Paper 369-2008

How to use SAS® to fit Multiple Logistic Regression Models

Anpalaki J. Ragavan, Department of Mathematics and Statistics, University of Nevada, Reno, NV 89557

ABSTRACT

When response outcomes are continuous error terms in models are normally distributed and a standard normal distribution function is adequate. The logistic distribution function which is very similar to the normal distribution function is required when the response variable is binary. Parameters of a logistic response function are often estimated using the method of maximum likelihood (ML). One of the problems with ML estimation is that, no closed-form solution exists for the values of the parameters that maximize the log-likelihood function. Hence sophisticated computer-intensive numerical search procedures (i.e: Newton Raphson) are required to find ML estimates of parameters.

This paper is a step by step guide to develop a multiple logistic regression model for data sets with binary response variable using PROC LOGISTIC in SAS®. Since PROC LOGISTIC requires uniform coding and does not accommodate missing data, data need be corrected for missing values and for outliers, those can reduce the efficiency of ML estimation. In addition, best subset among 25 predictor variables was selected. This step is useful to increase the speed of ML estimation. Complete model diagnostics tests, sensitivity and residual analyses were performed.

ML parameter estimation was significant at 5%level. P-value of Hosmer and Lemeshow goodness of fit test was 0.9545. Model had a maximum adjusted R-square value equal to 0.78, showed 93.2% sensitivity and 79.5% specificity. Residuals from the model did not show significant patterns and were normally distributed. Odds ratios of the predictor variables supported the conclusions drawn from ML estimation.

INTRODUCTION

One of the assumptions of regression is that the variance of the response variable (Y) is constant across values of independent variable (homoscedasticity). This cannot be the case when the response variable is binary since the variance is given simply by p*(1-p) (where p is the probability of success). When 50 percent of the people are = 1, then the variance is = 0.25, its maximum value. As we move to more extreme values, the variance decreases. When p = 0.10, the variance is 0.1*0.9 = 0.09, so as p approaches 1 or zero, the variance approaches zero. In addition in regression analysis the significance testing of the parameters relies upon the assumption that errors of prediction (Y-Y(hat)) are normally distributed. When response outcomes are continuous the error terms in the model are normally distributed and a standard normal cumulative distribution function is used for the data with the assumption of normally distributed errors. When Y is binary, because Y can take only the values 0 and 1, this assumption is hard to justify, even approximately. Therefore, the tests of the regression parameter estimates are unreliable when linear regression is used with a binary dependent variable. An alternative distribution function that is very similar to the normal distribution is the logistic distribution function. Estimates of the parameters of the logistic response function are estimated with the method of maximum likelihood. Maximum likelihood estimates of the parameters in the logistic regression model are those parameter values that maximize the log-likelihood function. This method deals pretty well with the problems associated with the response being binary. For large samples under generally applicable conditions maximum likelihood estimators for logistic regression are approximately normally distributed, with little or no bias. However, one of the problems with the maximum likelihood estimation is that, no closed-form solution exists for the values of the parameters that maximize the log-likelihood function. Sophisticated computer-intensive numerical search procedures (i.e: Newton Raphson) with computer software are required to find the maximum likelihood estimates of parameters (Kutner et. al., 2004).

In this paper, the several steps that should be taken when fitting a multiple logistic regression model with a data set with dichotomous indicator response variable to evaluate the future probabilities of events (sale of properties) using PROC LOGISTIC in SAS® is presented and discussed. The several options available with PROC LOGISTIC in SAS® to improve the efficiency of parameter estimation and the adequacy of the model are presented. Solutions to the various problems encountered when fitting such as model is presented as sample SAS® codes and results obtained from SAS outputs wherever necessary. The logit link function was considered with a first order multiple logistic regression model, which was fitted using the maximum-likelihood estimation method. Data were corrected for missing values, tested for the

presence of any influential observations or outliers that can reduce the accuracy of the maximum likelihood estimation. The best subset of predictor variables that fit the data adequately without loss of information was selected. Model diagnostic tests, residual and sensitivity analyses were performed to validate the model. The model thus built was used to estimate the variations in probabilities of selling properties in the near future as functions of several useful predictor variables (e.g.: sale price, age, # of bedrooms, # of bathrooms, unit size).

ABOUT MODEL BUILDING CAPABILITY OF PROC LOGISTIC

PROC LOGISTIC is specifically designed for logistic regression. A usual logistic regression model, proportional odds model and a generalized logit model can be fit for data with dichotomous outcomes, ordinal and nominal outcomes, respectively, by the method of maximum likelihood (Allison 2001) with PROC LOGISTIC. PROC LOGISTIC can perform a conditional logistic regression analysis (matched-set and case-controlled) for binary response data. Analyses can be performed separately by classifications with the CLASS statement, specified with independent CLASS variables in the model. LOGISTIC procedure has capabilities for including stepwise, forward, backward, and/or selection of best subset of independent variables among multiple independent variables. PROC LOGISTIC computes the sensitivity of the built model as the receiver operating characteristics (ROC) curve and the area beneath the curve, computes predicted values and a number of regression diagnostics. Output data sets containing these values can be created separately along with other statistics. For small data sets, PROC LOGISTIC can perform an exact conditional logistic analysis (Allison, 2001). However, PROC LOGISTIC requires uniform coding and does not allow missing values in response or predictor variables, hence the data need be corrected for missing values prior to using the procedure.

