Credit Risk Modelling

Tiziano Bellini

Università di Bologna

December 13, 2013
Credit Risk Modelling

- Introduction to credit risk.
- Probability of default overview.
- Probability of default through GLM.
- A practical estimation of PD through R GLM.
- Concluding remarks.
Exposure at Default (EAD)

EAD stands for the Exposure at Default. As a borrower goes towards default it will normally attempt to increase its leverage (lend more).

- The degree in which this is possible will be dependant on the type of products (facilities) the borrower has and the bank ability to prevent excessive draw down on facilities.
- The products can be separated into three main categories.

1. Loans.
2. Working capital facilities.
Financial Product Categories

- Loans are products where the money is made available at predetermined moments and the customer is required to repay at predetermined moments. Therefore there is very little the borrower can do to increase the debt.

- A working capital facility is used by a company to manage their liquidity. The facility allows the company to borrow money up to a pre-set limit. The customer is free to borrow and repay any amount at any time as long as the total exposure remains below the limit.

- Potential exposure products might lead to an exposure as in the case of a guarantee. The bank gives a guarantee for the customer to a third party. This guarantee will only translate into an exposure if this third party requests payment under the guarantee.
Working Capital and Potential Exposures

- It is necessary to specify the holding period for EAD estimation. Usually one year.
- Alternative approaches can be followed in order to estimate a $k$ product factor to apply to the borrower working capital. It is useful to distinguish between:

1. Descriptive model. A cluster analysis is carried out and the mean $K$ factor is applied to the borrower according to its cluster.
2. Econometric model. A regression analysis is carried out considering dimensions such as exposure amount, geographic area and so on.
Loss Given Default (LGD)

- Loss given default (LGD) represents the percentage of the EAD which is expected to lose if a counterparty goes into default.
- There are many scenarios of events which may occur after a company goes into default. The two most extreme are as follows:
  1. The counterparty recovers without any loss to the bank.
  2. Sale of assets and collateral is required.
- Because the definition of default is rather strict (90 days overdue) many defaults will fall in the first category. Most companies who are 90 days overdue simply recover. Often even without intervention by your bank.
- The sale of assets and collateral occurs less frequently but leads to higher losses. It can be assumed that this scenario only occurs when a company goes bankrupt. Note that bankruptcy is a lot worse than default (minimally 90 days overdue). Generally you can separate the returns in two types:
  - Return on collateral
  - Return on unpledged assets
LossCalc\(^1\) Mechanics

- The combination of the predictive factors is a linear weighted sum, derived using regression techniques without an intercept term. The model takes the additive form

\[
\hat{r} = \beta_1 X_1 + \ldots + \beta_k X_k. \tag{1}
\]

- \(\hat{r}\) is the normalized recovery.
- \(x_l\) is the transformed value or mini-model.
- The final step is to apply a Beta-distribution transformation.

\(^1\)Moody's KMV (2005). *LossCalc v2: Dynamic Prediction of LGD.*
# LossCalc Factors

