Follow-up study of home purchase intentions

Introduction

PRA Inc., on behalf of Canada Mortgage and Housing Corporation (CMHC), conducts an annual survey asking Canadians about their home purchase and home renovation intentions for the coming year.

In 2002, PRA re-contacted respondents to the 2000 and 2001 surveys (who had mentioned that they were intending to buy a home in the next year) to conduct a follow-up survey, one of the goals of which was to see whether those respondents who had stated that they intended to buy a house in the next 12 months had actually made the purchase.

The following sections summarize the results obtained from the analysis of the data gathered.

Methodology

The data

The data gathered by the 2000 and 2001 annual surveys conducted for CMHC serve as a baseline for this analysis.

In particular, the annual surveys gathered information on respondents' intentions and preferences with respect to purchasing a home in the coming year. Basic demographic data was also gathered.

In 2002 (April 26th to May 8th), the follow-up survey was conducted. An attempt was made to contact all respondents from the 2000 and 2001 surveys who indicated that they were thinking of purchasing a home in the next 12 months (1,526 respondents in 2000; 2,648 in 2001). Due to factors such as change in phone numbers and refusal to participate in the follow-up survey, only 1,452 of the previous respondents successfully completed the follow-up survey (469 from 2000; 983 from 2001).

The follow-up survey asked respondents whether they did indeed buy a home within the last year or two. This is the variable of interest – whether those who said that they were intending to buy a home, did buy a home. The response will be either "yes" or "no."

Binomial logistic regression

The variable of interest (the dependent variable) from the follow-up survey is f1 (whether or not the respondent bought a home in the past twelve/24 months). This variable is a simple dichotomous yes/no response. Hence, a normal distribution cannot be assumed.

A logistic regression does not assume that the dependent variable is normally distributed. It also does not assume a linear relationship between the dependent and independent variables, nor does it assume homoscedasticity of the dependent variable at different levels of the independent variables.

Overall, the logistic regression has less stringent requirements than the usual ordinary least squares regression. For these reasons, the data was analysed using logistic regression.

Specification of the logistic model

A response of "yes" (did buy a home in the past twelve/24 months) was coded as a 1 for the purpose of the logistic regression. "No" was coded as a 0.

Several logistic models were run, using different independent variable combinations, from the 2000 and 2001 data. The remainder of this discussion will focus on results obtained by running a backward stepwise logistic regression, with the following independents: region, gender, own/rent current home, chance of buying in next year, age, household size, and income.

For the purpose of the logistic regression, the independent variables were redefined:

▸ Region (six categories) is broken down into five dummy variables: dreg1 (Halifax), dreg2 (Montreal), dreg3 (Toronto), dreg4 (Edmonton), and dreg5 (Calgary). If all five dummy variables are 0, the region is defined as Vancouver.

▸ Gender (ngend2): A code of 1 refers to male, and 0 refers to female. The undetermined are dropped.

▸ Own/rent (q1a): 1 refers to own current home; 0 refers to rent current home.

▸ Chance of buying a home in next 12 months (q2b) is split into two dummy variables: d2b1 (high chance of buying) and d2b2 (50/50 chance of buying). If both dummy variables equate to 0, this refers to a low chance/don't know response.

▸ Age and household size (n26): The "no responses" are dropped, and these are entered as continuous variables.

▸ Income (n31): Since the income variable is ordinal, the "no responses" are dropped, and it is treated as a continuous variable.

The above dependent and independent variables were entered into the logistic regression model, and the backward stepwise method was used to drop "insignificant" variables at each step. In other words, variables that, when removed, were found to have minimal effect on the model were dropped.

Cases with missing information were excluded from the analysis, so that the final sample size used to determine the model was 1,223. (The initial sample size was 1,452.)

The results

The resulting model retained dreg5 (Calgary), d2b1 (high chance of buying), d2b2 (50/50 chance of buying), and n31 (income) as predictors to estimate the probability of actually buying a home.

The model chi-square used to assess the overall logistic model was found to be significant, indicating that the independent variables do have an impact on predicting the dependent variable.

Individually, all the remaining variables were found to be significant, though d2b2 (50/50 chance of buying) was found significant to a lesser degree.

The following table summarizes the beta coefficients and other statistics associated with the remaining independent variables.

TABLE 1
Logistic regression results (sample n=1,223)

Variables	B	SE	Wald	Sig
dreg5	.8839	.2157	16.79	.0000
d2b1	1.5282	.2409	40.24	.0000
d2b2	.7729	.2529	9.34	.0022
n31	.1734	.0496	12.24	.0005
Constant	-3.0411	.2811	117.07	.0000

The beta coefficients, along with the mean of the dependent variable (.22), can be used to determine the marginal effect of each variable on the probability of buying a house.[1]

TABLE 2
Estimated marginal effect on the probability of buying a house[2]

Variable	Marginal effect (change in probability)*
dreg5	0.186
d2b1	0.345
d2b2	0.159
n31	0.030
* Marginal effects for dummy variables are calculated using the difference in probabilities between the dummy variable group and the omitted group. The change in probability for income (n31) is based on the partial derivative.	