EXPLORATORY DATA ANALYSIS

DESCRIPTION OF DATA

Data set (Reno) used for this paper contains information on properties for sale in the city of Reno in the state of Nevada. Data set is large and consists of 200 property units (units) that were listed or came to the market approximately at the same time, 117 of which were sold before completing 3 months in the market, and another 83 units that were not sold and are still waiting in the market to be sold, and multiple independent variables (p=25). The properties were selected randomly to have condos, town homes, single and multiple family residences from 4 areas (North (N), South(S), East(E), and West(W) in the city of Reno close to the Reno campus of the University of Nevada in the state of Nevada. Independent variables are of two types: a) continuous independent variables (e.g: # of bedrooms, # of bathrooms, size of the property), and b) indicator independent variables that provide supporting information about the unit in the form of an item or facility that is either present (1) or not present (0) in the unit (e.g: built in dish washer (DW), refrigerator (fridge), laundry facilities (WD)). There were 6 continuous independent variables namely, i) number of bedrooms (BED), ii) number of bathrooms (BATH), iii) square footage or size (SIZE), iv) sale price (PRICE), v) age of the property (AGE), and v) size of the yard (LOT) and 25 indicator independent variables in the original data. In addition there are two class variables in the data namely: i) Type (condo/town home or single family residences), ii) Area (N, S, E, or W). Selected properties are mostly located close to University of Nevada at Reno. The response variable is an indicator variable (variable with two outcomes: E(Y) = 1 when a property/unit is sold out and E(Y) = 0 when the property unit is not sold out).

MISSING VALUES

PROC LOGISTIC does not predict the estimated linear predictor and its standard error estimate, the fitted probabilities and confidence limits, and the regression diagnostic statistics for any observation with missing explanatory variable values. By default any observation with missing values for the response, or explanatory variables is excluded from the analysis. However, missing values are valid for the class variables if the MISSING option is specified with the CLASS statement. In this study data were corrected for missing values using PROC MI in SAS® (SAS® CODE 1). The regression method was used with the REG option in PROC MI to impute (Schafer, 1999) the continuous variables (AGE, SIZE etc.) and the discriminant function method (Schafer, 1999) with the DISCRIM option was used to impute the binary (AC, HEAT etc.) variables specified in the CLASS statement (SAS® CODE 1). MISSING option was specified with the CLASS statement in PROC LOGISTIC to correct for any missing observations in the class variables (AREA, TYPE).

SAS® CODE 1

```
PROC MI DATA=Reno SEED=7545417 NIMPUTE=10 OUT=outmi;
    CLASS GARAGE PATIO WD CABLE;
     MONOTONE REG(AGE SIZE PRICE BED BATH LOT)
     DISCRIM( GARAGE PATIO WD CABLE= AGE SIZE PARICE BED BATH
                    LOT/ DETAILS);
    VAR AGE SIZE PRICE BED BATH LOT GARAGE PATIO WD CABLE;
  RUN;
```

DESCRIPTIVE STATISTICS OF CONTINUOUS VARIABLES

Descriptive statistics (mean, standard deviation, minimum and maximum) for each continuous explanatory variable, and for each CLASS variable (TYPE, AREA), and the frequency counts of the classification levels were obtained by specifying the SIMPLE option with the LOGISTIC procedure (SAS® CODE 2). This option generates a breakdown of the simple descriptive statistics or frequency counts for the entire data set and also for individual response categories (not all data are shown here). Frequency of data in each property type is shown in Figure 1 for the entire data set. The frequency of the two categories of the response variable is shown in Figure 2 for the entire data set. Descriptive statistics of the continuous variables in the original data are shown in Table 1. Maximum and minimum number of bedrooms and bathrooms are 6 and 1 and 6.5 and 1 respectively. Standard deviations were high overall and are larger than means for the variables LOT and PRICE (Table 1). This is usually the case when there are influential observations and/or outliers in the data. Original data was checked for the presence of outliers and influential observations.

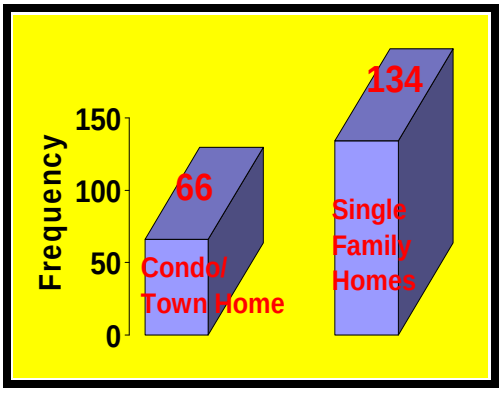

Figure 1: Frequency distribution of property type in the observed data

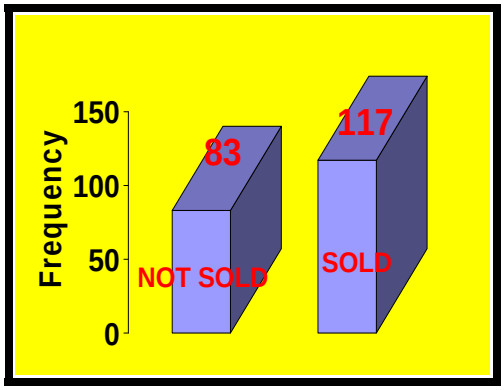

Figure 2: Frequency distribution of the response variable in the observed data

SAS® CODE 2

```
PROC LOGISTIC DATA=Reno SIMPLE OUTEST=betas COVOUT;
```

VARIABLE SELECTION

Since the data contained numerous indicator independent variables selection of variables were required to remove any insignificant independent variables before adding any significant independent variables. Four indicator independent variables (CABLE, WD, GARAGE, PATIO) that most significantly influenced the response variable out of the 25 indicator independent variables were selected using the stepwise selection method provided with PROC LOGISTIC in SAS®. Indicator independent variables thus selected most significantly influence the response variable hence reduce model variance. Stepwise selection method was

invoked in the PROC LOGISTIC procedure in SAS® by specifying the SELECTION=STEPWISE option (SAS® CODE 3).

Table 1: Descriptive Statistics of Continuous Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
# Bedrooms	2.8	1.06	1	6
# Bathrooms	2.0	1.10	1	6.5
Living Space (Ft^2)	1693.6	1151.8	427	6300
Lot Size (Acre)	0.19	0.27	0	2
Sale Price (X 1000 Dollars)	372.1	485.8	55	3995
Age (Years)	23.7	19.4	0	100

Best subset of variables was also obtained by specifying SELECTION =SCORE and the results compared. By default, SELECTION=NONE. A significance level of 0.3 (SLENTRY=0.3) was used to allow a variable into the model, and a significance level of 0.35 (SLSTAY=0.35) was used to keep a variable in the model. A detailed account of the variable selection process was requested by specifying the DETAILS option with the MODEL statement. The two class and the six continuous independent variables were included in the final model using the INCLUDE= n option (where n is the number of variables to be included) with the MODEL statement. The first n number of independent variables thus listed with the INCLUDE= option (SAS® CODE 3) in the MODEL statement will be included in all steps of the variable selection process thus in the final model.

