Logistic Regression

A Presentation To The Michigan SAS Users Group 2023 Conference

By

Mike Oliansky

MikeOliansky@gmail.com

Logistic Regression

Example
Start with database of customers and their purchases with some appended demographic information
Goal
Rank order the list from the most likely prospects to perform the behavior of interest to those least likely
Purpose
There are costs associated with contacting a customer by postal mail or email
We want to select the people who are most likely to respond to the communication

Types of models
- Likely to purchase any vehicle
- Likely to purchase a specific model (F-150, Accord)
- Likely to service at a dealership
- Likely to open email.

Logistic Regression

Why logistic?

We want to predict a probability so we can rank order a list
Probability is bounded by 0 and 1

An OLS multiple regression is unbounded and not suitable for probabilities

The "logit" model solves these problems:

$\ln[p/(1-p)] = \alpha + \beta X + e$

It's bounded by 0 and 1 and is easily converted to a probability

Dependent Variable · Population Y intercept · Population Slope Coefficient · Independent Variable · Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Logistic Regression

Logit model
$\ln[p/(1-p)] = \alpha + \beta X + e$

p/(1-p) is an odds ratio

Odds are just counts, 2 people out of 100 bought a vehicle, odds are 2:98 someone will buy

Odds ratio compares one group to another.

High income: 2 out of 100 purchased, odds are 2:98

Low income: 1 out of 100 purchased, odds are 1:99

Odds ratio $2{:}98/1{:}99$

So the higher income group is twice as likely to make a purchase as the lower income group

Logistic Regression

```
proc logistic data=develop_fix descending  plots=roc   namelen = 32  ;  /* descending models on 1 not 0*/
        *class total_bnsr_SUV(ref = "0");                                            /* class variable – models each value*/
        model &dv = &final_predictors   / stb      selection=forward lackfit   clodds=wald ;
                                                                              /* model statement, using macro variables  */
                                                                               /*stb – standardized regression coefficients */
                                                                               /*selection = forward, most predictive variable enters first
                                                                                     then other 'lessor' predictors if they are significant */
                                                                                /*clodds – gives odds ratio table    */
        score data=develop_fix      out=develop_fix_scored;      /*scores a dataset with the equation developed here*/
        score data=validate_fix      out= validate_fix_scored ;
        score data=validate_all_fix  out= validate_all_fix_scored;

        store out = MODATA3.scoring_BNSR_Seltos_29NOV2022; /contains info to score another dataset with the
                                                                                           predictors in the equation*/
run;
```

Scoring

```
 proc plm restore= MODATA3.scoring_bnsr_seltos_29nov2022;
         score data =  scoring_BNSR_dataset_&sysdate9    out = choice_file_scored    predicted / ilink;
run;
```

Logistic Regression

Some output

Top obs read – be sure all obs are used - listwise deletion

Middle – Response profile

N of events and non-events

Summary of forward selection

variables with sufficient p values to enter the model

Number of Observations Read	66781
Number of Observations Used	66781

Response Profile

Ordered Value	purchase_SUV2022	Total Frequency
1	1	5119
2	0	61662

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Newest BNSR MSRP	1	1	701.2065	<.0001
2	Total BNSR Small SUV	1	2	335.2454	<.0001
3	Lead 30 Days	1	3	165.6434	<.0001
4	Newest BNST Other Small SUV	1	4	136.0767	<.0001
5	Newest BNST Other Small SUV	1	5	116.0599	<.0001
6	Lead 60 Days	1	6	38.2545	<.0001
7	Multiple OEM BNSR	1	7	30.5000	<.0001
8	Age 30-39	1	8	23.3430	<.0001
9	Total BNSR Car	1	9	20.9666	<.0001
10	Female In HH	1	10	14.4349	0.0001
11	Age 50-59	1	11	16.3496	<.0001
12	Historical New OEM Vehicles	1	12	12.6367	0.0004

Logistic Regression

Assoc pred and obs
Left side all observations are paired
Concordant – higher value is 1, lower is 0
Discordant – reverse of above
Tied – same logit score

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	65.7	Somers' D	0.315
Percent Discordant	34.3	Gamma	0.315
Percent Tied	0	Tau-a	0.045
Pairs	315647778	c	0.657