## Table 1: Explanatory Factors in the LossCalc Models

This is a summary of the factors applied in Moody’s LossCalc model to predict LGD. The table highlights the five broad categories of predictive information: collateral, instrument, firm, industry, and broad economic environment. These factors have little inter-correlation and join to make a powerful LGD prediction.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collateral and Other Support</td>
<td>The proportion of coverage of the exposure by Cash, ‘All Assets’ or Property, Plant and Equipment. Support from subsidiaries takes a yes/no flag rather than a coverage ratio.</td>
<td>Collateral</td>
</tr>
<tr>
<td>Debt Type and Seniority Class</td>
<td>LGD, controlling for debt-type (loan, bond, and preferred stock) and seniority classes (senior, junior, secured, unsecured, subordinate, etc.).</td>
<td>Historical Averages</td>
</tr>
<tr>
<td>Firm-level Information</td>
<td>Seniority standing of debt within the firm’s overall capital structure; this is the relative seniority of a claim. This is different from the absolute seniority stated in Debt Type and Seniority Class above. For example, the most senior obligation of a firm might be a subordinate note if no claim stands above it.</td>
<td>Seniority Standing</td>
</tr>
<tr>
<td>Cycle Adjusted Firm Leverage</td>
<td>All Corp. Default Rate, interacted with the default probabilities directly implied by book leverage.</td>
<td>Leverage</td>
</tr>
<tr>
<td>The firm’s Distance-to-Default (for public firms only)</td>
<td>The firm’s Distance-to-Default for public firms only</td>
<td>Firm Distress</td>
</tr>
<tr>
<td>Industry</td>
<td>Historical normalized industry recovery averages after controlling for seniority class.</td>
<td>Industry Experience</td>
</tr>
<tr>
<td>The industry’s Distance-to-Default (tabulated by country/region)</td>
<td>The industry’s Distance-to-Default (tabulated by country/region)</td>
<td>Industry Distress</td>
</tr>
<tr>
<td>Macroeconomic and Geographic</td>
<td>The country/region’s Distance-to-Default (tabulated by industry)</td>
<td>Region Distress</td>
</tr>
<tr>
<td>Country/region shifts in mean expectation</td>
<td>Shift</td>
<td></td>
</tr>
</tbody>
</table>

All the factors in LossCalc are highly statistically significant individually and, in all cases, signs were in the expected direction. Figure 5 below shows the contributions of each broad factor category towards the prediction of the one-year and immediate LGD forecasts. Each of each color add up to 100%. This shows the relative significance across the predictive factors.

---

**Figure:** LossCalc Moody’s KMV explanatory factors².

---

²Source: LossCalcV2.

---

Tiziano Bellini (Università di Bologna)  
Credit Risk Modelling  
December 13, 2013  
8 / 55
LossCalc Beta Transformation

- Mathematically, a Beta distribution is as follows

\[
\text{Beta}(x, \alpha, \beta, \text{Min} = 0, \text{Max}) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x}{\text{Max}}\right)^{\alpha-1} \left(1 - \frac{x}{\text{Max}}\right)^{\beta-1} \left(\frac{1}{\text{Max}}\right),
\]

where \(x\) is the estimated recovery rate \(\hat{r}\).

- The shape parameters can be derived in a variety of ways. For example, the following give them in terms of population mean and standard deviation

\[
\alpha = \frac{\mu}{\text{Max}} \left[\frac{\mu(\text{Max} - \mu)}{\text{Max} \cdot \sigma^2} - 1\right],
\]

\[
\beta = \alpha \left[\frac{\text{Max}}{\mu} - 1\right].
\]
LossCalc Drawbacks and Realized LGD

- LossCalc is not directly applicable to commercial banks because of the lack of market information.
- Alternative ways can be followed considering the economic LGD. The most general framework is as follows

\[
\text{LGD}_{\text{realized}} = 1 - \frac{\sum_i \text{Recovery}_i - \sum_i \text{Cost}_i}{\text{Exposure}},
\]

where the recovery is the actual value of cash flows.

- Starting from the above described equation, further developments have been carried out in practice\(^3\).

Introduction to PD Modelling

- The probability of default (PD), also indicated as expected default frequency, is the likelihood that a loan will not be repaid and will fall into default.

- There are many alternatives for estimating the probability of default. Default probabilities may be estimated exploiting:
  2. Formal (statistical) models.
  3. Integration of alternative approaches.

- From the source of data we distinguish among model based on:
  1. Non market data (balance sheet and other firm information).
  2. Market data (share quotations, bond spreads and so on).
Statistical modelling

- We consider as statistical those models based on data analysis carried out through statistical tools.
- Many approaches have been developed. We focus on the most widely used in practice:
  1. Discriminant analysis. Developed by Altman (1968), this has been one of the first approaches to PD estimation, but it has quickly been quitted.
  2. Logit (and GLM) regression. This is the most widely used approach exploited in commercial banks.
  3. Distance to default. (This approach will be analyzed dealing with regulatory economic capital).
- There are other models such as, for example, classification and regression threes, data envelopment analysis, neural networks, ... which are used in some area, but on the one hand they are difficult to be interpreted and, on the other, they are not easy to be used in practice.
Introduction to discriminant analysis

