**ACADEMIC PERFORMANCE INDICATOR PROJECT IN SAS**

*Linear Regression Analysis*

**TABLE OF CONTENTS:**

1. Setting up the Library in SAS
2. Outliers Checking and Treatment
3. Missing Value Checking and Treatment
4. Correlation Checking of Dependant with Independent variables
5. Regression procedure to check different statistics
6. Homoscedasticity Check for Error term
7. Normality Check for Error term
8. MAPE Calculation
9. Output file creation

**1. SETTING UP THE LIBRARY IN SAS:**

/\*Setting Library\*/

LIBNAME D "C:\SAS\_101\SAS Class Sessions\SESSION 9";

**RUN**;

1. *Cleaning/Sanity Check*
   1. *Outliers – Valid Value (Maximum 1% of Total data) – Floor Them/Cap them, if same observation have outliers across multiple variables remove the observation*
      1. *Extremely Value list from Proc Univariate (2-3% on either side)*
      2. *Scatter Plot*
      3. *Min and 1% Values/Max and 99% Values is very different*
      4. *Mean and Median is very different*
      5. *Thumb Rules* 
         1. *Normal Distribution (Mean ± 3\* Standard Deviation)*
         2. *Not Normal (Q1 – 1.5\*(Q3-Q1), Q3 + 1.5\*(Q3-Q1)*
   2. *Invalid Values – Make them Valid*
   3. *Missing Values – NMISS Option*
      1. *Mean of Non Missing Values*
      2. *Grouped Means*
      3. *Regression*

**2. OUTLIERS TREATMENT:**

Check 15% of either side (Low and High) for Outliers. Check if the difference between Mean and Median is substantial. Also, use scatter plot to find the outliers. Outlier treatment by using *PROC MEANS* and *PROC UNIVARIATE*

/\*Checking outliers\*/

ODS HTML FILE = "C:\SAS\_101\SAS Class Sessions\SESSION 9\OutliersChecking.XLS";

**PROC** **MEANS** DATA = API.ELEMAPI MIN MAX

MEAN MEDIAN Q1 Q3 STD;**RUN**;ODS HTML CLOSE;

/\*Univariate Test to check outliers exist or not\*/

ODS HTML FILE = "C:\SAS\_101\SAS Class Sessions\SESSION 9\Univariate.XLS";

**PROC** **UNIVARIATE** DATA = API.ELEMAPI NEXTROBS = **15**;

**RUN**;ODS HTML CLOSE;

Select the data points, which are outliers and replace them with next valid values (Flooring/Capping). The variables ‘*acs\_k3*’,’AVG\_ED’ and ‘full’ have invalid data, replace them with valid ones (absolute of data points)

/\*Outlier Treatment\*/

**DATA** API.A;SET API.A;

ACS\_K3 = ABS(ACS\_K3); /\*-ve to +ve\*/

IF AVG\_ED = **.** THEN AVG\_ED = **0**; /\*Missing value to 0\*/

IF FULL <=**1** THEN FULL = FULL\***100**; /\*Invalid Ratios to %\*/

/\*OUTLIERS\*/

IF ENROLL>**1264** THEN ENROLL=**1264**;

IF API00>**918** THEN API00=**918**;**RUN**;

**3. MISSING VALUES TREATMENT:**

Check for missing values and replace those with relevant values as shown below:

/\*Missing Values Checking\*/

**PROC** **MEANS** DATA = API.A MEAN NMISS;**RUN**;

/\*Missing Values Treatment\*/

**DATA** API.A;SET API.A;

IF ACS\_K3 = **.** THEN ACS\_K3 =**19.1608040**;

IF ACS\_46 = **.** THEN ACS\_46 =**29.6851385**;

IF MOBILITY = **.** THEN MOBILITY =**18.2531328**;

IF MEALCAT = **1** AND MEALS = **.** THEN

MEALS = **28.36**;

IF MEALCAT = **2** AND MEALS = **.** THEN

MEALS = **66.0468750**;**RUN**;

**4. CORRELATION CHECKING:**

Check for correlation between independent continuous variables with dependent variable.

/\*Correlation with Continuous variables\*/

ODS HTML FILE = "C:\SAS\_101\SAS Class Sessions\SESSION 9\CorrelationContinuous.XLS";

**PROC** **CORR** DATA = API.A;

WITH API00;**RUN**;ODS HTML CLOSE;

Checks for correlation:

/\*

1. Sign OK Significance OK =>Include

2. Sign Not OK SIgnificance Not OK =>Exclude

3. Sign OK Significane Not OK =>Scatter Plot and Study Non Linear Pattern, Correlation between some transformed variables

4. Sign Not OK Significance OK =>Report

\*/

Check for correlation between independent category variables with dependent variable.

/\*Mean Test with Category Variables to check the Correlation of Mealcat\*/

**PROC** **MEANS** DATA = API.A MEAN;VAR API00;

CLASS MEALCAT;**RUN**;

/\*ANOVA to check Significance of Mealcat Category Variable\*/

**PROC** **ANOVA** DATA = API.A;CLASS MEALCAT;

MODEL API00=MEALCAT;**RUN**;

/\*Mean Test with Category Variables to check the Correlation of yr\_rnd variable\*/

**PROC** **MEANS** DATA = API.A MEAN;VAR API00;

CLASS YR\_RND;**RUN**;

/\*Mean of other variables to validate yr\_rnd data\*/

**PROC** **MEANS** DATA = API.A MEAN;

VAR MOBILITY FULL MEALS;

CLASS YR\_RND;**RUN**;

/\*ANOVA to check Significance of yr\_rnd variable\*/

**PROC** **ANOVA** DATA = API.A;CLASS YR\_RND;

MODEL API00=YR\_RND;**RUN**;

**5. REGRESSION PROCEDURE TO CHECK DIFFERENT STATISTICS:**

**PROC REG** to check the overall significance, R2 and Adjusted R2 values. Also, need to check Parameter Significance, P-Value (Significant, <0.0001) and VIF(<=1.5) of each variable separately. Need to remove variables one by one w.r.t the VIF to reduce Multicollinearity. We should check the variable with highest VIF. It’s a process to remove variables and reduce Multicollinearity.