The Hosmer and Lemeshow (Kutner et. al., 2004) goodness-of-fit test and R-square for the final selected model were requested by specifying the LACKFIT and RSQ options with the MODEL statement. The OUTEST= and COVOUT options were used to create a data set that contains parameter estimates and their covariance for the final selected model. The response variable option EVENT= '1' (sale of property is the event coded as =1) was used to model the probability of sale of property (SAS® CODE 3). In PROC LOGISTIC, the response with ordered value= 1 is regarded as the event. The variables selected by the STEPWISE method agreed well with the best subset of variables selected by the SCORE method. Sixteen out of the twenty five indicator independent variables were selected by the STEPWISE method hence were included along with the class and the continuous independent variables in the final model.

SAS® CODE 3

```
PROC LOGISTIC DATA=Reno OUTEST=betas COVOUT;
 CLASS AREA TYPE / PARAM=GLM;
     MODEL SOLD(EVENT='1') = AGE LOT PRICE SIZE BED BATH AC HEAT
           CABLE FRIDGE DW WD DISP WATER GAS CARPET DRAPES
           ELECTRIC VIEW PATIO DECK RECREATION STORAGE POOL
           TENNIS GOLF CARPORT SHOPPING SCHOOL GARAGE LANDSCAPE
          /SELECTION= STEPWISE INCLUDE= 8 SLENTRY=0.3
           SLSTAY=0.35 DETAILS LACKFIT RSQ;
RUN;
```

OUTLIERS AND INFLUENTIAL OBSERVATIONS

The INFLUENCE, and the IPLOTS options with PROC LOGISTIC were specified to obtain the regression diagnostics and the index plots (SAS® CODE 4). The INFLUENCE option displays columns for each diagnostic produced, and the case number which represents the sequence number of the observation. INFLUENCE option also produces line-printer plots (not shown here) where the vertical axis represents the

case number and the horizontal axis represents the value of the diagnostic statistic. The IPLOTS option produces the index plots of the diagnostic statistics on a more refined scale. The following influence plots were requested with the IPLOTS option: i) H (Hat diagonal matrix plot), ii) DFBETA plots for intercept and all independent variables, iii) C Displacement CBAR plot, iv) Delta Deviance plot, and v) Delta chi-square plot.

SAS® CODE 4

```
ODS LISTING CLOSE;
ODS graphics ON;

PROC LOGISTIC DATA=Reno SIMPLE OUTEST=betas COVOUT;
 CLASS AREA TYPE / PARAM=GLM;
     MODEL SOLD(EVENT='1') = AGE LOT PRICE SIZE BED BATH CABLE WD
           PATIO GARAGE / LACKFIT INFLUENCE IPLOTS;
     OUTPUT OUT=pred H=hat XBETA=xbeta DFBETAS=dfbetas P=phat
         LOWER=lcl UPPER=ucl PRED=pred C=c CBAR=cbar DIFDEV=difdev
         DIFCHISQ=difchisq predprob=(individual crossvalidate);
RUN;

ODS graphics OFF;
ODS HTML CLOSE;
```

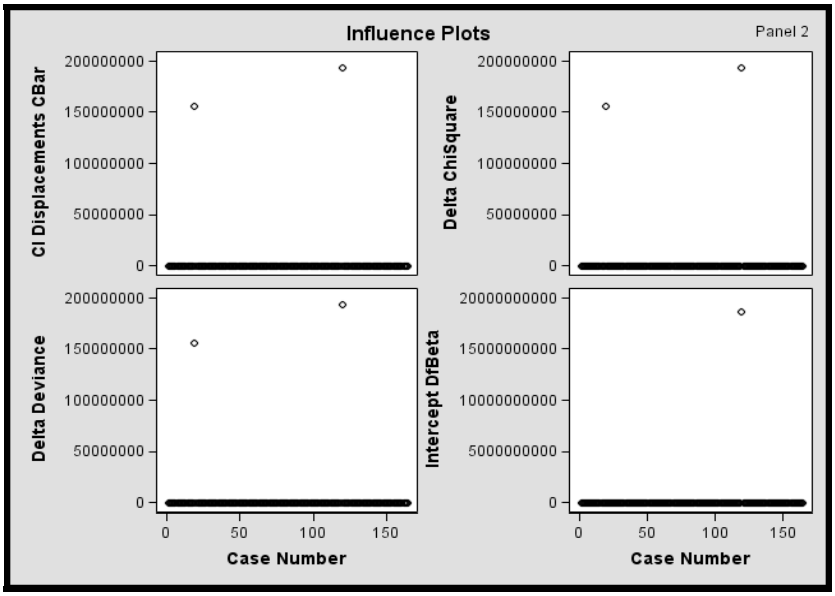

Figure 3: Influence plots of the observed data with outliers

FITTING THE MULTIPLE LOGISTIC REGRESSION MODEL

THE MULTIPLE LOGISTIC REGRESSION MODEL

The multiple logistic regression model was fitted with the previously selected independent variables against the binary indicator response variable Y (sold=1). Maximum likelihood parameter estimates, diagnostic and goodness of fit statistics, residuals, and odds ratios were obtained from the final fitted logistic regression model. The fitted model was then used to predict the probability of response by selected independent variables. The model to be fitted is shown below (Eq. 1).

$$E(Y_i) = \pi_i = \{ 1+ \exp(-\mathbf{X}'\boldsymbol{\beta})\}^{-1} \qquad (1)$$