- The main idea of discriminant analysis is to divide observations in groups.
- From a theoretical point of view, there are many approaches to discriminant analysis. Following Fisher approach \(^4\), we assume that each observation belong to one of the \(k\) multivariate samples with the same covariance matrix.
- We estimate the group \(g\) mean from the sample and assuming \(\sum_1 = \ldots = \sum_k = \sum\) we use \(S\) to estimate \(\sum\).
- We search for the linear combination \(Z_g = a'X_g\) which maximizes the separation of groups. We consider the ratio

\[
F = \frac{SSB(a)/(k - 1)}{SSW(a)/(n - k)}
\]

where \(SSB\) is the variance between groups while \(SSW\) is the variance within groups.

Discriminant Analysis and PD Estimation

- The ratio $F$ is maximized when $a$ is the eigenvector associated to the highest eigenvalue of $S^{-1}_W S_B$. In the case where there are many groups we consider many eigenvectors, while in our analysis, distinguishing between default and no default we consider only one eigenvector.

- This approach is not currently used to estimate PD because:
  1. It can be used only in the case of numerical variables.
  2. It assumes $\sum_1 = \ldots = \sum_k = \sum$ which is not the case in our analysis.
  3. It does not immediately supply a PD output.

- The use of GLM analysis allow to overwhelm these drawbacks.
Generalized Linear Model

1. **Random component.** The random component of a GLM consists of a response variable \( Y \) with independent observations \((y_1, \ldots, y_N)\).

2. **Systematic component.** The systematic component of a GLM relates a vector \((\eta_1, \ldots, \eta_N)\) to the explanatory variables through a linear model.

3. **Link function.** The third component of a GLM is a link function that connects the random and systematic components.
Random Component

- The random component of a GLM consists of a response variable \( Y \) with independent observations \((y_1, \ldots, y_N)\) from a distribution in the natural exponential family.
- This family has probability density function or mass function of form

\[
f(y_i; \theta_i) = a(\theta_i) b(y_i) \exp[y_i Q(\theta_i)]. \tag{7}
\]
Systematic Component

- The systematic component of a GLM relates a vector \((\eta_1, \ldots, \eta_N)\) to the explanatory variables through a linear model.
- Let \(x_{ij}\) denote the value of predictor \(j \in (1, \ldots, p)\) for subject \(i\). Then for all \(i \in (1, \ldots, N)\)

\[
\eta_i = \sum_j \beta_j x_{ij}. \tag{8}
\]
Link Function

- The third component of a GLM is a link function that connects the random and systematic components.
- Let $\mu_i = E(Y_i)$. The model links $\mu_i$ to $\eta_i$ by $\eta_i = g(\mu_i)$, where the link function $g$ is a monotonic, differentiable function. Thus, $g$ links $E(Y_i)$ to explanatory variables through the formula

$$g(\mu_i) = \sum_j \beta_j x_{ij}. \quad (9)$$
Default vs Non-Default

- Let $Y$ denote a binary response variable. In our analysis it denotes the default or non-default of a counterparty. Each observation has the outcomes denoted by 0 and 1, binomial for a single trial.
- The mean $E(Y) = P(Y = 1)$. We denote $P(Y = 1)$ by $\pi(x)$, reflecting its dependence on values $x = (x_1, \ldots, x_p)$ of predictors.
- The variance of $Y$ is

$$\text{var}(Y) = \pi(x)[1 - \pi(x)]$$

which corresponds to the variance of a Bernoulli random variable.
Linear Probability Model

- For a binary response variable, the regression model

\[ \pi(x) = \alpha + \beta x \]  

(11)

is called linear probability model. With independent observations it is a GLM with binomial random component and identity link function.
Logistic Regression Model

- Usually, binary data result from a nonlinear relationship between $\pi(x)$ and $x$
Logit Link Function

- The most important curve with the above described shape is the logistic regression model which is specified as follows:

\[ \pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \]  \hspace{1cm} (12)

- The link function for the logistic regression is as follows:

\[ \frac{\pi(x)}{1 - \pi(x)} = \exp(\alpha + \beta x) \]  \hspace{1cm} (13)

where the log odds has the linear relationship:

\[ \ln \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x \]  \hspace{1cm} (14)
Likelihood Equations

- We now turn our attention to details such as likelihood equations and methods for fitting them.