Hosmer-Lemeshow Test
Divide sample into groups – attempt to divide into deciles if model supports
Chi – Sq test – do the counts from the known results square up with the counts from the predicted p values from the model
Conservative test

Partition for the Hosmer and Lemeshow Test

		purchase_small_suv_2022 = 1		purchase_small_suv_2022 = 0	
Group	Total	Observed	Expected	Observed	Expected
1	6678	180	164.23	6498	6513.77
2	6677	212	240.30	6465	6436.70
3	6680	277	306.44	6403	6373.56
4	6681	361	381.48	6320	6299.52
5	6678	449	447.19	6229	6230.81
6	6678	515	507.83	6163	6170.17
7	6679	596	568.45	6083	6110.55
8	6680	692	637.96	5988	6042.04
9	6678	798	782.13	5880	5895.87
10	6672	1039	1082.99	5633	5589.01

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
18.2764	8	0.0192

Logistic Regression

Lift is a common metric in direct marketing

Lift is related to how well your model separates buyers from non buyers and gives them higher scores which puts them into the upper deciles

$Lift = Percent\ buy\ rate\ in\ a\ decile / Buy\ rate\ for\ all\ observations\ \times 100$

Logistic Regression

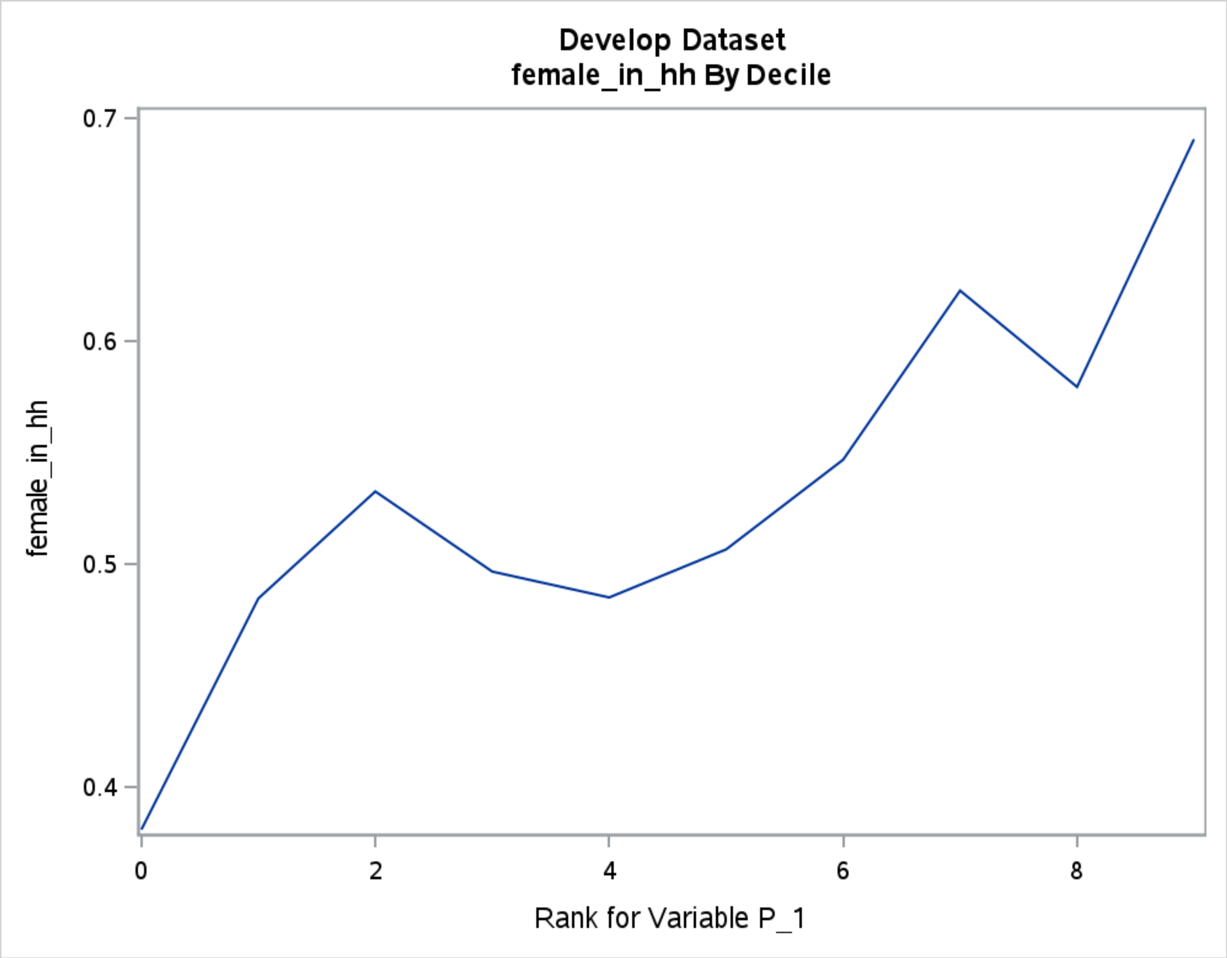
Decile 9 is higher scores

This tables has lift, predicted and observed buy rates, as well as averages for the predictor values across the deciles.

		Rank for Variable P_1									
		0	1	2	3	4	5	6	7	8	9
Obs Pct Buy Rat	Mean	2.70	3.17	4.15	5.41	6.72	7.71	8.93	10.4	12.0	15.6
Lift	Mean	35.00	41.00	54.00	71.00	88.00	101.00	116.00	135.00	156.00	203.00
Cumulative BR	Mean	7.67	8.22	8.85	9.52	10.20	10.90	11.70	12.62	13.76	15.56
Cumulative Lift	Mean	100.00	107.00	115.00	124.00	133.00	142.00	153.00	165.00	180.00	203.00
Predicted Pct Buy Rate	Mean	2.46	3.60	4.59	5.71	6.70	7.60	8.51	9.55	11.71	16.23
