Statistics in Retail Finance

Chapter 2: Statistical models of default
Overview

We consider how to build statistical models of default, or delinquency, and how such models are traditionally used for credit application scoring and decision making.

Topics covered are:-

- The problem of modelling default;
- Probability of default and log-odds scores;
- Logistic regression;
- Application scoring and the accept/reject decision;
- Risk grades.
Default, or delinquency, is an event that may occur during the duration of a loan or credit card.

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

<table>
<thead>
<tr>
<th>$Y = 1$</th>
<th>$Y = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>Non-default</td>
</tr>
<tr>
<td>Delinquency</td>
<td>Non-delinquency</td>
</tr>
<tr>
<td>Positive event</td>
<td>Negative event</td>
</tr>
<tr>
<td>Bad customer</td>
<td>Good customer</td>
</tr>
</tbody>
</table>
For any particular loan, we will have a vector of variables which may allow us to model default. We call these predictor variables. We use $X = (X_1, ..., X_m)^T$ to represent a vector of $m$ random variables.

Typical types of data available from which these variables can be drawn:

<table>
<thead>
<tr>
<th><strong>Personal details</strong></th>
<th>Details about the individual taking the credit, such as employment status, profession, income, residential status, record of court judgements and number of dependents. Often this data is available from the credit application form, but some from credit bureaus.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Past credit history</strong></td>
<td>Length of credit history, number and value of past loans, number and value of past <em>delinquent</em> loans. Usually provided by credit bureaus.</td>
</tr>
<tr>
<td><strong>Behavioural data</strong></td>
<td>Past history of use of credit on previous products (eg spend amount, repayment patterns).</td>
</tr>
</tbody>
</table>
The default modelling problem can be framed as an attempt to model outcome $Y$ on predictor variables $X$. As such, this is a **classification** problem.

Typically, we require the probability of the event, rather than just a point estimate of outcome. Therefore we are looking to develop a model to estimate

$$P(Y = 1|X = x)$$

This the probability of default (PD), conditional on characteristics $x$. 
A typical way to do this is to use a logistic regression model. For reasons that become apparent soon, we actually model \( Y = 0 \). The logistic regression model is then

\[
P(Y = 0|\mathbf{X} = \mathbf{x}) = f_L(\beta_0 + \mathbf{\beta}^T \mathbf{x})
\]

where \( f_L \) is the logistic link function

\[
f_L(s) = \frac{1}{1 + e^{-s}}
\]

and \( \beta_0 \) is an intercept and \( \mathbf{\beta} = (\beta_1, \ldots, \beta_m)^T \) is a vector of coefficients, one for each predictor variable.
The function \( s(x) = \beta_0 + \beta^T x \) is then called the *log-odds score* since
\[
s(x) = f^{-1}_L(P(Y = 0|X = x)) = \log \left( \frac{P(Y = 0|X = x)}{P(Y = 1|X = x)} \right)
\]
The log-odds score is typically the basis of the *credit score* used by banks and credit bureaus to rank people.

- Credit scores are usually presented nicely for the general public as positive integers within an interval (e.g., Experian’s score is between 0 and 999), but this is typically based on a “raw” log-odds score.

- It is conventional for high scores to represent good risk (i.e., low PD) and low scores to represent bad risk (i.e., high PD) and it is for that conventional reason that \( Y = 0 \) is modelled, rather than \( Y = 1 \).
  
  - Of course, it easy to use \( P(Y = 1|X = x) = 1 - P(Y = 0|X = x) \)
Estimating the logistic regression model

The parameters that need to be estimated in logistic regression are $\beta_0$ and $\beta$.
Denote estimates as $\hat{\beta}_0$ and $\hat{\beta}$, respectively.

Compute estimates based on maximum likelihood estimation (MLE) and a training data set of $n$ observations:

$$D = [(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)]$$

and assume independence between events for each observation.
Then the log-likelihood function is given by

$$\log L(\beta_0, \beta | D_n) = \sum_{i=1}^{n} \log P(Y_i = y_i | X_i = x_i, \beta_0, \beta)$$
But \( P(Y_i = y_i | X_i = x_i, \beta_0, \beta) = P(Y_i = 0 | X_i = x_i, \beta_0, \beta) \) if \( y_i = 0 \),
and \( P(Y_i = y_i | X_i = x_i, \beta_0, \beta) = 1 - P(Y_i = 0 | X_i = x_i, \beta_0, \beta) \) if \( y_i = 1 \)

so \( P(Y_i = y_i | X_i = x_i, \beta) = P(Y_i = 0 | X_i = x_i, \beta)^{1-y_i} (1 - P(Y_i = 0 | X_i = x_i, \beta))^{y_i} \)

and therefore, substituting \( P(Y_i = 0 | X_i = x_i, \beta) = f_L(\beta_0 + \beta^T x_i) \),

\[
\log L(\beta_0, \beta | D_n) = \sum_{i=1}^{n} (1 - y_i) \log \left( \frac{1}{1 + e^{-(\beta_0 + \beta^T x_i)}} \right) + y_i \log \left( \frac{1}{1 + e^{\beta_0 + \beta^T x_i}} \right)
\]

