**APPENDIX**

SAS Program Code

Housing Affordability Dataset

Group Members:

Kathy Chowaniec

Li Zhang

Bo Zhang

CS 593: Data Mining II

May 4, 2017

Market Value Linear Regression

SAS Code:

-------------------------------------------------------------------------;

\* Project : CS593 ;

\* Developer(s) : Kathy Chowaniec, Li Zhang, Bo Zhang ;

\* Comments : Final Project: Market Value Prediction ;

\* ;

\* Dependencies : libnames.sas ;

\*-------------------------------------------------------------------------;

libname sasdata "C:\CS 593\SAS\_Data";

**run**;

**proc** **copy** in=sasdata out=work;

select hads;

**run**;

\*remove unreasonable values (age less than 18 and value less than 1000);

**data** housing value\_exptn ;

set hads;

if VALUE <= **1000** or AGE1 < **18** then output value\_exptn ;

else output housing;

**run**;

\*transform data;

**data** housing2 exptn;

set housing;

lnarincome=log(LMED);

lnFMR=log(FMR);

sqhouseincome=sqrt(ZINC2);

sqmonthlycost=sqrt(ZSMHC);

squtility=sqrt(UTILITY);

lnother=log(OTHERCOST);

lnvalue=log(VALUE);

lnunits=log(NUNITS);

lnpeople=log(PER);

**run**;

\*Create indicator variables;

**data** housing3 exptn;

set housing2;

if OWNRENT = **1** then v\_owner = **1**;

else v\_owner = **0**;

if FMTREGION = "Northeast" then v\_northeast=**1**;

else v\_northeast=**0**;

if FMTREGION = "South" then v\_south=**1**;

else v\_south=**0**;

if FMTREGION = "West" then v\_west=**1**;

else v\_west=**0**;

if STATUS = **3** then v\_occupied=**0**;

else v\_occupied=**1**;

if FMTBURDEN = "1 Less than 30%" then v\_affordable=**1**;

else v\_affordable=**0**;

if METRO3='1' then v\_city=**1**;

else v\_city=**0**;

if FMTBUILT = 'After 2010' or FMTBUILT = '2000-2009' then v\_y2000=**1**;

else v\_y2000=**0**;

if FMTBUILT = ' Pre 1940' then v\_y1940=**1**;

else v\_y1940=**0**;

if STRUCTURETYPE = **1** then v\_singlefam=**1**;

else v\_singlefam=**0**;

**run**;

title "market value : univariate distribution";

\*show distribution of transformed variables;

**proc** **univariate** data=housing3 normaltest plot;

var lnvalue lnother squtility sqmonthlycost sqhouseincome lnFMR lnarincome AGE1 NUNITS;

**run**;

ods graphics on;

\*MODEL 2;

\*predict market value;

title "market value : regression without any variables removed";

**proc** **reg** data=housing3 outest=out\_housing ;

model lnvalue= v\_occupied v\_owner v\_y1940 v\_y2000 v\_northeast v\_south v\_city squtility lnother

sqmonthlycost lnFMR lnarincome sqhouseincome ROOMS AGE1 lnunits PER / VIF;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**run**;

**quit**;

\*\*FORWARD SELECTION;

title "market value : forward selection";

**proc** **reg** data=housing3 outest=est\_housing;

model lnvalue= v\_occupied v\_owner v\_y1940 v\_y2000 v\_northeast v\_south v\_city squtility lnother

sqmonthlycost lnFMR lnarincome sqhouseincome ROOMS AGE1 lnunits PER

/ selection=forward SLENTRY=**0.05**;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

\*\*BACKWARD SELECTION;

title "market value : backward selection";

**proc** **reg** data=housing3 outest=est\_housing;

model lnvalue= v\_occupied v\_owner v\_y1940 v\_y2000 v\_northeast v\_south v\_city squtility lnother

sqmonthlycost lnFMR lnarincome sqhouseincome ROOMS AGE1 lnunits PER

/ selection=backward ;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

\*\*STEPWISE SELECTION;

title "market value : stepwise selection";

**proc** **reg** data=housing3 outest=est\_housing;

model lnvalue= v\_occupied v\_owner v\_y1940 v\_y2000 v\_northeast v\_south v\_city squtility lnother

sqmonthlycost lnFMR lnarincome sqhouseincome ROOMS AGE1 lnunits PER

/ selection=stepwise SLENTRY=**0.05**;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

\*\*MAXR SELECTION;

title "market value: MAXR selection";

**proc** **reg** data=housing3 outest=est\_housing;

model lnvalue= v\_occupied v\_owner v\_y1940 v\_y2000 v\_northeast v\_south v\_city squtility lnother

sqmonthlycost lnFMR lnarincome sqhouseincome ROOMS AGE1 lnunits PER

/ selection=MAXR;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

title "market value: rsquare selection";

