A Study on the Default Determination of Residential Mortgages: 
The Application of Bayes’ Theorem on Classification Adequacy

Hsien-Chueh Peter Yang 1, Tsoyu Calvin Lin 2, Tsung-Hao Chen 3

1 Department of Risk Management and Insurance, National Kaohsiung First University of Science and Technology, Kaohsiung 811, Taiwan
2 Corresponding author, Department of Public Finance and Taxation, National Taichung Institute of Technology, 404, Taiwan. Address: #129, San-Min Rd., Sec. 3, Taichung, 404, Taiwan. E-mail: calvinlin168@gmail.com, Tel: 886-4-22196173, Fax: 886-4-22196171.
3 Doctoral Program in Management, National Kaohsiung First University of Science and Technology, Kaohsiung 811, Taiwan.

Abstract

Many banks lending activities, from retail to small business financing, requires some collateral. Banks’ loss resulting from default can also be evaluated through the collateral value. On the one hand, banking industries need to promote the lending business to increase revenue; on the other, they seek to lend secured loans to avoid loss. Therefore, the prediction of default behavior plays an important role in mortgage lending activities for banking industries.

This study attempts to explore the significant factors affecting default of residential mortgages, and find out the appropriate cutoff point in the binary logistic regression model (as compared with traditional cutoff point of 0.5). Empirical results show that the mortgage contract rate, LTV, low education level, the numbers of cash card and unemployment rate have significant positive effects on default. The economic growth rate has significant negative impact on default.

Through the screening tests (sensitivity, specificity, positive predictivity, and negative predictivity, respectively), this study chose an optimal cutoff point of 0.8 where there is crossover between positive predictivity and negative predictivity curves, instead of traditional sensitivity and specificity approaches. We finally conclude that the ability of classification accuracy by the positive predictivity and negative predictivity approaches are higher than traditional sensitivity and specificity approaches.

Key words: default, residential mortgages, sensitivity; specificity; positive predictivity; negative predictivity, Bayes’ Theorem
1. Introduction

Residential mortgage loans differ from other types of loans in several respects. First, mortgage loans are usually long term. Second, mortgage loans are usually secured by the real estate as collateral. Third, mortgage loans are relatively large monetary amounts. Many factors influence the default behavior of residential mortgages. Lawrence et al. (1992) concluded that mortgagors’ credit history and age, contract maturity, loan to value (LTV) ratio and the ratio of mortgage payments to family income are crucial to the likelihood of default. Deng et al. (1996, 1997, and 2000) showed that the present value of mortgage payments, characteristics of family, LTV ratio, home equity, unemployment rates, and divorce rates are significant to mortgage default. Kau and Keenan (1999) concluded that LTV and market housing price are significantly factors affecting the default likelihood of mortgage default. Lin (2004) showed LTV ratio, payment to income (PTI) ratio, regional variation, unemployment, divorce rates, distinctive decline of macro economy (e.g., financial crisis), changes in interest rates and housing prices are critical to mortgage default in the empirical experiences in Taiwan.

This study attempts to explore the significant factors affecting default of residential mortgages, and find out the appropriate cutoff point in the binary logistic regression model (as compared with the traditional cutoff point of 0.5). The remaining part of this paper is organized as follows. Section 2 introduces theoretical background for understanding the screening test of logistic regression model. Section 3 is devoted to modify the traditional analytical methods of sensitivity and specificity applied by the positive predictivity and negative predictivity formulas from the Bayes’ theorem. Finally, the fitness of the data and a more complete description of classification accuracy are assessed on the default determination of residential mortgages.
2. Research Framework

Logistic regression is a commonly used procedure when analyzing data with a binary target variable. Binary data are generally the most common categorical data form. The response variable for a typical logistic regression is dichotomous (Walker, 2002). The binary logistic regression model allows many observed factors to impact dependent variables.

