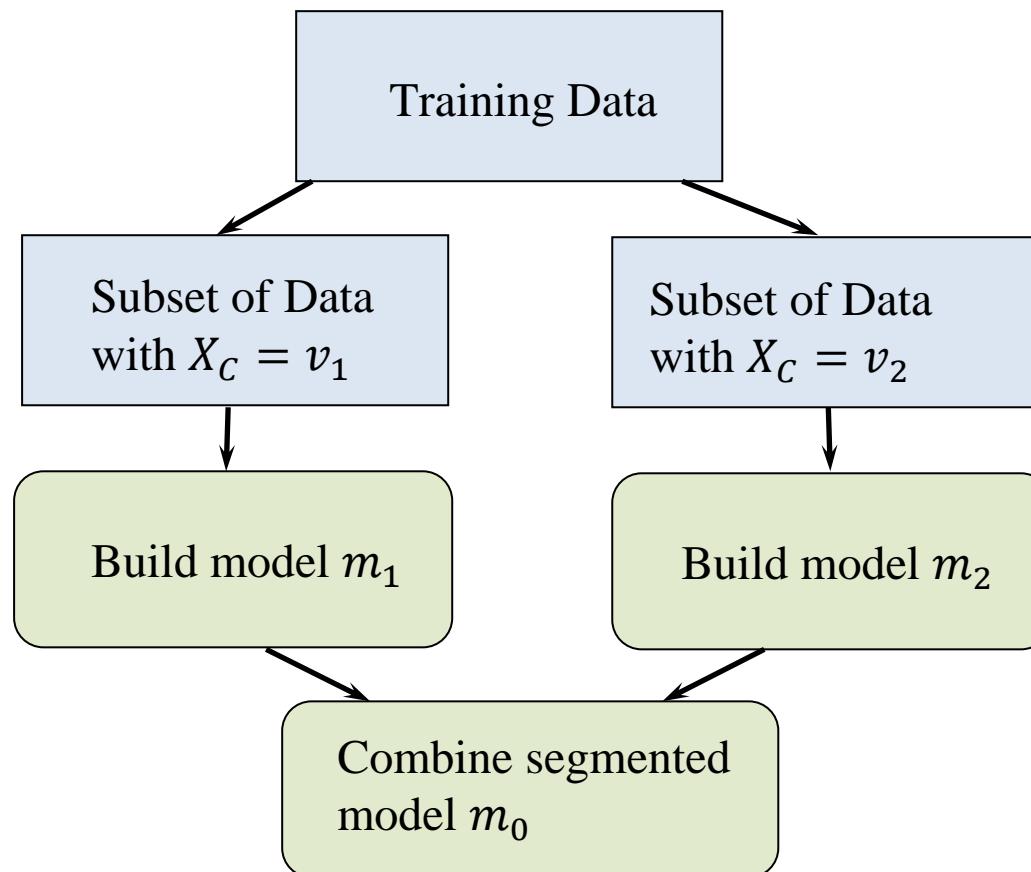
Then an overall scorecard model is constructed from the segments as

$$m_0(\mathbf{x}) = \begin{cases} m_1(\mathbf{x}) & \text{if } x_C = v_1 \\ \vdots & \vdots \\ m_K(\mathbf{x}) & \text{if } x_C = v_K \end{cases}$$

where  $\mathbf{x} = (x_1, \dots, x_C, \dots, x_m)$ .

Note: This definition relies on the separate segment scorecards to be calibrated/scaled in the same way (otherwise it is like taking measurements in both meters and feet).

### **Illustration of building a segmented model with two segments.**



## **Segmenting in more than one variable**

- We can segment on multiple categorical variables.
- However, the more variables that we use, the more segments we have.
- If we have too many segments, then the sample size in each segment may become too small to build useful models.

## Measuring performance of segmentation

- Ultimately it is the discrimination performance gain of the segmented model which is valuable.
  - Therefore, test the performance of the overall scorecard  $m_0$ , rather than the individual segment models  $m_j$  for  $j \in \{1, \dots, K\}$ .
  - In particular, we want it to outperform a scorecard built on the whole data set (that is, without segmentation).
- If some understanding of performance improvements within each segment is required, then AUC can be computed for specific segments of data.

### Example 11.5

A lender has a portfolio of a credit card called *WhizzCard*. In the past, *WhizzCard* has been marketed fairly equally to all age groups. However, for next year, aggressive marketing will be targetted to the under 25 age group. They expect that the proportion of under 25s applicants to over 25s will rise to 2:1.

They will build a scorecard segmented on age to reflect this anticipated change.

Characteristics available for scorecard build are: age, income, residence status (rents, lives with parents, other) and months in residence.

Since most over 25s in this population do not live with their parents, this characteristic is not included in their scorecard.

*Example 11.5 continued*

The two scorecards for each segment, built using logistic regression, are shown below.

Characteristic	Age<25		Age>=25	
	Coefficient estimate	P-value	Coefficient estimate	P-value
Intercept	0.978	<0.001	0.952	<0.001
Age	0.116	<0.001	0.0351	<0.001
Income (log)	0.0254	0.48	0.0186	0.40
Rents home?	0.0585	0.67	-0.303	0.0017
Lives with parents?	0.26	0.02	n/a	
Home owner (base)	0		0	
Months in current residence	0.000106	0.02	-0.00054	0.024

*Example 11.5 continued*

The scorecards show differences in risk factors between the two age groups:

- The association of age (within the age range) is much higher for the younger sample;
- Being a renter (ie not owning one's own home) is only a major risk factor for the older group.

How does the segmented model perform on an independent test set?

- We set up a test set with numbers in each age group as 2:1 to reflect the expected distribution of future applicants.
- Overall AUC on the test set for the unsegmented and segmented model are given below.

	AUC
Unsegmented model	0.615
Segmented model	0.621

*Example 11.5 continued*

We see a small improvement in predictive power (as measured by AUC) for the segmented model.

*(However, we may ask: Is the difference statistically significant?...)*

Given that the segmented model performs no worse than the unsegmented model and given the business requirements, it would make sense to adopt the segmented model.

## Overview of chapter 11

We have considered methods in model structure development that are important for credit scoring:

1. Interaction terms
2. Segmentation of the population

Further reading:

- Bijak K and Thomas LC, *Does segmentation always improve model performance in credit scoring?*, Expert Systems with Applications 39 (2012) 2433–2442.

# Consumer Credit Risk Modelling

## Chapter 12: Segmentation and Decision Trees

## Overview

We will see how to represent segmented models as decision trees.

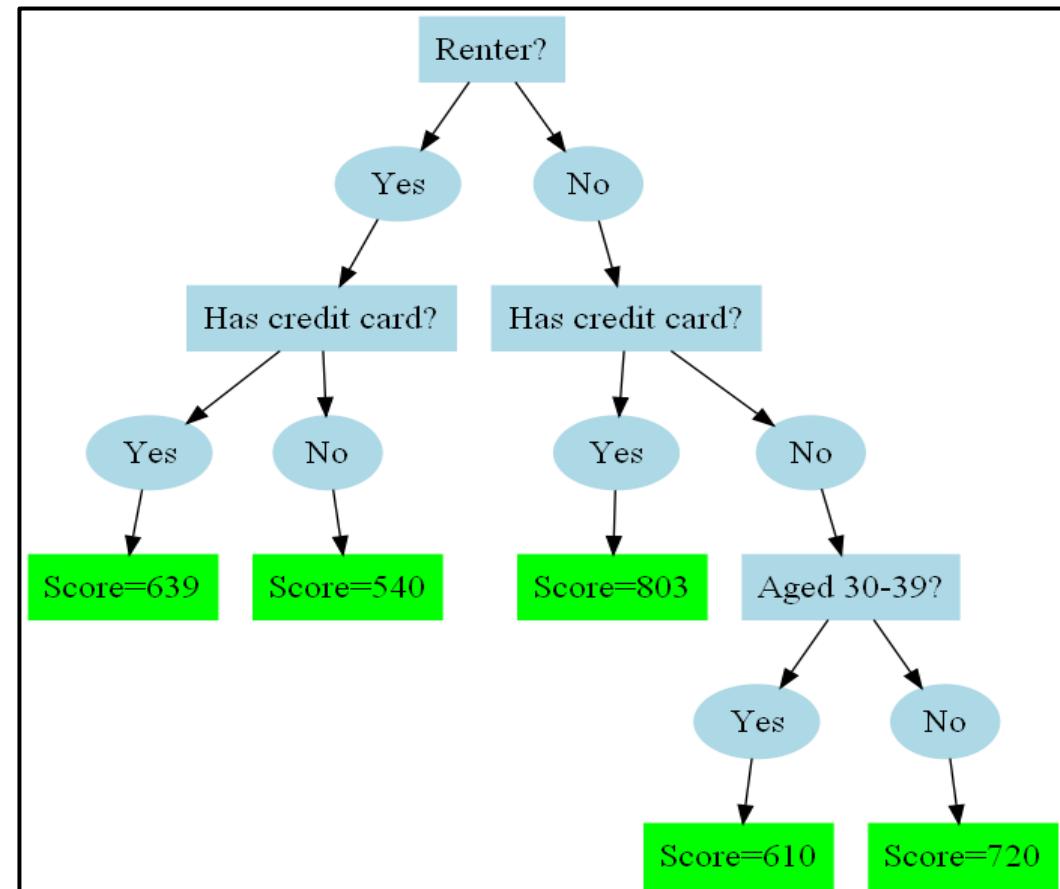
1. Decision trees
2. General decision tree models for segmentation
3. Automated decision tree models: CART
4. Logistic regression tree

## Decision trees and segmentation

We can think of a segmentation on several variables as a decision tree with separate models built at each of the leaf nodes of the decision tree.

A decision tree is a tree structure that is traversed by starting at a root node, asking questions of the data at each node and taking a different branch for each possible answer, until a leaf node is reached (ie no further branches) and a decision is made.

Here is a credit scoring example (adapted from Thomas 2010, fig. 1.10.2).



## Terminology for decision trees.

Node	A point in the tree.
Level	Nodes exist at different levels of the decision tree: 1, 2, etc.
Root node	The starting point in the tree on level 1.
Branch	A directed path from a node on a level $x$ to a node on level $x + 1$ .
Leaf node	A terminating node; ie a node with no branches down to the next level. An estimate for the outcome variable is given here.
Non-leaf node	A node with further branches; this is a decision node based on the value of a predictor variable.
Partition	The population represented at each node of the decision tree. This is synonymous with a <i>segment</i> in credit scoring terminology.

Because decision trees deal with a finite number of choices for each decision node, only categorical variables can be included *directly* for decisions.

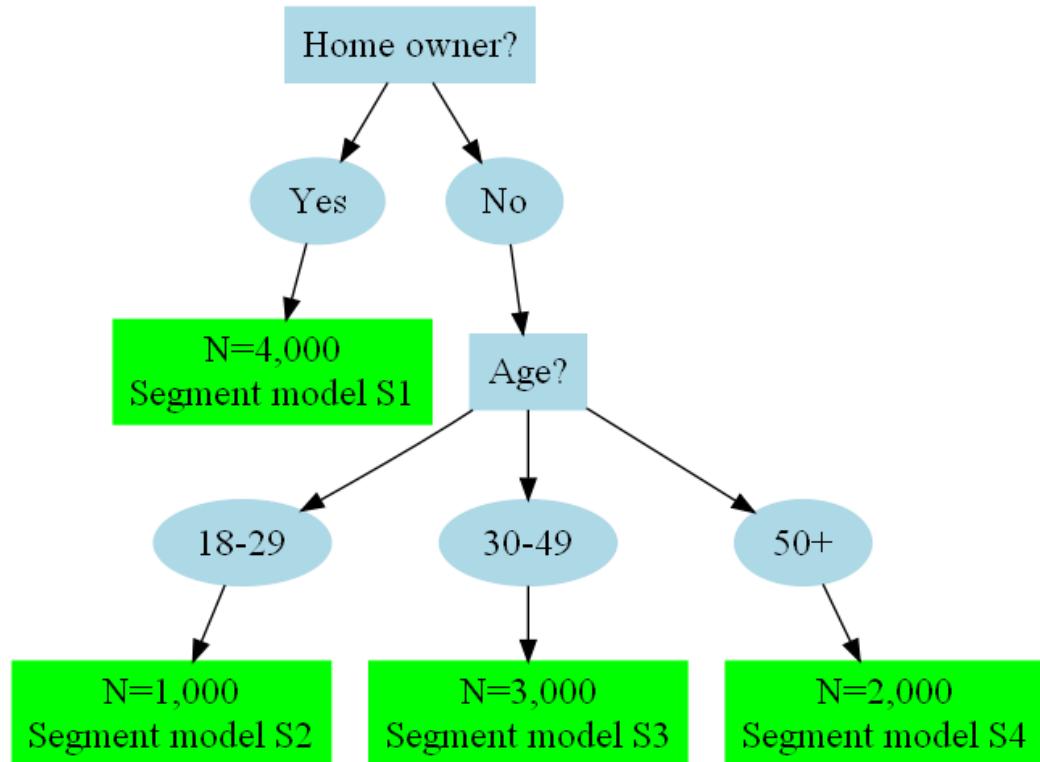
- Continuous variables need to be converted to categorical variables (eg age ranges).
- We have seen how to do this with discretization techniques.

## General decision tree model

- The leaf nodes may be point estimates of outcome or score for each segment. This is the typical use of decision trees.
- However, the leaf node may be a model of outcome for the specific segment represented by the leaf node.
- The leaf node model can be built using a standard method such as ordinary least squares (OLS) or logistic regression.
- This therefore leads to a general modelling approach for credit scoring.

### Example 12.1

- We have a training data set with  $N=10,000$  observations.
- We segment the population first on whether they are a home owner and then, if not, on three levels of age. This gives four segments.
- The decision tree below illustrates this segmentation.



Notice that the tree segments the population into four partitions and a separate scoring model is built for each one: S1 to S4.

How do we build a segmented decision tree model?

- **Expert knowledge.** This is usually a decision made by a credit risk expert, based on business needs and prior knowledge of the financial product and the borrowers.
- **Automation.** However, it is possible to automate the process by searching for variables to segment on; or, to augment expert judgement with statistical analysis. We can proceed as follows:
  - Examine all possible decision trees given all variables in the data set?  
This would produce the best fitting decision tree (in terms of minimizing loss). However, it is not a computationally feasible approach.
  - Therefore, we need a *heuristic* approach to search the space of possible decision trees. This is the approach of several related machine learning algorithms: ID3, C4.5 and CART ("classification and regression trees").

## Decision tree tasks

Decision trees are used for a variety of regression tasks. We will study the following:

1. Classification trees
2. Regression trees
3. Regression tree for segmentation
4. Logistic regression tree

Tasks (1) and (2) are well-known in the literature, but for credit scoring we will need to define tasks (3) and (4).

All four tasks use the same basic decision tree algorithm: the *recursive partitioning algorithm*.

In the remainder of this chapter, this algorithm will be introduced, followed by each of the four tasks.

## Recursive partitioning algorithm

We use a *recursive partitioning algorithm* which builds the decision tree in a top-down manner:

1. Firstly, the most informative variable is taken for the root node.
2. Then for each branch, the next best variable is selected within the data set restricted to the answer from the previous node.
3. This is repeated until no further useful variables can be added.

What we mean by “most informative” or “no further useful variable” depends on the problem the decision tree is trying to solve.

Formally, we initialize the tree build with a single root node  $N^0$ , training data set  $D^0$  and a set of discrete-valued variables  $S_X^0$  which we want to consider for inclusion in the tree. Then call a function Build Tree ( $N^0, D^0, S_X^0$ ).

The function Build Tree is defined recursively as:

Build Tree ( $N, D, S_X$ ) for some given node  $N$ , data set  $D$  and variables  $S_X$ :

If stopping criteria is *not* met, then do:

1. Select  $X_{\min} = \arg \min_{X \in S_X} P(X, D)$  where  $P$  is some penalty measure for using variable  $X$ , measured on data  $D$ .
2. Assign  $X_{\min}$  to node  $N$ .
3. Let  $v_1, \dots, v_K$  be the set of all possible discrete values of  $X_{\min}$ .
4. For  $j=1$  to  $K$ , do:
  - a. Create a new  $j$ th branch from  $N$ , labelled with value  $v_j$  and linked to a new node  $N'$ .
  - b. Call Build Tree ( $N', D[X_{\min} = v_j], S_X \setminus \{X_{\min}\}$ ).
5. Return node  $N$ .

