Diane Nguyen

MSBA 635- Data Analytics 2

Homework 2

1) Use the las vegas data for the following:

a) Obtain the frequency distribution of *delinquent* using SAS 9.4 and proc freq.

```
* print data;
proc print data=tmp1.Lasvegas;
run;
* display data attributes;
proc contents data=tmp1.Lasvegas;
run;

* produce frequencies;
proc freq data=tmp1.Lasvegas;
tables delinquent;
run;
```

The following is a frequency distribution of the number of loans that are delinquent which is defined as payments that are past 90 days late. 80.10% of all payments are not delinquent, while 19.90% of loans in this data set are classified as delinquent.

The SAS System 16:35 Saturday, January 26, 2019 316

The FREQ Procedure

= 1 if payment late by 90+ days

DELINQUENT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	801	80.10	801	80.10
1	199	19.90	1000	100.00

b) Estimate the following linear probability model using SAS 9.4 and proc reg:

$delinquent = \beta_1 + \beta_2 lvr + \beta_3 ref + \beta_4 insur + \beta_5 rate + \beta_6 amount + \beta_7 Ccredit + \beta_8 term + \beta_9 arm + \varepsilon$

```
proc reg data=tmp1.Lasvegas;
model Delinquent=lvr ref insur rate amount credit term arm;
output out=lpmout p=phat_lpm;
run;
quit;
```

The SAS System 16:35 Saturday, January 26, 2019 317

The REG Procedure
Model: MODEL1
Dependent Variable: DELINQUENT = 1 if payment late by 90+ days

Number of Observations Read 1000
Number of Observations Used 1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F

Model	8	53.60627	6.70078	62.77	<.0001
Error	991	105.79273	0.10675		
Corrected Total	999	159.39900			

Root MSE	0.32673	R-Square	0.3363
Dependent Mean	0.19900	Adj R-Sq	0.3309
Coeff Var	164.18671		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.68849	0.21125	3.26	0.0012
LVR	loan amount to value of property, percent	1	0.00162	0.00078456	2.07	0.0387
REF	= 1 if for a refinance, 0 if for purchase	1	-0.05932	0.02383	-2.49	0.0130
INSUR	= 1 if borrower has mortgage insurance	1	-0.48158	0.02364	-20.37	<.0001
RATE	initial interest rate	1	0.03438	0.00860	4.00	<.0001
AMOUNT	loan amount in $100,000 units	1	0.02377	0.01267	1.88	0.0610
CREDIT	credit score	1	-0.00044190	0.00020181	-2.19	0.0288
TERM	loan term in years	1	-0.01262	0.00354	-3.57	0.0004
ARM	= 1 if adjustable rate mortgage, 0 if fixed	1	0.12832	0.03189	4.02	<.0001

Show some of the predicted probabilities lie outside of the [0,1] interval hence invalidating the use of proc reg for this type of data generation process (hint: use proc univariate for displaying order statistics like minimum and maximum).

```
proc Univariate data=work.lpmout;
run;
quit;
```

Once I obtained the maximum and minimum values through running a proc univariate procedure I was able to obtain some of the predicted probabilities by looking at the extreme observations for phat_lpm. Here you can see that some of the values are negative on the lowest extreme observations.

The SAS System 16:35 Saturday, January 26, 2019 337

The UNIVARIATE Procedure
Variable: phat_lpm (Predicted Value of DELINQUENT)

Extreme Observations

------Lowest------		------Highest-----	
Value	Obs	Value	Obs
-0.203935	193	0.752697	416
-0.184994	151	0.754550	17
-0.179662	857	0.769252	442
-0.178536	949	0.782932	382
-0.177114	809	0.792123	4

c) Estimate the following probit model using SAS 9.4 and proc qlim:

$delinquent = \beta_1 + \beta_2 lvr + \beta_3 ref + \beta_4 insur + \beta_5 rate + \beta_6 amount + \beta_7 Ccredit + \beta_8 term + \beta_9 arm + \varepsilon$

```
proc qlim data=tmp1.Lasvegas;
model Delinquent= lvr ref insur rate amount credit term arm/discrete;
output out=probitout xbeta marginal;
run;
quit;
```


The QLIM Procedure

Discrete Response Profile of DELINQUENT

Index	Value	Total Frequency
1	0	801
2	1	199

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	DELINQUENT
Number of Observations	1000
Log Likelihood	-332.79661
Maximum Absolute Gradient	0.0000698
Number of Iterations	18
Optimization Method	Quasi-Newton
AIC	683.59322
Schwarz Criterion	727.76302

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	332.43	2 * (LogL - LogL0)
Upper Bound of R (U)	998.03	- 2 * LogL0
Aldrich-Nelson	0.2495	R / (R+N)
Cragg-Uhler 1	0.2828	1 - exp(-R/N)
Cragg-Uhler 2	0.4479	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.3326	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.3145	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.3331	R / U
Veall-Zimmermann	0.4995	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.4576	

N = # of observations, K = # of regressors

Algorithm converged.