- It is helpful to extend the notation for a GLM so that it can handle many distributions that have a second parameter. The random component of the GLM specifies that the $N$ observations $(y_1, \ldots, y_N)$ on $Y$ are independent, with probability mass or density function for $y_i$ of form

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \quad (15)$$

that is called the exponential dispersion family and $\phi$ is called the dispersion parameter. The parameter $\theta_i$ is the natural parameter.
Mean and Variance for the Random Component 1/3

We start from the contribution of $L_i = \ln f(y_i; \theta_i, \phi)$ where the log-likelihood is $L = \sum_i L_i$. Then, considering equation (15), we obtain what follows

$$L_i = [y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi).$$

Therefore

$$\frac{\partial L_i}{\partial \theta_i} = [y_i - b'(\theta_i)]/a(\phi),$$

$$\frac{\partial^2 L_i}{\partial \theta_i^2} = -b''(\theta_i)/a(\phi).$$
Mean and Variance for the Random Component 2/3

- We now apply the general likelihood results

\[
E \left( \frac{\partial L}{\partial \theta} \right) = 0, \tag{19}
\]

\[
- E \left( \frac{\partial^2 L}{\partial \theta^2} \right) = E \left( \frac{\partial L}{\partial \theta} \right)^2, \tag{20}
\]

which hold under regularity conditions satisfied by the exponential family.
From equation (19) we obtain what follows

$$\mu_i = E(Y_i) = b'(\theta_i). \quad (21)$$

From equation (20) we obtain what follows

$$b''(\theta_i)/a(\phi) = E[(Y_i - b'(\theta_i)/a(\phi)]^2 = \text{var}(Y_i)/[a(\phi)]^2, \quad (22)$$

which implies

$$\text{var}(Y_i) = b''(\theta_i)a(\phi). \quad (23)$$
Mean and Variance for the Logit Model 1/2

Next, suppose that \( n_i Y_i \) has a \( \text{bin}(n_i, \pi_i) \) distribution. In this context, \( y_i \) is the sample proportion of successes, so \( E(Y_i) \) is independent on \( n_i \). Let \( \theta_i = \ln[\pi_i/(1 - \pi_i)] \). Then \( \pi_i = \exp(\theta_i)/[1 + \exp(\theta_i)] \) and \( \ln(1 - \pi_i) = -\ln[1 + \exp(\theta_i)] \), therefore we can show what follows

\[
f(y_i; \pi_i, n_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} = \nonumber \]

\[= \exp \left[ y_i \theta_i - \ln[1 + \exp(\theta_i)] \right] + \ln \left( \binom{n_i}{y_i} \right). \quad (24)\]

This has exponential dispersion from equation (15) with \( b(\theta_i) = \ln[1 + \exp(\theta_i)], a(\phi) = 1/n_i \) and \( c(y_i, \phi) = \ln \left( \binom{n_i}{y_i} \right) \).
According to what we stated with reference to equation (24), we can highlight what follows

\[ E(Y_i) = b'(\theta_i) = \exp(\theta_i)/[1 + \exp(\theta_i)] = \pi_i, \]  
\[ \text{var}(Y_i) = b''(\theta_i)/a(\phi) = \exp(\theta_i)/[1 + \exp(\theta_i)]^2 n_i = \pi_i(1 - \pi_i)/n_i. \]
For $N$ independent observations, from equation (16), the log likelihood is

$$L(\beta) = \sum_i L_i = \sum_i \ln(f(y_i; \theta_i, \phi)) = \sum_i \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i, \phi).$$

(27)

The likelihood equations are

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_i \frac{\partial L(\beta)}{\partial \beta_j} = 0,$$

(28)

for all $j$. 