Else (*leaf node*)

Report partition or some summary statistic  $f_{\text{summary}}(D)$ .

Specifically, for binary classification or credit scoring (although what is reported can be different):

6. Let  $n_0 = |D[y = 0]|$  and  $n_1 = |D[y = 1]|$ .

7. Report probability  $p_1 = \frac{n_1}{n_0+n_1}$  of positive event.

8. Report decision  $y=0$  if  $n_0 \geq n_1$ .

9. Report decision  $y=1$  if  $n_0 < n_1$ .

where

- $|D|$  denotes the number of observations in data set  $D$ .
- $D[\text{condition}]$  denotes the partition formed from only observations in data set  $D$  where *condition* is true.

We now just need to specify the penalty measure  $P$  and the stopping criteria. This depends on the problem.

## Recursive functions: some background

- Recursive functions are functions that are defined in reference to themselves.
- They are used in computer science and can be well-defined.
- For example, the factorial function can be properly defined as follows:

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ xf(x - 1) & \text{if } x > 0 \end{cases} \text{ for } x \in \mathbb{N}_0$$

## Classification Tree

Suppose  $Y \in \{0,1\}$  (or can be generalized to any size finite range).

### ***Penalty measure***

To build a *classification tree* (also known as a *decision tree classifier*), where we wish to make a classification decision, any of the discrimination measures we have already specified can be used (eg chi-square or divergence).

However, it is typical to use the information gain measure

$$P(X, D) = -I_G(X)$$

where information gain is calculated across data partition  $D$ .

Recall from Chapter 7:-

$$I_G(V) = \sum_{j=1}^K \left( P(V = v_j | Y = 0) - P(V = v_j | Y = 1) \right) w(v_j)$$

where  $P(V = v | Y = 0)$  is frequency of value  $v$  given a positive outcome,  
 $P(V = v | Y = 1)$  is frequency of value  $v$  given a negative outcome,  
and  $w(v) = \log\left(\frac{P(V=v|Y=0)}{P(V=v|Y=1)}\right)$  is the weights of evidence (WOE) of  $v$ .

- When used in the recursive partitioning algorithm, the conditional frequencies are taken as the empirical frequencies within the data set (partition)  $D$ .
- Note that  $P(X, D) = -I_G(X)$  is negative since we want to use a performance measure as a penalty measure.

## ***Stopping criteria (classification tree)***

The original stopping criteria (from ID3) was:

- Stop when all observations have the same class; ie  $n_0=0$  or  $n_1=0$ .

However, this can lead to large trees that are not good estimators, because each leaf node partition has too small a sample size.

Therefore, *additional* conditions have been suggested to avoid this problem. The obvious one is to control the sample size at each node:

- Stop when number of observations is less than some minimum threshold  $n_{\min}$ ; ie  $|D| < n_{\min}$ .

## Regression Tree

If the task is to estimate a real number value,  $Y \in \mathbb{R}$ , then we will construct a regression tree.

### **Penalty measure**

We can use the penalty measures which are usual with standard regression models, such as that based on square error.

Thus, for the recursive partitioning algorithm, the mean square error (MSE) penalty measure is

$$P(X, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} (y_i - d(x_i))^2$$

where

- $y_i$  is the outcome for observation  $i$  in  $D$  for  $i = 1$  to  $|D|$ ;
- $x_i$  is the value of  $X$  for observation  $i$  in  $D$ ;
- $d(x_i)$  is the estimate for outcomes given for  $X = x_i$ .

For the CART algorithm, the estimate is taken as the mean value within the partition:

$$d(x_i) = \frac{1}{|E|} \sum_{j \in E} y_j$$

where  $E = \{j : x_j = x_i, j \in \{1, \dots, |D|\}\}$ .

This leads to a decision tree where the leaf nodes have some fixed estimate of a real number outcome variable.

- Slide 24 has an example if we had Score as outcome variable,

Other penalty measures could also be used, such as least mean absolute difference:

$$P(X, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |y_i - d(x_i)|$$

## ***Stopping criteria (regression tree)***

This is easier to specify for regression because we can use the penalty measure directly, stopping when the total penalty on the segments is greater than for the whole data set:

- Stop if  $P(X, D) \geq \lambda P_D$  where  $P_D$  is the penalty on the whole data set  $D$ , where  $0 < \lambda \leq 1$ .

For MSE, this is

$$P_D = \frac{1}{|D|} \sum_{i=1}^{|D|} \left( y_i - \left( \frac{1}{|D|} \sum_{j=1}^{|D|} y_j \right) \right)^2 = \sigma_y^2$$

where  $\sigma_y^2$  is the (biassed) sample variance of  $y$  in  $D$ .

As with classification trees, there may be a problem with this criterion in terms of sample size, therefore the same *additional* criterion can be added to control sample size:

- Stop when number of observations is less than some minimum threshold  $n_{\min}$ ; ie  $|D| < n_{\min}$ .

## Regression Tree for Segmentation

- For the general segmentation problem, we need a *regression model* at each leaf node.
- Each possible segmentation needs to be tested using the penalty measure  $P$ :
  - Suppose  $X$  has  $K$  discrete values  $\{v_1, \dots, v_K\}$ .
  - Then estimate  $K$  regression models  $f_j$  where  $\hat{y} = f_j(\mathbf{x})$ 
    - using training data  $D[X = v_j]$
    - and where  $\mathbf{x}$  is a vector of values for all variables in  $S_X$ .
  - Typically, OLS regression will be used.
    - In which case,  $\hat{y} = \hat{f}_j(\mathbf{x}) = \hat{\beta}_{0,j} + \hat{\beta}_j^T \mathbf{x}$ .
  - Then use the estimator  $d(x_i) = \hat{f}_j(\mathbf{x}_i)$ 
    - where  $x_i = v_j$  and  $\mathbf{x}_i$  is the vector of values for observation  $i$  in  $D$ ,
    - in place of the simple mean of  $y$  within the partitions of  $D$ .
- Note: The mean value can be interpreted as the estimate given by the within-partition *null model* (ie a model with no variables) using OLS.

- This will give us a decision tree with segmented models at each leaf node as we require (see example on slide 7).
- Stopping criteria for segmentation must emphasize large sample size since we expect each leaf node regression to be based on a large sample size. Hence  $n_{\min}$  should be set to a high value.
- When reporting partition at the leaf node, report the regression model  $\hat{y} = \hat{f}(\mathbf{x})$  built using training data  $D$ .

## Logistic Regression Tree

- The regression tree formulation is that the leaf node models are regression models.
- If we are using logistic regression to build scorecards, we replace the penalty measure with a logistic regression penalty.
- The equivalent penalty measure commonly used with logistic regression is the **deviance** which measures the extent to which variation in data  $D$  is explained by a model estimated using MLE:

$$\Delta_C(D) \triangleq -2 \log \left( \frac{L(\hat{\boldsymbol{\theta}}_C; D)}{L(\hat{\boldsymbol{\theta}}_S; D)} \right)$$

where

- $L(\hat{\boldsymbol{\theta}}_C|D)$  is the maximized likelihood for a model  $C$  with estimates  $\hat{\boldsymbol{\theta}}_C$
- and  $S$  is the saturated model (ie it fits all training data as well as possible given the model structure).

For logistic regression, and  $n$  observations in  $D$ , we have

$$\log L(\hat{\beta}_0, \hat{\beta}; D) = \sum_{i=1}^n y_i \log \hat{p}_{1i} + (1 - y_i) \log(1 - \hat{p}_{1i}) \text{ and } L(\hat{\theta}_S; D) = 1$$

where  $\hat{p}_{1i}$  is the estimated probability of  $y_i = 1$  given model  $C$  (recall MLE from Chapter 5).

Therefore

$$\Delta_C(D) = -2 \sum_{i=1}^n y_i \log \hat{p}_{1i} + (1 - y_i) \log(1 - \hat{p}_{1i})$$

Larger values of deviance indicate worse model fit to data, hence deviance is a penalty measure. Therefore use

$$P(X, D) = \sum_{j=1}^K \Delta_C(D[X = v_j])$$

where  $\{v_1, \dots, v_K\}$  is the set of discrete values in category variable  $X$ .

## Pruning

There is some evidence that simply using the stopping criteria to control the size of the decision tree and its accuracy as an estimator is not sufficient.

Decision tree modelling algorithms such as CART also include a pruning stage:

1. Build a full decision tree using the recursive partition algorithm,
2. Prune the full decision tree back to generate a parsimonious tree.

One way to prune is to include a penalty term for the size of the decision tree and modify the penalty measure accordingly:

$$P_\alpha(X, D) = P(X, D) + \alpha |\tilde{T}|$$

where

- $\tilde{T}$  indicates the decision tree built up to the segmentation on  $X$ ;
- $|\tilde{T}|$  is some measure of the size of that tree (the number of leaf nodes in CART); and
- $\alpha > 0$  is a parameter that controls the relative size of the tree size penalty.

This is called an ***error-complexity measure***, since it combines penalties for model fit error and model complexity.  
Notice the similarity with the AIC measure introduced in Chapter 8.

### *Example 12.2*

We have another data set of personal loans from Itovia with 10,000 observations. We consider several segmentations on geographical location North/South and whether the individual was a previous customer.

The table below lists the square error for regression models built on each possible segment (partition) of data.

<b>Segment / partition</b>	<b>Number of observations</b>	<b>Square error</b>
All data	10,000	906.0
Location=South	7,400	500.2
Location=North	2,600	203.8
Previous customer	2,800	189.3
Not previous customer	7,200	564.5
Location=South and Previous customer	2,200	160.4
Location=South and Not previous customer	5,200	208.0
Location=North and Previous customer	600	40.6
Location=North and Not previous customer	2000	96.8

*Example 12.2 continued*

Use the information in this table to build a decision tree for segmentation using the square error penalty function and a minimum sample size stopping criterion with  $n_{\min}=3,000$ .

*Solution*

**Root node:**

- Sum of square error for Location is  $500.2+203.8=704.0$ ;
- Sum of square error for Previous customer indicator is  $189.3+564.5=753.8$ ;
- $704.0 < 753.8 < 906.0$  (all data) so split on Location.

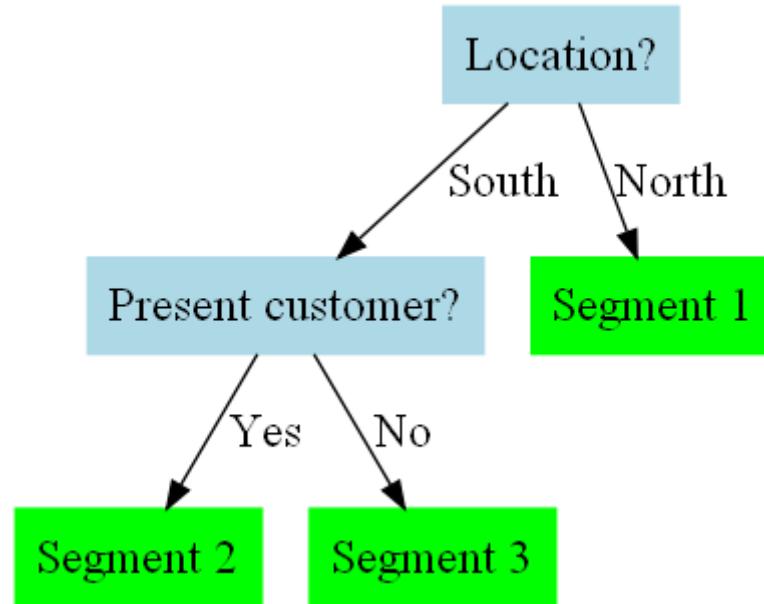
**Level 2: Segment on Location=North:**

- Sample size= $2,600 < n_{\min} = 3,000$ , therefore stop.

**Level 2: Segment on Location=South:**

- Sample size= $7,400 > n_{\min} = 3,000$ , so continue.
- Sum of square error for Previous customer indicator is  $160.4+208.0=368.4$ ;

- $368.4 < 500.2$  (for all Location=South) so split on Previous customer indicator.
- No further candidate variables so stop.



## Further references

To read more about automated processes for building decision trees see:-

- Breiman, Friedman, Olshen and Stone (1984). Classification and Regression Trees (Chapman and Hall).

For more about logistic regression trees, see

- Chan, K.-Y. and Loh, W.-Y. (2004), "LOTUS: An algorithm for building accurate and comprehensible logistic regression trees," *Journal of Computational and Graphical Statistics*, **13**(4): 826-852.

This paper is available online. These authors also discuss a potential problem with biased estimates if using the deviance measure and suggest using a chi-square test instead.

## Overview of chapter 12

We have seen how to represent and build segmented models as decision trees:

1. Decision trees
2. General decision tree models for segmentation
3. Automated decision tree models: CART
4. Logistic regression tree

# Consumer Credit Risk Modelling

## Chapter 13: Sampling and Testing

## Overview

What is the best way to use our data to sample and build our credit scoring models?

How should we test that the credit scorecard is performing well?

In this chapter we look at sampling and training. We investigate the following topics.

- Overfitting
- Training and Testing
- Forecasting
- Population drift
- Back-testing
- Champion/Challenger

## Under- and Overfitting

Scorecards are usually based on existing data.

We have already seen how to build a model given a data set of existing observations.

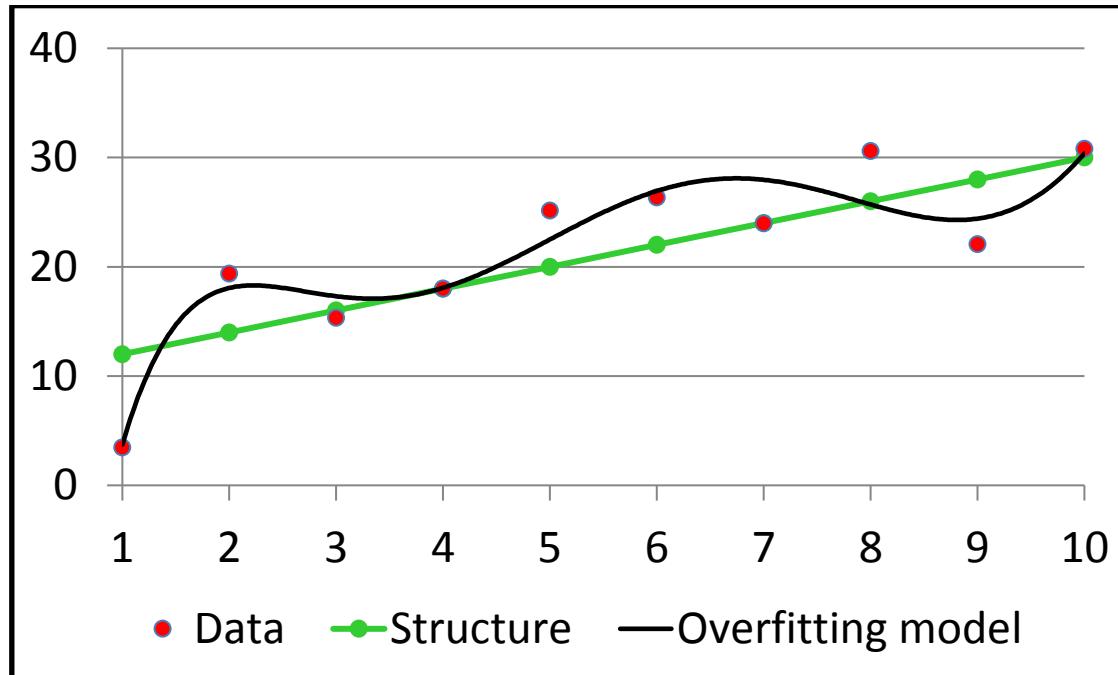
If the model does not explain the outcome well we have goodness of fit measures that will reveal this. We say that the model **underfits** the data.

However, it is also possible for a model to **overfit** the data. If we think of the data as including structure that we are interested in modelling, plus some random variation, then a model overfits if it fits the random variation along with the structure of interest.

Data components		
Fit	Structure	Random variation
Underfit	Unexplained	Unexplained
Fit is good	Explained	Unexplained
Overfit	Explained	Explained

### Example 13.1

This graph shows some simulated data with real value outcome.



