International Journal of Strategic Property Management
ISSN: 1648-715X / eISSN: 1648-9179
2019 Volume 23 Issue 2: 130–141
https://doi.org/10.3846/ijspm.2019.7064

A COMPARISON OF THE PREDICTIVE POWERS OF TENURE CHOICES BETWEEN PROPERTY OWNERSHIP AND RENTING

Chun-Chang LEE[1,*], Chih-Min LIANG[2], Yang-Tung LIU[1]

[1] *Department of Real Estate Management, National Pingtung University, No. 51, Ming Sheng East Road, Pingtung, Taiwan*
[2] *Department of Public Finance and Tax Administration, National Taipei University of Business, No. 321, Sec. 1, Jinan Rd., Zhongzheng District, Taipei City, Taiwan*

Received 14 August 2018; accepted 26 October 2018

Abstract. This paper compares the predictive powers of hierarchical generalized linear modeling (HGLM), logistic regression, and discriminant analysis with regard to tenure choices between buying property and renting property by sampling the residents of the Greater Taipei area. The results imply that the hit rate and other indicators included in HGLM have better predictive power with regard to tenure choices than the binary logistic regression model and the discriminant analysis model. That is, using HGLM to process nested data can increase prediction accuracy regarding household tenure choices. Furthermore, cross-validation is performed to analyze hit rate stability. The hit rate sequencing from this cross-validation is found to be consistent with the HGLM results, implying that the comparison of the three models in terms of hit rate performance prediction in this study is stable and reliable.

Keywords: tenure choice, hit rate, hierarchical generalized linear modeling, logistic regression model, discriminant analysis, cross-validation.

Introduction

Due to the high prices of residential properties, households take a variety of factors into account in their tenure choices vis-à-vis owning or renting such property. The tenure choice between buying and renting is largely concerned with the gap between the given household's living requirements and utility satisfaction and, understandably, constitutes an important topic in the study of housing demands and investments. The majority of studies on tenure choices focus on determinants such as the characteristics of the heads of households (Spalkova & Spalek, 2012; Kim & Jeon, 2012), property attributes (Henley, 1998; Subhan & Ahman, 2012), funding sources for property purchases (Chambers, Garriga, & Schlagenhauf, 2009), social statuses (Tao, Hui, Wong, & Chen, 2015), neighborhood factors (Vera-Toscano & Amestoy, 2008, Lee, Ho, & Chiu, 2016) and policy effects (Carter, 2011; Lang & Hurst, 2013). There are few studies, however, seeking to compare the accuracy of distributions (i.e., the predictive power regarding tenure decisions) between different regression models. At the same time, the construction of an effective predictive model for tenure choices should be an important topic for real estate agents and researchers alike.

The majority of studies on real estate and regression models focus on the predictive power of those models regarding housing prices. Feng and Jones (2015), for example, use hierarchical linear modeling (HLM) and artificial neural networks to predict residential property prices. Their empirical findings suggest that, compared to artificial neural networks, HLM yields better predictive results. There are also studies seeking to predict defaults on the part of mortgage customers (Steenackers & Goovaerts, 1989). There are few studies, however, that compare the predictive powers of different models with regard to tenure choices. Nonetheless, such information could be of great value to governments in formulating policies and forecasting which household types would prefer buying and which household types would prefer renting. It is also essential for households themselves to take into account all the factors that influence their tenure choices. Furthermore, banks can use forecasting tools to determine trends in tenure choices and devise strategies for their mortgage businesses accordingly.

*Corresponding author. E-mail: *lcc@mail.nptu.edu.tw*

The key contributions of this paper are the construction of forecast models for tenure choices and the validation of the predictive accuracy of such models. These issues have not been touched upon much in prior studies. Distinct from the existing literature, this paper compares the predictive results of hierarchical generalized linear modeling (HGLM), the binary logistic regression model (a frequently used forecasting model), and the discriminant analysis model with regard to tenure choices.

1. Literature review

Most of the past studies on the use of forecasting models for real estate are focused on the prediction of housing prices and financial credit ratings. Kontrimas and Verikas (2011) apply ordinary least square (OLS), multilayer perceptron (MLP), and support vector machines (SVM) to estimate property values. Their empirical results suggest that the estimated errors based on SVM are smaller than those for OLS and MLP as far as the use of MAPE, MAE, and unacceptable valuations (UVs) as valuation indicators are concerned. This implies that SVM has superior predictive power compared with OLS and MLP. Feng and Jones (2015) apply HLM and artificial neural networks to forecast housing prices. Their empirical findings indicate that HLM generates better forecasts than artificial neural networks. Tung, Lee, Chen, and Wu (2016) use SVM to forecast housing prices in Taipei City. Their empirical results suggest that SVM has higher predictive power than OLS.

The majority of the studies on tenure choices deal with the explanatory variables that influence the tenure choices of households. These variables include the characteristics of the heads of households, property attributes, funding sources for property purchases, social statuses, neighborhood factors, and policy effects (Lee et al., 2016, 2018). In fact, most of these studies conduct analyses with logistic regression models. Carter (2011) estimates tenure choices with probits, and the results suggest that if endogenous variables are ignored, there will be a bias in the estimated coefficient. Wagner (2014) applies a logistic regression to data sourced from the Household Finance and Consumption Survey (HFCS) in Austria. The results indicate that the probability of becoming a homeowner increases significantly, i.e., by 31 percentage points, for a household with homeowning parents. Chen (2016) uses logits to estimate the heterogeneity of tenure choices. The results show that housing tenure choices depend on subprocesses and socioeconomic differentiation and suggest a need to create housing policies tailored for specific housing groups.

If the research data is nested, the traditional regression approach (without hierarchy) cannot address the coefficient differences of the explanatory variables on the micro level in the characteristic variables on the macro level. In other words, the traditional approach does not take into account the heterogeneity among different administrative zones. This tends to breach the presumption that all the explanatory variables are mutually independent and may result in analytical errors (Raudenbush & Bryk, 2002). Subsequently, some studies have begun to use HLM to analyze hierarchical data that is embedded or nested. Huang and Clark (2002) examine tenure choices in China during the transition into a liberalized market in the 1990s. They employ HGLM as the analytical tool and define the selling prices of residential, business, and commercial properties as the characteristic variables on the second level. They contend that housing prices have an adverse impact on the increase in homeownership. Lee et al. (2016) apply HGLM to explore the influence of the level of regional commercialization, the percentage of park area, and the percentage of school areas on tenure choices. According to their empirical results, a high level of regional commercialization and a high percentage of park areas gear the tenure choices toward purchases, whereas a high percentage of school areas push the tenure choices toward renting. Only a few studies apply HLM to explore the influence of neighborhood characteristics as a variable affecting tenure choices. Such neighborhood characteristics include the level of locals' satisfaction with environmental qualities and with leisure and sports facilities.