X is the (p x 1 where p=22) vector of independent variables and **β** is the (p x 1 where p=22) vector of coefficients of the predictor variables. π is the E(Y), which is the Probability that Y=1. The log likelihood function of the above multiple logistic regression model is shown below (Eq. 2):

$$\text{Log}_e L(\boldsymbol{\beta}) = \Sigma (Y_i (\mathbf{X}'_i\boldsymbol{\beta}) - \Sigma \log_e [1+ \exp(\mathbf{X}'_i\boldsymbol{\beta})] \qquad (2)$$

where the summation extends from i = 1 to n, where n is the number of units which is 163. Numerical search procedures were used to find values of $\beta_o, \beta_1, \ldots, \beta_{p-1}$ that maximize $L(\boldsymbol{\beta})$. The maximum likelihood estimates of **β** denoted by **b** (p x 1 vector where p=22) is used to predict π_i (π_i(hat)). The logit response function (Eq. 3) was used to fit the model which is the inverse of the cumulative logistic distribution function.

$$\text{Logit}(\pi) = \text{Log}(\pi / [1-\pi]) \qquad (3)$$

The above specified model was fitted using PROC LOGISTIC in SAS® (SAS® CODE 5). Since the event category has a value equal to 1, the response level modeled was explicitly stated with the response variable option EVENT= '1' in the MODEL statement. By default, PROC LOGISTIC models the probability of response level with the lowest ordered value (0 in this case).

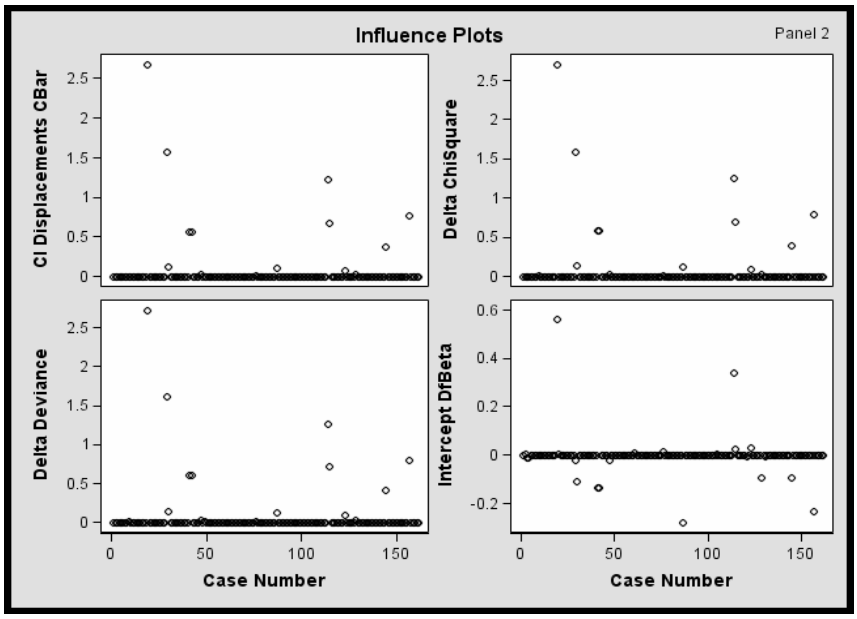

Figure 4: Influence plots of the outlier removed observed data

CLASS VARIABLE PARAMETERIZATION

Since there were two class variables (AREA, TYPE) with unbalanced replication of levels across variables the GLM parameterization option was specified with the CLASS statement. Otherwise since there are 4 levels in the class variable 'AREA' and two levels in the class variable ' TYPE', the fourth level of 'AREA' only occurs with the first level of 'TYPE'. Since the model contains the 'AREA(TYPE)' EFFECT, the design for 'AREA' within the second level of 'TYPE' will not be a differential effect. For this problem GLM or REFERENCE parameterization is more appropriate (Christensen, 1997). By default SAS® uses EFFECT parameterization. The GLM coding scheme estimates the difference in the effects of each level compared to the last level. The GLM parameterization of the class variables was specified with the PARAM= option available with the CLASS statement (SAS® CODE 5).

METHOD OF MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

Maximum likelihood estimation method was used to estimate the model parameters. Efficiency of parameter estimation was enhanced at every step of model fitting. Newton-Raphson estimation technique (Allison, 2001; Pampel, 2000) was specified with the TECHNIQUE= NEWTON option available with the MODEL statement to improve the efficiency of maximum likelihood estimation. By default SAS® uses the

FISHER estimation technique. To suppress SAS® from checking for infinite parameters the NOCHECK option available with the MODEL statement was used. This increases the speed of estimation of parameters with the maximum likelihood method in SAS®. Estimation process was further improved by ridging the log-likelihood function absolutely by specifying RIDGING=absolute in the MODEL statement. R-square and lack of fit parameters of the final model were requested through the LACKFIT and RSQ options. WALD test was selected for the confidence intervals for the odds ratios (CLODDS=WALD) and for the confidence intervals for the parameter estimates (CLPARM=WALD) (SAS® CODE 5). Error in the estimation as false positive and false negative values, and the sensitivity and specificity of the final model were computed using the CTABLE option available with the MODEL statement. Although log-scaling can handle numerically larger values in the data set and variables than normal scaling; however, computations in the log-scale are slower than computations in normal-scale. Therefore normal scaling was used to fit the model. By defaults SAS® uses the log-scaling which can be changed to normal by specifying the NOLOGSCALE computational option available with the MODEL statement (SAS® CODE 5).

The EXACT statement was used to perform exact tests on the continuous model variables. When the keyword INTERCEPT is specified followed by any effects in the MODEL statement, inference on the parameters of the specified effect is performed by conditioning on the sufficient statistics of all the other model parameters including the intercept (SAS® CODE 5).