The density function for a logistic regression is as follows (Myers, 2002):

\[ f(z) = \frac{\exp\left(\frac{z - \mu}{\tau}\right)}{c[1 + \exp\left(\frac{(z - \mu)}{\tau}\right)]}, \quad -\infty < z < \infty, \text{ with mean of } \mu \text{ and variance } \frac{\pi^2 \tau^2}{3} \] (1)

The probability of logistic regression function is the cumulative distribution function:

\[ P = \int_{-\infty}^{x} f(z)dz = \frac{\exp\left(\frac{x - \mu}{\tau}\right)}{1 + \exp\left(\frac{(x - \mu)}{\tau}\right)}. \] (2)

which can also be listed as follows (Collect, 2003):

\[ P = \frac{\exp[\beta_0 + \beta_1 x]}{1 + \exp[\beta_0 + \beta_1 x]}, \text{ where } \beta_0=-\mu/\tau \text{ and } \beta_1=1/\tau \] (3)

The binary logistic regression model can thus be written as

\[ \pi_i = E(Y_i) = p = \frac{e^{f(x)}}{1 + e^{f(x)}} \] (4)

\[ Z_i = f(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \] (5)

For a binary response (Y) and the explanatory variables (X) = (x1, x2, x3, ..., xk), π1 (X) represents “success” probability. To simplify the notation, this study uses πi(X)=E(Yi |x) to represent the conditional mean of Y given x when a logistic distribution is utilized.

\[ \pi_i = E(Y_i) = P = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}, \quad \frac{1}{1 + e^{-Z_i}} \] (6)

\[ 1 - \pi_i = 1 - E(Y_i) = 1 - P = 1 - \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}, \quad \frac{1}{1 + e^{-Z_i}} = \frac{e^{-Z_i}}{1 + e^{-Z_i}} \] (7)
such that

\[
\frac{\pi_i}{1-\pi_i} = e^{Z_i} = e^{(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})}
\]  \hspace{1cm} (8)

A transformation of \(\pi(x_i)\), which is central to this study of logistic regression, is the logit transformation. This transformation is defined as

\[
Z_i = \ln\left(\frac{E(Y_i)}{1 - E(Y_i)}\right) = \ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}
\]  \hspace{1cm} (9)

### 2.1 Screening tests

Logistic regressions are frequently employed for predicting for new observations. A classification table uses the logistic regression model to classify observations for occurrence of an event and the measurement of its predictive accuracy. Two basic indices to evaluate diagnostic procedures for the accuracy of the classification are “sensitivity” and “specificity”. “Sensitivity” is the proportion of event responses that were predicted to be events. “Specificity” is the proportion of nonevent responses that were predicted to be nonevents. Limitations of this classification table are: It transforms continuous predictive values into binary one. The choice of \(\pi_0\) is arbitrary, and it is highly sensitive to the relative numbers of times \(y=1\) and \(y=0\) (Agresti, 2002).

Le (1998) proposed the sample of sensitivity and specificity:

\[
\text{Sensitivity} = \frac{\text{number of diseased individuals who screen positive}}{\text{total number of diseased individuals}} = P_i(\text{Test } = + | \text{Diseased } = +)
\]  \hspace{1cm} (10)

whereas

\[
\text{Specificity} = \frac{\text{number of healthy individuals who screen negative}}{\text{total number of healthy individuals}} = P_i(\text{Test } = - | \text{Disease } = -)
\]  \hspace{1cm} (11)
2.1.1 The Risk of Using 0.5 as the Cutoff Point

The difficulty in predicting the category of a dichotomous variable is in determining the cutoff point $\pi_0$, and then compares each estimated probability to $\pi_0$. If the estimated probability exceeds $\pi_0$, then we assume the derived variable to be 1; otherwise it is assumed 0. The most commonly used value for $\pi_0$ is 0.5 (Hosmer and Lemeshow, 2000).

This approach is acceptable while there is equal likelihood of outcome of 0 and 1 in the population. The costs of incorrectly predicting 0 and 1 are approximately the same (Neter et al., 1999). If there is different occurrence likelihood of both sides, however, then using 0.5 as the cutoff point may lead to mis-prediction.

2.1.2 Searching for the Best Cutoff Point

The classification results of both “sensitivity” and “specificity” relies on a single cutoff point for determination. Therefore, it is critical to choose an appropriate cutoff point. A more complete description of classification accuracy is to choose an optimal cutoff point with the lowest probability of incorrect prediction. In other words, a good cutoff point should be able to maximize both the “sensitivity” and “specificity”. This selection procedure can be explained through the graph shown in Figure 1. An optimal choice for a cutoff point should approximately lie on where the sensitivity and specificity curves cross over. This approach is reasonable when the data set is selected from a random sample in the relevant population, which can reflect the proper proportions of 0s and 1s in
the population. The costs of incorrectly predicting 0 and 1 would approximately be the same (Neter et al., 1999).