Prior studies mostly focus on the influence of explanatory variables by deploying traditional regression techniques. Discriminant analysis is often used to examine the banking industry's capability to review the credit levels of customers. Espahibodi (1991) posits that if the explanatory variables are not in compliance with the assumption for normal distributions, logistic regressions report a higher accuracy than discriminant analysis. Most of the empirical comparisons of the models forecasting mortgage defaults presume no particular limitations on the distribution of explanatory or predictor variables in logistic regressions. If there are both discrete and continuous explanatory variables, the forecast models built on logistic regressions tend to demonstrate higher accuracy levels. In addition to the examination of the factors that influence tenure choices, this paper goes a step further by comparing the predictive accuracy levels of HGLM, logistic regression, and discriminant analysis with regard to tenure choices.

2. Research method

2.1. HGLM

The application of the HGLM method can accurately analyze the fixed effects on dependent variables and estimate intercepts and variances. Meanwhile, it is possible to review whether the influence on dependent variables from the explanatory variables on the first level and the characteristic variables on the second level is statistically significant. This paper defines the dependent variable as binary. The data is organized into a Bernoulli distribution for analysis in the HGLM model as follows:

$$\eta_{ij} = \log(\frac{\varphi_{ij}}{1-\varphi_{ij}}), \tag{1}$$

where: i denotes each interviewed household; j is the administrative zone code; φ_{ij} is the probability of housing purchases; $1-\varphi_{ij}$ is the probability of housing rentals; η_{ij} is the natural logarithm of the odds ratio. If the φ_{ij} probability is 0.5, the natural logarithm of the odds ratio would be $\varphi_{ij} / \left(1-\varphi_{ij}\right) = 0.5/0.5 = 1$. In other words, log (1) = 0. If the φ_{ij} probability is lower than 0.5, the odds ratio would be smaller than 1. If $\eta_{ij} < 0$, it means the households have an inclination toward renting properties. If the φ_{ij} probability is higher than 0.5, the odds ratio is greater than 1. If $\eta_{ij} > 0$, it means the households tend to purchase properties. This paper designs its HLM as an intercepts-as-outcomes model.

An intercept-as-outcomes model uses the intercept estimated with the regression model in the first level as the outcome variable for the second level. The dependent variables and the explanatory variables in the first level are micro in nature, as the first level deals with the relationship between explanatory variables and dependent variables. The characteristic variables on the second level express the direct and cross-level impact on the intercept of the first level. The explanatory variables on the first level include the gender of household head (GENDER), age of household head (PAGE), squared age of household head (PAGES), household head with a senior high school education (HIGHT), household head with a college education (COLLEGE), household head with a post-graduate degree (UNI), age of the property (HAGE), living space per person (PAREA), number of rooms per person (PROOM), number of permanent residents in household (FMSZ), loan-to-value ratio (LOANR), personal borrowing 1 (from NT$ 10,000 to NT$ 500,000) (PMORTGAGE1), personal borrowing 2 (NT$ 500,000 and above) (PMORTGAGE2), and government subsidized loans (PRELOAN). The independent variables on the first level are processed with group mean centering. The characteristic variables on the second level include the level of satisfaction with environmental qualities (ENVI) and the level of satisfaction with leisure and sports facilities (LEIS) (see Table 1 for detailed definitions of the variables). The model is specified as follows:

Level 1:

$$\begin{aligned}\eta_{ij} = {} & \beta_{0j} + \beta_{1j}GENDER + \beta_{2j}PAGE + \beta_{3j}PAGES + \\ & \beta_{4j}HIGHT + \beta_{5j}COLLEGE + \beta_{6j}UNI + \\ & \beta_{7j}HAGE + \beta_{8j}PAREA + \beta_{9j}PROOM + \\ & \beta_{10j}FMSZ + \beta_{11j}LONAR + \beta_{12j}PMORTGAGE_1 + \\ & \beta_{13j}PMORTGAGE_2 + \beta_{14\mathrm{j}}PRELOAN,\end{aligned} \tag{2}$$

where: i denotes each interviewed household; j is the administrative zone code; η_{ij} is the natural logarithm of the odds ratio of the i-th household in the j-th administrative zone; β_{0j} is the mean of the natural logarithm of the odds ratio of all the households in the j-th administrative zone; $\beta_{1j} \sim \beta_{14j}$ is the coefficient of the independent variable on the first level.

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}ENVI + \gamma_{02}LEIS + \mu_{0j}; \tag{3}$$

$$\beta_{pj} = \gamma_{p0}\ , \mathrm{p} = 1,....,14, \tag{4}$$

where: γ_{00} is the natural logarithm of the odds ratio regarding the tenure choices of all the households in a given administrative zone; γ_{01}, γ_{02} is the coefficient of the characteristic variable on the second level; μ_{0j} is the error term for administrative zones, which is assumed to be in a normal distribution with a mean of 0 and a variance of τ_{00}.

2.2. Binary logistic regression and discriminant analysis

The basic mechanism of a logistic regression model is similar to that of traditional linear regressions. However, the dependent variables in a logistic model have to be converted into a probability value between 0 and 1. Hence, a given dependent variable may not conform with the assumption of a normal distribution. This is the major difference between logistic models and traditional linear models. The logistic model in this paper is specified as follows:

$$\varphi_i = F(Z_i); \tag{5}$$

$$\eta_i = \log(\frac{\varphi_i}{1-\varphi_i}) = \frac{1}{1+e^{-Z_i}}; \tag{6}$$

$$\begin{aligned}Z_i = {} & \beta_0 + \beta_1 GENDER_i + \beta_2 PAGE_i + \beta_3 PAGES_i + \\ & \beta_4 HIGHT_i + \beta_5 COLLEGE_i + \beta_6 UNI_i + \\ & \beta_7 HAGE_i + \beta_8 PAREA_i + \beta_9 PROOM_i + \\ & \beta_{10} FMSZ_i + \beta_{11} LONAR_i + \beta_{12} PMORTGAGE_{1i} + \\ & \beta_{13} PMORTGAGE_{2i} + \beta_{14} PRELOAN_i + \\ & \beta_{15} ENVI_i + \beta_{16} LEIS_i,\end{aligned} \tag{7}$$

where: φ_i is the probability of the i-th household purchasing a property; $1\text{-}\varphi_i$ is the probability of that household renting a property; $F(.)$ is the cumulative density function in the logistic model. Eq. (7) is transformed into Eq. (8) as follows:

$$\begin{aligned}\eta_i = {} & \log(\frac{\varphi_i}{1-\varphi_i}) = Z_i = \beta_0 + \beta_1 GENDER_i + \\ & \beta_2 PAGE_i + \beta_3 PAGES_i + \beta_4 HIGHT_i + \\ & \beta_5 COLLEGE_i + \beta_6 UNI_i + \beta_7 HAGE_i + \\ & \beta_8 PAREA_i + \beta_9 PROOM_i + \beta_{10} FMSZ_i + \\ & \beta_{11} LONAR_i + \beta_{12} PMORTGAGE_{1i} + \\ & \beta_{13} PMORTGAGE_{2i} + \beta_{14} PRELOAN_i + \\ & \beta_{15} ENVI_i + \beta_{16} LEIS_i.\end{aligned} \tag{8}$$