Chain Rule

- To differentiate the log likelihood of equation (27) it is useful to exploit the chain rule

\[
\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. 
\]  

(29)

- In the above equation, \( \frac{\partial L_i}{\partial \theta_i} = (y_i - \mu_i)/a(\phi) \), \( \frac{\partial \mu_i}{\partial \eta_i} = b''(\theta) = \text{var}(Y_i)/a(\phi) \), considering that \( \eta_i = \sum_j \beta_j x_{ij} \) we can state \( \frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \) and \( \frac{\partial \mu_i}{\partial \eta_i} \) depends on the link function.
Likelihood Equations

- Exploiting the above described chain rule we obtain what follows:

\[
\frac{\partial L_i}{\partial \beta_j} = y_i - \mu_i \frac{a(\phi)}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}.
\] (30)

- The likelihood equations are

\[
\sum_i \left( y_i - \mu_i \right) x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \mu_i}{\partial \eta_i} = 0.
\] (31)

Although the vector of $\beta_j$ does not appear, it is there implicitly through $\mu_i$, since $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$. 

Likelihood Equations for Binomial GLM 1/2

- Considering that \( n_i Y_i \) has a \( \text{bin}(n_i, \pi_i) \) distribution, \( y_i \) is the sample proportion of successes for \( n_i \) trials. In the case of several predictors we have what follows

\[
\pi_i = \Phi\left( \sum_j \beta_j x_{ij} \right),
\]

(32)

where \( \Phi \) is a generic cdf of some class of continuous distributions.

- Since \( \pi_i = \mu_i = \Phi(\eta_i) \) with \( \eta_i = \sum_j \beta_j x_{ij} \), we can highlight that

\[
\frac{\partial \mu_i}{\partial \eta_i} = \phi(\eta_i) = \phi \left( \sum_j \beta_j x_{ij} \right)
\]

(33)
Since $\text{var}(Y_i) = \pi_i(1 - \pi_i)/n_i$, the likelihood equation (31) simplify to

$$\sum_i n_i (y_i - \pi_i) x_{ij} \frac{\pi_i(1 - \pi_i)}{\pi_i(1 - \pi_i)} \phi \left( \sum_j \beta_j x_{ij} \right) = 0. \quad (34)$$

where $\pi_i = \Phi \left( \sum_j \beta_j x_{ij} \right)$.

For the logit link, $\eta_i = \log[\pi_i/(1 - \pi_i)]$, so $\partial \eta_i / \partial \pi_i = 1/[\pi_i(1 - \pi_i)]$ and $\partial \mu_i / \partial \eta_i = \pi_i(1 - \pi_i)$. Then, the likelihood of equation (31) and (34) simplify to

$$\sum_i n_i (y_i - \pi_i) x_{ij} = 0, \quad (35)$$

where $\pi_i$ satisfies equation (32) with $\Phi$ the standard logistic cdf.
Fitting GLM through Newton-Rapson and Fisher Scoring Methods

The Newton-Rapson approach is an iterative method for solving nonlinear equations. In more detail, in this context, Newton-Rapson method is exploited to obtain the value $\hat{\beta}$ at which the function $L(\beta)$ is maximized.

Let $u' = (\partial L/\partial \beta_1, \ldots, \partial L/\partial \beta_p)$ and considering the Hessian matrix $H$, we use the notation $u_t$ and $H_t$ to consider the $t$ evaluation for $\hat{\beta}$.

Considering the Taylor series expansion

$$L(\beta) = L(\beta_t) + u'_t(\beta - \beta_t) + \frac{1}{2}(\beta - \beta_t)'H(\beta - \beta_t).$$ (36)

Fisher scoring differs from Newton-Rapson because of the use of the expected information (Hessian matrix) instead of the observed information matrix.
## Dataset