Figure 1  Plot of Sensitivity and Specificity versus all Possible Cutoff Points

2.1.3 Positive Predictivity and Negative Predictivity

Conditional probability ("sensitivity" and "specificity") takes into account of the information under the occurrence of one event to predict the probability of another event. Therefore, “sensitivity” is a ratio of the number correctly classified to the total number of events. However, given the model for predicting, what is the probability that the event will occur? The concept of conditional probability can be extended to revise probability with new information and to determine the probability of a particular effect resulted from a specific case. The procedure for revising the “sensitivity” and “specificity” is called Bayes’ Theorem, which is a valid tool for assessing how probable evidence makes some hypothesis (Swinburne, 2005). Bayes' Theorem shows the procedures to update the probability in light of new evidence, which relates the conditional and marginal probability distributions of random variables.
Two formulas, “positive predictivity” and “negative predictivity”, from the Bayes’ Theorem are derived from sensitivity, specificity and default prevalence. Positive predictivity and negative predictivity can be written as:

\[
\text{Positive Predictivity} = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid \bar{D})P(\bar{D})} = P(D \mid Test = diseased)
\]  

and

\[
\text{Negative Predictivity} = \frac{P(\bar{T} \mid \bar{D})P(\bar{D})}{P(\bar{T} \mid \bar{D})P(\bar{D}) + P(\bar{T} \mid D)P(D)} = P(D \mid Test = healthy)
\]

As discussed in section 2.1.2, two traces of the positive predictivity and negative predictivity versus all possible cutoff points of the test are shown in Figure 2. One may select a cutoff point that maximizes both positive predictivity and negative predictivity. Namely, an optimal choice for a cutoff point might be the exact point where positive predictivity and negative predictivity curves cross over.
Figure 2 Plot of Positive Predictivity and Negative Predictivity versus all Possible Cutoff points

3 Analytical Results

Mortgage data collected in this study are all individual residential loans originated in 1985 with maturities of 20 years. Data are collected from a local bank in Taiwan, including 277 defaults, and 2381 normal-performing records during the observance period. The censoring time is the end of 2005, with all mortgages been terminated (either paid off or default).

3.1 Binary logistic regression

For a binary logistic regression model, the dependent variables are classified into two groups, default and normal-performed records.

3.1.1 Parameter estimate

There are many factors influencing the default of residential mortgages. As fitting a model, several problems may affect the lack-of-fit, such as a model short of important explanatory variables (Collect, 2003). Allison (1999) suggests that lack-of-fit may cause “lack of independence of the observations.” To overcome these shortcomings, mortgage contract rate($X_1$), loan to value (LTV) ($X_2$), education ($X_3$), regional variation($X_4$), household income($X_5$), the numbers of cash card($X_6$),
economic growth rate \((X_7)\) and unemployment rate \((X_8)\) are considered in the regression model according to the aforementioned literature conclusions and practical experiences.

Table 1 presents analytical results of maximum likelihood estimates. Most of them are consistent with expected influence on default. The mortgage contract rate, LTV, low education level, the numbers of cash card and unemployment rate have significant positive effects on default. The economic growth rate has a significant negative relation to default. The regional variation and household income are not significant.

Table 1 also shows the coefficient of the reduced model without significant factors, in which all of the explanatory variables have significant impact on default of residential mortgages.