The independent variables in Eq. (8) are the same as those in the aforementioned HGLM model. The total of 16 variables include the characteristics of the property, the characteristics of the individual, the funding sources for property purchases, the level of satisfaction with environmental qualities (ENVI), and the level of satisfaction with leisure and sports facilities (LEIS). This paper also takes into account the variances among administrative zones. Songshan District in Taipei is used as the benchmark. A total of 23 sets of dummy variables are set up for all the administrative zones, with 1 used as the value for one of the specific 23 administrative zones and 0 used for the others.

Similar to multiple regressions, discriminant analysis is also one of the most frequently used classification methods. The linear combination of a set of forecasting (discriminant) variables is referred to for the reclassification of a set of variables, so as to inspect the accuracy of grouping.

The focus of the analysis is the construction of a discriminant equation, in order to effectively discriminate different groups of the dependent variable (e.g., buying and renting). Discriminant analysis is based on the presumption that the observed values of the discriminant variables conform with the presumption of multivariate normal distributions. Meanwhile, the significance tests on discriminant equations assume that variances and covariances of discriminant variables for independent variables are homogeneous. It is usually necessary to establish a linear discriminant (or classification) function to calculate the fraction of data units. The discriminant function is specified as follows:

$$\begin{aligned}Z = {} & w_0 + w_1 GENDER_i + w_2 PAGE_i + w_3 PAGES_i + \\ & w_4 HIGHT_i + w_5 COLLEGE_i + w_6 UNI_i + \\ & w_7 HAGE_i + w_8 PAREA_i + w_9 PROOM_i + \\ & w_{10} FMSZ_i + w_{11} LONAR_i + w_{12} PMORTGAGE_{1i} + \\ & w_{13} PMORTGAGE_{2i} + w_{14} PRELOAN_i + \\ & w_{15} ENVI_i + w_{16} LEIS_i ,\end{aligned} \quad (9)$$

where: Z denotes the discriminant function for all types of data; w_i is the weight of individual discriminant variables; w_0 is the constant. The purpose of the discriminant analysis is to identify the individual weights w_i in the function to express the relevant importance or influence of the respective discriminant variables. The discriminant analysis seeks to maximize the ratio of between-group sum of squares (SS_b) to within-group sum of squares (SS_w), i.e., the maximization of the Λ value expressed below:

$$\Lambda = \frac{SS_b}{SS_w} . \quad (10)$$

Discriminant analysis serves the same purposes as regression analysis, i.e., to explain and forecast. It is possible to input all the variable values observed from the trained data in the discriminant function to derive forecasts. This is followed by classifications in order to validate, on an ex post basis, the accuracy of the forecasts.

2.3. Selection and set-up of variables

The explanatory variables in the first level are established on the basis of a literature review. These variables include the gender of the household head, squared age of household head, education background of household head, age of the property, living space per person, number of rooms per person, number of permanent residents in household, loan-to-value ratio, personal borrowing, and government loans at an incentive rate (see Table 1).

Table 1. Variable selections and definitions

Variable	Definition	Expected sign
Explanatory variable on the first level:		
TENU	Tenure choice as a dummy variable, defined as 1 for a decision to buy and as 0 for a decision to rent	
GENDER	Gender of household head as a dummy variable, defined as 1 if male and 0 if female	+
PAGE	Age of household head, a continuous variable	+
PAGES	Squared age of household head, a continuous variable	–
HIGHT	Education of household head in four categories, i.e., junior high school or below, senior high school or vocational school, college, and post-graduate degrees. Junior high school or below referred to as the benchmark for three dummy variables: 1 for senior high school or vocational school and 0 for others	+
COLLEGE	Dummy variable defined as 1 if household head has a college education and as 0 if not	+
UNI	Dummy variable defined as 1 if household head has a post-graduate education and as 0 if not	+
HAGE	Age of the property where the interviewee lives, as a continuous variable	–
PAREA	Living space per person, i.e., number of *ping* (1 *ping* equals 35.58 sq. ft.) divided by number of permanent residents in the property, a continuous variable	+
PROOM	Number of rooms per resident in the property in which the interviewee resides, a continuous variable	+
FMSZ	Number of permanent residents in the property, a continuous variable	+
LOANR	Loan-to-value ratio (loan divided by property value and times 100%) as a dummy variable defined as 1 if greater than 0.5 and as 0 if not	+
PMORTGAGE1	Three categories for personal borrowing in home purchases (including rotating savings): no personal borrowing, personal borrowing between NT$ 10,000 and NT$ 500,000, and personal borrowing greater than NT$ 500,000. PMORTGAGE1 as a dummy variable defined as 1 for personal borrowing between NT$ 10,000 and NT$ 500,000 and as 0 for other categories	+
PMORTGAGE2	PMORTGAGE2 as a dummy variable defined as 1 for personal borrowing of NT$ 500,000 and above and as 0 for other categories	+
PRELOAN	Government subsidized loans offered for home purchases as a dummy variable defined as 1 if available and as 0 if not	+
Characteristic variables on the second level:		
ENVI	Satisfaction with environmental qualities, including five variables (i.e., air pollution, noise, hygiene, garbage pick-up, and drinking water quality). Mean value estimated for each administrative zone	+
LEIS	Satisfaction with leisure and sports facilities, including five variables (i.e., parks, sports centers, libraries or cultural venues, local landscapes, and community beautifications). Mean value estimated for each administrative zone	+