**proc** **reg** data=housing3 outest=est\_housing;

model lnvalue= v\_occupied v\_owner v\_y1940 v\_y2000 v\_northeast v\_south v\_city squtility lnother

sqmonthlycost lnFMR lnarincome sqhouseincome ROOMS AGE1 lnunits PER

/ selection=rsquare;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

title "market value: regression with 12-variable model";

\*regression with 12 variables recommended;

**proc** **reg** data=housing3 outest=est\_housing ;

model lnvalue= v\_y2000 v\_northeast v\_south v\_city lnother

sqmonthlycost lnFMR sqhouseincome ROOMS AGE1 lnunits PER / VIF;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

title "market value: regression with 12-variable model and residual plots";

\*regression with 12 variables recommended (with residual plots);

**proc** **reg** data=housing3 outest=est\_housing PLOTS(MAXPOINTS=NONE);

model lnvalue= v\_y2000 v\_northeast v\_south v\_city lnother

sqmonthlycost lnFMR sqhouseincome ROOMS AGE1 lnunits PER / VIF;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

\*regression with 6-variable model from MAXR (no residual plots);

title "market value using MAXR for 6-variable model";

**proc** **reg** data=housing3 outest=est\_housing;

model lnvalue=lnother sqmonthlycost lnFMR sqhouseincome AGE1 ROOMS/VIF;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

\*regression with 6-variable model from MAXR (with residual plots);

title "market value using MAXR for 6-variable model";

**proc** **reg** data=housing3 outest=est\_housing PLOTS(MAXPOINTS=NONE);

model lnvalue=lnother sqmonthlycost lnFMR sqhouseincome AGE1 ROOMS/VIF;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**quit**;

\*\*Model using PCA;

\*\*\* Normalize the data \*\*\*;

**PROC** **STANDARD** DATA=housing3 MEAN=**0** STD=**1**

OUT=housing\_z;

VAR v\_y2000 v\_northeast v\_south v\_city lnother

sqmonthlycost lnFMR sqhouseincome ROOMS AGE1 lnunits PER ;

**RUN**;

title "PCA : Correlation between components";

\*principal component analysis;

**proc** **princomp** data=housing\_z out=pca\_housing;

var v\_y2000 v\_northeast v\_south v\_city lnother

sqmonthlycost lnFMR sqhouseincome ROOMS AGE1 lnunits PER ;

**run**;

title "PCA : market value regression model";

\*regression using principal components;

**proc** **reg** data=pca\_housing outest=out\_housing;

model lnvalue=prin1 prin2 prin3 prin4 prin5 prin6 prin7 prin8 prin9/ STB VIF dwprob;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**run**;

**quit**;

\*regression using principal components (with residual plots);

**proc** **reg** data=pca\_housing outest=out\_housing PLOTS(MAXPOINTS=NONE);

model lnvalue=prin1 prin2 prin3 prin4 prin5 prin6 prin7 prin8 prin9/ STB VIF dwprob;

OUTPUT OUT=reg\_housingOUT PREDICTED= RESIDUAL=Res L95M=C\_l95m U95M=C\_u95m L95=C\_l95 U95=C\_u95

rstudent=C\_rstudent h=lev cookd=Cookd dffits=dffit

;

**run**;

**quit**;

ods graphics off;

\*\*EVALUATION;

\*training and test data set;

**data** training test;

set housing3;

\*determine if record number is even (divisible by 2);

id=**1000**+\_n\_;

if mod(id,**2**)=**0** then output training;

else output test;

**run**;

\*logistic regression of odd records;

**proc** **reg** data=training outest=training\_model;

market\_value: model lnvalue=lnother sqmonthlycost lnFMR sqhouseincome AGE1 ROOMS ;

**quit**;

\*logistic regression of even records;

**proc** **reg** data=test outest=test\_model;

market\_value: model lnvalue=lnother sqmonthlycost lnFMR sqhouseincome AGE1 ROOMS ;

**quit**;

\*score the data;

title "market value test dataset";

**proc** **freq** data=test;

table lnvalue/out=prior\_dist2(rename=count=\_prior\_ drop=percent);

**run**;

**proc** **score** data=test score=training\_model type=parms predict

out=test\_score ;

var lnvalue lnother sqmonthlycost lnFMR sqhouseincome AGE1 ROOMS;

**run**;

\*RMSE and MAE for lnvalue (actual) and market\_value (predicted) in dataset test\_score;

**%macro** mae\_rmse(

dataset /\* data set which contains the actual and predicted values \*/,

actual /\* variable which contains the actual or observed valued \*/,

predicted /\* variable which contains the predicted value \*/

);

%global mae rmse; /\* make the scope of the macro variables global \*/

proc sql noprint;

select count(**1**) into :count from &dataset;

select mean(abs(&actual-&predicted)) format **5.10** into :mae from &dataset;

select sqrt(mean((&actual-&predicted)\*\***2**)) format **5.10** into :rmse from &dataset;

quit;

**%mend**;

%***mae\_rmse***(test\_score, lnvalue, market\_value);

%put NOTE: Evaluation for market price: MAE=&MAE ;