The fitted logistic regression model can be listed as follows.

\[
\ln \left( \frac{\hat{p}}{1 - \hat{p}} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8
\]

\[
= -11.1922 + 1.5891 x_5 + 0.0653 x_3 + 1.1254 x_5 + 0.0456 x_6 - 0.0915 x_7 + 0.5143 x_8
\]

**Table 1**  Results of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficients</th>
<th>P-value</th>
<th>Coefficients</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-10.7955</td>
<td>&lt;.0001***</td>
<td>-11.1922</td>
<td>&lt;.0001***</td>
</tr>
<tr>
<td>Flexible Mortgage contract rate</td>
<td>1.5949</td>
<td>&lt;.0001***</td>
<td>1.5891</td>
<td>&lt;.0001***</td>
</tr>
<tr>
<td>LTV</td>
<td>0.0639</td>
<td>&lt;.0001***</td>
<td>0.0653</td>
<td>&lt;.0001***</td>
</tr>
<tr>
<td>Education</td>
<td>1.1234</td>
<td>&lt;.0001***</td>
<td>1.1254</td>
<td>&lt;.0001***</td>
</tr>
<tr>
<td>Regional variation</td>
<td>0.00134</td>
<td>0.9823</td>
<td></td>
<td></td>
</tr>
<tr>
<td>household income</td>
<td>-0.2558</td>
<td>0.1304</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The numbers of cash card</td>
<td>0.0423</td>
<td>0.0672*</td>
<td>0.0456</td>
<td>0.0454**</td>
</tr>
<tr>
<td>Economic growth rate</td>
<td>-0.0976</td>
<td>0.0310**</td>
<td>-0.0915</td>
<td>0.0387**</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>0.5115</td>
<td>&lt;.0001***</td>
<td>0.5143</td>
<td>&lt;.0001***</td>
</tr>
</tbody>
</table>

Note * means significance of 10%; ** significance of 5%; *** significance of 1%
Figure 3 shows the graph of fitted logistic regression model (equation 14). The logistic transformation of a success probability $p$ is $\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right)$, which is written as $\logit(\hat{p})$. The vertical axis is labeled as:

$$
\logit(\hat{p}) = \ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6.
$$

This equation contains estimates of transformation values, and the horizontal axis labels all probabilities of predictive values $\hat{p}$. The graph indicates all the values of $p$ in the range $(0, 1)$ corresponding to the value of $\logit(\hat{p})$ in $(-\infty, \infty)$.

![Figure 3](image_url)

**Figure 3** Plot of Individual Explanatory Variables from $X_1$ to $X_6$ versus All Probabilities $\hat{p}$

### 3.1.2 Goodness-of-Fit Statistics

Using the counterparts of Pearson Chi-square and deviance Chi-square tests, Pearson Chi-square is distributed as Chi-square with the degree of freedom of $(r-1)(s-1)-t$, where $t$ is the number of explanatory variables, $r$ is the number of response levels, and $s$ is the number of subpopulations. Table 2 presents the goodness-of-fit statistics. The estimate of deviance, labeled as Value/DF, contains a dispersion parameter (value/DF) of 0.6404 and a Pearson Chi-square dispersion parameter of 0.9814. The statistic values of Pearson Chi-square and deviance Chi-square are 1134.7353 and 1739.1269, respectively, with 1772 ($=(2-1) \times (1779-1)-7$) degrees of freedom.
freedom. The deviance and Pearson Chi-square are smaller than the degrees of freedom; meanwhile, the P-values for deviance and Pearson chi-square are all greater than 0.05 (1.0000 and 0.7069, respectively). Therefore, this model seems to have an acceptable goodness of fit for the data.

Table 2  Goodness-of-Fit Statistics: Explanatory Variable X₁

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
<th>Pr &gt; Chi-Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>1772</td>
<td>1134.7353</td>
<td>0.6404</td>
<td>1.0000</td>
</tr>
<tr>
<td>Pearson</td>
<td>1772</td>
<td>1739.1269</td>
<td>0.9814</td>
<td>0.7069</td>
</tr>
</tbody>
</table>

Number of unique profiles: 1779

3.2 Screening Tests

3.2.1 Using 0.5 as the Cutoff Point

Table 3 shows the classification results of data through the use of 0.5 cutoff point in the logistic regression in this study. The overall rate of correct classification is estimated as 91.16% (=[(75+2348)/2658]), with only 27.08 % (75/277) of the default group (sensitivity) and 98.61 % (2348/2381) of the normal-performing group (specificity) being correctly classified. That is, using 0.5 as the cutoff point results in the test of high specificity (98.61 %) but low sensitivity (27.08%); there is more than half (72.92%) false default.

Table 3  Classification Result of the Logistic Regression Model in Table 1

---using a Cutoff Point of 0.5.