SAS® CODE 5

```
PROC LOGISTIC DATA=Reno OUTEST=betas COVOUT EXACTONLY;
 CLASS AREA TYPE / PARAM=GLM;
 MODEL SOLD(EVENT='1') = AGE LOT PRICE SIZE BED BATH CABLE WD PATIO
       GARAGE
   /LACKFIT RSQ TECHNIQUE=NEWTON OUTROC=roc CLPARM=WALD CLODDS=WALD
       SCALE=none PPROB=0.5 CTABLE NODUMMYPRINT NOLOGSCALE NOCHECK;
 EXACT  'AGE, PRICE, SIZE' INTERCEPT AGE PRICE SIZE/ESTIMATE=both;
     TEST BED=BATH;
 OUTPUT OUT=pred  RESDEV=resdev RESCHI=reschi H=hat
       p=phat lower=lcl upper=ucl PRED=pred PREDPROB=(individual
       crossvalidate;
RUN;
```

RESULTS AND DISCUSSION

ESTIMATED PROBABILITIES

PROC LOGISTIC computes event probability for each observation and saves in the output data set specified with the OUTPUT OUT= statement. If the estimated event probability exceeds some cut-point value specified with the PPROB=statement the observation is predicted to be an event observation; otherwise, it is predicted as a nonevent observation.

ESTIMATED PROBABILITY PLOTS

Plots of the estimated probabilities along with observed probabilities and 95% confidence intervals as functions of the continuous independent variables were requested by specifying the OUTROC= and the GRAPHICS ESTPROB statements in the MODEL statement with the LOGISTIC procedure (SAS® CODE 6) Estimated probability of selling a unit decreases as the age of the property increases from 0 to 100 years (Figure 5). Older properties are harder to sell. Estimated probabilities decrease significantly as the price of the unit increases from $50,000 to $4000,000 (Figure 6). More expensive units are harder to sell. The decrease in the probability with price is steeper after $500,000 (Figure 6). Estimated probabilities increase significantly as the size or the living space of the property increases (Figure 7). People prefer larger properties in the college area. Probabilities of selling a property are not affected by the size of the lot or yard in the college area (Figure 8). The estimated probabilities of selling a property unit increases as the number of bedrooms and number of bathrooms increase (Figures 9 and 10). The ESTPROB option with the LOGISTIC procedure is experimental to SAS®. Estimated probability plots can only be obtained for continuous independent variables. This option also provides such plots for only one variable at a time.

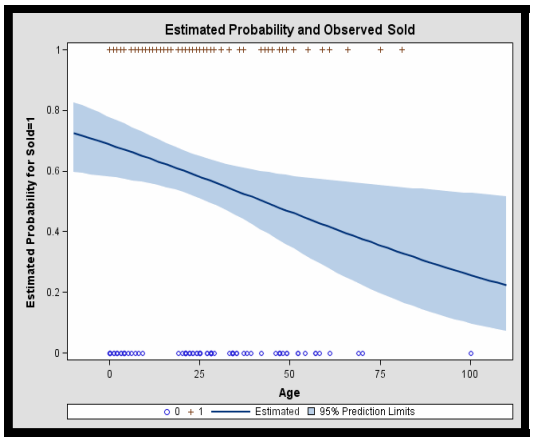

Figure 5: Estimated sale probabilities versus property age in years

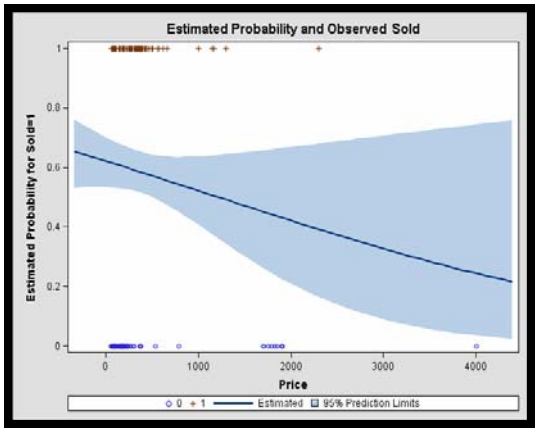

Figure 6: Estimated sale probabilities versus property sale price in (X $1000)

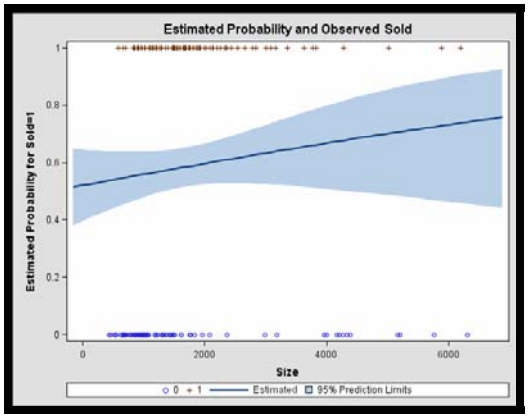

Figure 7: Estimated sale probabilities versus living space (size) of property in sq feet