%put NOTE: Evaluation for market price: RMSE=&RMSE;

Housing Affordability Logistic Regression

SAS Code:

-------------------------------------------------------------------------;

\* Project : CS593 ;

\* Developer(s) : Kathy Chowaniec, Li Zhang, Bo Zhang ;

\* Comments : Final Project: Affordability Prediction ;

\* ;

\* Dependencies : libnames.sas ;

\*-------------------------------------------------------------------------;

libname sasdata "C:\CS 593\SAS\_Data";

**run**;

**proc** **copy** in=sasdata out=work;

select hads;

**run**;

\*remove unreasonable values (age less than 18 and value less than 1000);

**data** housing value\_exptn;

set hads;

if AGE1 < **18** then output value\_exptn ;

else output housing;

**run**;

\*transform data;

**data** housing2 exptn;

set housing;

lnarincome=log(LMED);

lnFMR=log(FMR);

sqhouseincome=sqrt(ZINC2);

sqmonthlycost=sqrt(ZSMHC);

squtility=sqrt(UTILITY);

lnother=log(OTHERCOST);

lnvalue=log(VALUE);

lnunits=log(NUNITS);

lnpeople=log(PER);

**run**;

\*Create indicator variables;

**data** housing3 exptn;

set housing2;

if OWNRENT = **1** then v\_owner = **1**;

else v\_owner = **0**;

if FMTREGION = "Northeast" then v\_northeast=**1**;

else v\_northeast=**0**;

if FMTREGION = "South" then v\_south=**1**;

else v\_south=**0**;

if FMTREGION = "West" then v\_west=**1**;

else v\_west=**0**;

if STATUS = **3** then v\_occupied=**0**;

else v\_occupied=**1**;

if FMTBURDEN = "1 Less than 30%" then v\_affordable=**1**;

else v\_affordable=**0**;

if METRO3='1' then v\_city=**1**;

else v\_city=**0**;

if FMTBUILT = 'After 2010' or FMTBUILT = '2000-2009' then v\_y2000=**1**;

else v\_y2000=**0**;

if FMTBUILT = ' Pre 1940' then v\_y1940=**1**;

else v\_y1940=**0**;

if STRUCTURETYPE = **1** then v\_singlefam=**1**;

else v\_singlefam=**0**;

**run**;

\*MODEL 3;

\*logistic regression;

title "logistic regreesion data for affordability ";

**proc** **logistic** data=housing3 descending;

class v\_city(ref='0')/param=ref;

class v\_south(ref='0')/param=ref;

class v\_northeast(ref='0')/param=ref;

class v\_owner(ref='0')/param=ref;

class v\_west(ref='0')/param=ref;

class v\_singlefam(ref='0')/param=ref;

model v\_affordable= v\_owner v\_y2000 v\_y1940 v\_northeast v\_west sqhouseincome sqmonthlycost v\_singlefam

v\_south v\_city squtility lnother lnFMR lnarincome AGE1 lnunits ROOMS PER / selection=stepwise;

**quit**;

\*MODEL 3;

\*logistic regression;

title "logistic regression data for affordability";

**proc** **logistic** data=housing3 descending;

class v\_city(ref='0')/param=ref;

class v\_south(ref='0')/param=ref;

class v\_west(ref='0')/param=ref;

class v\_northeast(ref='0')/param=ref;

class v\_owner(ref='0')/param=ref;

class v\_singlefam(ref='0')/param=ref;

model v\_affordable= v\_owner v\_y2000 v\_y1940 v\_northeast v\_west sqhouseincome sqmonthlycost v\_singlefam

v\_south v\_city squtility lnother lnFMR lnarincome AGE1 lnunits ROOMS PER ;

**quit**;

\*MODEL 3;

\*logistic regression;

title "logistic regression data for affordability ";

**proc** **logistic** data=housing3 descending;

model v\_affordable= sqhouseincome sqmonthlycost;

**quit**;

\*\*EVALUATION;

\*training and test data set;

**data** training test;

set housing3;

\*determine if record number is even (divisible by 2);

id=**1000**+\_n\_;

if mod(id,**2**)=**0** then output training;

else output test;

**run**;

\*logistic regression of odd records;

title "logistic regression for affordability (training)";

**proc** **logistic** data=training descending outmodel=training\_model;

model v\_affordable=sqhouseincome sqmonthlycost;

**quit**;

\*logistic regression of even records;

title "logistic regression for affordability (test)";

**proc** **logistic** data=test descending outmodel=test\_model;

model v\_affordable=sqhouseincome sqmonthlycost;

**quit**;

**proc** **freq** data=test;

table v\_affordable/out=prior\_dist2(rename=count=\_prior\_ drop=percent);

**run**;

\*use training to score test;

**proc** **logistic** inmodel=training\_model;

score data=test prior=prior\_dist2

out=test\_score fitstat;

**run**;

**proc** **freq** data=test\_score;

table v\_affordable\*prediction;

**run**;

**quit**;