The two characteristic variables on the second level are the level of satisfaction with environmental qualities (ENVI) and the level of satisfaction with leisure and sports facilities (LEIS). Satisfaction has been referred to in a variety of studies as an expression of opinions regarding products, work, living quality, and community or outdoor leisure quality. It is a very useful indicator (Cardozo, 1965). This paper uses the measurements and statistics released by the Construction and Planning Agency, Ministry of the Interior, Taiwan, regarding the levels of satisfaction with environmental qualities and the levels of satisfaction with leisure and sports facilities in the 2006 housing property survey. The subjective perceptions of the interviewees are examined in order to highlight the influence of different levels of satisfaction on tenure choices in individual administrative zones. The survey reviews the levels of satisfaction with five environmental qualities (such as air pollution, noise, hygiene, garbage pick-up, and drinking water quality). The leisure and sport facilities include parks, sports centers, libraries or cultural venues, local landscapes, and community beautifications. The related measurements are based on a Likert scale with five values: extremely satisfied (5 points), very satisfied (4 points), no opinion (3 points), unsatisfied (2 points), and very unsatisfied (1 point). To explore whether the levels of satisfaction affect tenure choices in different administrative zones, this paper calculates the mean of these values for different administrative zones. This paper anticipates that a high level of satisfaction with environmental qualities and a high level of satisfaction with leisure and sports facilities will have positive influences on tenure choices.

2.4. Five indicators to assess forecasting effectiveness

This paper uses five frequently seen indicators to evaluate the predictive ability of the three models. These indicators express the difference between actual values and forecasted values.

2.4.1. Hit rates

A hit rate represents the percentage of a given sample that is being accurately classified. This paper establishes a classification rule function or a set of discriminating rules on the basis of the prior characteristics of the tenure choice variable. The known set of rules is used to validate the tenure choices of all households before they are employed to forecast whether the tenure choices of individual households have been accurately classified.

2.4.2. Receiver operating characteristic curve

A receiver operating characteristic curve (ROC) is an indicator used to evaluate the classification and discrimination power of a binary variable. The size of the area under the ROC serves as an indicator of forecast accuracy (Mossman, 1994; Rice & Harris, 1995). ROC is often used as a scoring system by medical institutions to determine whether a disease cause occurs or not, by transport authorities to forecast traffic flows and journey times, by police and social workers to forecast the repeat of violent crimes or domestic violence, and by financial systems to forecast loan defaults (Quinsey, Harris, Rice, & Cormier, 1998). There are no past studies, however, that have used ROC to examine tenure choices. However, this paper believes that it can serve as a forecasting indicator that is both theoretically sound and effective if the principle is applied to the discrimination and classification of tenure choice as a binary variable.

An ROC is a curve formed with the false alarm rate (1-specifity) on the X axis and the hit rate (sensitivity) on the Y axis to validate the quality of a model. It can serve as a cut-off for home buyers and property renters as two distinctive groups. The X axis defines the false alarm rate (FAR) as the percentage of home buyers erroneously classified as property renters and vice versa. The Y defines the hit rate (HR), i.e., the percentage of home buyers accurately classified as home buyers and the percentage of property renters accurately classified as property renters. The coordinates (X = false alarm rate and Y = hit rate) formed by a sample of households can be depicted as an ROC as in Figure 1. The greater the ROC, the better the cut-off between home buyers and property renters is.

A diagonal line is referred to as the accuracy benchmark for the ROC. If the ROC of the tenure choice model falls on the diagonal line, it implies that the model has no discriminating power over the classification of buyers and renters. The further to the upper left the ROC, the greater the model sensitivity, the lower the false alarm rate, and, hence, the greater the discriminating power of the model. Meanwhile, the area under curve (AUC) can also be used to determine the discriminating power of the model. The value of the AUC is between 0 and 1. If the AUC is smaller than 0.5, it means the classification has no discriminating effectiveness. If the AUC is equal to 0.5, it means the discriminating power is no better than a random guess. An AUC of greater than 0.5 indicates good discriminating power, and an AUC of 1 suggests 100% accuracy in classification and forecasting accuracy.

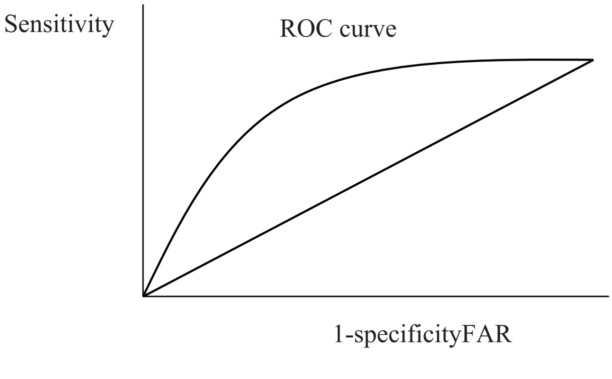

Figure 1. ROC for tenure choices

2.4.3. MAE

Mean average error (MAE) measures the gap between each forecasted value and the actual value. The absolute values of these differences are aggregated and averaged with the number of observations. MAE is expressed with a percentage. The higher the value, the poorer the forecasting power of the model and the greater the dispersion. The formula is expressed as follows:

$$MAE = \frac{1}{N}\sum_{k=1}^{N}\left|y(k) - y'(k)\right| \times 100\%, \tag{11}$$

where: $y(k)$ is the actual value; $y'(k)$ is the estimated value; $y(k) - y'(k)$ are the random errors ε_k; N is the number of observations. The closer MAE is to zero, the better the forecasting power of the model.

2.4.4. MAPE

Mean absolute percentage error (MAPE) measures the gap between each forecasted value and actual value, and such gaps are expressed in percentages. The absolute values of these differences are aggregated and averaged with the number of observations. MAPE, expressed with a percentage, serves as an indicator to the predictive power of the model. The higher the MAPE value, the greater the dispersion and, hence, the weaker the forecasting effectiveness is. MAPE as a relative measure is an objective measurement for the difference between estimated values and actual values. It is calculated as follows:

$$MAPE = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{y(k) - y'(k)}{y(k)}\right| \times 100\%. \tag{12}$$

A MAPE value of less than 10% means high accuracy in forecasting. The closer it is to zero, the better the estimating power of the model.

2.4.5. Four accuracy indicators of binary result forecasts

The four statistical indicators of binary result forecasts, Gamma, Somers' D, Tau-a, and C (the concordance index), are the probability expressions of forecasting accuracy when the observed variables turn into a binary one (e.g., where the actual value of 0 indicates renting and 1 indicates buying). These statistical indicators measure the correlation between the probability value forecasted and the actual value. The stronger the correlation, the greater the forecasting accuracy is. Assuming there is N number of observations (i.e., actual values at 0 or 1), these observations consist of m number of the actual values at 0 and n number of the actual values at 0. The number of concordant pairs (N_c), the number of discordant pairs (N_d), and the number of tie pairs (T_{ie}) can be calculated from the t pairs (m x n), i.e., all the possible pairs comprised of a zero and a one. These three types of pairs are defined as follows:

(1) Number of concordant pairs (N_c):
This refers to the number of pairs where the forecasted probability for each sample when the actual value is zero is smaller than the forecasted probability for each sample when the actual value is one.