- The x-axis represents the value of a predictor variable and the y-axis plots the outcome. The red points are data points.
- The green line shows genuine *structure* within data.
- The vertical distance between the data points and the green line represents *random variation*.
- The black line represents a model which is overfitting since it explains the random variation along with the structure.

## Overfitting

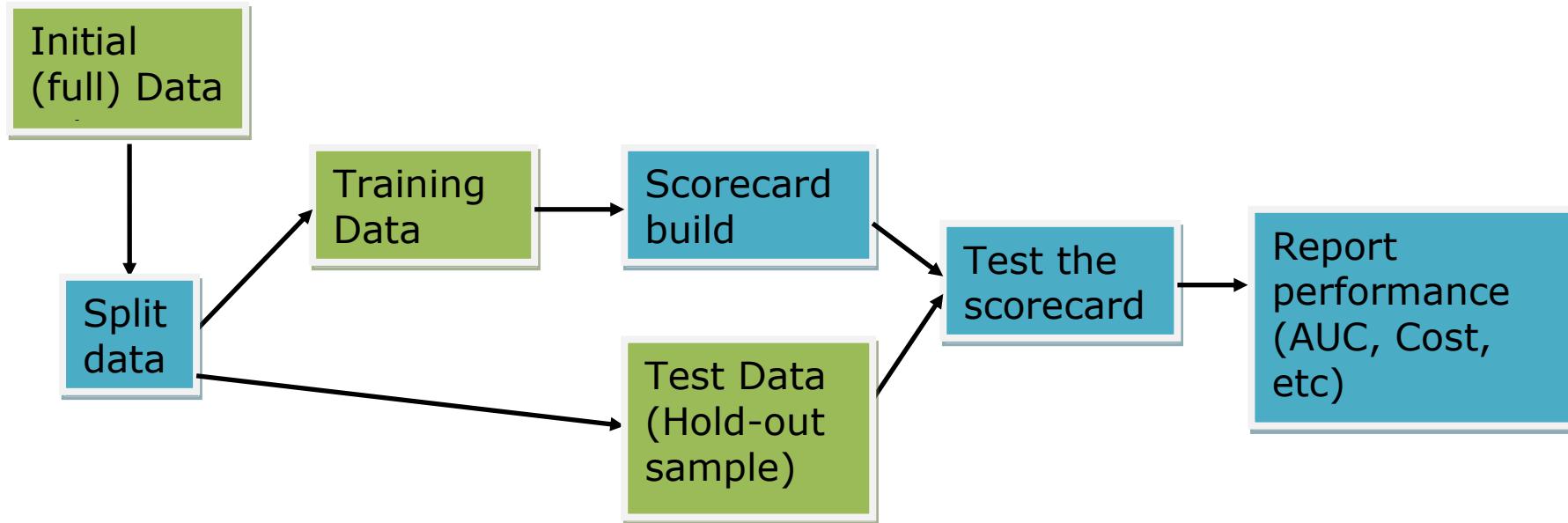
- The consequence of overfitting is that better performance is achieved on training data than could be expected on an independent test data set, ***on average***.
  - The “on average” here is important. Random variation in performance makes this vary for specific examples.
- The expected improvement in performance is referred to as the **optimism** of the model.
- The problem of overfitting worsens with smaller sample size.
- The problem of overfitting worsens with more complex scorecards, in terms of number of predictor variables included and model structure.
  - This is because more complex scorecards have more parametric freedom to fit random variation.

## Training and Testing Process

To avoid problems with overfitting, an independent data sample is used as a validation data set to test the model performance. This is called the **test data set** or **hold-out sample**.

Since the test data is independent of the training data, then it gives an *unbiased* estimate of performance.

We therefore consider the following training and testing process.



When we split the original, given data set, we need to ensure that we have sufficient observations to train a good model and also sufficient to get a good estimate of performance.

- This is not a problem if the initial data set is large with a good representation of events (eg default).
- If this is the case, typically, the initial data set is split randomly into a training data set comprising 2/3 of observations and a test data set comprising 1/3 of observations.

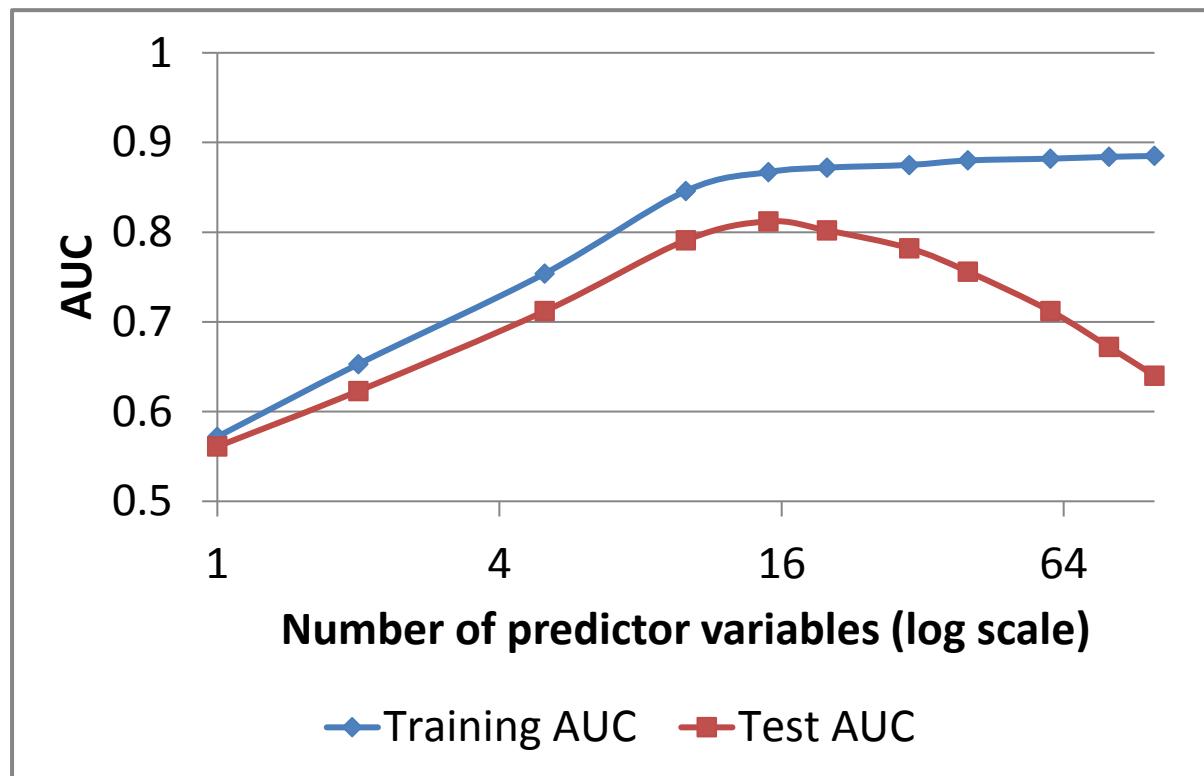
Because of overfitting, we usually find performance on the test data set is worse than on the training data.

- The extent of the difference gives an indication of the overfitting problem.

### Example 13.2

Several models are built with different numbers of predictor variables (up to 100 available) selected using Information Gain.

The training and test results are shown in the following graph as AUC measures for each of the models.



This shows monotonically increasing performance for training. However, performance on test data begins to degrade when too many parameters are included.

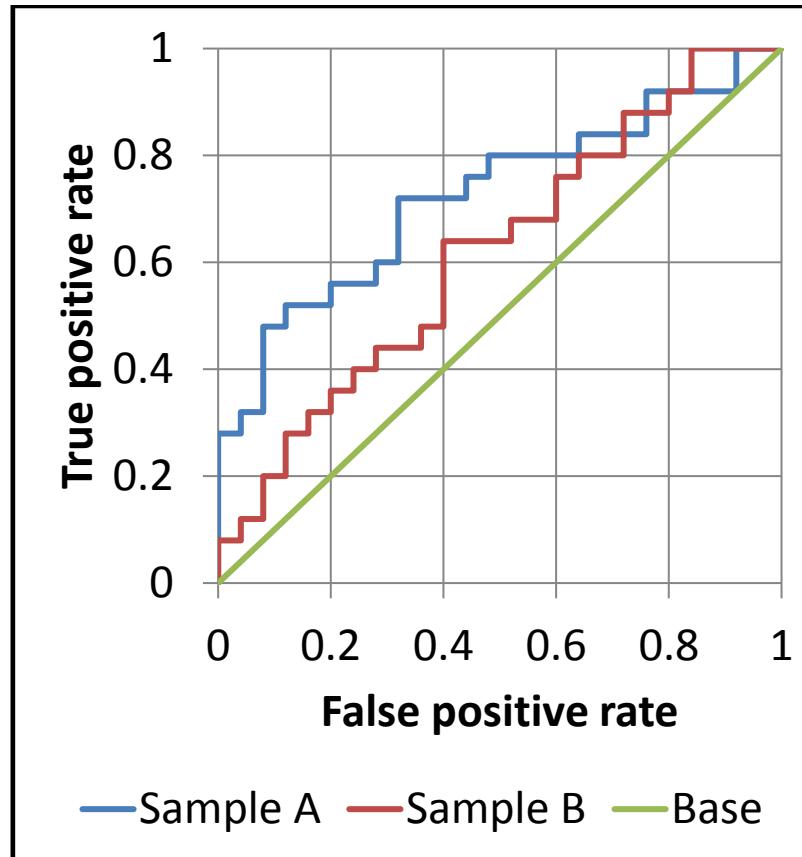
This is a demonstration of the effect of overfitting.

*Example 13.3*

A small sample A of 500 borrowers (400 goods and 100 bads) is used to build a model using logistic regression using variables for age, residency status, income and months in job.

A new (independent) sample B of 250 borrowers (210 goods and 40 bads) is now selected and used as a validation data set.

ROC curves for performance on the two samples are computed.



Performance on Sample A looks good and suggests a promising model. However, the performance for Sample B is poor, giving a ROC curve close to the diagonal.  
This suggests the model is not so good.

## Training, Testing and Sample Size

Sometimes the initial data set is too small to take a test sample:

- there would either be too few cases for training,
- or too few cases to get good estimates of performance from testing

This may occur in credit scoring: eg

- a new loan product, or
- a small segment of a set of loan accounts.

In such circumstances we need to use special testing procedures such as ***cross-validation*** or ***bootstrap*** testing.

- Both methods make efficient use of the data for training and testing.

## Forecasting

So far we have only considered test procedures that split a data set completely randomly. However, we usually want to use a scorecard to **forecast** for events in the future.

- An **application scorecard** is built on *past* application data, where outcome to default is known, but is required to forecast *future* default on new applications.
- A **behavioural scorecard** uses *past* behavioural data of current customers to predict *future* behaviour or default.

Therefore, if we want to test expected **operational** performance, it makes sense to take the training data from older data than the hold-out sample.

This becomes more important when we consider that the distributions of the credit data can change over time. This is known as **population drift**.

## Population Drift

This is when the distribution of the data changes over time.

Population drift can occur through gradual change or sudden change (eg economic crisis).

There are several reasons for population drift in retail credit.

1. It may be that the risk factors associated with the positive event change over time  
*(eg 20 years ago in the UK, not having a home telephone was a risk factor for default, but with the rise in use of mobile phones this is less the case).*
  
2. The character of the population may have changed  
*(eg perhaps more young people are applying for loans now than several years ago).*

3. The credit product characteristics may have changed  
*(eg several years ago, the lender may have charged higher interest rates).*
4. The lenders decision process may have changed  
*(eg the lender may be accepting less applicants today than in the past, to reduce risk).*
5. The general economic or social conditions at different times will affect behaviour of borrowers  
*(eg a sharp rise in bank interest rates will have put pressure on mortgage holders).*

## Formalizing Population Drift

- Each observation can be characterized by  $(\mathbf{X}, Y)$  and the time  $t$  at which it is observed.
- (Actually, there is a delay between origination and default, but we will simplify the issue for now to one observation time).
- So the population can be represented by the conditional density  $f(\mathbf{X}, Y|t)$ .
- If  $f(\mathbf{X}, Y|t) = f(\mathbf{X}, Y|s)$  for all  $(\mathbf{X}, Y)$  for any  $t < s$ , then there is no population drift between  $t$  and  $s$ .
- Population drift can be *measured* by the *divergence* between densities  $f(\cdot|t)$  and  $f(\cdot|s)$ , at least theoretically.
- Even if  $P(Y|\mathbf{X}, t) \approx P(Y|\mathbf{X}, s)$  for all  $(\mathbf{X}, Y)$ , population drift of the density of  $\mathbf{X}$  may still affect model development; ie the efficiency of coefficient estimates or regions of predictor space that are under-represented.

## Population Stability Index (PSI)

- As a special case of the divergence option, *information gain* can be used to monitor changes in categorical variables over time.
- This uses the information gain formula introduced in Chapter 7 but replacing outcome  $Y$  with change in time  $T$ :

$$PSI(V) = \sum_{j=1}^K \left( P(V = v_j | T = t) - P(V = v_j | T = s) \right) \log \left( \frac{P(V = v | T = t)}{P(V = v | T = s)} \right)$$

where  $P(V = v | T = t)$  is probability of value  $v$  at some time  $t$ .

- Plot PSI comparing development of distribution over time from some initial fixed time.

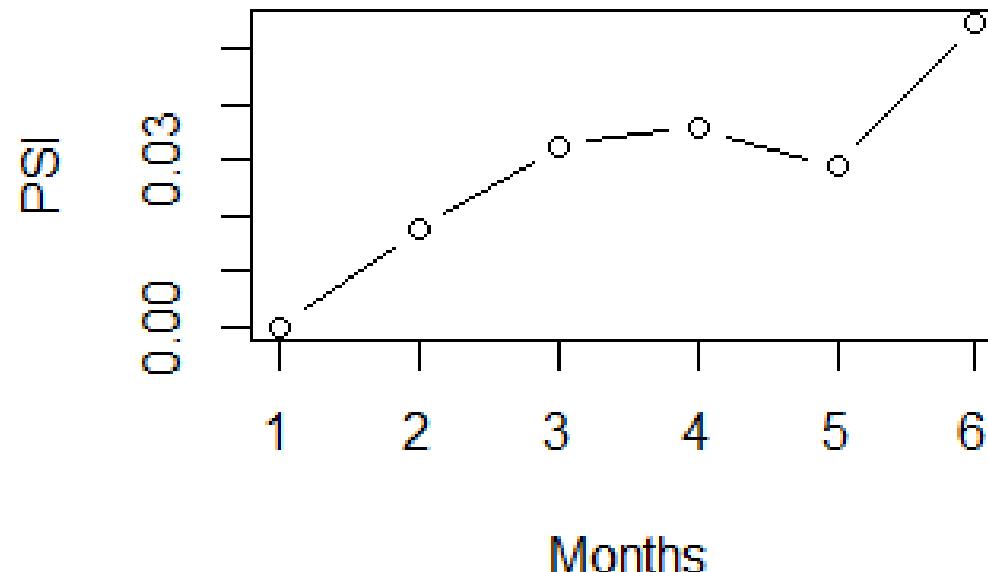
*Example 13.4*

This table gives change in distribution of risk grade over time for new loans.

Time (months)	Low risk	Medium risk	High risk
1	1267	3467	1456
2	1567	3567	1209
3	1734	3678	1123
4	1734	3823	1123
5	1767	4234	1234
6	1923	3722	1088

*Example 13.4 continued*

Now plot PSI, comparing time 1 against later months:



How to use this plot?

- Higher values express more risk from population drift.
- Rule of thumb: “ $\text{PSI} > 0.1$  indicates a need to re-evaluate the model”.

**Population drift implies that the performance of a scorecard will degrade over time.**

This means that:

1. Scorecards should be built on the most recent data.
2. Scorecards need to be rebuilt on a regular basis.
3. Forecasting is an important testing procedure to get accurate performance measures in light of population drift.

## Event window

We can no longer talk about an event happening. We have to give some time limitation, or window, for the measurement of the event.

For example, it is typical to take a default event to mean:

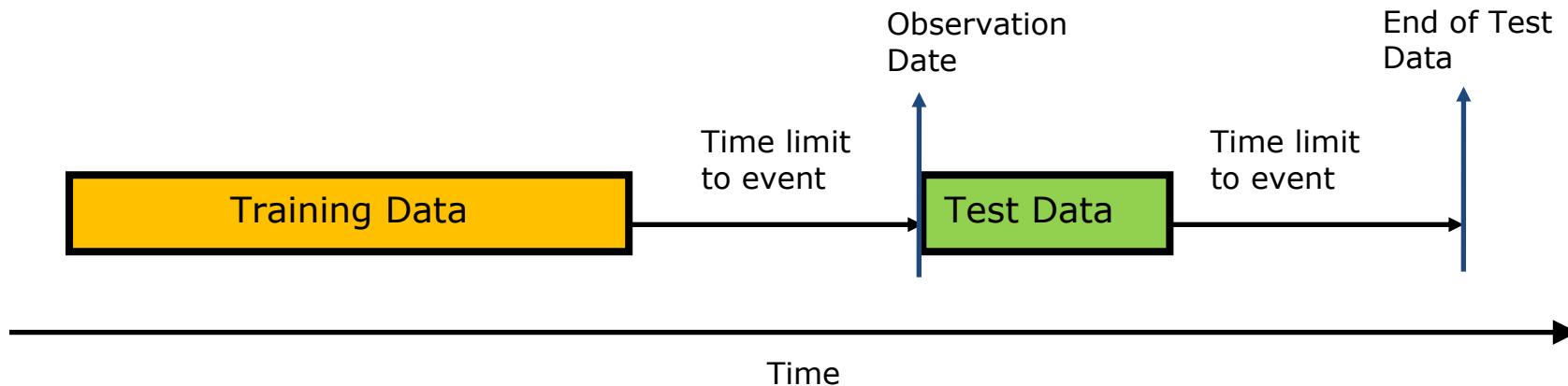
Borrower fails to make a repayment in three consecutive months  
***within a 12 month period.***

This is for two key reasons.

1. It ensures we measure the same event for each observation;
  - That is, borrowers will be on the lenders books for different periods of time (eg Bob may have had his credit card for four years, whereas Sue has had hers for only a year);
  - The probability of default is higher over longer periods, so defining the event over a fixed period ensures the probability is measuring the same thing.
2. It enables forecasting by setting a time limit to an event.

## Forecasting procedure

This leads to the following forecasting framework.



Note: blocks of data refer to availability of predictor variables  $X$ , whereas outcome  $Y$  can be measured up to "time limit to event".

- Only observations that begin after the observation date are included in the test data set.
  - For application scorecards, this is the application date.
- The observation date is controlled to allow for a sufficiently large test data set.
- Training data includes only observations that begin prior to the observation date.
- The longer the time limit to the occurrence of the event, the older the training data.

*Example 13.5*

Suppose you have credit card data running from January 2010 to December 2015 with 10,000 accounts opened in each year. You want to build a scorecard and use forecasting for model validation. You will model default within a 12 month period and require at least 15,000 observations for testing.

What is the best split of the data into training and test data sets?

### *Solution*

1. We start at the most recent date available in the entire sample, December 2015, and work back.
2. We require 12 months to measure default. Therefore the end date of the test data set must be December 2014 (ie accounts must have been open on or before that date).
3. 15,000 test observations requires 18 months of data so beginning of test data set must be July 2013.
4. We require 12 months to measure default for the training data set. Therefore the end date of the training data set must be June 2012.

Therefore:

- Training data set runs from January 2010 to June 2012.
- Test data set runs from July 2013 to December 2014.

## Back-Testing

The forecasting procedure given above is an “ideal” scenario. But the main disadvantage it has is that it pushes back the date range used for the training data set.

- The training data set could be two or three years old within this framework!
- Given the problem of population drift, this may make matters much worse: out-of-date data will lead to an out-of-date model.

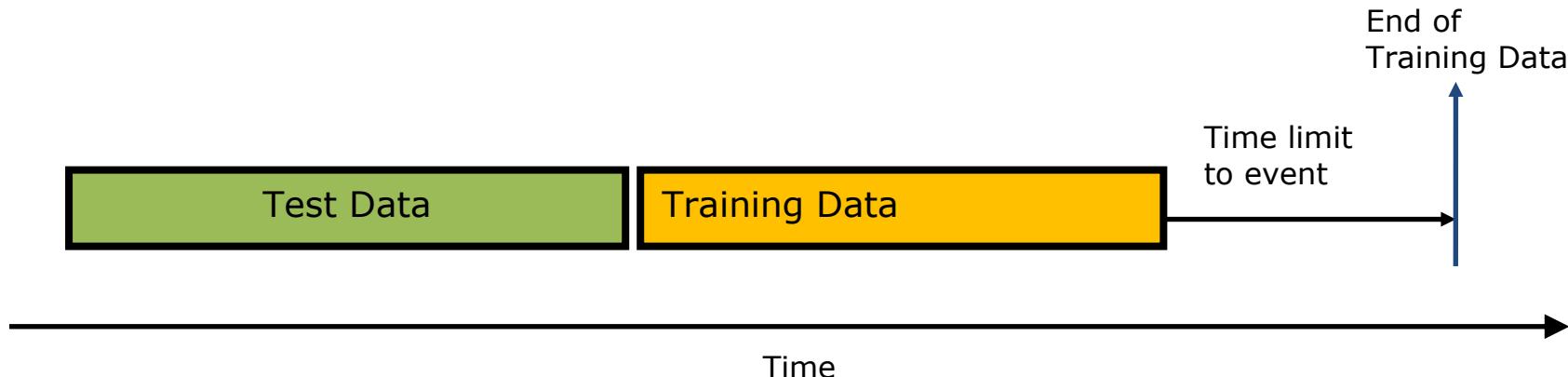
One solution is to loosen the constraints on the forecasting framework.

- For example, it may be beneficial to remove the time gap between training and test data.

However, an alternative solution which allows us to use the most recent data for modelling is **back-testing**.

Also known as **retrodiction** (predicting past events).

Back-testing uses the most recent data for training and old data for validation.



The issue of population drift for validation is dealt with since validation is on a different time frame.

- A problem is that testing on old data may not give a good indication of performance in the future.
- *Which is better?  
Training on recent data or measuring performance on recent data?*

## Champion and Challengers

*How do lenders decide when a model needs to be updated?*

For day-to-day business, the lender will have a key scorecard that they use for real decision making (such as application processing).

This scorecard is called the **Champion** and is usually the model that had the best predictive performance on historic data.

However, a series of alternative models build with somewhat different structure are also maintained and run in parallel with the Champion on the real data.

These alternative models are called **Challengers**.

The Challengers are not used to make business decisions; they are used for comparison with the Champion.

There are two main reasons for having Challengers:

1. To allow for alternative models of default: the Champion may be best on historic data, but over time, the assumptions built into the model structure may no longer hold (population drift) and so alternatives need to be on hand that are more appropriate, or inform us of changes.
2. If a Challenger is observed as giving good performance over a sustained period of time (eg 3 monthly cycles or more), then it is a candidate to replace the Champion.

The Challenger/Champion approach is good practice for keeping scorecards up to date.

## Overview of Chapter 13

We have considered the following sampling and testing issues.

- Overfitting
- Training and Testing
- Forecasting
- Population drift
- Back-testing
- Champion and Challengers

# Consumer Credit Risk Modelling

## Chapter 14: Selection bias and Reject inference

## Overview

Each step in lending process where applicants are selected or rejected generates a selection bias.

In particular we explore the two common problems of selection bias in application scorecard development.

- Reject bias
- Adverse selection

We will explore reject inference methods to help resolve these issues.

## Selection bias in the application process

If we consider scorecard development as a statistical study, it is an ***observational study***.

This is because as model builders we have little or no control of the sample and risk factors of interest.

The sample generates a potential bias whenever applicants are selected or drop out of the application process.

Recall from Chapter 3, there are three critical stages in the lending process where this happens.

1. Individuals' initial application for a financial product – this may be in response to marketing.
2. Lender's accept/reject decision.
3. Applicant's decision to accept/reject the loan (or credit card), if offered by the lender.

## Selection bias: introduction

Each of these decision points could lead to **selection bias** in the data set used to build and test the model.

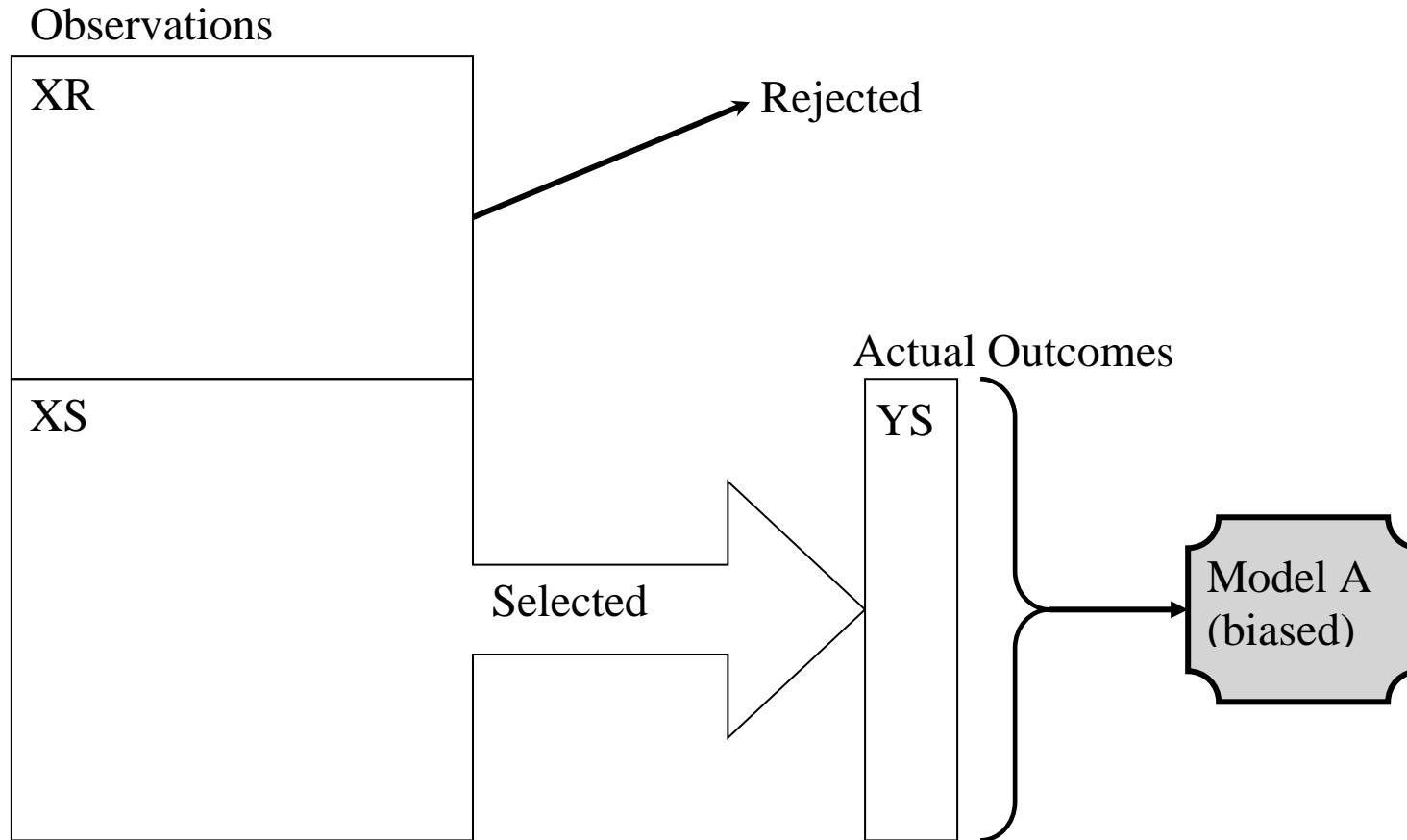
This is because an *application scorecard* is intended to be used to assist in the decision to accept/reject applicants at stage 2. However, the only data that is available about borrower behaviour is from those people who got beyond stage 3. Therefore the model can only be built on a sample which does not represent the whole population.

We will look at the selection bias effect at each stage of the process and consider remedies.

Note that this is not a problem for behavioural models since the sample reflects the whole population.

## Illustration of Selection bias

Note: Critically there are no outcomes for rejects (YR) in this illustration:



Notice that selection bias is a special case of the missing value problem (refer back to Chapter 8):

- Values of the response variable are missing.
- These are clearly non-ignorable (NI) missing values, since the non-response is dependent on the type of person they are and the rejection rule, based on the application data.
- Importantly, if the missing responses were MCAR then there would be no selection bias.

## 1. Marketing bias

If the process by which applicants are drawn to a financial product changes then this represents a potential bias. This occurs if the way the product is marketed changes considerably.

This is not a well-understood problem, and the extent of the problem is uncertain.

However, a risk manager will be aware of marketing strategies in the past and could adjust model build appropriately.

*Example 14.1*

If a bank targeted a credit card to young people over an extended period with extensive marketing, and then abruptly changed its strategy to target all age groups equally with little marketing, this is likely to affect a change in the sample distribution.

A model builder would then need to be wary about applying a model built on the older data on new applicants.

## 2. Lender selection bias

The rejection of applications by the lender is the most serious issue of selection bias for scorecard development, since it is the stage which reduces sample size the most.

- For forecasts, we want to have  $p(Y = 1|\mathbf{X} = \mathbf{x})$  for given characteristics  $\mathbf{x}$ .
- What we actually have from model build is  $p(Y = 1|A, \mathbf{X} = \mathbf{x})$  where  $A$  denotes acceptance of a past application, since only evidence of accepted applications is available.
- We do not know  $p(Y = 1|\mathcal{A}, \mathbf{X} = \mathbf{x})$  where  $\mathcal{A}$  denotes rejection (ie *not* accepted) of a past application (since they have not been given an account, it makes no sense to ask if they defaulted).
- However, we expect that calibration of scores may be somewhat too optimistic given the scorecard was only built on previously accepted applicants.

- Therefore, we expect

$$p(Y = 1|\bar{A}, \mathbf{X} = \mathbf{x}) > p(Y = 1|A, \mathbf{X} = \mathbf{x})$$

From which it follows that

$$\begin{aligned} p(Y = 1|\mathbf{X} = \mathbf{x}) &= p(Y = 1|\bar{A}, \mathbf{X} = \mathbf{x})p(\bar{A}|\mathbf{X} = \mathbf{x}) + p(Y = 1|A, \mathbf{X} = \mathbf{x})p(A|\mathbf{X} = \mathbf{x}) \\ &> p(Y = 1|A, \mathbf{X} = \mathbf{x}) \end{aligned}$$

## Reject inference

A range of techniques have been developed under the heading of **reject inference**.

- The basic idea is to ask,  
*if the rejected applicant had been given an account, would they have defaulted?*
- That is, reject inference methods attempt to estimate  $p(Y = 1|\bar{A}, \mathbf{X} = \mathbf{x})$ .

We will review three reject inference methods.

- Parcelling
- Augmentation
- Experimentation

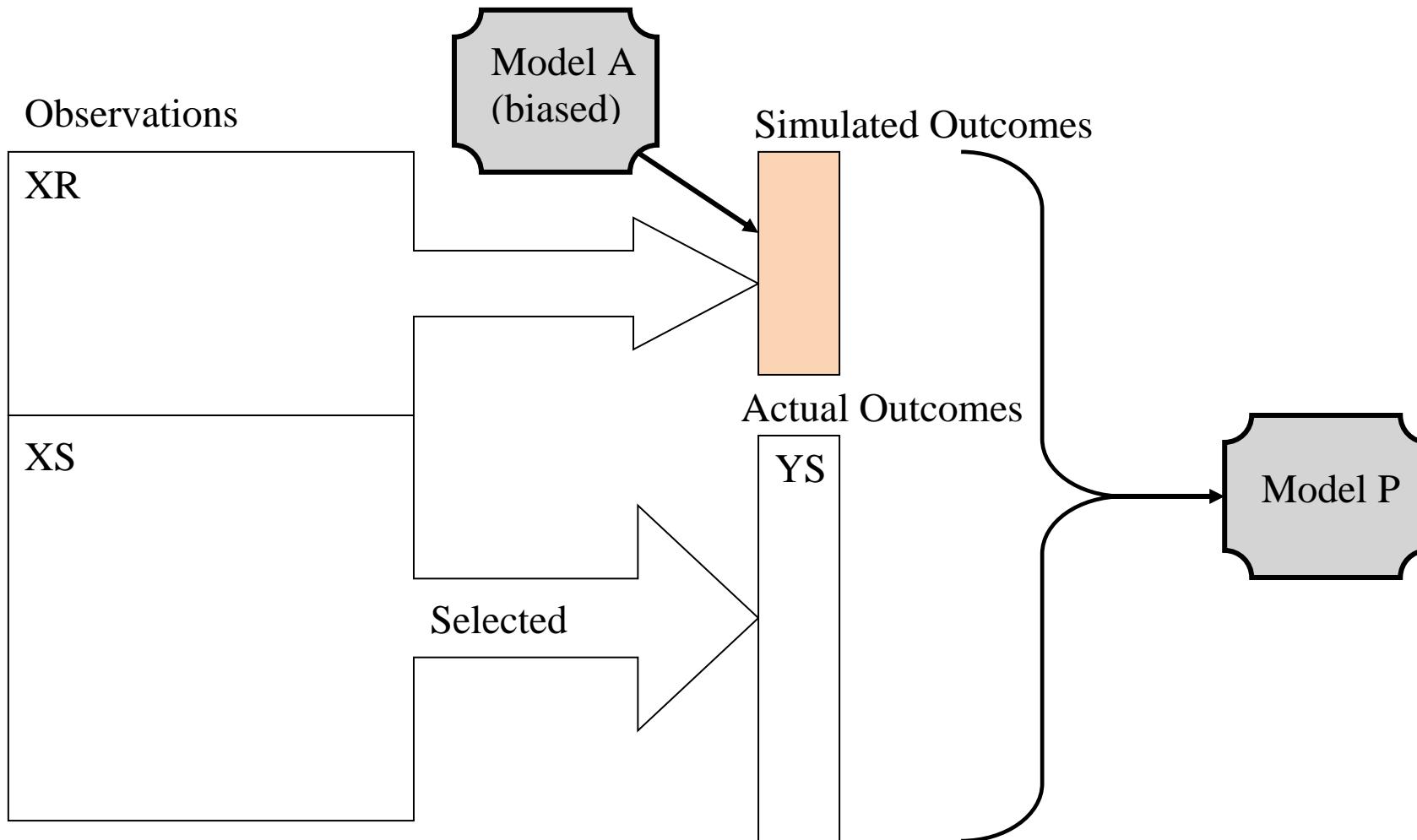
## Reject inference: Parcelling

Parcelling is an approach based on regression imputation (see Chapter 8), regressing responses where they are missing, based on a regression model built on the accepted applications.

It is a three-step approach which allows us to simulate values for the missing outcomes for rejected applications (YR).

1. Build a model on just the accepted applicants as usual to get the scorecard A.
2. *Simulate* outcomes for rejected applications .
3. Build a new model with accepted applicants and simulated outcomes for rejects to get scorecard P.

## Parcelling illustration



## Parcelling: How to simulate outcomes

There are several ways to simulate outcomes.

Typically, three approaches are used.

- Polarised: Use a score given by scorecard A to rank the rejected, and then assign those above and below a given cut-off probability to 0 and 1 respectively.
- Random: The assignment is done at random. Simulate a 0/1 outcome for each rejected applicant:
  - For each reject, generate a uniformly distributed random number  $r \in [0,1]$ .
  - Then a reject with  $\mathbf{X} = \mathbf{x}$ , computed with model A, is simulated with  $y = 1$  if and only if  $r < p(Y = 1|A, \mathbf{X} = \mathbf{x})$ .

- Fuzzy: Simulate 0/1 outcomes with probabilities:
  - Include each reject twice:
    - once with  $y = 1$  and weight  $p(Y = 1|A, \mathbf{X} = \mathbf{x})$ , and
    - once with  $y = 0$  and weight  $1 - p(Y = 1|A, \mathbf{X} = \mathbf{x})$ .
  - Real outcomes (from accepted applicants) are given weight 1.

Note that the overall weight for each applicant is always 1, but for rejects this weight is shared between the possible good and bad outcomes.

## Including weights on observations

What is meant by putting a weight on an observation?

It means that as part of the estimation process, an observation will have a weight on the estimate, relative to other observations.

*Example.*

Normally an observation will have a weight 1, but if it is given a weight of 2 then it will be treated as though it were two observations.

Weights are easily implemented in maximum likelihood estimation as given weight values on each observation.

Suppose there are  $n$  independent observations with log-likelihood function given by

$$l(\boldsymbol{\theta}; D_1, \dots, D_n) = \sum_{i=1}^n \log P(D_i | \boldsymbol{\theta})$$

Include given weights  $w_i$  on each observation  $i$  by generalizing this log-likelihood function to

$$l_w(\boldsymbol{\theta}; D_1, \dots, D_n) = \sum_{i=1}^n w_i \log P(D_i | \boldsymbol{\theta})$$

Most statistical packages allow the inclusion of weights (including the `glm` function in R).

## Reject inference: Augmentation

A popular approach is to adjust the distributions amongst the selected samples, by considering the distribution amongst the rejects. This is augmentation.

The idea is that categories of observations that are under-represented within the accepted applicants are given greater weight in the estimation. This would mean that the distribution of rejects is better represented during the estimation.

If  $p(A|\mathbf{x}_i \in C)$  is the probability an application is accepted given it is in some category  $C$ , then it is reweighted by

$$w_i = \frac{p(A)}{p(A|\mathbf{x}_i \in C)}$$

within the estimation. Note that the numerator is a constant to control the scale of weights (in particular,  $w_i = 1$  when  $p(A|\mathbf{x}_i \in C) = p(A)$ ).

*Example 14.2.*

Suppose a data set of 1000 applications has the following numbers within each employment category.

	Employed	Unemployed	Self-employed	Retired	Student	Total
Rejects	100 (25%)	200 (50%)	40 (10%)	20 (5%)	40 (10%)	400
Accepted	330 (55%)	30 (5%)	60 (10%)	60 (10%)	120 (20%)	600

It is clear that amongst the accepted applications, the unemployed are under-represented.

Overall,  $P(A) = 0.6$ .

Therefore, get the probability of being accepted within each category and re-weight observations in each category accordingly.

	Employed	Unemployed	Self-employed	Retired	Student
$p(A \mathbf{x}_i \in C)$	$\frac{330}{100 + 330}$	$\frac{30}{200 + 30}$	$\frac{60}{40 + 60}$	$\frac{60}{20 + 60}$	$\frac{120}{40 + 120}$
	=0.77	=0.13	=0.6	=0.75	=0.75
$w_i$	0.78	4.6	1	0.8	0.8

Therefore estimating with these weights will emphasize the unemployed observations within the accepted data set to compensate for their under-representation.

## Using the Accept/Reject model >

Example 14.2 illustrates the augmentation approach for a given age category. But why employment status? How do we decide which category to use?

We do not have to! We can build an AR model and base probability of accept on that.

In particular an AR model will give us the probability  $p(A|\mathbf{X} = \mathbf{x}_i)$  which is all we need. Then,

$$w_i = \frac{p(A)}{p(A|\mathbf{X} = \mathbf{x}_i)}.$$

Note, the AR model can be built using logistic regression just like the scorecard model.

## Derivation of augmentation method

Our goal is to adjust (re-weight) observations in our training data so that their distribution is closer to the unbiased population.

Hence our goal is to compute weights  $w_i$  such that the conditional density is transformable to the unconditional density:

$$f(\mathbf{x}_i) = w_i f(\mathbf{x}_i | A)$$

$$\Rightarrow f(\mathbf{x}_i) = w_i \frac{P(A|\mathbf{x}_i)f(\mathbf{x}_i)}{P(A)}$$

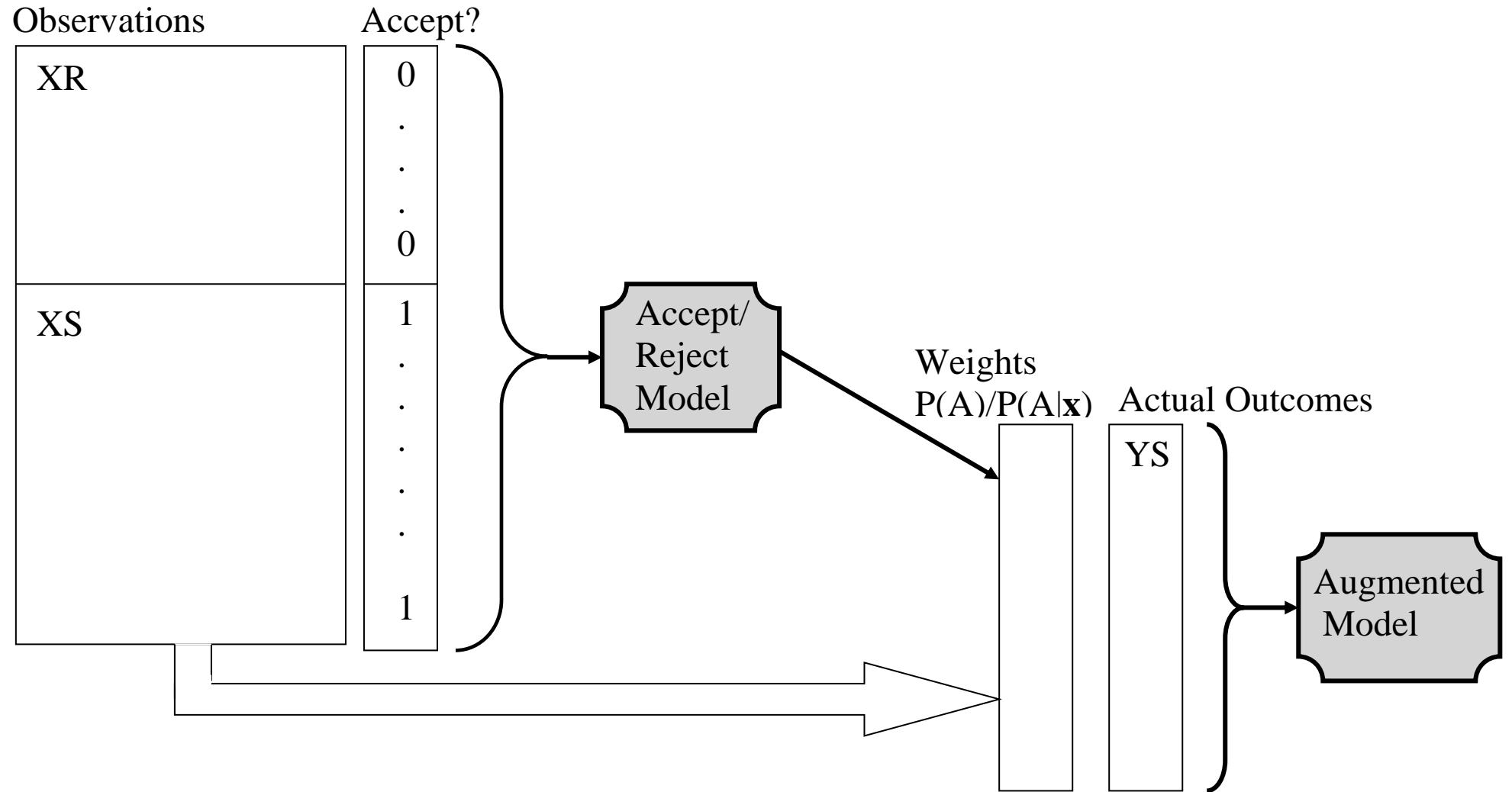
using Bayes' rule

$$\Rightarrow w_i = \frac{P(A)}{P(A|\mathbf{x}_i)}$$

(assuming  $f(\mathbf{x}_i) > 0$ )

which is the re-weighting rule for augmentation.

## Illustration of Augmentation using the AR model



## Example of augmentation

Do these methods of parcelling and augmentation actually work? There is some debate about this. One problem is that in the real world, they are not testable!

However, simulation studies show that *in some circumstances* (**but not all!**) they can be effective.

*Example 14.3.*

Simulate 50,000 applications with two predictor variables:

- income and
- ndel = number of delinquencies on record from other loans;

and, also, an outcome  $Y \in \{0,1\}$  is also generated for each application:  
1=default; 0=non-default.

Because we are simulating, we have YR (remember, in the real world, this is not available).

In our simulation, we arrange that a previous scorecard was used to reject 40% of them and accept 60% (it does not matter too much how this was done).

Then, a scorecard model S1 is built with just XS and YS, the accepted accounts, using logistic regression. The outcome event is non-default.

<i>Variable</i>	<i>Estimate</i>	<i>Std. err.</i>	<i>Z</i>	<i>P-value</i>
Intercept	-2.47	0.16	-16.2	<0.001
Income (log)	+1.57	0.05	29.9	<0.001
ndel	-0.32	0.05	-6.14	<0.001

How good an estimate is this model, against the whole data set?

Normally, *in the real world*, we could not answer this question. However, since this is a simulation, we have simulated outcomes for the rejects, YR, and so can build an unbiased model S2 using XR,YR and XS,YS.

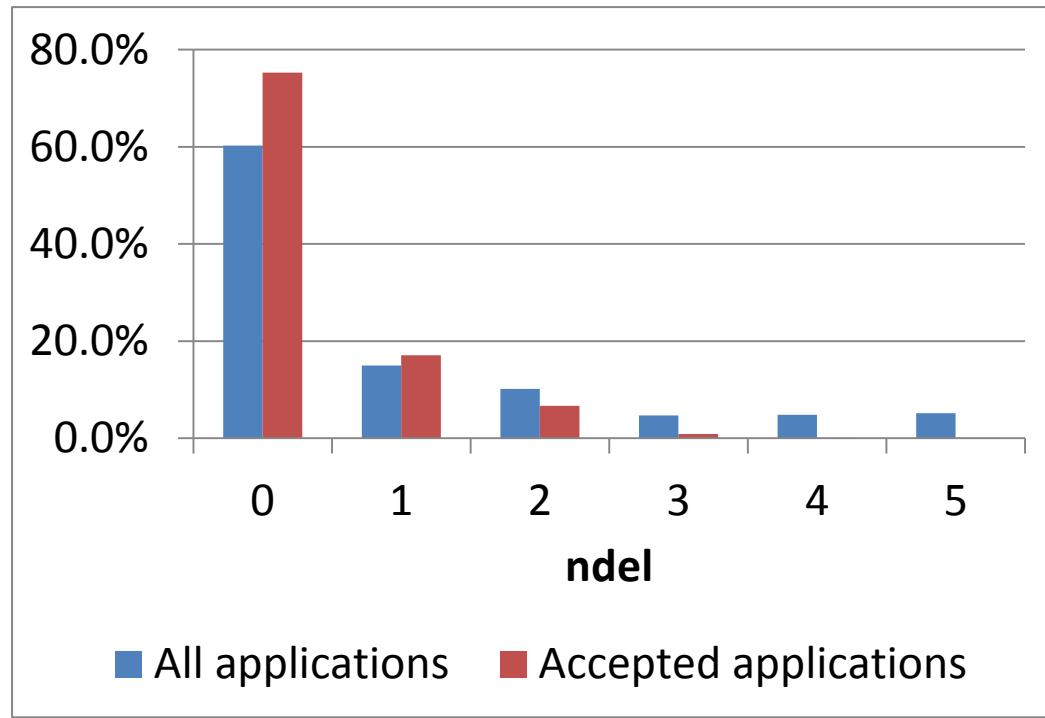
The result is:

Variable	Estimate	Std. err.	Z	P-value
Intercept	-2.13	0.046	-46.0	<0.001
Income (log)	+1.49	0.018	82.2	<0.001
ndel	-0.64	0.009	-74.6	<0.001

The results show that in S1, coefficient estimates on the intercept and Income (log) variable are good, but the estimate on the number of past delinquencies (ndel) variable is poor. The estimate is almost half what we should expect without bias.

*Why is this happening?*

This graph shows the distribution of ndel across all applicants and across accepts only.



It shows that those with high numbers of past delinquencies are under-represented.

Hence, the association of past delinquencies is underestimated in model S1.

This could have serious consequences, especially if past delinquency is generally higher in future populations of applicants.

How could we improve the model without “cheating” (ie by looking at YR)?

*Try augmentation*

Using augmentation weights based on this AR model has an effect on the coefficient estimates in the resulting model S3:

Variable		Std. err.	Z	P-value
Intercept	-2.36	0.07	-21.8	<0.001
Income (log)	+1.54	0.03	37.2	<0.001
ndel	-0.38	0.03	-10.6	<0.001

This has now increased the effect size of ndel (from -0.32), which is what we want.

Ultimately we are interested in predictions on an independent test set.  
*Can augmentation improve prediction results?*

To test this, another 25,000 applications are simulated with outcome, following the same procedure. Test results using each model are given below in AUC.

Model	Description	AUC
S1	Biassed model	0.828
S2	Benchmark unbiassed model	0.842
S3	Model with augmentation	0.833

- These results show that the biassed model underperforms the unbiassed model as we might expect.
- The model with augmentation goes some way to improving on the biassed model, but is not as good as the unbiassed model.

But remember, *in the real world*, we could never build S2.  
Only S1 and S3 are really possible.

*Additional note:*

Augmentation is related to the Weighting Method used in sample surveys to deal with nonresponse. For more reading see:

- Little RJA and Rubin DB (1987), *Statistical analysis with missing data*, Wiley, Chapter 4  
*(available in the library)*

## Reject inference: Experimentation

The methods for parcelling, augmentation, and other similar methods, are not ideal since they all make some assumptions about the rejects.

A statistically more rigorous approach is to run an experiment.

- In the extreme case, a lender would not reject any applicants.
  - Then clearly,  $p(Y = 1|X = \mathbf{x}) = p(Y = 1|A, X = \mathbf{x})$ .
  - Unfortunately, this would prove expensive and irresponsible since it would result in many bad debts.
- However, a lender can run a limited experiment by:
  - Either allowing some applicants through with a slightly lower score than the cut-off score;
  - Or, allowing a small proportion of those that would have been categorized as rejects through as accepts.

This approach allows the lender to monitor cases of those that were categorized as rejects and therefore estimate  $p(Y = 1|\bar{A}, s)$  on a small sample, whilst controlling the cost of this experiment.

- Indeed, a costs analysis can be conducted to determine the optimal proportion of “rejects” to allow through, in order to provide sufficient and useful information to improve the scorecard calibration.
- There is a serious ethical issue with this practice, since lenders would be deliberately providing loans to people they know are high risk. This would be viewed as highly irresponsible, even if it were done in a highly controlled way and may attract consumer complaints.

### 3. Applicant's accept/reject selection bias

Although the scorecard will tell us something about the riskiness of an applicant, the applicant will know *more* about their riskiness than the lender. Therefore there will be a discrepancy between the lender's assessment and the individual's own assessment. This is a case of ***asymmetric information***.

The bank will price a product according to its riskiness.

Therefore:

- A low risk loan will have a low interest rate.
- A high risk loan will have a high interest rate.

Economics suggests that:

- If an individual believes a lender has underestimated their riskiness, and therefore has offered a lower interest rate, they will be more attracted to the loan.
- Conversely, if an individual believes a lender has overestimated their riskiness, and therefore has offered a higher interest rate, they are less likely to take the loan.

This is an example of **Gresham's Law**.

The consequence is that there is an **adverse selection** of applicants at take-up of the loan. That is:

*Those who take the offer of a loan are likely to be a higher risk than the scorecard may suggest.*

This is a very similar problem to reject inference, since we want to estimate  $p(Y = 1|s)$ , but what we have is  $p(Y = 1|T, r, s)$  where  $r$  is the offered interest rate and  $T$  is the event that the individual takes the offer of a loan.

Similar solutions exist as with reject inference.

In particular, experimentation can be used to estimate  $p(Y = 1|\bar{T}, r, s)$ .

How can this be done, given this selection bias is based on the applicants decision, not the lenders?

Although the lender cannot directly control whether an individual takes a loan, they can *influence* the decision by offering preferential interest rates (less than they would have offered based just on the scorecard) or other loan features.

## Review of Chapter 14



Each step in lending process where applicants are selected or rejected generates a **selection bias**.

In particular we looked at the two common problems of selection bias in application scorecard development and considered possible solutions.

- Reject bias
- Adverse selection

We developed a family of methods to help resolve these problems in reject inference.

# Consumer Credit Risk Modelling

## Chapter 15: Scorecard Performance, Part 2

## Overview

We have already looked at error rates and ROC curves.

We now continue to look at alternative measures of discrimination commonly used in the credit scoring industry and consider the issue of model comparison.

In this chapter we will cover:

- Gini coefficient
- Cumulative accuracy profile (CAP)
- Kolmogorov-Smirnov statistic

We also consider links to the AUC measure that has already been introduced in Chapter 7.

## The Gini Coefficient

The Gini Coefficient is a measure of discrimination, developed independently to the ROC curve, but related as:

$$\text{Gini} = 2 \times \text{AUC} - 1.$$

It is sometimes preferred, since it gives a value 1 for perfect discrimination and 0 for models no better than random.

## The cumulative accuracy profile (CAP)

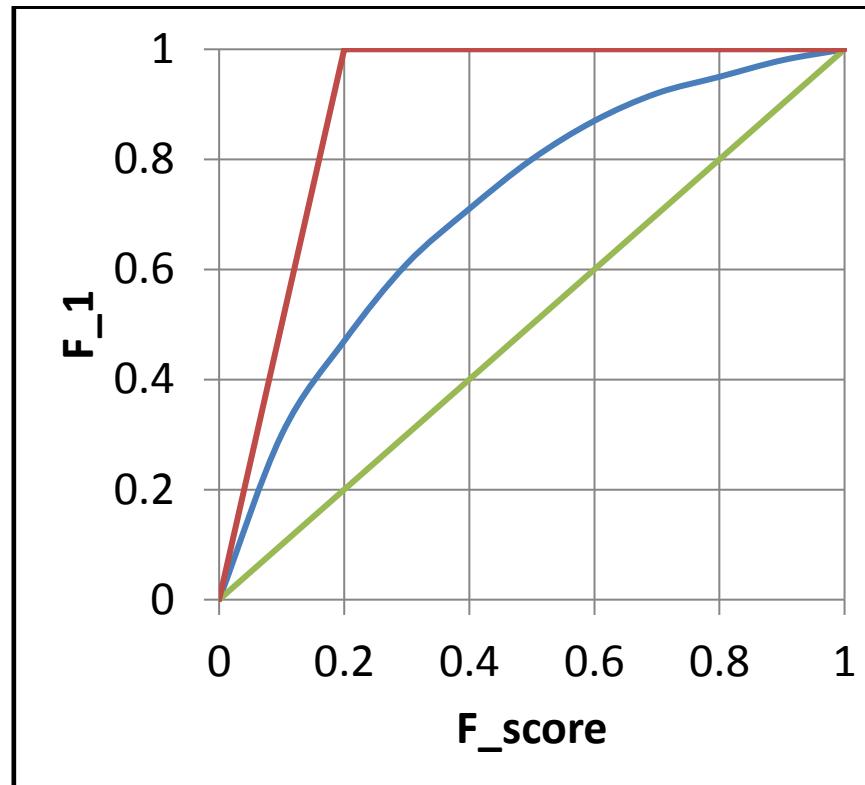
The cumulative accuracy profile (CAP) is widely used in marketing and banking. It is also known as a *lift curve*.

They are similar to ROC curves except they plot  $F_1$  against  $F_{\text{score}}$  where  $F_{\text{score}}(c) = P(S \leq c)$  is the CDF of the scores.

It follows that  $F_{\text{score}}(c) = F_0(c)(1 - p_1) + F_1(c)p_1$ . Therefore,

- The least discriminative model is such that  $F_{\text{score}}(c) = F_0(c) = F_1(c)$  which represents the diagonal  $(0,0)$  to  $(1,1)$  on the CAP.
- The most discriminative model, with no errors,  $F_0(c) = 0$  and  $1 - F_1(c) = 0$  is the line passing through  $(p_1, 1)$ .
- Notice that, as with ROC, all CAPs must go through  $(0,0)$  and  $(1,1)$ .

An example of the CAP is given below.



The blue line shows the CAP.

The green line is the model with no discriminative power.

The red line shows a model with perfect discriminatory power (where  $p_1=0.2$ ).

## Accuracy Rate

As with the ROC curve, a measure of discrimination is given using the area of the CAP.

Specifically, the accuracy rate (AR) is defined as the area between the CAP and the undiscriminating diagonal divided by the area between the perfect model and the undiscriminating diagonal. That is,

$$AR \triangleq \frac{\int_c F_1(c)F_{\text{score}}'(c)dc - \frac{1}{2}}{\frac{1}{2}(1 - p_1)}$$

AR is related to AUC as follows:

Where  $A$  is AUC,

$$\begin{aligned} \text{AR} &= \frac{2}{(1-p_1)} \left[ \int_c F_1(c) (F_0'(c)(1-p_1) + F_1'(c)p_1) dc - \frac{1}{2} \right] \\ &= 2A + \frac{2}{(1-p_1)} \left[ p_1 \int_c F_1(c) F_1'(c) dc - \frac{1}{2} \right] \\ &= 2A + \frac{2}{(1-p_1)} \left[ \frac{1}{2}p_1 - \frac{1}{2} \right] = 2A - 1 \end{aligned}$$

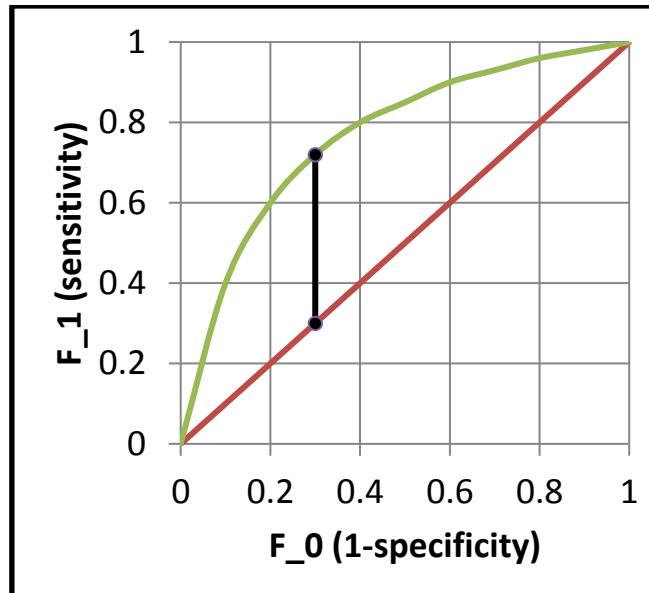
That is, AR is *exactly* the same measure as Gini.

## Kolmogorov-Smirnoff statistic

The Kolmogorov-Smirnoff (KS) statistic is another measure of how different the distributions of scores are for positives and negatives.

$$KS = \max_c |F_1(c) - F_0(c)|$$

This is equivalent to the maximum length of the vertical line between the ROC curve and the base line, shown as the black line below.



- Note that KS is the basis of a statistical test to show that  $F_0$  and  $F_1$  are different. However, this use does not concern us here.

## Review of Chapter 15

We have considered several different measures of classification performance:

- Gini coefficient
- Cumulative accuracy profile (CAP)
- Kolmogorov-Smirnoff statistic

# Consumer Credit Risk Modelling

## Chapter 16: Probability calibration

## Overview

In this chapter we consider how to assess probability estimates given by our credit scoring models.

We cover the following topics:

- Probability calibration
- Hosmer-Lemeshow test
- Brier Score

## Assessing probability estimates

We have seen that credit scores can be used to generate probability estimates.

In particular, we have used logistic regression to build a model that provides probability estimates using the log-odds link function.

It is important for us to measure how well a model estimates probabilities.

*This is a different question to the problem of measuring discriminatory power, which we looked at in the previous chapter.*

- Discrimination just relies on rank-ordering.
- However, assessment of probability estimates requires something more subtle.

Very roughly, if an event is *predicted* with a certain probability, then we expect the proportion of *observed* occurrences with that event to be approximately the same probability.

For instance, if I predict it will rain on 25% of days, then if I observe 100 days, I would expect to see it rain on about 25 of those days. The more the observed frequency deviates from the prediction, the more unreliable the probability estimator is.

### ***How can a predictor be wrong?***

- If the model is misspecified in any way then this can cause a bias in probability estimation.
- One way this can happen, eg, is if the logistic distribution does not accurately model the true probability distribution.
- A second way is that the model may degrade over time due to population drift in the observations over time.

## Approach to probability calibration

The problem is that, in general, a model will assign a specific probability to each observation and it is not possible to determine (empirically) if a probability is correct on a single case.

The solution is to group observations together and to assess their estimates together.

- An obvious way is to take all observations within the validation data set.
- However, that will not reveal any subtlety in the probability estimates.
- It is better to split the observations into several groups and assess probability estimates within each group.
- This is usually done by grouping observations by risk grade. This will mean observations in the same group will have similar probabilities.

## Probability calibration and grouping observations

A typical approach is to bin the observations into several discrete groups and determine the difference between the ***estimated probability*** and the ***observed probability*** of an event within each group.

This will then allow us to construct statistical tests and confidence intervals for the probability estimates.

Consider  $G$  groups expressed as sets of observations,  $C_1$  to  $C_G$ , such that for any observation with predictor variables  $\mathbf{X}$ ,

- $\mathbf{X} \in C_j$  for some group  $j$ ; and
- $C_j \cap C_k = \emptyset$  for any two groups  $j \neq k$ .

## Grouping by Risk Grade

It is usual to group by risk grade for probability calibration.

Risk grades are specified using scores from a scorecard  $s(\mathbf{X})$  which returns a score given an observation  $\mathbf{X}$ .

Let each risk grade  $j$  be specified by an interval  $(g_j, g_{j+1}]$  on the scores.

To ensure that the full range of scores are included, we set  $g_1 = -\infty$  and  $g_{G+1} = +\infty$ .

Then use the following risk group

$$C_j = \{\mathbf{X}: g_j < s(\mathbf{X}) \leq g_{j+1}\}.$$

## Probability calibration

Then the probability of the event conditional on an observation being from group  $j$  is given by

$$\begin{aligned}
 P_{1j} \triangleq P(Y = 1 | \mathbf{X} \in C_j) &= \frac{P(Y = 1, \mathbf{X} \in C_j)}{P(\mathbf{X} \in C_j)} \\
 &= \frac{\int_{\mathbf{x}} f(Y = 1, \mathbf{x} \in C_j, \mathbf{X} = \mathbf{x}) d\mathbf{x}}{P(\mathbf{X} \in C_j)} \\
 &= \frac{\int_{\mathbf{x}} P(Y = 1 | \mathbf{X} = \mathbf{x}) I(\mathbf{x} \in C_j) f(\mathbf{X} = \mathbf{x}) d\mathbf{x}}{P(\mathbf{X} \in C_j)} \\
 &= \frac{E_{\mathbf{x}}(P(Y = 1 | \mathbf{X} = \mathbf{x}) I(\mathbf{x} \in C_j))}{P(\mathbf{X} \in C_j)}
 \end{aligned}$$

where  $I(\cdot)$  is the indicator function.

Note that the integral is a multiple integral. If one of the variables in  $\mathbf{X}$  is discrete, then the integral should be a sum. However, the same result follows. (*Exercise: Check this*).

## Empirical probability estimates

Given a validation data set of  $n$  observations  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ ,  $P_{1j}$  is approximated using the sample expected value and empirical frequency for numerator and denominator respectively:

$$\begin{aligned}\tilde{P}_{1j} &= \frac{\frac{1}{n} \sum_{i=1}^n P(Y = 1 | \mathbf{X} = \mathbf{x}_i) I(\mathbf{x}_i \in C_j)}{n_j/n} \\ &= \frac{1}{n_j} \sum_{i=1}^n P(Y = 1 | \mathbf{X} = \mathbf{x}_i) I(\mathbf{x}_i \in C_j)\end{aligned}$$

where  $n_j$  is the number of observations in group  $j$ :

$$n_j = |\{i : i \in \{1, \dots, n\}, \mathbf{x}_i \in C_j\}|.$$

This estimate can be computed using a model that gives  $P(Y = 1 | \mathbf{X} = \mathbf{x})$ , eg logistic regression.

## Observed probabilities

Similarly, if we know the outcome for each observation  $y_1, y_2, \dots, y_n$ , then the corresponding **observed probability** for group  $j$  is:

$$Q_{1j} \triangleq \frac{1}{n_j} \sum_{i=1}^n y_i I(\mathbf{x}_i \in C_j)$$

## Probability calibration graph

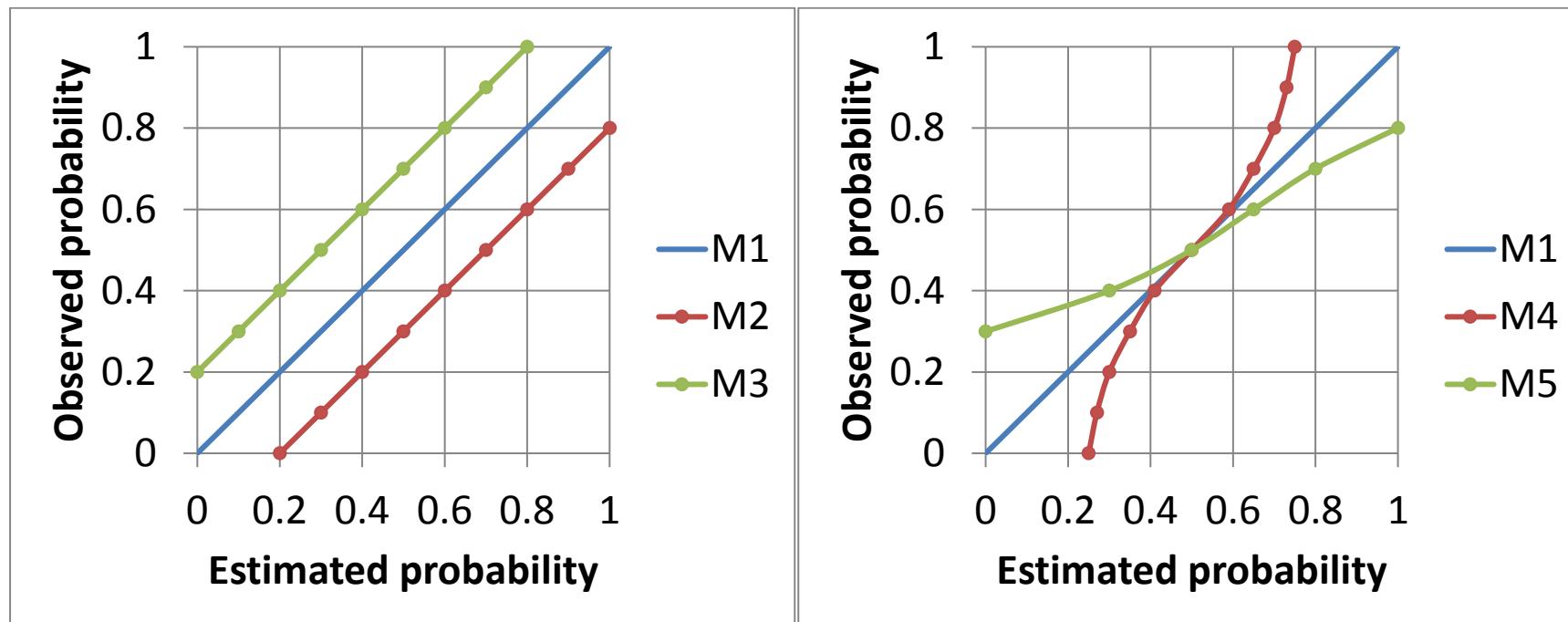
We can compare expected probabilities with observed probabilities graphically on a probability calibration graph:

$\tilde{P}_{1j}$  is plotted on the x-axis against  $Q_{1j}$  on the y-axis for each  $j \in \{1, \dots, G\}$ .

The scorecard is **well-calibrated** if estimated probabilities are similar to observed frequencies. On the probability calibration graph, this means points sit close to the diagonal from  $(0,0)$  to  $(1,1)$ .

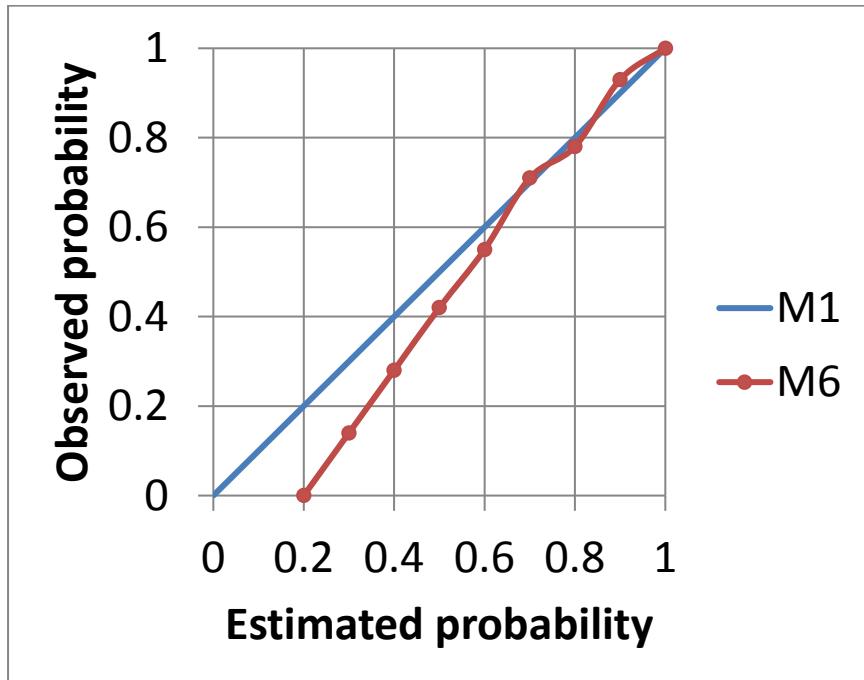
*Example 16.1*

Interpret each of the six models M1 to M6 shown in the following probability calibration graphs.



*... continued ...*

*Example 16.1 continued*



*Solution*

- M1. Perfect calibration of estimated probability with observation.
- M2. Consistently overestimates probability.
- M3. Consistently underestimates probability.
- M4. Generally fine for estimating mid-range probabilities (around 0.5) but underestimates extreme probabilities  
(ie estimates are too conservative).
- M5. Produces too many extreme probability estimates.
- M6. Good probability calibration for high risk cases (ie for probability>0.7) but overestimates probability for low risk cases.

## Probability recalibration

- If a model is producing scores that are not well-calibrated, but we do not want to update the model, then it is possible to recalibrate the scores.
- This can be achieved by a simple logistic regression:
  1. Suppose we have a training/validation data set of observations  $(\mathbf{x}_i, y_i)$  for  $i \in \{1, \dots, n\}$  for which the model produces scores  $s(\mathbf{x}_i)$ .
  2. Build logistic regression with a single predictor variable equal to  $s(\mathbf{x}_i)$  and outcome variable  $y_i$ .
  3. This will generate a new score, linear on the first:
$$s'(\mathbf{x}_i) = \beta_0 + \beta_1 s(\mathbf{x}_i)$$
- It may be that the recalibration requires a non-linear transformation of the original model scores, in which case non-linear (eg polynomial) terms can be included in the recalibration model.

### Example 16.2

Suppose we have a validation data set of 16 borrowers.

This table shows log-odds scores and estimated probabilities from a model (1) along with outcomes ( $Y = 1$  means default).

$s(\mathbf{x}_i)$	$P(Y = 1 \mathbf{x}_i)$	$y_i$	$s(\mathbf{x}_i)$	$P(Y = 1 \mathbf{x}_i)$	$y_i$
0.696	0.333	1	3.648	0.025	0
0.824	0.305	0	3.696	0.024	0
1.784	0.144	1	4.096	0.016	0
1.8	0.142	0	4.176	0.015	0
2.008	0.118	1	4.784	0.008	1
2.456	0.079	1	5.16	0.006	0
2.832	0.056	0	5.616	0.004	0
3.384	0.033	0	6.288	0.002	0

Given the following risk grades, calculate estimated probability  $\tilde{P}_{1j}$  and observed probabilities  $Q_{1j}$  within each grade  $j$ .

Grade C: Scores less than 2.2.

Grade B: Scores between 2.2 and 4.

Grade A: Scores greater than 4.

*Solution*

Compute mean probability for model (1) for each risk grade:

Grade	Probability estimate $\tilde{P}_{1j}$	Outcome frequency $Q_{1j}$
C	0.208	3/5
B	0.043	1/5
A	0.0085	1/6

*Example 16.2 (part ii):* Now, consider log-odds scores and estimated probabilities from a new model (2).

$s'(\mathbf{x}_i)$	$P(Y = 1 \mathbf{x}_i)$	$y_i$	$s'(\mathbf{x}_i)$	$P(Y = 1 \mathbf{x}_i)$	$y_i$
-0.735	0.676	1	1.11	0.248	0
-0.655	0.658	0	1.14	0.242	0
-0.055	0.514	1	1.39	0.199	0
-0.045	0.511	0	1.44	0.192	0
0.085	0.479	1	1.82	0.139	1
0.365	0.410	1	2.055	0.114	0
0.6	0.354	0	2.34	0.088	0
0.945	0.280	0	2.76	0.060	0

Notice that the scores are just a linear rescaling of those of model (1).

This is a score **recalibration**:  $s'(\mathbf{x}_i) = 0.625s(\mathbf{x}_i) - 1.17$ .

The following risk grades will select the same observations into each grade as the one used for model (1).

Grade C: Scores less than 0.205.

Grade B: Scores between 0.205 and 1.33.

Grade A: Scores greater than 1.33.

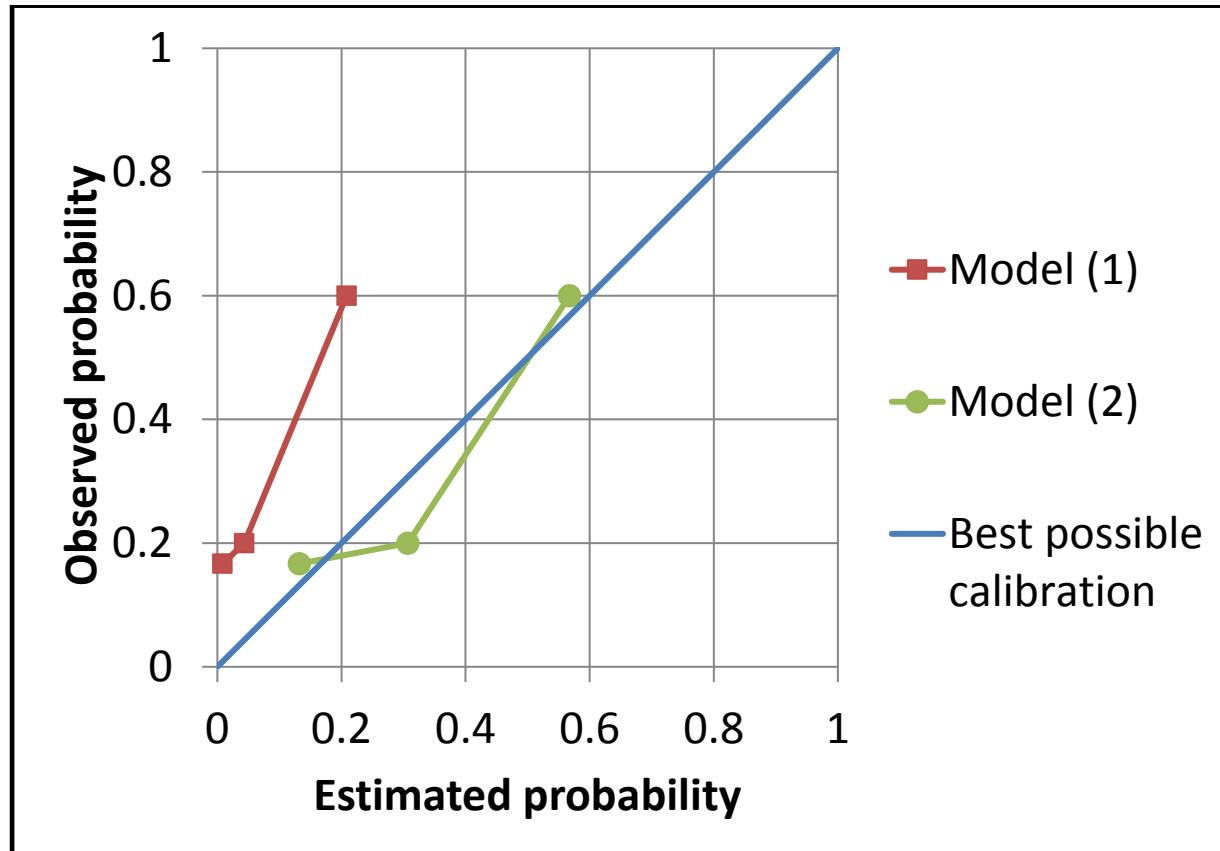
Calculate estimated probability  $\tilde{P}_{1j}$  and observed probabilities  $Q_{1j}$  within each grade  $j$  for model (2). Then plot the probability calibration graphs for both models and compare.

*Solution*

1. Compute mean probability for model (2) for each risk grade:

Grade	Probability estimate $\tilde{P}_{1j}$	Outcome frequency $Q_{1j}$
C	0.568	3/5
B	0.307	1/5
A	0.132	1/6

## 2. Plot probability calibration graph:



## 3. Interpretation:

Model (1) underestimates probabilities of default, whilst model (2) gives much better estimates.

## Hosmer-Lemeshow Test

It is useful to have a test of probability calibration across all risk grades. We can use the Hosmer-Lemeshow Test to do this.

The Hosmer-Lemeshow Test is a form of Chi-square test. The null hypothesis is that the observed probabilities are not different from the estimated probabilities and the alternative hypothesis is that there is a difference.

- Null hypothesis:  $H_0: P_{1j} = Q_{1j}$  for all  $j \in \{1, \dots, G\}$ .
- Alternative hypothesis:  $H_1: P_{1j} \neq Q_{1j}$  for any  $j \in \{1, \dots, G\}$ .

## Chi-square Test of Association

Recall that the Chi-square test is based on observed and expected frequencies,  $O_i$  and  $E_i$  respectively, falling into  $N$  groups,  $\in \{1, \dots, N\}$ . It tests the null hypothesis against alternative hypothesis:

- Null hypothesis:  $H_0: E_i = O_i$  for all  $i \in \{1, \dots, N\}$ .
- Alternative hypothesis:  $H_1: E_i \neq O_i$  for any  $i \in \{1, \dots, N\}$ .

The chi-square statistic is calculated as

$$\chi^2 = \sum_{i=1}^N \frac{(E_i - O_i)^2}{E_i}$$

And, under the null hypothesis, this follows a chi-square distribution with  $N - p$  degrees of freedom, where  $p$  is the reduction in degrees of freedom within the groups (usually 1).

The chi-square test is used to test probability calibration by testing within each group how many expected goods were really good ( $Y = 0$ ) and how many expected bards were really bad ( $Y = 1$ ). As such we consider two sets of expectations/observations: the goods and the bards.

The following frequencies are used for each risk grade  $j$ :

- Number of expected goods =  $n_j(1 - \tilde{P}_{1j})$
- Number of observed goods =  $n_j(1 - \tilde{Q}_{1j})$
- Number of expected bards =  $n_j\tilde{P}_{1j}$
- Number of observed bards =  $n_j\tilde{Q}_{1j}$

Then the chi-square statistic is:

$$\begin{aligned}
 \chi^2 &= \sum_{j=1}^G \frac{(n_j(1 - \tilde{P}_{1j}) - n_j(1 - \tilde{Q}_{1j}))^2}{n_j(1 - \tilde{P}_{1j})} + \sum_{j=1}^G \frac{(n_j\tilde{P}_{1j} - n_j\tilde{Q}_{1j})^2}{n_j\tilde{P}_{1j}} \\
 &= \sum_{j=1}^G \frac{n_j [\tilde{P}_{1j}(\tilde{P}_{1j} - \tilde{Q}_{1j})^2 + (1 - \tilde{P}_{1j})(\tilde{P}_{1j} - \tilde{Q}_{1j})^2]}{\tilde{P}_{1j}(1 - \tilde{P}_{1j})} \\
 &= \sum_{j=1}^G \frac{n_j(\tilde{P}_{1j} - \tilde{Q}_{1j})^2}{\tilde{P}_{1j}(1 - \tilde{P}_{1j})}
 \end{aligned}$$

Since  $N = 2G$ , the number of degrees of freedom is  $2G - p$ , but what is  $p$ ? The series of good/bad observations are highly dependent, so  $p$  will be large. Simulation studies have shown that an optimal value is given by

Degrees of freedom =  $G - 2$ .

### Example 16.3

Use the grading system and probabilities from Example 16.2 to conduct a Hosmer-Lemeshow Test for each model.

			Model (1)		Model (2)	
Grade	$n_j$	$\tilde{Q}_{1j}$	$\tilde{P}_{1j}$	$\chi^2$	$\tilde{P}_{1j}$	$\chi^2$
A	6	1/6	0.0085	17.8	0.132	0.06
B	5	1/5	0.0434	3.0	0.307	0.27
C	5	3/5	0.2083	4.7	0.568	0.02
Sum				25.4		0.35
P-value *				<0.001		0.55

\* Chi-square tests are at 1 degree of freedom.

- These results suggest that model (1) does not calibrate observations well (null hypothesis is rejected at the 1% significance level).
- However, the null hypothesis is not rejected for model (2). Hence, the observations could have been generated by the estimated probabilities.

Further information can be found about the Hosmer-Lemeshow test, especially in relation to logistic regression, in

Dobson AJ and Barnett AG (2008). An Introduction to Generalized Linear Models (CRC press), pp.135-137  
*(available in the library).*

## Brier Score

The methods we have used so far rely on dividing the validation data set into risk grades and so, to some extent, the results depend on *how* the data set is divided.

The Brier Score is a measurement of the skill of a probability forecaster. It does not rely on dividing the scores into risk grades. It compares probability of an event against observed outcome as a mean square error (MSE).

If we have  $n$  observations  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  and corresponding outcomes  $y_1, y_2, \dots, y_n$ , then the Brier Score is

$$B \triangleq \frac{1}{n} \sum_{i=1}^n (y_i - P(Y=1|\mathbf{x}=\mathbf{x}_i))^2$$

1. Brier Score is always between 0 and 1.
2. The larger the Brier Score, the more the discrepancy between prediction and outcome:
  - 0 = perfect predictions,
  - 0.25 = random predictions,
  - 1 = worst possible predictions.
3. Brier Scores should not be interpreted as an absolute measure.  
They are best used as comparative measures for different models on the same validation data set.

To see how the Brier score behaves, consider a simplified scenario where the probability of an event is  $q$  and an uninformative model that gives the same probability  $p = P(y = 1|s)$  for all scores. Then,

$$B \approx q(1 - p)^2 + (1 - q)p^2 = q - 2qp + p^2.$$

The minimum possible value of the Brier score in this case is when  $p = q$ :

$$B \approx q(1 - q).$$

This demonstrates that the Brier score is a different measure for each problem.

*Example 16.4*

Consider the log-odds scores and outcomes from example 5.2 for models (1) and (2). Calculate Brier Scores for the two models.

Recall for model (1):

Score $s_i$	$P(Y = 1 s_i)$	Outcome $y_i$	Score $s_i$	$P(Y = 1 s_i)$	Outcome $y_i$
0.696	0.333	1	3.648	0.025	0
0.824	0.305	0	3.696	0.024	0
1.784	0.144	1	4.096	0.016	0
1.8	0.142	0	4.176	0.015	0
2.008	0.118	1	4.784	0.008	1
2.456	0.079	1	5.16	0.006	0
2.832	0.056	0	5.616	0.004	0
3.384	0.033	0	6.288	0.002	0

Then the Brier Score is the mean squared difference between estimated probability and outcome = 0.244.

And for model (2):

Score $s_i$	$P(Y = 1 s_i)$	Outcome $y_i$	Score $s_i$	$P(Y = 1 s_i)$	Outcome $y_i$
-0.735	0.676	1	1.11	0.248	0
-0.655	0.658	0	1.14	0.242	0
-0.055	0.514	1	1.39	0.199	0
-0.045	0.511	0	1.44	0.192	0
0.085	0.479	1	1.82	0.139	1
0.365	0.410	1	2.055	0.114	0
0.6	0.354	0	2.34	0.088	0
0.945	0.280	0	2.76	0.060	0

Then the Brier Score is the mean of squared difference between estimated probability and outcome = 0.177.

Model (2) has a lower Brier score than model (1),  
hence we conclude it is a better probability forecaster.

## Review of Chapter 16

In this chapter we looked at how to assess probability estimates.

We covered the following topics:

- Probability calibration
- Hosmer-Lemeshow test
- Brier Score

# Consumer Credit Risk Modelling

## Chapter 17: Cost-based measures

## Overview

In this chapter we will consider a profit measure of performance based on our scorecard model.

This will lead to:

- Cost-based measures
- Finding optimal values for cut-off scores
- Cost measure and the ROC curve

## Profitability models

- Ultimately the bank is interested in the profit that can be derived from borrowers and avoiding any losses.
- So when assessing a credit risk model, we would like to use a measure of profit/loss and choose the model that maximizes expected profit.
- Ideally, models are required that incorporate multiple aspects of loan, including fees, administration rates, loan value, real interest rates, probabilities of default and costs of default, amongst other things.
- Such models would be a mixture of financial accounting and statistical models. **They are still being developed** (*interested students can read Chapter 4 of Lyn Thomas's text book for details of this approach*).
- Nevertheless, making some simplifying assumptions, we can estimate profit based on the standard scorecard model we have already studied.

## Loss and gain on a scorecard

For an applicant, we can build a table of outcomes as shown below, where Positive indicates Default as usual.

		Actual outcome	
		Positive	Negative
Prediction	Positive / Reject	0	0
	Negative / Accept	$-l$	$g$

The number in each cell refers to the expected profit from each outcome:

- Clearly, if the application is rejected then there is no profit (or loss).
- If the application is accepted then there is a gain  $g$  if the borrower does not default and a loss  $l$  if they do default.
- Suppose these amounts can be treated as a constant across all cases.

## Cost-based measure

The values in the above table correspond to the probabilities in the confusion matrix (Chapter 7, slide 8).

Together they give a formula for *expected profit*:

$$\begin{aligned}\text{Profit} &= -n(1 - F_1(c))p_1l + n(1 - F_0(c))(1 - p_1)g \\ &= ng(1 - p_1) - ng[F_0(c)(1 - p_1) + \gamma(1 - F_1(c))p_1]\end{aligned}$$

where  $\gamma = l/g$ .

The positive component ( $ng(1 - p_1)$ ) is fixed, relative to the model.

Therefore, we only need consider the term in square brackets which represents the **relative cost measure** (per account):

$$e_\gamma(c) = F_0(c)(1 - p_1) + \gamma(1 - F_1(c))p_1.$$

- Notice that this is very similar to the simple error rate (Chapter 7, slide 10), except for the relative weight  $\gamma$  between the two error types.

But the unknown in this formula is the relative weight  $\gamma = l/g$ .

What is a good value for  $\gamma$ ?

- Typically, the cost of a default is higher than the profit on a single account, therefore  $\gamma > 1$ . It is likely to be much greater than 1, but how much greater?
- Some suggest 5, others 10 or 20. This is of course something of a subjective decision which requires expert advice and knowledge of the loan product, interest rates and general product risks.
- Nevertheless, this performance measure does have the advantage of having financial meaning, whereas other measures such as AUC or Gini lack this quality.
- Also, the cost-based measure gives a quantitative basis for choosing the cut-off.

## Optimal cut-off score

The cost measure allows us to choose an optimal cut-off score.

Remembering that the lender may have a threshold  $p_{\text{accept}}$  for the proportion of applications that need to be accepted, the optimal cut-off score is given by

$$\hat{c} \triangleq \underset{c: P(S \leq c) < 1 - p_{\text{accept}}}{\operatorname{argmin}} e_\gamma(c)$$

In particular,

- As  $\gamma$  increases, so the optimal cut-off increases (reflecting the increased cost of default risk).
- As  $\gamma$  decreases, so the optimal cut-off decreases.
- Note that there is not necessarily a unique optimal cut-off score (there could be several).
- In general, the relationship between the cut-off score and cost measure is *not* monotonic.

This is expressed mathematically as follows:

- Let  $\gamma_A$  and  $\gamma_B$  be two alternative relative cost weights on a score distribution, such that  $\gamma_A < \gamma_B$ .
- Let  $\hat{c}_A$  and  $\hat{c}_B$  be the optimal cut-off scores for  $\gamma_A$  and  $\gamma_B$ , respectively.

Then,  $\hat{c}_A \leq \hat{c}_B$ .

### Proof.

From the definition of the relative cost measure, for any  $c$ ,

$$e_{\gamma_B}(c) - e_{\gamma_A}(c) = (\gamma_B - \gamma_A)(1 - F_1(c))p_1 \quad [1]$$

Since  $\hat{c}_A$  and  $\hat{c}_B$  are optimal, for any  $c \in \{c: P(S \leq c) < 1 - p_{\text{accept}}\}$ ,

$$e_{\gamma_A}(\hat{c}_A) \leq e_{\gamma_A}(c), \text{ and} \quad [2]$$

$$e_{\gamma_B}(\hat{c}_B) \leq e_{\gamma_B}(c) \quad [3]$$

$$\Rightarrow e_{\gamma_B}(\hat{c}_B) = e_{\gamma_A}(\hat{c}_B) + (\gamma_B - \gamma_A)(1 - F_1(\hat{c}_B))p_1 \quad \text{from [1]}$$

$$\geq e_{\gamma_A}(\hat{c}_A) + (\gamma_B - \gamma_A)(1 - F_1(\hat{c}_B))p_1 \quad \text{from [2]}$$

$$= [e_{\gamma_B}(\hat{c}_A) - (\gamma_B - \gamma_A)(1 - F_1(\hat{c}_A))p_1] + (\gamma_B - \gamma_A)(1 - F_1(\hat{c}_B))p_1 \quad \text{from [1]}$$

$$\Rightarrow (\gamma_B - \gamma_A)(F_1(\hat{c}_B) - F_1(\hat{c}_A))p_1 \geq e_{\gamma_B}(\hat{c}_A) - e_{\gamma_B}(\hat{c}_B) \geq 0 \quad \text{from [3]}$$

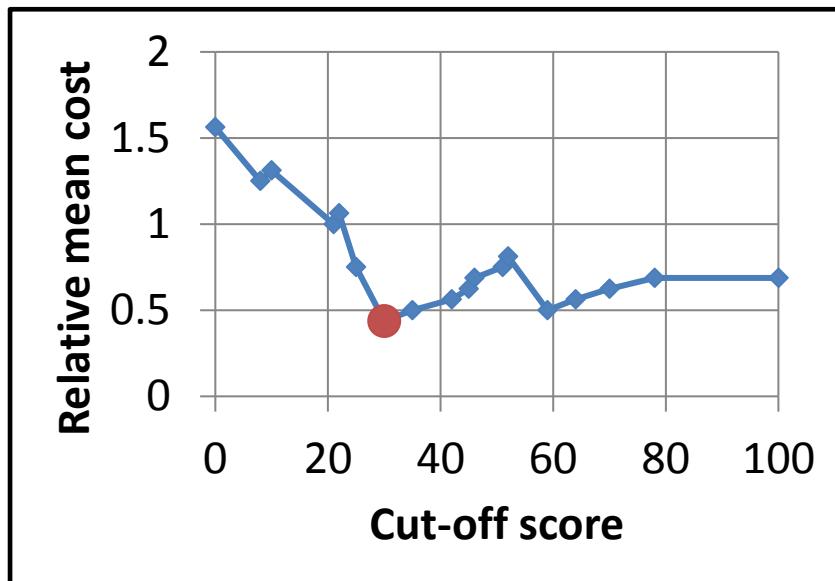
$$\Rightarrow F_1(\hat{c}_B) \geq F_1(\hat{c}_A) \Rightarrow \hat{c}_A \leq \hat{c}_B \text{ since } \gamma_A < \gamma_B, p_1 > 0 \text{ and } F_1 \text{ is a CDF.}$$

*Example 17.1*

Again, consider the 16 applicants from Example 4.1.

Score	8	10	21	22	25	30	35	42	45	46	51	52	59	64	70	78
Outcome	1	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0

If we set  $\gamma=5$ , then the relative mean cost for each value of  $c$  is shown in this graph.



We see that the expected cost is minimized at a cut-off score of 30 (the red spot).

## Cost measure and the ROC curve

Remember that  $F_0$  and  $F_1$  are plotted respectively on the x and y axes of the ROC curve.

It follows that all straight lines in the ROC curve space with slope  $\frac{1-p_1}{\gamma p_1}$  are lines of equal cost.

Proof.

Set a fixed cost  $k$ :  $e_\gamma(c) = F_0(c)(1 - p_1) + \gamma(1 - F_1(c))p_1 = k$ .

Then differentiate by  $c$ :  $F'_0(c)(1 - p_1) - \gamma F'_1(c)p_1 = 0$ ,

The slope of the ROC curve is  $F'_1(c)/F'_0(c) = \frac{1-p_1}{\gamma p_1}$ .

Therefore the optimal cut-off score can be found as the tangent to a convex ROC curve which has this slope.

- Note: This is not necessarily a unique point.

## Analytic solution

If  $F_0$  and  $F_1$  are differentiable, then an analytic solution is possible,

A minimum to  $e_\gamma(c)$  is found when

$$\frac{de_\gamma(c)}{dc} = f_0(c)p_0 - \gamma f_1(c)p_1 = 0$$

$$\Rightarrow \frac{f_0(c)}{f_1(c)} = \frac{\gamma p_1}{p_0} \Rightarrow \frac{f(S = c|Y = 0)}{f(S = c|Y = 1)} = \frac{\gamma p_1}{p_0}$$

$$\Rightarrow \frac{P(Y = 0|S = c)p_1}{P(Y = 1|S = c)p_0} = \frac{\gamma p_1}{p_0} \Rightarrow \frac{P(Y = 0|S = c)}{1 - P(Y = 0|S = c)} = \gamma$$

by Bayes theorem (assuming  $f(S = c) > 0$ ), working with densities

$$\Rightarrow P(Y = 0|S = c) = \frac{\gamma}{1 + \gamma}$$

Since  $S = s(\mathbf{X})$  for a score link-function  $s$  and vector of predictor variables  $\mathbf{X}$ ,

$$P(Y = 0 | \mathbf{X} = \mathbf{x}') = \frac{\gamma}{1 + \gamma}$$

for all  $\mathbf{x}'$  such that  $s(\mathbf{x}') = c$ .

For a general link function  $F_L$ ,  $P(Y = 0 | \mathbf{X} = \mathbf{x}') = F_L(s(\mathbf{x}'))$

so  $s(\mathbf{x}') = c = F_L^{-1}\left(\frac{\gamma}{1+\gamma}\right)$  is a general solution.

In particular, for the log-odds score,  $F_L(s(\mathbf{x}')) = \frac{1}{1+e^{-s(\mathbf{x}')}}$  (see Chapter 5), therefore

$$c = \log \left[ \frac{\frac{\gamma}{1+\gamma}}{1 - \frac{\gamma}{1+\gamma}} \right] = \log \gamma$$

Need to show that this solution is indeed a unique minimum.

*Proof.*

Suppose  $c^*$  is a stationary point: ie  $\frac{de_\gamma(c^*)}{dc} = 0$ .

Using Bayes Theorem again, for any  $c$ ,

$$\begin{aligned}\frac{de_\gamma(c)}{dc} &= P(Y = 0|S = c)f(S = c) - \gamma P(Y = 1|S = c)f(S = c) \\ &= [(1 + \gamma)P(Y = 0|S = c) - \gamma]f(S = c)\end{aligned}$$

Since probability of  $Y = 0$  increases with score (*property of the score*),

$$P(Y = 0|S = a) < P(Y = 0|S = b) \Leftrightarrow a < b$$

Then, it follows that

- $\frac{de_\gamma(c)}{dc} < \frac{de_\gamma(c^*)}{dc} = 0$  for  $c < c^*$  and
- $\frac{de_\gamma(c)}{dc} > \frac{de_\gamma(c^*)}{dc} = 0$  for  $c > c^*$

Hence  $e_\gamma(c)$  is always decreasing down to  $c^*$  and increasing from  $c^*$ .

Hence  $c^*$  must be a unique minima.

## Review of Chapter 17

In this chapter we have followed the following topics:

- Profit formula
- Cost-based measure
- Finding optimal values for cut-off scores
- Cost-based measure and ROC curve