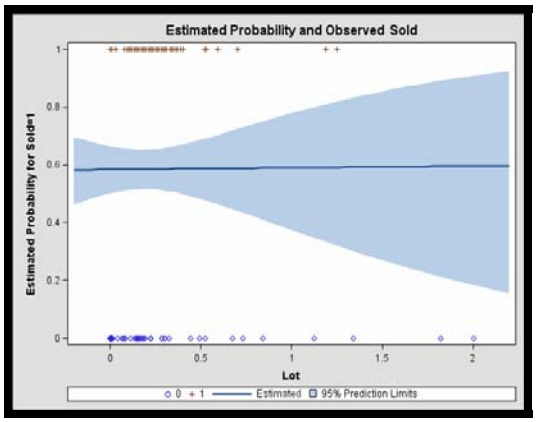

Figure 8: Estimated sale probabilities versus property lot size in acres

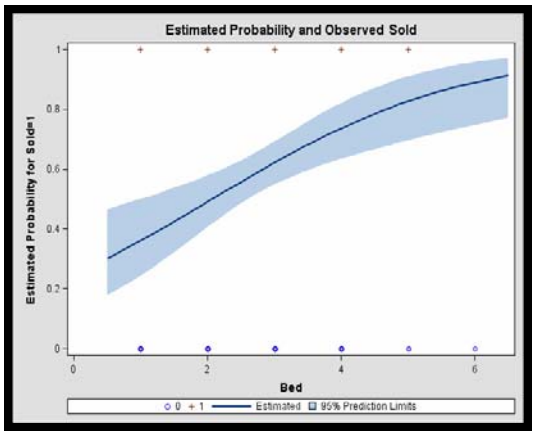

Figure 9: Estimated sale probabilities versus number of bedrooms

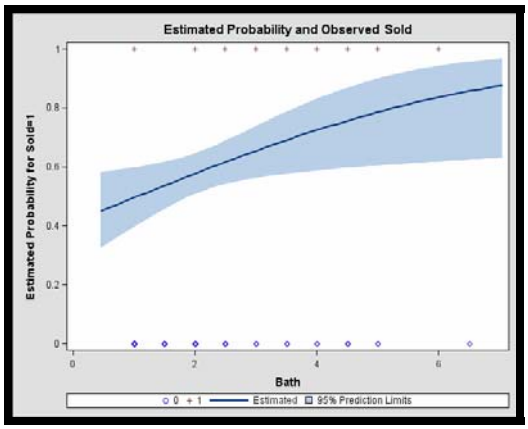

Figure 10: Estimated sale probabilities versus number of bathrooms

SAS® CODE 6

```
PROC LOGISTIC DATA=Reno;
   MODEL SOLD(EVENT='1')=BATH
   /   SCALE=none CLPARM=wald
       CLODDS=pl  RSQUARE
       OUTROC=roc1;
   UNITS BATH=1;
   GRAPHICS ESTPROB
RUN;
```

SAS® CODE 7

```
PROC UNIVARIATE DATA=pred;
 HISTOGRAM resdev reschi /
    NORMAL(COLOR=red W=3
    PERCENT=20 40 60 80
    MIDPERCENTS)CFILL= blue
      CFRAME = yellow;
RUN;
```

ESTIMATES OF PARAMETERS

Estimates for the parameters obtained through the maximum likelihood estimation method with 95% Wald's confidence limits for the final model are shown in Table 2. Negative parameter estimates were obtained for the variables AGE, PRICE, BED, BATH, and CABLE. These variables negatively influence the probability of selling the property units in the area. The interaction between selling price and the size of a property unit negatively influence the sale of the unit. Larger odds ratios were obtained for LOT, WD and GARAGE (Table 3). Odds ratios were close to or greater than 1 for most of the variables except for CABLE. These variables increase the probability of selling units more than other variables with smaller odds ratios.

Table 2: Wald Confidence Intervals for Parameter Estimates

Parameter	Estimate	95% Confidence Limits	
Intercept	4.6171	-4.0048	13.2390
Age	-0.0590	-0.1837	0.0656
Lot	-0.2351	-5.4205	4.9503
Price	-0.00365	-0.0181	0.0108
Size	0.00149	-0.00087	0.00386
Bed	-0.5304	-2.4228	1.3620
Bath	-0.5907	-2.0017	0.8204
Cable	-5.1703	-6.9076	-3.4330
WD	2.5084	1.2252	3.7916
Patio	0.5248	-0.8668	1.9165
Garage	1.4395	-0.2887	3.1678
Price*Bed	0.000631	-0.00279	0.00405
Price*Size	-6.72E-7	-3.37E-6	2.03E-6
Age*Bed	0.0110	-0.0354	0.0574
Age*Price	0.000071	-0.00005	0.000195

Table 3: Wald Confidence Intervals for Odds Ratios

Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.990	0.957	1.023
Lot	3.241	0.084	125.522
Price	0.998	0.995	1.001
Size	1.000	0.999	1.002
Bed	1.027	0.424	2.490
Bath	0.811	0.227	2.892
Cable	0.006	0.001	0.033
WD	14.155	3.962	50.565
Patio	1.975	0.512	7.613
Garage	4.473	1.030	19.428

Table 4: Linear Hypotheses Testing of BED vs BATH

Label	Wald Chi-Square	DF	Pr > Chi Sq
BED vs BATH	0.0160	1	0.8995

The parameter estimate for variable BED (-0.5304) is almost equal to that of variable BATH (-0.5907). WALD chi-square test was performed with the TEST option in PROC LOGISTIC to test the equality of influence of these two variables on sale probability. Results indicate that people prefer to have equal number of bathrooms to go with the bedrooms in the college area (Table 4).

MODEL GOODNESS OF FIT STATISTICS

All of the goodness of fit tests suggest that the model is significant and adequate. The AIC, SBC and -2Log likelihood (lower the better) values indicate that the model with the selected covariates is superior to the model with intercept only (Table 5). As in linear multiple regression, covariates can control for possible

confounding effects, and account for other sources of variation in multiple logistic regression models. Hosmer and Lemeshow goodness-of-fit test (Hosmer and Lemeshow, 2000; Kutner et. al., 2004) performs the Hosmer and Lemeshow goodness-of-fit test for the binary response logistic model. In this test the subjects are divided into approximately ten groups of roughly the same size based on the percentiles of the estimated probabilities. The discrepancies between the observed and expected number of observations in these groups are summarized by the Pearson chi-square statistic, which is then compared to a chi-square distribution with k degrees of freedom, where k is the number of groups (=10) minus n (=2). A large p-value (>0.05) usually suggests that the fitted model is an adequate model. In this study the test's large p-value (0.9545) indicates that the model predicts the data very well (Table 6). R-square and the maximum rescaled R-square values of the model are high enough (Table 7). The overall logistic regression model was highly significant at the 5% level as indicated by the Likelihood ratio, Wald and Score tests of the global null hypothesis that the model parameters are significant (Table 8).