(2) Number of discordant pairs (N_d):
This refers to the number of pairs where the forecasted probability for each sample when the actual value is zero is larger than the forecasted probability for each sample when the actual value is one.

(3) Number of tie pairs (T_{ie}):
This refers to the number of pairs where the forecasted probability for each sample when the actual value is zero is equal to the forecasted probability for each sample when the actual value is one.

The four indicators for the correlation between actual values and forecasted probabilities of a binary variable are calculated by using the above three pair types as follows:

$$Gamma = \frac{N_c - N_d}{N_c + N_d}; \tag{13}$$

$$Somer'D = \frac{N_c - N_d}{t}; \tag{14}$$

$$Tau\text{-}a = \frac{N_c - N_d}{0.5 \times N \times (N-1)}; \tag{15}$$

$$C = \frac{N_c + 0.5(t - N_c - N_d)}{t} \quad \text{or} \quad C = \frac{Somer'\,D}{2} + 0.5\cdot \tag{16}$$

The greater these indicator values are, the better the forecasting accuracy. A value for Gamma must lie between +1 and −1. A Gamma value of 1 indicates that all the pairs are concordant, i.e., complete correlation between the forecasted values and actual values and 100% accuracy of the forecasts. A Gamma value of −1 indicates that all the pairs are discordant, i.e., a complete lack of correlation between the forecasted values and the actual values and the complete inaccuracy of the forecasting model. Somer' D and Tau-a also represent the correlation intensity. The greater the values are, the higher the forecast accuracy. The values for both must lie between 0 and 1. A value for C (concordance index) must be between 0.5 and 1, with 0.5 indicating no correlation and 1 indicating complete correlation. Similarly, the higher the C value is, the greater the forecast accuracy. It is worth noting that the C value is a close approximation of the AUC value (Agresti, 2002; Uno, Cai, Pencina, Agostino, & Wei, 2011).

3. Source and sample statistics description

3.1. Source

This paper samples data from a housing survey conducted by the Construction and Planning Agency, Ministry of the Interior, from January 1, 2006, through February 15, 2006. The population consisted of the most updated household registrations and address numbers in all the neighborhoods and villages archived by the Ministry of the Interior. Data was sampled in two stages based on the stratified method. The data on the city/county level made up the sub-population. Data was sampled for each township from the sub-population with the stratified method. The first-stage sampling was conducted on the village and neighborhood level, and the second-stage sampling was conducted according to the address numbers of each village and neighborhood. The survey conducted by the Construction and Planning Agency covered the entire main island of Taiwan as well as Kinmen County and Lienchiang County which are in Fuijian Province. This paper, however, only focuses on the data collected for Taipei City and New Taipei City. A total of 7,594 data points were collated. After eliminating the data concerning properties with dual purposes (i.e., properties used for both residential and business purposes, properties used for both residential and industrial purposes, and properties used for both residential and service purposes) and incomplete data points (incomplete answers), this paper established a total of 3,031 data points for the empirical analysis.

3.2. Sample statistics description

Table 2 indicates that a total of 2,500 survey respondents chose to purchase properties (82.5% of the total sample), while 531 respondents (17.5%) chose to rent properties. The majority of the household heads were male (2,386 respondents, or 78.7%, of the sample). A total of 645 household heads, or 21.3%, of the sample were female. The average age of the household heads was 48.12 years. A total of 766 household heads (25.3% of the sample) had an education at the junior high school level or below, 871 (28.7%) graduated from a senior high school or vocational school, 616 (20.3%) had a college education, and 193 (25.7%) had a post-graduate degree. The average age of the sampled properties was 22.62 years. The average living space per person was 9.7 *ping*. The average number of rooms per person was 1.06, and the number of permanent residents per property was 4.04.

A total of 1,354 of the respondents (44.7% of the sample) were made on a loan-to-value ratio of less than 0.5. The remaining 55.3% of the sample (or 1,677 respondents) reported a loan-to-value ratio of lower than 0.5. Some household heads resorted to personal borrowing in order to make home purchases. A total of 640 purchases (19.9%) were supported with a personal borrowing of NT$ 10,000~500,000, while 106 purchases (3.5%) were financed partly with a personal borrowing of NT$ 500,000 and more. A total of 2,282 purchases (76.6%) were made without personal borrowing. A total of 73.3% purchases were not made with the use of a government-subsidized loan (vs. 26.7% of purchases that were supported with government subsidized loans). As shown in Table 3, the level of satisfaction with environmental qualities had an average of 3.68 and the level of satisfaction with leisure and sports facilities had an average of 3.41 for all the administrative zones. The majority of the respondents indicated satisfaction.

Table 2. Descriptive statistics of first-level variables (N = 3031)

Variable	Classification		N	Percentage
TENU	Buying		2500	82.5
	Renting		531	17.5
GENDER	Male		2386	78.7
	Female		645	21.3
EDU	Below junior high school		766	25.3
	Senior high school or vocational school		871	28.7
	College		616	20.3
	Post-graduate degree		193	25.7
LOANR	Loan-to-value ratio >0.5		1354	44.7
	Loan-to-value ratio <0.5		1677	55.3
PMORTGAGE	No personal borrowing		2282	76.6
	Personal borrowing between NT$ 10,000 and NT$ 500,000		640	19.9
	Personal borrowing higher than NT$ 500,000		106	3.5
PRELOAN	Without subsidized loans from government		2222	73.3
	With subsidized loans from government		809	26.7
Continuous variable	Mean	SD	Min	Max
PAGE	48.12	12.26	20	93
HAGE	22.62	11.94	1	94
PAREA	9.70	6.35	2	95
PROOM	1.06	0.50	0.07	8
FMSZ	4.04	1.69	1	16

Table 3. Descriptive statistics of second-level variables (N = 24)

Variable	Mean	SD	Min	Max
ENVI	3.68	0.35	3.17	4.89
LEIS	3.41	0.36	2.93	4.75

4. Empirical results and discussion

This paper compares the model fit with the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The smaller these indicator values, the better the model fit. According to the empirical findings in Tables A1~A3, the AIC values for HGLM, the binary logistic regression model, and the discriminant analysis are –7130.19, –7020.47, and –6889.93, respectively. The BIC values for HGLM, the binary logistic regression model, and the discriminant analysis are 889146.7, 907527.6, and 946397.2, respectively. HGLM has the best model fit, followed by the binary logistic regression, and then by the discriminant analysis.

