```
*********************************************************;
* Program: H:\Spring2016\Bios591P\saaber_labAssignPart2.sas *;
* Programmer: Sadaf Saaber *;
* Creation Date: 3/27/2016*;
* *;
* Purpose: This program is used for the completion of Lab Assignment*;
*part 2 where I was building an associative model for SBP with*;
*different variables using the JHS dataset*;
*********************************************************;


libname bios591 'H:\Spring2016\Bios591P';

libname library 'H:\Spring2016\Bios591P';

*Determining/evaluating the dataset through looking at the summary of
the contents of the whole dataset;

proc contents data=bios591.jhs position;
run;

data stress1;
set bios591.jhs;

*Creating an indicator variable from the stress variable to compare all
the values to the reference of no stress;
if rstsa1 eq . then do;
mild=.;
mod=.;
very=.;
end;

else if rstsa1 eq 1 then do;
mild=0;
mod=0;
very=0;
end;

else if rstsa1 eq 2 then do;
mild=1;
mod=0;
very=0;
end;

else if rstsa1 eq 3 then do;
mild=0;
mod=1;
very=0;
end;

else if rstsa1 eq 4 then do;
mild=0;
mod=0;
very=1;
end;
```

```
label
mild="Mildly Stressed"
mod="Moderately stressed"
very="Very Stressed";

*Changing the categorical variable of sex into a numerical value I can
use to run a Simple linear regression model;

if sex eq " " then sex1=.;
else if sex eq "Female" then sex1=0;
else if sex eq "Male" then sex1=1;

label sex1="Sex 0=fem 1=Male";

run;

*Checking to make sure that I have correctly created my new stress
indicator variable and sex variable;

proc freq data=stress1;
tables rstsa1*mild*mod*very/list missing;
run;

proc freq data=stress1;
tables sex*sex1/list missing;
run;

   /*Stress analysis: conducting an analysis just looking at the newly
                  created stress indicator variables*/

*SLR for each of the variables of stress;

proc reg data=stress;
model sbp=mild;
run;

proc reg data=stress;
model sbp=mod;
run;

proc reg data=stress;
model sbp=very;
run;



*MLR for the stress;
proc reg data=stress1;
model sbp= mild mod very;
run;
quit;


   /*FULL MLR MODEL: looking at smoker status, BMI, alcohol drinks per
                           week, and stress*/
```

```
*Running a preliminary regression model to see all variables in the
model;

proc reg data=stress1;
model sbp= currentsmoker eversmoker BMI alcw diabetes sex1 mild mod
very;
run;
quit;

*Step 1: To build my model I started to looking at the correlations
between the outcome variable of sbp and the independent variables;

proc corr data=stress1 plots=matrix;
var sbp;
with BMI alcw sex1 currentsmoker eversmoker diabetes mild mod very;
run;

*Step 2: After looking at the correlation coefficients saw that
diabetes had the highest correlation coefficient so started with that
variable first to run an SLR and determined that it was significant;
proc reg data=stress1;
model sbp=diabetes;
run;
quit;

*Step 3: Using diabetes status then added other variables in the model
to determine which needed to come next;
proc reg data=stress1;
model sbp=diabetes currentsmoker;
model sbp=diabetes eversmoker;
model sbp=diabetes BMI;
model sbp=diabetes alcw;
model sbp=diabetes sex1;
model sbp=diabetes mild mod very;
run;
quit;

*Step 4: Finding the next variable to be included in the model with
diabetes and gender (found out gender in step 3);

proc reg data=stress1;
model sbp= diabetes sex1 eversmoker;
model sbp=  diabetes sex1 currentsmoker;
model sbp=diabetes sex1 BMI;
model sbp=diabetes sex1 alcw;
model sbp= diabetes sex1 mild mod very;
run;
quit;
run;

*Step 5: looking at the stress categories (as only some of them were
significant in step 4) specifically and seeing if overall significant
in a model with diabetes and sex;
proc reg data= stress1;
model sbp= diabetes sex1 mild mod very;
coinc: test mild, mod, very;
run;
```

```
quit;

*Step 6: Determining the next variable to include in a model that
already includes diabetes, gender and stress (as in step 5 saw that
stress was significant and needed to be in the model);

proc reg data=stress1;
model sbp = diabetes sex1 mild mod very bmi/vif;
*vif tests for multicollineaity;
model sbp = diabetes sex1 mild mod very alcw/vif;
model sbp = diabetes sex1 mild mod very currentsmoker/vif;
model sbp = diabetes sex1 mild mod very eversmoker/vif;
run;
quit;

*FINAL OUTCOME: So my final model includes diabetes, gender and stress
as all other factors are not significant at the 5% level;


      /*ANOTHER ANALYSIS: Conducting automative selection method
analysis on the data to double check final outcome*/

*Backwards elimination;
PROC REG DATA= stress1;
MODEL sbp= currentsmoker eversmoker BMI alcw diabetes sex1 mild mod
very
/ SELECTION=BACKWARD STAY=.025;
RUN;

*forwards selection;
PROC REG DATA=stress1;
MODEL sbp= currentsmoker eversmoker BMI alcw diabetes sex1 mild mod
very
/ SELECTION=FORWARD SLENTRY=0.025 ;
RUN; QUIT;

*stepwise selection method;
PROC REG DATA=stress1;
      MODEL sbp= currentsmoker eversmoker BMI alcw diabetes sex1 mild
mod very
/ SELECTION=STEPWISE SLENTRY=.025 SLSTAY=.025 ; RUN;
QUIT;

*FINAL OUTCOME OF AUTOMATIVE METHODS: these methods stated that
diabetes and sex belong in the model;


*Adding a template for future graphs of checking MLR assumptions;
ods graphic on;

       proc template;
 Define style styles.mystyle;
  Parent=styles.default;
   Style GraphTitleText /  fontsize=18pt fontfamily="arial"
fontweight=bold;
   Style GraphValueText / fontsize=16pt fontfamily="arial"
fontweight=bold;
```

```
   Style GraphLabelText / fontsize=16pt fontfamily="arial"
fontweight=bold;
   end;
 run;

 ods html style=mystyle;

  /*Regression analysis: Checking assumptions after determining which
                variables to include in the model */

*Linear Assumption Check;
TITLE 'Linearity Assumption';
PROC CORR DATA=stress1 PLOTS=MATRIX;
VAR sex1 diabetes mild mod very;
WITH sbp;
RUN;

title "MLR Model With Diagnostics";
      proc reg data= stress1;
            title "The Regression of SBP on 5 Identified Variables";
            model sbp = diabetes sex1 mild mod very/ partial pcorr2
influence R VIF;
            *partial = requesting partial plots; *pcorr2 = produces
squared partial correlations; *influence R = produces tables of outlier
diagnostic statistics; *vif = variance inflation factors (VIF>0.10
indicates multicollinearity);
            output out = outresid r = resid p = yhat rstudent = jackres
cookd = cooks h = leverage;
            *create a dataset called outresid, in which the residuals,
predicted values, jacknife residuals, Cooks Distances, and leverage
statstics will be saved in variables name resid, yhat, jackres, cooks,
and leverage, respectively;
            run; quit;


      *Outlier Diagnostics;
title "Outliers and Influential Observations";
data outliers;
set outresid (keep = sbp sex1 diabetes mild mod very yhat resid cooks
leverage jackres);
            *each of these new variable will either be 0 or 1,
depending if the observation satisfies the criterion for being an
outlier based on each diagnostic;

            outlier_jk = (abs(jackres)) > 2;
            outlier_cooks = (cooks > 4/1000); *cooksd value greater
than 4/n is potential outlier;
            outlier_leverage = (leverage > 12/1000); *leverage value
greater than 2(k+1)/n is outlier where k is number of variables and n
is sample size;
            obs_num = _N_; *creates a variable with the observation
number from the original dataset;
            IF sum (outlier_jk, outlier_cooks, outlier_leverage) > 0;
*tells sas to only keep pbservations that are an outlier by at least
one of the listed methods;
run;
```

```
proc print data = outliers;
run;

proc freq data=outliers;
tables  outlier_jk outlier_cooks outlier_leverage obs_num;
run;

*Determining the distributions of different types of outliers based on
a specific set of criteria;
proc freq data=outliers;
tables outlier_jk*outlier_cooks*outlier_leverage/list missing;
run;
```