/\*Y = A+B1X1+B2X2+….+BKXK + U\*/

/\*Regression Analysis\*/

ODS HTML FILE = "C:\SAS\_101\SAS Class Sessions\SESSION 9\RegressionWithVIFCollin.XLS";

**PROC** **REG** DATA = API.A;

MODEL API00 =

meals

ell

yr\_rnd

mobility

acs\_k3

acs\_46

not\_hsg

hsg

some\_col

col\_grad

grad\_sch

avg\_ed

full

emer

enrol

mealcat

/VIF COLLIN;

**RUN**;

ODS HTML CLOSE;

If after comparing we find both the variables are significant, search for 3rd variable which is correlated with both of them.

*Variables removed:*

*meals*

*yr\_rnd*

*acs\_k3*

*acs\_46*

*not\_hsg*

*hsg*

*some\_col*

*col\_grad*

*grad\_sch*

*emer*

*enroll*

*mealcat*

Once we get the model which is satisfying the basic criteria’s (ALPHA = 0.01%, VIF <=1.5, P <0.0001, R2 > 65%), we can check the assumptions of the model.

**6. HOMOSCEDASTICITY CHECK FOR ERROR COMPONENT:**

*SPEC* option in *PROC* *REG* is used to check WHITE’s TEST, which is used to check the Homoscedasticity of the ERROR Term*. If P-Value is > Alpha (0.0001 or 0.01%), then the model is fine.*

/\*HOMOSCEDASTICITY - i.e. THE\*/

/\*VARIANCE OF THE ERROR COMPONENT \*/

/\*MUST BE CONSTANT ACROSS THE CROSS SECTION\*/

/\*WHITE'S TEST\*/

/\*H0: MODEL IS HOMOSCEDASTIC\*/

/\*H1: MODEL IS HETEROSCEDASTIC;\*/

/\*P < ALPHA => HETEROSCEDASTICITY WHICH\*/

/\*CAN BE REDUCED BY TRANSFORMATION OF X\*/

/\*VARIABLES PARTICULARLY LOG OR\*/

/\*SQUARE ROOT\*/

ODS HTML FILE = "C:\SAS\_101\SAS Class Sessions\SESSION 9\ModelChecking.XLS";

**PROC** **REG** DATA = API.A;

MODEL API00 =

/\*meals\*/

ell

/\*yr\_rnd\*/

mobility

/\*acs\_k3\*/

/\*acs\_46\*/

/\*not\_hsg\*/

/\*hsg\*/

/\*some\_col\*/

/\*col\_grad\*/

/\*grad\_sch\*/

avg\_ed

full

/\*emer\*/

/\*enroll\*/

/\*mealcat\*/

/SPEC;

**RUN**;

ODS HTML CLOSE;

**7. NORMALITY CHECK OF ERROR TERM:**

*PROC UNIVARIATE* procedure with option HISTOGRAM RES/NORMAL to check the Normality of error term.

/\*NORMALITY OF RESIDUAL\*/

**PROC** **UNIVARIATE** DATA = API.O NORMAL;

VAR RES;HISTOGRAM RES/NORMAL;**RUN**;

**8. MAPE CALCULATION:**

It is used to calculate Mean of the absolute percentage error in the model.

/\*MAPE\*/

/\*MEAN ABSOLUTE PERCENTAGE ERROR\*/

/\*ERROR = ACTUAL - PREDICTED\*/

/\*ABS(ERROR/ACTUAL)\*100\*/

/\*MEAN OF THE ABOVE = MAPE\*/

**DATA** API.O;SET API.O;

ERROR = ABS(RES/API00)\***100**;**RUN**;

**PROC** **MEANS** DATA = API.O MEAN;VAR ERROR;**RUN**;

MAPE below 10% is acceptable.

**9. OUTPUT FILE CREATION:**

Create the output file to store the PREDICTED and ERROR values along with the original dataset.

/\*CREATE THE OUTPUT FILE\*/

**PROC** **REG** DATA = API.A;

MODEL API00 =

/\*meals\*/

ell

/\*yr\_rnd\*/

mobility

/\*acs\_k3\*/

/\*acs\_46\*/

/\*not\_hsg\*/

/\*hsg\*/

/\*some\_col\*/

/\*col\_grad\*/

/\*grad\_sch\*/

avg\_ed

full

/\*emer\*/

/\*enroll\*/

/\*mealcat\*/

;

OUTPUT OUT = API.O

P = PRED R = RES;**RUN**;**QUIT**;

ODS HTML FILE = "C:\SAS\_101\SAS Class Sessions\SESSION 9\OutputFile.xls";

**PROC** **PRINT** DATA = API.O; **RUN**;

ODS HTML CLOSE;

**OVERALL CONDITIONS CHECKING**

/\*

1. OVERALL SIGNIFICANCE P < ALPHA

2. MULTICOLLINEARITY - VIF <=1.5

3. INDIVIDUAL SIGNIFICANCE - P < ALPHA

4. HOMOSCEDASTICITY CHECK - P > ALPHA

5. NORMALITY CHECK - P > APLHA

6. MAPE <=10%

7. R - SQUARE >= 65%

\*/