As shown in Table 4, HGLM has the highest hit rate of 87.4%, followed by the 86.1% rate of the binary logistic regression model and the 85.7% rate of the discriminant analysis. This is possibly due to the nested nature of the data. Also, the binary logistic regression model overlooks the scenario (contextual) variables and inter-class variances across administrative zones and, hence, may be biased in the estimated coefficients it produces (Wen, 2006). The single-level processing of nested data fails to incorporate the influence on intercepts or slopes from variables on another level. The estimation, which is only based on fixed effects, may thus lead to biased results. To determine the reason for a lower hit rate produced by the discriminant analysis, this paper conducts one-sample Kolmogorov-Smirnov (K-S) tests to validate whether the discriminant variables are in compliance with the presumption for normal distributions. The results indicate that not all the discriminant variables in the discriminant function are in a normal distribution. The breach of the normal-distribution presumption undermines the forecasting accuracy of the discriminant analysis.

The ROC serves to validate a given model's discriminating capability with regard to tenure choices. The effectiveness of the cut-off between the buying group and the renting group is measured with the AUC. An AUC of greater than 0.5 indicates strong discriminating capability. Figure 2 indicates that the AUC in HGLM is the largest, at 0.918, followed by that of the binary logistic regression model at 0.893 and then that of the discriminant analysis at 0.879. In sum, HGLM has a better discriminating capability with regard to tenure choices than the binary logistic regression model and the discriminant analysis.

Both MAE and MAPE measure the gap between the forecasted values of buying and renting and the actual values of buying and renting. The empirical results indicate that HGLM reports the lowest MAE at 13.3%, followed by the discriminant analysis (at 14.5%) and the binary logistic regression model (at 15.6%). The lower the MAE values, the greater the model accuracy. In sum, HGLM is superior to the other two models. This is also evidenced by its lowest MAPE of 19.1%, which is followed by that of the binary logistic regression model at 23.0% and that of the discriminant analysis at 23.7%. The smaller the MAPE values, the better the forecasting accuracy. Lewis (1982) believes that a MAPE value of 10~20% suggests a good predicting power and that a MAPE value of 20~50% shows a reasonable forecasting power. The closer a MAPE

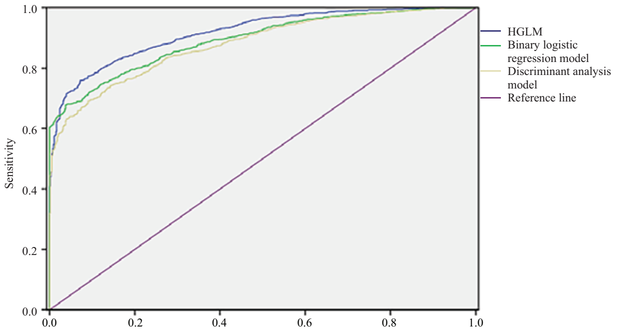

Figure 2. ROCs for all the three models

Table 4. Comparison of three models

	HGLM	Binary logistic regression	Discriminant analysis
Hit rate	87.40%	86.08%	85.65%
AUC	0.918	0.893	0.879
MAE	0.133	0.156	0.145
MAPE	0.191	0.230	0.237
Gamma	0.884	0.826	0.802
Somers’ D	0.826	0.780	0.745
Tau-a	0.239	0.226	0.215
C	0.912	0.890	0.873
Cross validation	0.846	0.838	0.833

value is to zero, the better the forecasting results. In other words, HGLM has a good forecasting power (19.1%), while the binary logistic regression model (23.0%) and the discriminant analysis (23.7%) can achieve reasonable forecasting. According its MAPE value, HGLM is two times more accurate than the other two models.

MAE and MAPE are frequently used and appropriate indicators for measuring the gaps between forecasted values and actual values when the actual values are continuous. However, if the actual values are binary (0, 1), the MAPE calculations cannot capture all the data points if the denominator of the actual value is zero. This affects the accuracy and interpretation of the estimated variances. As such, this paper also refers to the four accuracy indicators of binary result forecasts to measure the correlations between actual values and forecasted values when the results are binary. These four indicators, Gamma, Somers' D, Tau-a, and C, serve as objective measurements of the gap between the actual values and forecasted values of a binary variable. These accuracy metrics are also applicable to different models. As shown in Table 4, the Gamma value is the highest for HGLM (0.884), followed by the value for the binary logistic regression model (0.826) and then that for the discriminant analysis (0.802). The Gamma value is between +1 and –1. A Gamma value of 1 indicates that all the pairs are concordant (N_c) and that there is a complete correlation between the forecasted and actual values. In contrast, a Gamma value of –1 indicates that all the pairs are discordant (N_d) and that there is a complete lack of correlation between the forecasted and actual values. The empirical results indicate that HGLM has the best predictive power (as evidenced by its highest Gamma value), followed by the binary logistic regression model. Somers' D and Tau-a (with values between 0 and 1) also indicate the strength of correlations between the forecasted and actual values. The higher the values, the better the forecasting accuracy is. According to Table 4, HGLM has higher Somers' D and Tau-a values than the other two models, a testimony to its superior forecasting accuracy.

The C (concordance index) value is between 0.5 and 1. A value of 0.5 indicates a 0% correlation, and a value of 1 indicates a 100% correlation. The higher the C value, the better the forecasting accuracy is. Table 4 shows that HGLM yields the highest C value (0.912), followed by that produce by the binary logistic regression model (0.890) and that of the discriminant analysis (0.873). In sum, HGLM produces the best correlation between forecasted and actual values. Table 4 indicates that the AUC and C values of the three models are similar. In fact, the C value is a close approximation of AUC. This is consistent with Uno et al. (2011).

This paper sums up the above findings in its conclusion regarding the predictive power of the different models by referring to AIC, BIC, ROC, MAE, MAPE, and the four indicators of the correlations between forecasted and actual values, i.e., Gamma, Somers' D, Tau-a, and C. HGLM is superior to the other two models, as evidenced by the model fit indicators AIC and BIC, the complete cut-off between buying and renting (ROC as a metric for discriminating power), model accuracy indicators (MAE and MAPE), and the correlations between forecasted and actual values. Meanwhile, the binary logistic regression model outperforms the discriminant analysis in terms of all the metrics other than MAE.