<table>
<thead>
<tr>
<th>Real Result</th>
<th>Test result</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>default</td>
<td>normal</td>
</tr>
<tr>
<td>Default</td>
<td>75</td>
<td>202</td>
</tr>
<tr>
<td>Normal</td>
<td>33</td>
<td>2348</td>
</tr>
<tr>
<td>Total</td>
<td>108</td>
<td>2550</td>
</tr>
</tbody>
</table>
3.2.2 In Search of the Optimal Cutoff Point

3.2.2.1 Sensitivity and Specificity

In order to choose an optimal cutoff point for classification, one may select a cutoff point approximately where the sensitivity and specificity curves cross over. Table 4 shows the default classification results with the range of probabilities of 0.05 increments. The trend is plotted in Figure 4.

From Table 4 and Figure 4, the optimal cutoff point may be found as 0.10, where there is both the highest sensitivity (74%) and specificity (78.50%), respectively.

Figure 4  Plot of Sensitivity and Specificity versus all Possible Cutoff Points

3.2.2.2 Positive Predictivity, and Negative Predictivity

Since the cost of incorrectly predicting default variables (positive predictivity) differs substantially from the cost of incorrectly predicting normal-performing variables (negative predictivity) for residential mortgages. The Bayes’ Theorem shows how to revise the classification of sensitivity
and specificity. Furthermore, as the cutoff point ($\pi_0$) is fixed, Le (1998) indicated that if a test is applied to a target population of low default prevalence, the positive predictivity will be low. Hosmer and Lemeshow (2000) further proposed that classification is sensitive to the relative sizes of the two component groups, and the large group is always preferred.

To solve these problems, this study employs an optimal approach by combining Bayes’ Theorem and the optimal cutoff point to maximize both positive predictivity and negative predictivity. As shown in Table 4, an optimal cutoff point of 0.8 is chosen, where the positive predictivity and negative predictivity curves cross over. The classification leads to both high positive predictivity (89.66%) and negative predictivity (90.45%), as shown in Figure 5.

Using a cutoff point of 0.1 as shown in Table 4, where the curves of sensitivity and specificity meet, the result of test is 74.00% sensitive, and 78.50% specific, with high negative predictivity (96.29%) but low positive predictivity (28.63%). The conclusion is clear---the point where the slopes of positive predictivity curve and negative predictivity curve are equal is preferred to be the interception point between sensitivity and specificity curves.
Figure 5  Plot of Sensitivity, Specificity, Positive Predictivity, and Negative Predictivity versus all Possible Cutoff Points
Table 4  Summary of Sensitivity, Specificity, Positive Predictivity, and Negative Predictivity for Classification, based on the Logistic Regression Model using a Cutoff Point of 0.01 to 1.00 with Increments of 0.05