<table>
<thead>
<tr>
<th>Risposta</th>
<th>Liquid</th>
<th>gg_credito</th>
<th>ROA</th>
<th>utiliz_accord</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.96</td>
<td>148</td>
<td>4.07</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1.45</td>
<td>73</td>
<td>1.29</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.94</td>
<td>185</td>
<td>3.80</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.52</td>
<td>125</td>
<td>3.07</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.88</td>
<td>70</td>
<td>4.97</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.75</td>
<td>109</td>
<td>6.45</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.23</td>
<td>54</td>
<td>6.23</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>456</td>
<td>1</td>
<td>1.07</td>
<td>145</td>
<td>6.25</td>
</tr>
<tr>
<td>457</td>
<td>1</td>
<td>0.54</td>
<td>88</td>
<td>9.71</td>
</tr>
</tbody>
</table>
Logit Regression - One Regressor

def<-as.data.frame(read.csv("110207_logit.csv", header = TRUE, sep = ";", dec="."))
#
log.regr.Liquid<- glm(formula= Risposta ~ Liquid, data=def, family=binomial(link=logit))
summary(log.regr.Liquid)
A Practical Estimation of PD Through R GLM

GLM Regression with R

Output of Logit Regression - One Regressor 1/2

Deviance Residuals:

    Min 1Q Median 3Q Max
-1.0199 -0.6164 -0.4866 -0.3741 2.6198

Coefficients:

    Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3008  0.3035  -0.991 0.322
Liquid        -2.0381  0.4437  -4.593 4.36e-06 ***
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
Null deviance: 394.75 on 456 degrees of freedom  
Residual deviance: 370.45 on 455 degrees of freedom  
AIC: 374.45  

Number of Fisher Scoring iterations: 5  

str(log.regr.Liquid)  
# in order to know all about the structure of log.regr.
ANOVA

- This looks very much like the summary of a (non-generalised) linear oneway ANOVA model, except there is no goodness-of-fit (R2).

- Instead, we are given the null deviance, which measures the variability of the dataset, compared to the residual deviance, which measures the variability of the residuals, after fitting the model.

- These deviances can be used like the total and residual sum of squares in a linear model to estimate the goodness of fit; this is sometimes referred to as the D2 (by analogy with R2)

\[
D2 <- \text{function(mod)} \\
\{1-(\text{deviance(mod)}/\text{mod}$null.deviance)\}
\]

D2.log.regr.Liquid <- D2(log.regr.Liquid)

0.06156855
Model Fitting - One Regressor

Success of logistic model

Figure: Logit plot for log.regr.Liquid.
Logit Regression - Multiple Regressors

```r
log.regr.reduced <- glm(formula = Risposta ~ Liquid + gg_credito + ROA + utiliz_accord, data = def, family = binomial(link = logit))
#
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.95341 -0.34985 -0.14296 -0.04294  3.28233

Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) -8.448299   1.240919  -6.808  9.89e-12 ***
    Liquid  -2.447493   0.806733  -3.034  0.002415 **
gg_credito   0.012330   0.003527   3.496  0.000473 ***
      ROA   -0.128378   0.035866  -3.579  0.000344 ***
utiliz_acco  9.668063   1.354553   7.137  9.51e-13 ***
#
D2.log.regr.reduced = 0.4840329
```
Model Fitting - Multiple Regressors

Success of logistic model

Model: Risposta ~ Liquid + gg_credito + ROA + utiliz_accord
AIC: 214
Null deviance: 395

Figure: Logit plot for log.regr.reduced.
Thresholds

- In the present example we have 457 observations, with predicted probabilities of change from almost zero to 1. If we select a threshold of $p = 0.5$ (change equally likely or not), 56 (of 457) are predicted to default. If the threshold is raised to 0.65, only 36 are predicted to default. In fact, 71 defaulted:

```r
length(log.regr.reduced$fitted)
457
summary(log.regr.reduced$fitted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
7.417e-07 4.782e-03 2.899e-02 1.554e-01 1.810e-01 1.000e+00
sum(log.regr.reduced$fitted > 0.5)
56
sum(log.regr.reduced$fitted > 0.65)
36
sum(def$Risposta)
71
```
Sensitivity

- At any threshold we can compute the sensitivity and specificity, by comparing the predicted with actual change.
- The sensitivity is defined as the ability of the model to find the *positives* as follows

\[
\text{Sensitivity} = \frac{\text{True positives}}{\text{Total positives}}.
\] (37)