FREQ	Mean	6678.00	6678.00	6679.00	6677.00	6678.00	6680.00	6677.00	6678.00	6677.00	6679.00
Decile Count Percent	Mean	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
Newest BNSR MSRP	Mean	42804.62	35480.86	32207.61	28684.19	26135.41	24059.08	22877.03	22728.41	21431.87	21390.61
Total BNSR Small SUV	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.31
Lead 30 Days	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Newest BNST Other Small SUV	Mean	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.10	0.58	0.66
Newest BNST Other Small SUV	Mean	0.02	0.03	0.06	0.14	0.23	0.30	0.34	0.48	0.27	0.08
Lead 60 Days	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Multiple OEM BNSR	Mean	0.51	0.31	0.29	0.27	0.28	0.24	0.17	0.18	0.26	0.19
Age 30-39	Mean	0.30	0.20	0.19	0.20	0.18	0.16	0.08	0.05	0.12	0.05
Total BNSR Car	Mean	0.14	0.21	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Female In HH	Mean	0.38	0.48	0.53	0.50	0.49	0.51	0.55	0.62	0.58	0.69
Historical New OEM Vehicles	Mean	0.28	0.31	0.39	0.38	0.38	0.32	0.29	0.46	0.46	0.44
Age 50-59	Mean	0.22	0.27	0.31	0.28	0.30	0.30	0.38	0.44	0.35	0.40

Logistic Regression

Besides the table, graphs are often helpful



Logistic Regression

One last thing, check the correlations of the predictors

Name1	Name2	corr	abs_corr
newest_bnsr_is_other_SUV	newest_bnsr_msrp	-.41582	.41582
age_p1_50_59	age_p1_30_39	-.29502	.29502
age_p1_50_59	female_in_hh	-.22625	.22625
newest_bnsr_is_other_SUV	newest_bnsr_is_other_SUV	-.17028	.17028
garage_new_oem	newest_bnsr_is_other_SUV	.16043	.16043
historical_new_oem	garage_new_oem	.14837	.14837
newest_bnsr_is_other_SUV	newest_bnsr_msrp	.11202	.11202
total_bnsr_car	newest_bnsr_msrp	.09276	.09276
total_bnsr_car	garage_new_oem	.09082	.09082