In addition, this paper examines the stability of the hit rates mentioned above with 10-fold cross-validation. The k-fold cross-validation consists of dividing the number of observations (N) into a large group comprised of N (k-1/k) number of observations and a small group of N (1/k) number of observations. The large group serves as the training group while the small group serves as the test group. The training group, i.e., the N (k-1/k) number of observations, is used to estimate the forecast function and then the test group, i.e., the N (1/k) number of observations, serves as the input of explanatory variables to the forecast function to derive the forecasted absolute values or probability values. The distribution of accurate forecasts (in relation to actual values) serves as an empirical test of hit rates by using the k-fold cross-validation technique. Assuming the observation is not to be repeated, this process continues for k times in the k number of tests, in order to verify the k number of hit rates with the k-fold validation. The k number of hit rates for the k fold is aggregated and then divided by k to derive the average hit rate on the basis of k-fold cross-validation. The average hit rate derived with cross validation is a further validation of the reliability of the hit rates produced by the whole sample. The processing of k folding yields refined and objective hit rates and, hence, is a more convincing method for comparing the forecasting power of different models. This paper conducts the k-fold cross-validation, as a robust metric to examine whether the ranking of hit rates yielded by different models on the complete sample is biased and whether the hit rates are consistent in direction and stable in accuracy. The more folds indicated by cross validation, the more refined the sample processing, the smaller the error in forecasting accuracy measurement and, understandably, the more objective the forecasting accuracy measurement is. Based on the sample size and the nature of the data, this paper adopts the most commonly used 10-fold cross-validation.

According to the empirical results of the 10-fold cross-validations, HGLM yields the highest hit rate of 84.6%, followed by that of the binary logistic regression model (83.8%) and that of the discriminant analysis (83.3%). The ranking is the same as the one produced with the total sample without folding. The hit rates on the basis of the 10-fold cross-validation still show that HGLM is superior to the binary logistic regression model and the discriminant analysis.

Conclusions and suggestions

The tenure choice theory helps consumers to review their own conditions, property characteristics, and the macroeconomic environment in order to reach the optimal decision with regard to buying or renting. Real estate companies can also apply the tenure choice theory when starting development projects and in marketing. For instance, if the business development and sales personnel can have a solid understanding of the factors (i.e., explanatory variables) that influence tenure choices, they can focus their efforts on sourcing the most highly sought properties in order to shorten the average number of days per transaction and boost their rates of success in selling properties. This also avoids the sourcing of properties that are difficult to place or that do not cater to consumers' tenure requirements.

An overview based on hit rates, ROC, MAE, MAPE, and the four indicators of the correlation between the forecasted values and actual values of a binary variable (Gamma, Somers' D, Tau-a, and C) suggest that HGLM yields higher forecasting accuracy than the binary logistic regression model and discriminant analysis. The 10-fold cross-validation of hit rates also suggests that HGLM is superior to the binary logistic regression model and the discriminant analysis. In conclusion, the processing of nested structure data with HGLM can enhance the forecasting accuracy with regard to tenure choices.

Most of the studies on tenure choices apply logistic regressions. However, this approach ignores the scenario variables (contextual variables) as explanatory variables between groups or nests in the data. It also breaches the presumption of the mutual independence of error terms and a normal distribution of error terms with a mean of zero and a variance of σ^2 . The derived logistic regression coefficient is often not without bias (Wen, 2006). On the other hand, the use of HGLM to process nested structure data can avoid the shortcomings of inflated coefficients and significance result biases often seen in traditional logistic regressions. As a result, the hit rates with regard to tenure choices can be improved.

This study on tenure choices and hit rates is only focused on the Greater Taipei Area (i.e., Taipei City and New Taipei City). Therefore, as far as the influence of the independent variables on tenure choices and the ranking of hit rates yielded by the three models in question goes, the empirical findings are only applicable to the Greater Taipei Area. The results may vary due to different data sources, different research methods, and different purposes. Follow-up studies may explore tenure choices with national data or data sourced from different cities or counties. To enhance forecasting accuracy, homeownership costs, property taxation rates, and any increase in transaction costs and spending may be included in the list of factors that influence tenure choices.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New Jersey: Wiley-Interscience. https://doi.org/10.1002/0471249688

Cardozo, R. N. (1965). An experimental study of consumer effort, expectation, and satisfaction. *Journal of Marketing Research*, *2*(3), 244-249. https://doi.org/10.2307/3150182

Carter, S. (2011). Housing tenure choice and the dual income household. *Journal of Housing Economics*, *20*(3), 159-170. https://doi.org/10.1016/j.jhe.2011.06.002

Chambers, M., Garriga, C., & Schlagenhauf, D. (2009). The loan structure and housing tenure decisions in an equilibrium model of mortgage choice. *Review of Economic Dynamics*, *12*(3), 444-468. https://doi.org/10.1016/j.red.2009.01.003

Chen, G. (2016). The heterogeneity of housing-tenure choice in urban China: a case study based in Guangzhou. *Urban Studies*, *53*(5), 957-977. https://doi.org/10.1177/0042098015571822

Espahibodi, P. (1991). Identification of problem banks and binary choice models. *Journal of Banking and Finance*, *15*(1), 53-71. https://doi.org/10.1016/0378-4266(91)90037-M

Feng, Y., & Jones, K. (2015). *Comparing methods: using multilevel modelling and artificial neural networks in the prediction of house prices based on property, location and neighbourhood characteristics.* School of Geographical Sciences, University of Bristol.

Henley, A. (1998). Residential mobility, housing equity and the labour market. *The Economic Journal*, *108*(447), 414-427. https://doi.org/10.1111/1468-0297.00295

Huang, Y., & Clark, W. A. V. (2002). Housing tenure choice in transitional urban China: a multilevel analysis. *Urban Studies*, *39*(1), 7-32. https://doi.org/10.1080/00420980220099041

Kim, K., & Jeon, J. S. (2012). Why do households rent while owning houses? Housing sub-tenure. *Habitat International*, *36*(1), 101-107. https://doi.org/10.1016/j.habitatint.2011.06.005

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, *11*(1), 443-448. https://doi.org/10.1016/j.asoc.2009.12.003

Lang, B. J., & Hurst, E. H. (2013). The effect of down payment assistance on mortgage choice. *The Journal of Real Estate Finance and Economics*, *49*(3), 329-351. https://doi.org/10.1007/s11146-013-9432-1

Lee, C. C., Ho, Y. M., & Chiu, H. Y. (2016). Role of personal conditions, housing properties, private loans, and housing tenure choice. *Habitat International*, *53*, 301-311. https://doi.org/10.1016/j.habitatint.2015.11.016

Lee, C. C., Liang, C. M., Chen, J. Z., & Tung, C. H. (2018). Effects of the housing price to income ratio on tenure choice in Taiwan: Forecasting performance of the hierarchical generalized linear model and traditional binary logistic regression model. *Journal of Housing and the Built Environment, 33*(4), 657-694. https://doi.org/10.1007/s10901-017-9572-3

Lewis, C. D. (1982). *Industrial and business forecasting methods.* London: Butterworths Scientific.