Table 5: Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	273.451	129.499
SC	276.749	182.272
-2 Log L	271.451	97.499

Table 6: Hosmer and Lemeshow Goodness of Fit Test

Chi-Square	DF	Pr > ChiSq
2.6469	8	0.9545

Table 7: R-Square Values of the Model

General R-Square	0.5809
Max-rescaled R-Square	0.7823

PREDICTIVE ABILITY OF THE MODEL

PROC LOGISTIC provides four indices of rank correlation for assessing the predictive ability of a model: i) C, ii) Somer's D, iii) Goodman Kruskal Gamma (Gamma), and iv) Kendall's Tau (Tau-a). Values for the fitted model are shown in Table 9 along with the percent discordant and concordant pairs of event and nonevent observation. Values indicate the predictive ability of the model is adequate. 95% of the pairs are concordant. Values of Sommer's D, Gamma and C statistics are high enough. The statistic 'c' is the area under the Receiver Operating Characteristic (ROC) curve (Figure 11, discussed under model specificity and sensitivity).

The total number of concordant and discordant pairs are calculated by categorizing the predicted mean score for each observation into intervals of length k/500 (where k is the number of response levels) and accumulating the corresponding frequencies of observations where the predicted mean score of an observation is the sum of the ordered values minus one, weighted by the corresponding predicted probabilities for that observation. A pair is two observations with different response values. In a concordant pair of observations the observation with the lower ordered response (non-event) value will have lower predicted mean score than the observation with the higher ordered response value (event). In a discordant pair of observations the observation with the lower ordered response (non-event) value will have higher predicted mean score than the observation with the lower ordered response value (event). The pair is a tie if their predicted mean scores are equal.

Table 8: Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	173.9517	15	<.0001
Score	133.5400	15	<.0001
Wald	51.7990	15	<.0001

Table 9: Association of Predicted and Observed Responses

Percent Concordant	95.6	Somers' D	0.914
Percent Discordant	4.2	Gamma	0.915
Percent Tied	0.1	Tau-a	0.446
Pairs	9711	c	0.957

MODEL SPEICIFICITY AND SENSITIVITY

The classification table for the frequency of events (sold=1 or 0) of the fitted model is given in Table 10 The column "Prob Level" represents the cutoff probability value specified with the PPROB= option in the MODEL statement (a value of 0.5 was specified) for predicting whether an observation is an event (sold out

=1 is the event) or not (sold out =0 is non-event). The "Correct" columns list the numbers of subjects that are correctly predicted as events and nonevents, respectively, and the "Incorrect" columns list a) the number of nonevents incorrectly predicted as events, and b) the number of events incorrectly predicted as nonevents, respectively. For the probability of event = 0.5, the FALSE positive rate is 13.5% and the FALSE negative rate is 10.8% (Table 10). The FALSE positive is the proportion of predicted event responses that were observed as nonevents and the FALSE negative is the proportion of predicted nonevent responses that were observed as events. The sensitivity (the ability of the model to predict an event correctly) (93.2%) and specificity (the ability of the model to predict a nonevent correctly) (79.5%) of the model are high enough. The ROC curve for the fitted model was plotted and shown in Figure 11. ROC curve is a plot of model sensitivity versus 1- specificity of the model. One minus specificity is the proportion of non-event observations that are predicted to have an event response. Very large percent estimated area (95.70%, indicated by the 'c' statistic ('c' varies from 0 to 1)) under the ROC curve indicates adequate fit of the model (Figure 11).

Table 10: Classification Table for Predicted Event Frequencies

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.5	109	66	17	8	87.5	93.2	79.5	13.5	10.8

Table 11: Descriptive Statistics of Estimated Probabilities (pred)

N	Mean	Std Dev	Minimum	Maximum
200	0.585	0.4143	6.607E-8	0.9937

PREDICTED SALE PROBABILITIES

The MEANS procedure was used to obtain simple statistics for model predicted sale probabilities. Predicted probability for each unit was plotted against unit numbers**.** The overall mean probability of selling a unit is larger than 0.5 (0.585) (Table 11). According to the model 57.5% of the units have predicted sale probabilities larger than 0.75 (Figure 12) and 62% have predicted sale probabilities larger than 0.5.

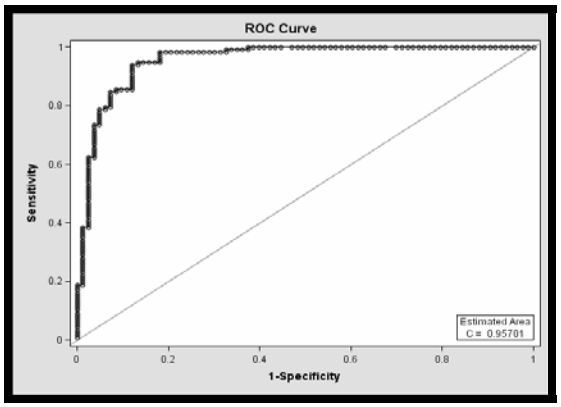

Figure 11: The Receiver Operating Characteristic Curve for the Final Model

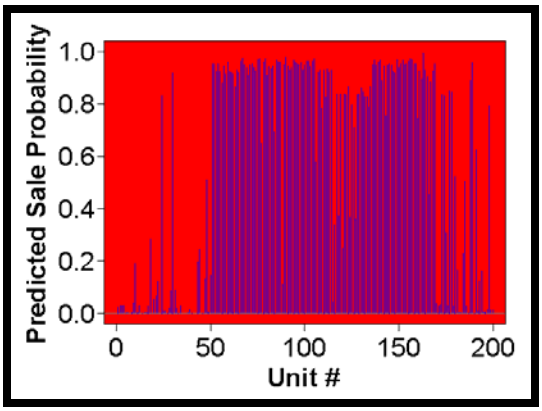

Figure 12: The model predicted sale probabilities for individual units.

RESIDUAL ANALYSIS

Pearson and Deviance residuals (Menard, 1995) versus case numbers for the final fitted multiple logistic regression model were requested with the RESCHI and RESDEV options in the MODEL statement in PROC LOGISTIC (SAS® CODE 5). PROC UNIVARIATE was used to plot histograms with normal plots superposed for the Pearson (Figure 13) and Deviance (Figure14) residuals obtained from the fitted multiple logistic regression model (SAS® CODE 7). The output dataset (pred) generated with the OUTPUT OUT= option was used as the input dataset for the UNIVARIATE procedure (SAS® CODE 7). The residual histograms indicate that the residuals from the final fitted multiple logistic regression model are normal. The fitted model represents the data adequately. Pearson (Figure 15) and Deviance (Figure 16) residuals were plotted against property unit numbers. These residual plots show no significant patterns and small enough values for the residuals. GPLOT option in SAS® along with the output dataset obtained (pred) from the fitted multiple logistic regression model were used to obtain these residual plots (SAS® CODE 8). This

dataset contains values of independent variables of the models, predicted probabilities, their confidence limits, model generated residuals, and influence statistics for all the observations.