<table>
<thead>
<tr>
<th>Cutoff point</th>
<th>correct sensitivity</th>
<th>sensitivity</th>
<th>specificity</th>
<th>positive predictive value</th>
<th>negative predictive value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>20.50</td>
<td>99.60</td>
<td>11.30</td>
<td>11.56</td>
<td>99.63</td>
</tr>
<tr>
<td>0.05</td>
<td>58.10</td>
<td>88.10</td>
<td>54.60</td>
<td>18.40</td>
<td>97.52</td>
</tr>
<tr>
<td>0.10</td>
<td>78.10</td>
<td>74.00</td>
<td>78.50</td>
<td>28.63</td>
<td>96.29</td>
</tr>
<tr>
<td>0.15</td>
<td>84.30</td>
<td>60.30</td>
<td>87.10</td>
<td>35.16</td>
<td>94.96</td>
</tr>
<tr>
<td>0.20</td>
<td>87.20</td>
<td>55.20</td>
<td>90.90</td>
<td>41.46</td>
<td>94.58</td>
</tr>
<tr>
<td>0.25</td>
<td>88.90</td>
<td>48.00</td>
<td>93.70</td>
<td>51.98</td>
<td>93.46</td>
</tr>
<tr>
<td>0.30</td>
<td>89.90</td>
<td>42.60</td>
<td>95.40</td>
<td>57.98</td>
<td>93.20</td>
</tr>
<tr>
<td>0.35</td>
<td>90.70</td>
<td>39.40</td>
<td>96.70</td>
<td>63.40</td>
<td>92.81</td>
</tr>
<tr>
<td>0.40</td>
<td>91.10</td>
<td>35.00</td>
<td>97.60</td>
<td>65.89</td>
<td>92.41</td>
</tr>
<tr>
<td>0.45</td>
<td>91.10</td>
<td>30.70</td>
<td>98.20</td>
<td>69.44</td>
<td>92.08</td>
</tr>
<tr>
<td>0.50</td>
<td>91.20</td>
<td>27.10</td>
<td>98.60</td>
<td>71.26</td>
<td>91.64</td>
</tr>
<tr>
<td>0.55</td>
<td>91.00</td>
<td>22.40</td>
<td>99.00</td>
<td>73.68</td>
<td>91.44</td>
</tr>
<tr>
<td>0.60</td>
<td>90.90</td>
<td>20.20</td>
<td>99.20</td>
<td>75.41</td>
<td>91.11</td>
</tr>
<tr>
<td>0.65</td>
<td>90.70</td>
<td>16.60</td>
<td>99.40</td>
<td>79.59</td>
<td>90.88</td>
</tr>
<tr>
<td>0.70</td>
<td>90.70</td>
<td>14.10</td>
<td>99.60</td>
<td>84.62</td>
<td>90.68</td>
</tr>
<tr>
<td>0.75</td>
<td>90.60</td>
<td>11.90</td>
<td>99.70</td>
<td>89.66</td>
<td>90.45</td>
</tr>
<tr>
<td>0.80</td>
<td>90.40</td>
<td>9.40</td>
<td>99.90</td>
<td>94.12</td>
<td>90.12</td>
</tr>
<tr>
<td>0.85</td>
<td>90.10</td>
<td>5.80</td>
<td>100.00</td>
<td>100.00</td>
<td>89.95</td>
</tr>
<tr>
<td>0.90</td>
<td>90.00</td>
<td>4.00</td>
<td>100.00</td>
<td>100.00</td>
<td>89.82</td>
</tr>
<tr>
<td>0.95</td>
<td>89.80</td>
<td>2.50</td>
<td>100.00</td>
<td>100.00</td>
<td>89.82</td>
</tr>
<tr>
<td>1.00</td>
<td>89.60</td>
<td>0.00</td>
<td>100.00</td>
<td>-</td>
<td>89.58</td>
</tr>
</tbody>
</table>

4. Conclusions

Residential mortgage loans differ from other types of loans in several aspects. They are usually long term, secured by the real estate as collateral, and relatively large monetary amounts. Many factors have significant impact on the default of residential mortgages, such as LTV, education level, economic growth rate and unemployment rate. The cost of incorrectly predicting default events differs substantially from the cost of incorrectly predicting normal-performing variables for
residential mortgages. This study intends to explore the optimal cutoff point in the Logistic regression model to reduce the cost of incorrect determination. Conclusions of this study are briefly discussed as follows.

(1) While analyzing the Maximum Likelihood Estimates, the mortgage contract rate, LTV, education level, the numbers of cash card and unemployment rate have negative significant effects on default. The economic growth rate is negatively related to default with significance.

(2) The goodness-of-fit of the model was examined by the deviance and Pearson Chi-square tests. The deviance and Pearson Chi-square are smaller than the degrees of freedom; meanwhile, the P-values for deviance and Pearson Chi-square are all greater than 0.05. The test results indicate that the model is applicable.

(3) We use a cutoff point of 0.1 where the curves of sensitivity and specificity meet to reduce the possibilities of incorrect classification. The test result is 74.00% sensitive and 78.50% specific, which leads to a high negative predictivity (96.29%) but a low positive predictivity (28.63%).

(4) We therefore advocates an optimal approach by combining Bayes’ theorem and the optimal cutoff point of 0.8, where positive predictivity and negative predictivity curves cross over. The classification result shows both high positive predictivity (89.66%) and high negative predictivity (90.45%). We thus conclude that the point where the slope of positive predictivity curve and negative predictivity curve meet is preferred to the optimal cutoff point for both sensitivity and specificity curves.
References:


