## Predicting the Flight Arrival Delays

/\*Create your own library in SAS like here it is libref and mention the path \*/

libname libref “/home/aro1260/deep”;

/\*Importing flight\_delay dataset \*/

PROC IMPORT DATAFILE= “/home/aro1260/data/flight\_delay.csv”

DBMS=CSV Replace

OUT=libref.flight\_delay;

GETNAMES=YES;

RUN;

/\*To check the contents of the data \*/

PROC CONTENTS DATA=libref.flight\_delay;

RUN;

/\*Descriptive statistics of the data\*/

proc means data = libref.flight\_delay;

vars Airport\_Distance Number\_of\_flights Weather

Baggage\_loading\_time Late\_Arrival\_o;

run;

/\*To check the Correlation of the dependent and independent variables in the data \*/

proc corr data = libref.flight\_delay nosimple;

var Arr\_Delay Airport\_Distance Number\_of\_flights Weather Support\_Crew\_Available Baggage\_loading\_time Late\_Arrival\_o Cleaning\_o Fueling\_o Security\_o;

run;

/\*Applying proc univariate for detailed summary statistics \*/

proc univariate data= libref.flight\_delay plot;

var Late\_Arrival\_o;

run;

/\*To check the relationship between the variables \*/

proc sgscatter data = libref.flight\_delay;

plot Arr\_Delay\*(Airport\_Distance Number\_of\_flights Weather Support\_Crew\_Available Baggage\_loading\_time Late\_Arrival\_o Cleaning\_o Fueling\_o Security\_o);

run;

/\*sorting the data set by strata variable (Arr\_Delay) \*/

proc sort data=libref.flight\_delay;

by Arr\_Delay;

run;

/\*splitting datasets into training (70%) and testing (30%)\*/

proc surveyselect data= libref.flight\_delay method=srs seed=1000 outall samprate=0.7 out= libref.flight\_delay\_subset;

strata Arr\_Delay ;

run;

proc print data=libref.flight\_delay\_subset;

run;

/\* Values of selected variable: 1 means selected for training set, 0 means testing set\*/

data libref.training;

set libref.flight\_delay\_subset;

if selected=1;

proc print;

data libref.testing;

set libref.flight\_delay\_subset;

if selected=0;

proc print;

/\* Program1: Building multiple linear regression model on training dataset\*/

ODS GRAPHICS ON;

PROC REG DATA = libref.training OUTEST=libref.Reg\_P\_Out;

MODEL Arr\_Delay = Airport\_Distance Number\_of\_flights Weather Support\_Crew\_Available Baggage\_loading\_time Late\_Arrival\_o Cleaning\_o Fueling\_o Security\_o/ SELECTION=stepwise VIF DW;

OUTPUT OUT = libref.Reg\_out\_train P=pred R=Residual RSTUDENT=r1 DFFITS=dffits COOKD=cookd;

RUN;

ODS GRAPHICS OFF;

/\*Score procedure on testing data \*/

PROC SCORE DATA= libref.testing SCORE=libref.Reg\_P\_Out TYPE = parms PREDICT OUT= libref.R\_Score\_Pred;

VAR Airport\_Distance Number\_of\_flights Weather Support\_Crew\_Available Baggage\_loading\_time Late\_Arrival\_oCleaning\_o Fueling\_o Security\_o ;

run;

proc print data= libref.R\_Score\_Pred;

title 'Predicted Scores for Regression on Testing Data';

run;

/\*Residual Scores from the Data Set Created by PROC SCORE, RESIDUAL option produces positive score \*/

PROC SCORE DATA= libref.testing SCORE=libref.Reg\_P\_Out RESIDUAL OUT= libref.R\_Score\_R TYPE = parms;

VAR Arr\_Delay Airport\_Distance Number\_of\_flights Weather Support\_Crew\_Available Baggage\_loading\_time Late\_Arrival\_oCleaning\_o Fueling\_o Security\_o ;

run;

proc print data = libref.R\_Score\_R;

title ' Residual Scores for Regression on Testing Data';

run;