Taylor Eckert

Homework 2

Data Analytics 2

1/29/19

A) 1) Use the lasvegas data for the following: a) Obtain the frequency distribution of delinquent using SAS 9.4 and proc freq.

The FREQ Procedure

Cumulative Cumulative

DELINQUENT Frequency Percent Frequency Percent

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

0 801 80.10 801 80.10

1 199 19.90 1000 100.00

The FREQ Procedure

a)

Cumulative Cumulative

DELINQUENT Frequency Percent Frequency Percent

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

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 = B1 + B2 lvr + B3 ref + B4 insur + B5 rate + B6 amount + B7 credit + B8 term + B9 arm + e

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).

The REG Procedure

Model: MODEL1

Dependent Variable: DELINQUENT

Number of Observations Read 1000

Number of Observations Used 1000

Analysis of Variance

Sum of Mean

Source DF Squares 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

Parameter Standard

Variable DF Estimate Error t Value Pr > |t|

Intercept 1 0.68849 0.21125 3.26 0.0012

LVR 1 0.00162 0.00078456 2.07 0.0387

REF 1 -0.05932 0.02383 -2.49 0.0130

INSUR 1 -0.48158 0.02364 -20.37 <.0001

RATE 1 0.03438 0.00860 4.00 <.0001

AMOUNT 1 0.02377 0.01267 1.88 0.0610

CREDIT 1 -0.00044190 0.00020181 -2.19 0.0288

TERM 1 -0.01262 0.00354 -3.57 0.0004

ARM 1 0.12832 0.03189 4.02 <.0001

The UNIVARIATE Procedure

Variable: phat\_lpm

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

The output shows probabilities under 0 (negative) which is not possible, so we cannot use this procedure. Instead we must use a logit/probit for more accurate results.

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

delinquent = B1 + B2 lvr + B3 ref + B4 insur + B5 rate + B6 amount + B7 credit + B8 term + B9 arm + e

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

The QLIM Procedure

Parameter Estimates

Standard Approx

Parameter DF Estimate Error t Value 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

LVR: 0.0978>0.05, so not significantly significant. As the loan amount to value increases/decreases, the probability of delinquency increases/decreases, ceteris paribus.

REF: 0.0220<0.05, so significantly significant. As the borrower refinancing increases/decreases, the probability of delinquency decreases/increases, ceteris paribus.

INSUR: <0.0001<0.05, so significantly significant. As the likelihood of the borrower to have mortgage insurance increases/decreases, the probability of delinquency decreases/increases, ceteris paribus.

RATE: <0.0001<0.05, so significantly significant. As the initial interest rate increases/decreases, the probability of delinquency increases/decreases, ceteris paribus.

AMOUNT: 0.0489<0.05, so significantly significant. As the amount borrowed increases/decreases, the probability of delinquency increases/decreases, ceteris paribus.

CREDIT: 0.0717>0.05, so not significantly significant. As the amount of credit score increases/decreases, the probability of delinquency decreases/increases, ceteris paribus.

TERM:<0.0001<0.05, so significantly significant. As the amount of terms borrowed increases/decreases, the probability of delinquency decreases/increases, ceteris paribus.

ARM: <0.0001<0.05, so significantly significant. As the adjustable rate mortgage increases/decreases, the probability of delinquency increase/decreases, ceteris paribus.

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

The MEANS Procedure

Variable N Mean Std Dev Minimum Maximum

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

Meff\_P1\_AMOUNT 1000 -0.0222696 0.0151608 -0.0483662 -0.000365236

Meff\_P2\_AMOUNT 1000 0.0222696 0.0151608 0.000365236 0.0483662

Ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

The MEANS Procedure

Variable N Mean Std Dev Minimum Maximum

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

Meff\_P1\_AMOUNT 1000 -0.0222696 0.0151608 -0.0483662 -0.000365236

Meff\_P2\_AMOUNT 1000 0.0222696 0.0151608 0.000365236 0.0483662

Ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

As the amount borrowed increases/decreases, the probability of delinquency increases/decreases by 2.22 percentage points, ceteris paribus.

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?