- For example, at \( p = 0.5 \), this model predicts 45 of the 71 defaults

```r
> sum((log.regr.reduced$fitted > 0.5) & def$Risposta)
45
sens.5 <- sum((log.regr.reduced$fitted > 0.5) & def$Risposta)/sum(def$Risposta)
0.6338028
```
Specificity

- There is another side to a model performance: the specificity, defined as the proportion of *negatives* that are correctly predicted.

\[
\text{Specificity} = \frac{\text{True negatives}}{\text{Total negatives}}. \tag{38}
\]

- \(\text{sum}(!\text{def}$\text{Risposta})\) \\
  386 \\
- \(\text{sum}(\text{log.regr.reduced$fitted < 0.5})\) \\
  401 \\
- \(\text{sum}((\text{log.regr.reduced$fitted < 0.5}) \& (\text{!def}$\text{Risposta}))\) \\
  375 \\
- \(\text{spec.5 <- sum}((\text{log.regr.reduced$fitted < 0.5}) \& (\text{!def}$\text{Risposta})/\text{sum}(!\text{def}$\text{Risposta})\)
  0.9715026
False Negative Rate and False Positive Rate

- The complement of the sensitivity is the false negative rate, that is, the proportion of incorrect predictions of no change to the total changed. This and the sensitivity must sum to 1.
- The complement of the specificity is the false positive rate, that is, the proportion of incorrect predictions of change to the total unchanged. This and the specificity must sum to 1.
Sensitivity vs Specificity: Threshold 0.2

Figure: Logit plot for log.regr.reduced. Sensitivity vs specificity.
Sensitivity vs Specificity: Threshold 0.5

Model success

False negatives: 26
True positives: 45
True negatives: 375
False positives: 11
Threshold = 0.5

Sensitivity: 0.6338 ; Specificity: 0.9715

Figure: Logit plot for log.regr.reduced. Sensitivity vs specificity.
Sensitivity vs Specificity: Threshold 0.8

Model success

False negatives: 46
True positives: 25
threshold = 0.8

True negatives: 385
False positives: 1

Sensitivity: 0.3521 ; Specificity: 0.9974

Figure: Logit plot for log.regr.reduced. Sensitivity vs specificity.
A graph of the sensitivity, i.e. true positive rate (on the y-axis) vs the false positive rate (on the x-axis) at different thresholds is called the Receiver Operating Characteristic (ROC) curve.

Ideally, even at low thresholds, the model would predict most of the true positives with few false positives, so the curve would rise quickly from (0, 0).

The closer the curve comes to the left-hand border and then the top border of the graph (ROC space), the more accurate is the model; i.e. it has high sensitivity and specificity even at low thresholds.

The closer the curve comes to the diagonal, the less accurate is the model. This is because the diagonal represents the random case.
AUC

- The ROC curve can be summarized by the area under the curve (AUC), computed by the trapezoidal rule (base times the median altitude)

\[ AUC = \sum_{i} [x_{i+1} - x_{i}][(y_{i+1} + y_{i})/2] \]  

(39)

where the \( i \) are the thresholds where the curve is computed.

- Note that the area under the diagonal is 0.5, so the ROC curve must define an area at least that large.

- The ROC area then measures the discriminating power of the model: the success of the model in correctly classifying sites that did and did not actually change.

- The closer the curve comes to the diagonal, the less accurate is the model. This is because the diagonal represents the random case.
ROC and AUC: One Regressor

Sensitivity: 0.507 ; Specificity: 0.7694

False negatives: 35 True positives: 36
False positives: 89 True negatives: 297

Threshold = 0.2

Area under ROC: 0.6879

Figure: One regressor analysis.
ROC and AUC: Multiple Regressor

Figure: Multiple regressors analysis.
Conclusions

- We introduced credit risk factors.
- We summarized how to estimate EAD and LGD.
- We analyzed in more detail PD considering alternative approaches.
- We introduced GLM and $R$ software to estimate PDs on real data.
References