The QLIM Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	0.964646	1.087393	0.89	0.3750
LVR	1	0.007601	0.004591	1.66	0.0978
REF	1	-0.288456	0.125898	-2.29	0.0220
INSUR	1	-1.772714	0.115765	-15.31	<.0001
RATE	1	0.171199	0.043839	3.91	<.0001
AMOUNT	1	0.121236	0.061546	1.97	0.0489

CREDIT	1	-0.001913	0.001062	-1.80	0.0717
TERM	1	-0.077577	0.019835	-3.91	<.0001
ARM	1	0.809111	0.207745	3.89	<.0001

Ascertain the statistical significance of each slope coefficient (i.e., skip intercept). Interpret each slope coefficient.

According to the QLIM Procedure above, of the 8 variables (lvr, ref, insur, rate, amount, credit, term, and arm) are above the .05 alpha level of significance.

- Lvr is significant at the .10 alpha level of significance. For every one unit increase there is a .076% increase in lvr.
- Ref is significant at the .05 alpha level of significance. For every one unit increase there is a 28.85% decrease in ref.
- Insur is significant at the .05 alpha level of significance. For every one unit increase there is a 177.27% decrease in insur.
- Rate is significant at the .05 alpha level of significance. For every one unit increase there is a 17.11% increase in rate.
- Amount is significant at the .05 alpha level of significance. For every one unit increase there is a 12.12% increase in amount.
- Credit is significant at the .10 alpha level of significance. For every one unit increase there is a .019% decrease in credit.
- Term is significant at the .05 alpha level of significance. For every one unit increase there is a 7.78% decrease in term.
- Arm is significant at the .05 alpha level of significance. For every one unit increase there is a 80.91% increase in arm.

d) Obtain and interpret the average marginal effect on the variable *amount*

```
proc means data=work.probitout;
var meff_p1_amount meff_p2_amount;
run;
quit;
```

For every one unit increase in the amount there is a 2.23% increase in the probability that the amount will be delinquent. For every one unit decrease in the amount there is a 2.23% decrease in the probability that the amount owed will be delinquent.

The MEANS Procedure

Variable	Label	N	Mean
Meff_P1_AMOUNT	Marginal effect of AMOUNT on the probability of DELINQUENT=1	1000	-0.0222696
Meff_P2_AMOUNT	Marginal effect of AMOUNT on the probability of DELINQUENT=2	1000	0.0222696

Variable	Label	Std Dev
Meff_P1_AMOUNT	Marginal effect of AMOUNT on the probability of DELINQUENT=1	0.0151608
Meff_P2_AMOUNT	Marginal effect of AMOUNT on the probability of DELINQUENT=2	0.0151608

Variable	Label	Minimum
Meff_P1_AMOUNT	Marginal effect of AMOUNT on the probability of DELINQUENT=1	-0.0483662
Meff_P2_AMOUNT	Marginal effect of AMOUNT on the probability of DELINQUENT=2	0.000365236

Variable	Label	Maximum
Meff_P1_AMOUNT	Marginal effect of AMOUNT on the probability of DELINQUENT=1	-0.000365236
Meff_P2_AMOUNT	Marginal effect of AMOUNT on the probability of DELINQUENT=2	0.0483662

e) Using a threshold of 50 percent, create a 2x2 table of actual versus predicted delinquent using SAS 9.4 and proc freq. What percent of total observations were correctly classified?

```
data LasVegas_e;
set work.probitout;
phat=probnorm(xbeta_Delinquent);
phat_classification=(phat>=0.50);
run;


proc print data=LasVegas_e;
run;
```


Obs	LVR	REF	INSUR	RATE	AMOUNT	CREDIT	TERM	ARM	DELINQUENT	Xbeta_ DELINQUENT	Meff_P1_LVR	Meff_P2_ LVR
995	90.0	1	0	8.650	2.38500	527	30	1	0	0.60386	-.002526877	.002526877
996	80.0	0	0	7.250	1.94800	624	30	1	1	0.33807	-.002863829	.002863829
997	20.0	0	0	10.875	0.48700	624	30	0	1	-0.48361	-.002697613	.002697613
998	80.0	0	0	8.720	1.80800	638	30	1	0	0.54598	-.002612375	.002612375
999	20.0	0	0	12.490	0.45200	638	30	0	0	-0.23815	-.002947475	.002947475
1000	88.2	1	0	7.650	2.91000	624	30	1	0	0.29705	-.002901380	.002901380

Obs	Meff_P1_ REF	Meff_P2_ REF	Meff_P1_ INSUR	Meff_P2_ INSUR	Meff_P1_ RATE	Meff_P2_ RATE	Meff_P1_ AMOUNT	Meff_P2_ AMOUNT	Meff_P1_ CREDIT
995	0.09590	-0.09590	0.58934	-0.58934	-0.056915	0.056915	-0.040305	0.040305	.000636024
996	0.10869	-0.10869	0.66793	-0.66793	-0.064505	0.064505	-0.045680	0.045680	.000720836
997	0.10238	-0.10238	0.62916	-0.62916	-0.060761	0.060761	-0.043028	0.043028	.000678999
998	0.09914	-0.09914	0.60928	-0.60928	-0.058841	0.058841	-0.041669	0.041669	.000657544
999	0.11186	-0.11186	0.68744	-0.68744	-0.066389	0.066389	-0.047014	0.047014	.000741890
1000	0.11011	-0.11011	0.67669	-0.67669	-0.065351	0.065351	-0.046279	0.046279	.000730287

Obs	Meff_P2_ CREDIT	Meff_P1_ TERM	Meff_P2_ TERM	Meff_P1_ ARM	Meff_P2_ ARM	phat	phat_classification
995	-.000636024	0.025791	-0.025791	-0.26899	0.26899	0.72703	1
996	-.000720836	0.029230	-0.029230	-0.30486	0.30486	0.63235	1
997	-.000678999	0.027533	-0.027533	-0.28717	0.28717	0.31433	0
998	-.000657544	0.026663	-0.026663	-0.27809	0.27809	0.70746	1
999	-.000741890	0.030083	-0.030083	-0.31376	0.31376	0.40588	0
1000	-.000730287	0.029613	-0.029613	-0.30886	0.30886	0.61679	1

```
*2x2 table of actual vs predicted delinquent using proc freq;
proc freq data=work.LasVegas_e;
tables Delinquent * phat_classification;
run;
quit;
```

```
The SAS System        16:35 Saturday, January 26, 2019 413

                              The FREQ Procedure

                    Table of DELINQUENT by phat_classification

       DELINQUENT(= 1 if payment late by 90+ days)
                 phat_classification

       Frequency,
       Percent  ,
       Row Pct  ,
       Col Pct  ,       0,       1,  Total
       ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
              0 ,    735 ,     66 ,    801
                ,  73.50 ,   6.60 ,  80.10
                ,  91.76 ,   8.24 ,
                ,  90.29 ,  35.48 ,
       ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
              1 ,     79 ,    120 ,    199
                ,   7.90 ,  12.00 ,  19.90
                ,  39.70 ,  60.30 ,
                ,   9.71 ,  64.52 ,
       ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
       Total         814      186     1000
                   81.40    18.60   100.00
```

According to the above classification matrix, 735 were correctly classified ad not delinquent and 120 were correctly classified as being delinquent. This means that 855 of the 1000 observations were correctly classified. This means that 85.5% were correctly classified.

f) Using your output from part (e), what percent of total observations were predicted to be delinquent but actually were not? What cost is incurred with these misclassified observations (hint: it is an opportunity cost)? How can the threshold be changed to reduce this type of cost?

According to the output from part (e), 66 of the 1000 observations were predicted to be delinquent but were actually not. This means that 6.6% of all the observations were incorrectly classified as delinquent but they actually were not.

In terms of the cost incurred this misclassification signals an opportunity cost. Due to this misclassification the individuals that were misclassified could not get a loan that they were seeking or a loan in the amount that they were seeking. The threshold can be changed to reduce this type of cost by perhaps adding more variables to train our model to make it more accurate by improving the R-square.

g) Using your output from part (e), what percent of total observations were not predicted to be delinquent but actually were? What cost is incurred with these misclassified observations assuming the delinquent accounts write-off and the outstanding balances cannot be recovered by the collections department? How can the threshold be changed to reduce this type of cost?

Using the output from part (e) 79 of the 1000 observations were predicted to not be delinquent but they actually were. This means that 7.9% of the observations were misclassified as not delinquent but they actually were. The cost that is incurred is in the form of opportunity cost. Since these delinquent accounts cannot be marked to be written off by the bank then they lose capital that they can use to invest in other ventures that may be profitable. The threshold can be changed to reduce this type of cost by including more variables of measure in the model to make it predict results more accurately.