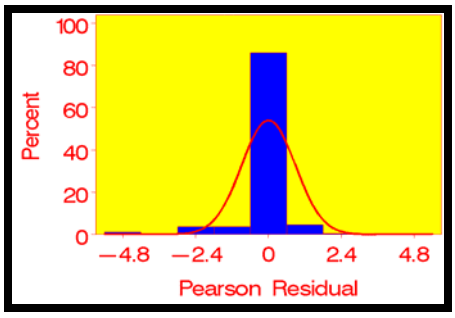

Figure 13: Pearson residual histogram of the final model

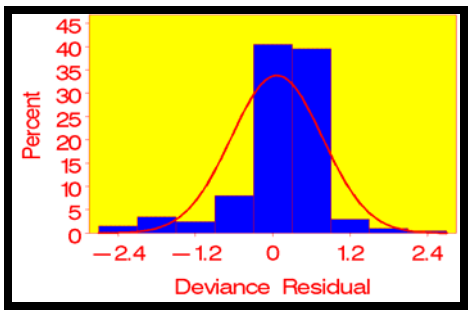

Figure 14: Deviance residual histogram of the final model

No significant model inadequacies are present in the residual plots. Small values of model residuals indicate that variability of the data described by the model is adequate.

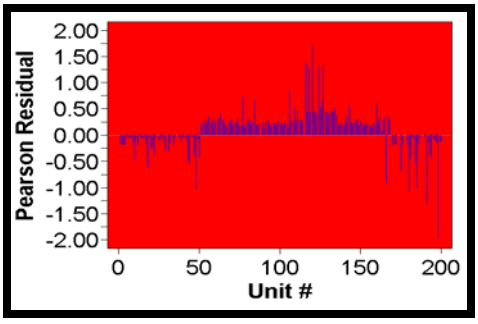

Figure 15: Pearson Residual Plot of the Final Model

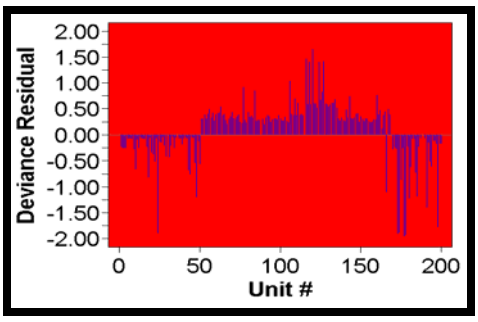

Figure 16: Deviance Residual Plot of the Final Model

SAS® CODE 8

```
  AXIS1 LABEL=(ANGLE=90 ROTATE=0 'Pearson Residual')
        MINOR=none ORDER=(-2 to 2 by 0.25);
  AXIS2 LABEL=('Unit #') MINOR=none ORDER=(0 to 200 by 50);
  SYMBOL1 VALUE=none c=blue h=0.35in i=needle r=19;

PROC GPLOT DATA=pred;
       PLOT reschi*id / FRAME CFRAME=red
       NOLEGEND VAXIS=axis1 HAXIS=axis2;
   RUN;
```

CONCLUSIONS

Ten out of twenty five independent variables were found significant and were used to fit the multiple logistic regression model. Maximum likelihood parameter estimates from the model are significant at 5% level. The fitted multiple logistic regression model predicts the data very well. Model diagnostics and residual plots indicate adequate fit of the multiple logistic regression model. P-value of Hosmer and Lemeshow goodness of fit test is 0.9545. Area under the model ROC curve is 0.9507 (95.07%). Sensitivity and specificity values of the model are 93.2% and 79.5% respectively. Odds ratios of the predicted variables of the logistic regression model support and confirm the conclusions drawn from the maximum likelihood parameter estimates. Estimated probability of selling a unit in the university area in Reno, Nevada increases as the price and age of the property unit decreases and as the number of bedrooms (up to 6), number of bathrooms (up to 6.5), and living space (up to 6000 square feet) increases. People in the university area also prefer to live in units with laundry facilities, garage and patio. People also prefer to have units with equal number of bathrooms as number of bedrooms in the university area.

REFERENCES

Kutner, M.H., Nachtsheim, C.J., and Neter, J. (2004) Applied Linear Regression Models, (4th ed.) Irwin, McGraw- Hill.

Schafer, J.L. (1999) Multiple Imputation: A Primer, Statistical Methods in Medical Research, 8, 3-15.

Allison, P.D. (2001) Logistic Regression using SAS Systems: Theory and Applications, (2nd ed.): SAS® Publishing, Books by Users. Press.

Christensen, R. (1997) Log-Linear Models and Logistic Regression, (2nd ed.) Springer-Verlag, New York.

Hosmer, D.W., and Lemeshow, S. (2000) Applied Logistic Regression (2nd ed.), New York, Wiley, 373pp.

Pampel, F.C. (2000) Logistic Regression: A Primer (1st ed.), London, Sage Publications Inc.

Menard, S. (1995) Applied Logistic Regression Analysis. Sage Publications Series: Quantitative Applications in the Social Sciences, No. 106

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Name: Anpalaki J. Ragavan M.S.
Enterprise: Department of Mathematics and Statistics, University of Nevada
Address: 3925, Clear Acre Lane, # 188,
Reno, NV 89512, USA.
Work phone: (775)-327-5260
Home phone: (775)-674-0397
Email: ragavan@unr.edu
Web: None

SAS and all other SAS Institute Inc. products or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration.

Other brand and product names are trademarks of their respective companies.