The FREQ Procedure

Table of DELINQUENT by phat\_classify

DELINQUENT phat\_classify

Frequency‚

Percent ‚

Row Pct ‚

Col Pct ‚ 0‚ 1‚ Total

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

0 ‚ 735 ‚ 66 ‚ 801

‚ 73.50 ‚ 6.60 ‚ 80.10

‚ 91.76 ‚ 8.24 ‚

‚ 90.29 ‚ 35.48 ‚

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

1 ‚ 79 ‚ 120 ‚ 199

‚ 7.90 ‚ 12.00 ‚ 19.90

‚ 39.70 ‚ 60.30 ‚

‚ 9.71 ‚ 64.52 ‚

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

Total 814 186 1000

81.40 18.60 100.00

e)

The FREQ Procedure

Table of DELINQUENT by phat\_classify

DELINQUENT phat\_classify

Frequency‚

Percent ‚

Row Pct ‚

Col Pct ‚ 0‚ 1‚ Total

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

0 ‚ 735 ‚ 66 ‚ 801

‚ 73.50 ‚ 6.60 ‚ 80.10

‚ 91.76 ‚ 8.24 ‚

‚ 90.29 ‚ 35.48 ‚

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

1 ‚ 79 ‚ 120 ‚ 199

‚ 7.90 ‚ 12.00 ‚ 19.90

‚ 39.70 ‚ 60.30 ‚

‚ 9.71 ‚ 64.52 ‚

ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ

Total 814 186 1000

81.40 18.60 100.00

Above, the table shows the cross tabulation of delinquency and phat\_classification. The top 0, 1 is the phat\_classify and the left 0,1 is the actual delinquency. The diagonals show that 73.5% of predicted zero and actual zero were correct. It also shows that 12% of predicted 1 and actual 1 were correct. Therefore, the model classified 1/1 and 0/0 85.5% of the time. The model classified it incorrectly 14.5% of the time.

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?

6.60% of total observations were predicted to be delinquent, but were not. Knowing that some borrowers proved to be trustworthy after all, providing variables such as income, previous credit scores, etc could improve the fit.

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?

7.90% of total observations were not predicted to be delinquent, but actually were according to the table. Fit could be improved by providing reasons/explanatory valuables for borrowers missing payments, including income brackets, etc.

Code:

\*Taylor Eckert

\*MSBA 635 Data Analytics II

\*Homework 2

\* print data;

**proc** **print** data=tmp1.lasvegas;

**run**;

\* display data attributes;

**proc** **contents** data=tmp1.lasvegas;

**run**;

\* a) Obtain the frequency distribution of Delinquent using SAS 9.4 and proc freq;

**proc** **freq** data=tmp1.lasvegas;

tables delinquent;

**run**;

\*b) Estimate linear probability model using proc reg

delinquent = B1 + B2 lvr + B3 ref + B4 insur + B5 rate +

B6 amount + B7 credit + B8 term + B9 arm + e;

options nolabel;

**proc** **reg** data=tmp1.lasvegas;

model delinquent = lvr ref insur rate amount credit term arm;

output out=lpmout p=phat\_lpm;

**run**;

**quit**;

\* print data;

**proc** **print** data=work.lpmout;

**run**;

\* list data showing max and min;

**proc** **univariate** data=lpmout;

**run**;

\* c) estimate probit using proc qlim;

options nolabel;

**proc** **qlim** data=tmp1.lasvegas;

model delinquent = lvr ref insur rate amount credit term arm / discrete;

output out=probitout xbeta marginal;

**run**;

**quit**;

\* print probitout data;

**proc** **print** data=work.probitout;

**run**;

\* list data showing max and min;

**proc** **univariate** data=work.probitout;

**run**;

\* produce means on variable amount;

options nolabel;

**proc** **means** data=work.probitout;

var meff\_p1\_amount meff\_p2\_amount;

**run**;

\* create predicted probabilities and their classification;

**data** lasvegas2;

set work.probitout;

phat = probnorm(xbeta\_delinquent);

phat\_classify = (phat >= **.5**);

**run**;

\* sort lasvegas2 data;

**proc** **sort** data=work.lasvegas2 out=lv2sorted;

by amount;

**run**;

\* print lv2sorted data;

**proc** **print** data=work.lv2sorted;

**run**;

\* list data showing max and min;

**proc** **univariate** data=work.lv2sorted;

**run**;

\* produce frequencies;

**proc** **freq** data=work.lv2sorted;

tables delinquent\*phat\_classify;

**run**;