Table 2 (previous page) shows that living in Calgary increases the probability of buying a house by 19%. If the respondent indicates that there is a high chance of buying a house in the next year, the likelihood of actually buying increases by .35, compared to a low/ no chance of buying in the next year. Similarly, the likelihood of buying a house increases by .16 for those who indicate a 50-50 chance of buying in the next year. A unit increase in income bracket3 increases the probability of buying a house by 3%.

All variables have a positive impact on the probability of buying a home (as was expected), with high chance of buying a home having the largest effect.

A simple correlation was also run between the dependent variable and the prevailing independent variables (in their original form).

TABLE 3
Correlation of dependent versus independent variables

Variable	Bought a home (f1)
Region	.076 (.004) n=1,452
Chances of buying in next 12 months – q2b	.238 (.000) n=1,441
Total household income – n31	.135 (.000) n=1,245

At the .005 level, both the logistic and correlation coefficients are significant.

The classification table indicates that the resulting model correctly classified 78.99% of the cases. This illustrates the predictive power of the model and verifies the value of the questions being used in the annual surveys.

The classification model also indicates that 20.7% of the cases that were predicted not to purchase a home actually did purchase a home. This demonstrates the influence of outside factors on the housing market. Factors such as the low interest rates that had been prevailing for the past few years may influence a person who initially was not considering purchasing a home, to reconsider and buy.

Summary

The results of the logistic regression demonstrate the importance of the questions being asked on the annual CMHC survey. In particular, they illustrate the power of a few key questions in predicting the likelihood of buying a home. Such information may be used to estimate housing demands; however, caution should be taken to account for external factors (such as low interest rates or, in present terms, rising interest rates).

[1] See Pampel (2000) for a convenient summary of logistic regression and determining probabilities from logit models.

[2] It is important to note that these marginal effects only have meaning at the mean (expected) probability of buying a house. A hypothetical maximum change in probability may be calculated by substituting .5 in place of the mean value; however, this will overstate the influence of the variable on the probability of buying a house.

[3] The income brackets are defined as <$20K, $20K-$40K, $40K-$60K, $60K-$80K, $80K-$100K, $100K-$200K, and $200K or more.

Bibliography

Pampel, F.C. (2000). *Logistic Regression.* Sage: Thousand Oaks, CA.

For additional information, please contact admin@pra.ca

APPENDIX 1

```
Estimation terminated at iteration number 4 because
Log Likelihood decreased by less than .01 percent.


-2 Log Likelihood 1185.161
Goodness of Fit 1222.321
Cox & Snell - R^2 .077
Nagelkerke - R^2 .119


                 Chi-Square  df    Significance
Model            98.421      4      .0000
Block            98.421      4      .0000
Step             -1.694      1      .1931


Note: A negative Chi-Square value indicates that the Chi-Square
value has decreased from the previous step.


Classification Table for F1
The Cut Value is .50


                          Predicted
                  Did not buy      Bought a home
                        D                 B
Observed
  Did not buy     D  946           10  98.95%
  Bought a home   B  247           20   7.49%
                                   Overall 78.99%


------------------- Variables in the Equation ----------------------


Variable        B           S.E.     Wald       df   Sig     R       Exp(B)
DREG5(1)        .8839       .2157    16.7868    1    .0000   .1073   2.4203
D2B1(1)         1.5282      .2409    40.2403    1    .0000   .1726   4.6097
D2B2(1)         .7729       .2529     9.3366    1    .0022   .0756   2.1659
N31             .1734       .0496    12.2440    1    .0005   .0893   1.1893
Constant        -3.0411     .2811   117.0705    1    .0000


---------------- Model if Term Removed -----------------


Term          Log                                    Significance
Removed       Likelihood    -2 Log LR     df          of Log LR
DREG5         -600.534      15.907        1            .0001
D2B1          -618.065      50.968        1            .0000
D2B2          -597.713      10.265        1            .0014
N31           -598.733      12.305        1            .0005


--------------- Variables not in the Equation ----------------


Residual    Chi Square  6.035 with  8 df  Sig = .6434
Variable        Score       df       Sig       R
DREG1(1)        .7559       1        .3846     .0000
DREG2(1)        .4103 1     .5218    .0000
DREG3(1)        1.6465 1    .1994    .0000
DREG4(1)        .5330 1     .4653    .0000
NGEND2(1)       1.5500 1    .2131    .0000
Q1A(1)          .1678 1     .6821    .0000
AGE             1.1407      1        .2855     .0000
N26             .5679       1        .4511     .0000
```