Mossman, D. (1994). Assessing predictions of violence: being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*(4), 783-792. https://doi.org/10.1037/0022-006X.62.4.783

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: appraising and managing risk.* Washington: American Psychological Association. https://doi.org/10.1037/10304-000

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks: CA: Sage.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63*(5), 737-748. https://doi.org/10.1037/0022-006X.63.5.737

Spalkova, D., & Spalek, J. (2012). *Factors of the tenure choice: the case of the Czech Republic.* Masaryk University.

Steenackers, A., & Goovaerts, M. J. (1989). A credit scoring model for personal loans. *Insurance: Mathematics and Economics*, *8*(1), 31-34. https://doi.org/10.1016/0167-6687(89)90044-9

Subhan, S., & Ahman, E. (2012). The economic and demographic effects on housing tenure choice in Pakistan. *American International Journal of Contemporary Research*, *2*(7), 15-24.

Tao, L., Hui, E. C. M., Wong, F. K. W., & Chen, T. (2015). Housing choices of migrant workers in China: beyond the Hukou perspective. *Habitat International*, *49*, 474-483. https://doi.org/10.1016/j.habitatint.2015.06.018

Tung, C. H., Lee, C. C., Chen, C. L., & Wu, Y. L. (2016). Application of support vector machines for the prediction of the residence price in Taipei city. *Journal of Housing Studies*, *25*(2), 31-51.

Uno, H., Cai, T., Pencina, M. J., Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, *30*(10), 1105-1117. https://doi.org/10.1002/sim.4154

Vera-Toscano, E., & Ateca-Amestoy, V. (2008). The relevance of social interactions on housing satisfaction. *Social Indicators Research*, *86*(2), 257-274. https://doi.org/10.1007/s11205-007-9107-5

Wagner, K. (2014). Intergenerational transmission: how strong is the effect of parental homeownership? Results of a survey on households in Austria. *Monetary Policy and the Economy Q*, *2*, 49-64.

Wen, F. H. (2006). *Principles, methods and applications of hierarchical linear modeling.* Taipei: Yeh Yeh Book Gallery.

Appendix

Table A1. Regression analysis with intercepts-as-outcomes model

Fixed effect	Coefficient	Standard error		*p*-value
Natural logarithm of relative probabilities of buying/renting γ_{00}	8.220***	1.475		0.000
ENVI γ_{01}	1.663*	0.927		0.087
LEIS γ_{02}	0.032	0.732		0.965
GRNDER γ_{10}	0.090	0.188		0.634
PAGE γ_{20}	0.135***	0.029		0.000
PAGES γ_{30}	−0.001***	0.000		0.006
HIGHT γ_{40}	0.504***	0.152		0.001
COLLEGE γ_{50}	0.953***	0.209		0.000
UNI γ_{60}	0.919***	0.220		0.000
HAGE γ_{70}	0.001	0.005		0.958
PAREA γ_{80}	0.031	0.023		0.180
PROOM γ_{90}	0.266	0.172		0.122
FMSZ γ_{100}	0.282***	0.049		0.000
LOANR γ_{110}	4.736***	0.266		0.000
PMORTGAGE1 γ_{120}	1.802***	0.502		0.001
PMORTGAGE2 γ_{130}	8.223***	2.540		0.002
PRELOAN γ_{140}	2.507***	0.384		0.000
Random effect	Variance	Df	Chi-square	*p*-value
Variance between groups τ_{00}	0.615	21	190.160	0.000
AIC	−7130.19			
BIC	889146.70			

Note: *** indicates P < 0.01, ** indicates P < 0.05, and * indicates P < 0.1.

Table A2. Coefficient estimates with binary logistic regression model

	Estimated coefficient	Standard error	Wald	*p*-value
Intercept	17.847	1631.843	0.000	0.991
GENDER	0.082	0.143	0.332	0.565
PAGE	0.133***	0.028	22.982	0.000
PAGES	–0.001***	0.000	7.623	0.006
HIGHT	0.499***	0.168	8.846	0.003
COLLEGE	0.935***	0.193	23.431	0.000
UNI	0.915***	0.197	21.677	0.000
HAGE	0.001	0.005	0.009	0.925
PAREA	0.032*	0.018	3.287	0.070
PROOM	0.256	0.233	1.200	0.273
FMSZ	0.283***	0.049	33.506	0.000
LOANR	19.387	900.329	0.000	0.983
PMORTGAGE1	17.014	1183.232	0.000	0.989
PMORTGAGE2	19.404	3241.508	0.000	0.995
PRELOAN	17.483	1047.169	0.000	0.987
ENVI	7.113	635.542	0.000	0.991
LEIS	–6.869	647.322	0.000	0.992
AIC	–7020.47			
BIC	907527.60			

Note: *** indicates P < 0.01, ** indicates P < 0.05, and * indicates P < 0.1. Dummy variables for administrative zones included in the estimates.

Table A3. Function coefficients estimated with discriminant analysis

	Estimated coefficient	Standard error	Wilks'… Lambda value	F test	*p*-value
Intercept	–4.670***	0.19581	0.996	10.812	0.001
GENDER	0.067	0.40432	0.999	2.042	0.153
PAGE	0.123***	12.05363	0.971	90.946	0.000
PAGES	–0.001***	1255.55424	0.977	70.426	0.000
HIGHT	0.286	0.44943	1.000	0.902	0.342
COLLEGE	0.567	0.39486	0.999	1.517	0.218
UNI	0.515	0.41852	1.000	0.580	0.446
HAGE	–0.002*	11.51496	0.999	3.592	0.058
PAREA	0.013	6.09419	1.000	0.147	0.702
PROOM	0.238	0.48454	1.000	0.512	0.475
FMSZ	0.165***	1.64095	0.991	28.083	0.000
LOANR	1.689***	0.48531	0.833	605.196	0.000
PMORTGAGE1	0.042***	0.38522	0.949	162.239	0.000
PMORTGAGE2	0.927***	0.18095	0.991	27.478	0.000
PRELOAN	0.389***	0.42675	0.922	255.627	0.000
ENVI	1.842	0.28461	1.000	1.310	0.252
LEIS	–0.664	0.32302	1.000	0.015	0.903
AIC	–6889.93				
BIC	946397.20				

Note: *** indicates P < 0.01, ** indicates P < 0.05, and * indicates P < 0.1. Dummy variables for administrative zones included in the estimates.