MLE requires this is maximized to compute estimates:

\[
(\hat{\beta}_0, \hat{\beta}) = \arg \max_{\beta_0, \beta} [\log L(\beta_0, \beta | D_n)]
\]

- The usual standard errors can also be computed on these estimates as with any other MLE.
- These then lead to the usual hypothesis tests for ML estimates.
Differentiating by each coefficient in $\beta$ and setting the derivatives equal to zero to find the maxima gives

$$\sum_{i=1}^{n} \left(1 - y_i - \left(\frac{1}{1 + e^{-(\vec{\beta}_0 + \vec{\beta}^T x_i)}}\right)\right) = 0$$

and

$$\sum_{i=1}^{n} x_{ij} \left(1 - y_i - \left(\frac{1}{1 + e^{-(\vec{\beta}_0 + \vec{\beta}^T x_i)}}\right)\right) = 0$$

for each variable $j=1$ to $m$.

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

**Exercise 2.1**
Perform this differentiation to derive this system of equations.
Example 2.1
The following logistic regression output was produced on a data set of 40,000 credit cards.

Likelihood Ratio = 1819 (p-value < 0.001)

| Variable                  | Coefficient | Estimate | Standard error | z    | P(>|z|) |
|---------------------------|-------------|----------|----------------|------|---------|
| Intercept                 | $\beta_0$  | -0.181   | 0.084          | 2.15 | 0.032   |
| Age                       | $\beta_1$  | +0.0353  | 0.0013         | 27.2 | <0.001  |
| Income (log)              | $\beta_2$  | -0.0164  | 0.0100         | 1.64 | 0.10    |
| Residential phone         | $\beta_3$  | +0.622   | 0.030          | 20.7 | <0.001  |
| Home owner *              | $\beta_4$  | 0        |                |      |         |
| Renter                    | $\beta_5$  | -0.155   | 0.039          | 3.97 | <0.001  |
| Lives with parents        | $\beta_6$  | +0.256   | 0.045          | 5.69 | <0.001  |
| Months in residence       | $\beta_7$  | -0.00025 | 0.00011        | 2.28 | 0.020   |
| Months in current job     | $\beta_8$  | +0.00210 | 0.00025        | 8.40 | <0.001  |

* Notice that the Home owner category is set as base residency category and so has no coefficient estimate.
Interpreting the default model

- The coefficient estimates tell us about the association between each predictor variable and the outcome variable.

- This is sometimes referred to as the “effect” but this does not necessarily imply a direct causal relationship between predictor variables and outcome (eg there could be a confounder causing them separately).

- The interpretation should always be made in the context of all variables in the model; ie the effect of one variable is in relation (controlled) by the others.
There are three aspects to interpreting the association:

1. Direction of association.
2. Magnitude of association.

**Direction of association**

This is given by the sign on the coefficient estimate:

- If +ve, then larger values of the variable are associated with non-default (less risky);
- If -ve, then larger values of the variable are associated with default (more risky).

**Magnitude of association**

- This is given by the magnitude of the coefficient estimate.
- This is dependent on the scale and unit of measure of the variable and it is not always meaningful to compare the magnitude on one variable to another with a different unit.
• The magnitude of the estimate gives the change on the log-odds score, not on the probability of outcome.

Categorical variables are usually included in the model as a series of indicator variables (0/1) for each level, excluding one as a base level. Direction and magnitude of association for each category level is then relative to the base level.

**Statistical significance of association**

• If \( \beta_j = 0 \), then variable \( j \) has no association with outcome.

• However we only have coefficient estimate \( \hat{\beta}_j \). If this is close to 0 then we may wonder if the true value is 0.

• A statistical test, based on the standard error (s.e.), can be used to test the hypotheses:
  - Null hypothesis: \( H_0: \beta_j = 0 \)
  - Alternative hypothesis: \( H_1: \beta_j \neq 0 \)

This is the Wald Test.
MLE has the property of asymptotic normality:

\[
\frac{(\hat{\beta}_j - \beta_j)}{\hat{s}_j} \rightarrow N(0,1) \text{ as } n \rightarrow \infty
\]

where \( \hat{s}_j \) is the standard error on \( \hat{\beta}_j \).

Set a significance level \( \alpha \).

Under \( H_0, \beta_j = 0 \), therefore reject \( H_0 \) if \( \frac{|\hat{\theta}_j|}{\hat{s}_j} > z_{\alpha/2} \)

where \( z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) \)

and \( \Phi \) is the CDF for the standard normal distribution.
Example 2.2

Review the model in Example 2.1 and interpret the association of each variable according to direction of association and statistical significance using a significance level of 1%.

Interpretation:

1. There is sufficient evidence, at 1% significance level, that there is an association between several variables (age, having a residential phone, housing type and months in current job) and default.

2. The direction of association for each of the statistically significant variables is:
   - age, having a residential phone and months in current job have a positive association with creditworthiness;
   - being a renter has a negative association with creditworthiness, relative to home owners;
• living with parents has a positive association with creditworthiness, relative to home owners.
Exercise 2.2

The following logistic regression scorecard model was built on a data set of 20,000 personal loans. The outcome variable was non-default.

Likelihood Ratio = 83.8 (p-value < 0.001)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Standard error</th>
<th>Wald chi-square</th>
<th>P &gt; chi-square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>+2.66</td>
<td>0.083</td>
<td>1028</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age 18-29*</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 30-47</td>
<td>+0.14</td>
<td>0.08</td>
<td>3.0</td>
<td>0.08</td>
</tr>
<tr>
<td>Age 48+</td>
<td>+0.47</td>
<td>0.10</td>
<td>20.5</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of credit cards</td>
<td>+0.05</td>
<td>0.03</td>
<td>2.4</td>
<td>0.12</td>
</tr>
<tr>
<td>Self-employed</td>
<td>-0.47</td>
<td>0.07</td>
<td>44.0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Months in current residence</td>
<td>+0.009</td>
<td>0.004</td>
<td>4.3</td>
<td>0.04</td>
</tr>
</tbody>
</table>

*Base category level

Interpret the association of each variable with default, in terms of direction of association and statistical significance, using a significance level of 1%.
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 \texttt{glm}.

\textit{Example 2.3}

\begin{verbatim}
train <- read.delim("train.txt")
attach(train)

glm.out <- glm(good ~ age + income2_log + res_phone +
res_type_A + res_type_C + months_in_res + months_in_the_job,
    family = binomial("logit"))
print(summary(glm.out))
\end{verbatim}
The default model is typically built as a credit scorecard to make automated decisions to reject or accept new credit applications. We consider this decision process in the next slides.

But since logistic regression provides an estimate of PD, this can feed into further estimates of profit or loss, as we will see in later chapters.

Notice that these uses just utilize the model as an estimator of credit score or probability of outcome. Nevertheless, the default model still requires interpretation since we need to ensure it is sensible. Additionally, the regulators require lending decisions to be transparent and open to scrutiny. In particular, there are certain legal requirements which differ from country to country (eg in UK, sex and age cannot be used in an application scorecard).
Application scorecards

- An application scorecard is a default model where all the predictor variables are available at time of application.
- When someone applies for credit, they will provide sufficient information for the lender to build up a profile of values $x$.
- Then, given an estimated model, a score $s(x) = \beta_0 + \beta^T x$ is assigned to the individual.
- Remember, high scores represent better risk, so only those with the highest credit scores are accepted.
- Therefore, given some cut-off score $c$, the following decision process is made:
  - If $s(x) \leq c$, then reject the credit application;
  - If $s(x) > c$, then accept the credit application.

- Because $s(x)$ is a log-odds score, this can be re-written as a decision based on PD, with probability cut-off $1 - f_L(c)$. 
Deciding the cut-off score

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

Indeed, the cut-off score is decided on the basis of a number of factors related to the lender’s goals. Typically:

1. Specify a cut-off score which maximizes expected profits: ie sufficient volume with minimal risk.
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.
Sometimes it is useful to group customers 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 products to be assessed in terms of summaries of cases in each group.

- Probabilities of outcome within group can be used to compute overall risk of default for the whole group.
Risk groups are defined by ranges of scores:

- Consider observations divided into $G$ groups with intervals $g_i$ for all $i \in \{1, \ldots, G - 1 \}$ such that $g_i < g_{i+1}$ for all $i \in \{1, \ldots, G - 2 \}$.

- The risk group for any observation with score $s$ is then given by the function

$$r_G(s) = \begin{cases} 
1 & \text{if } s \leq g_1 \\
i & \text{if } g_{i-1} < s \leq g_i \text{ for } i \in \{2, \ldots, G - 1 \} \\
G & \text{if } g_{G-1} < s 
\end{cases}$$
Example 4.1

The risk grades for the Experian generic credit score:

<table>
<thead>
<tr>
<th>Score</th>
<th>Creditworthiness assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 560</td>
<td>Very poor</td>
</tr>
<tr>
<td>561 - 720</td>
<td>Poor</td>
</tr>
<tr>
<td>721 - 880</td>
<td>Fair</td>
</tr>
<tr>
<td>881 - 960</td>
<td>Good</td>
</tr>
<tr>
<td>961 - 999</td>
<td>Excellent</td>
</tr>
</tbody>
</table>
Review of Chapter 2

Topics covered were:-

- The problem of modelling default;
- Probability of default and log-odds scores;
- Using logistic regression for default modelling;
- Application scoring and the accept/reject decision;
- Risk grades.