1. **Exploring Data**

First, we check the structure of the data and the datatypes of the variables.

ODS HTML File = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Contents.xls';

**Proc** **Contents** data = reg.elemapi;

**Run**;

ODS HTML CLOSE;

* Number of observations = 400
* Number of variables = 21 (including ID variables dnum and snum).
* All the variables are in numerical type, where **yr\_rnd** is a binary categorical variable and **mealcat** is an ordinal variable.

Next, check for outliers, invalid and missing values:

ODS HTML File = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Missing\_Outliers\_Check.xls';

ODS Trace on/Listing;

**Proc** **Univariate** Data = Reg.elemapi;

ODS Select MissingValues BasicMeasures Quantiles ExtremeObs;

**Run**;

ODS Trace Off;

ODS HTML Close;

The above code will only the select the required outputs (specified by ODS Select) and export the outputs in the specified Excel file.

***Note:*** *Since dnum (District Code) and snum (School Number) are irrelevant for this check, their outputs have been removed from the Excel file. Also,note that api00 (Average Performance Index 2000)is the dependent variable here, which means growth (Growth from 1999 to 2000) is to be removed from modelling as well.*

1. **Api00**: The values in Quantiles, ExtremeObs table and the closeness of Mean-Median suggest that there is no outlier. Also, there is no MissingValues table, indicating absence of missing values.
2. **Meals**: Similar outlier check suggests no outlier, but there are 85 missing values.
3. **Ell**: Looking at Quantiles table, it looks like a positively skewed data. But a smaller Mean-Median difference suggests no outliers as such. No MissingValues.
4. **Yr\_rnd**: It is a binary data with no missing values.
5. **Mobility**: No Outliers and 1 missing value.
6. **Acs\_k3**: It is the average class size, which contains negative values at the lowest percentiles. This suggests invalid data. Also, there are 2 missing values.
7. **Acs\_46**: No outliers, but 3 missing values.
8. **Not\_hsg**: No outliers or missing values.
9. **Hsg**: No outliers or missing values.
10. **Some\_col**: No outliers or missing values.
11. **Col\_grad**: No outliers or missing values.
12. **Grad\_sch**: No outliers or missing values.
13. **Avg\_ed**: No outliers but 19 missing values.
14. **Full**: Quantiles table and Mean-Median difference suggest outliers (or invalid values), but no missing values.
15. **Emer**: No outliers or missing values.
16. **Enroll**: Quantiles and Mean-Median-Mode suggest a positively skewed data but probably no outliers. No missing values.
17. **Mealcat**: Ordinal data [1,2,3] for 3 categories based on **meals**.
18. **Data Cleaning**

Missing Values/Outliers Treatments for the above highlighted variables:

* **Meals** has 85 missing values which is a huge number as compared to 400 observations. So, it can’t be imputed. We can check for the category wise frequency distribution in **Mealcat**:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Meals\_Missing\_Treatment.xls';

**Proc** **Means** Data = reg.elemapi mean nmiss;

Var Meals;

Class Mealcat;

**Run**;

ODS HTML CLOSE;

This code will compute mean and nmiss of **Meals** group by each value of Mealcat. As a result, we can impute the missing values corresponding to a particular **Mealcat** category by the category’s mean:

**Data** reg.data;

Set reg.elemapi;

If meals = **.** and mealcat = **1** then meals = **28.36**;

If meals = **.** and mealcat = **2** then meals = **66.046875**;

**Run**;

* **Mobility** has no outliers and only 1 missing value, which can be imputed using the mean of the column (obtained in Missing\_Outlier\_Check.xls):

**Data** reg.data;

Set reg.data;

If mobility = **.** then mobility = **18.25313**;

**Run**;

* **Acs\_k3** contains negative values which is invalid in this context. On a careful inspection using Quantiles and ExtremeObs tables, it makes sense to take the absolute values of the negative values to make the data valid. Also, the 2 missing values may be imputed using mean:

**Data** reg.data;

Set reg.data;

acs\_k3 = abs(acs\_k3);

If acs\_k3 = **.** then acs\_k3 = **18.54774**;

**Run**;

* **Acs\_46** has 3 missing values which can be imputed using mean:

**Data** reg.data;

Set reg.data;

If acs\_46 = **.** then acs\_46 = **29.68514**;

**Run**;

* **Full** has values in the form of percentages upto 25 Percentile i.e. 0.42, 0.45 and so on, while the others are whole numbers such as 87, 97,…,100. It looks like these anamoly is due to the variation in the expression of percentages by different people. It makes sense to multiply the number which as less than 1, by 100:

**Data** reg.data;

Set reg.data;

If full <=**1** then full = full\***100**;

**Run**;

* **Avg\_ed** has 19 missing values. We may impute these values using the statistics of the districts as well as the whole column. This makes sense since schools in a particular district can be expected to have somewhat similar ratings.

**Proc** **means** data = reg.elemapi;

var avg\_ed;

Class dnum;

output out = reg.a nmiss= miss median =median mean= dmean;

**Run**;

**/\* *Here we compute missing values, median and mode for each district*\*/**

**Proc** **SQL**;

Create table reg.a as

Select dnum, \_freq\_, miss,median, dmean

From reg.a

Where miss ne **0**;

**quit**;

/\* ***Here we filter the table to obtain only those district numbers which have missing values in Avg\_ed* \*/**

ODS HTML File ='C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Av\_ed\_Treatment\_New.xls';

**Proc** **Print** Data = reg.a;

**Run**;

ODS HTML Close;

/\* ***The missing value statistics table can be explored*** \*/

Next, we try to impute the missing values:

**Data** reg.data;

Set reg.data;

If dnum = **41** and avg\_ed = **0** then avg\_ed = **1.7049999833**;

else If dnum = **135** and avg\_ed = **0** then avg\_ed = **2.4942856857**;

else If dnum = **259** and avg\_ed = **0** then avg\_ed = **2.4899999857**;

else If dnum = **316** and avg\_ed = **0** then avg\_ed = **3.5519999504**;

else If dnum = **401** and avg\_ed = **0** then avg\_ed = **2.2799999714**;

else If dnum = **473** and avg\_ed = **0** then avg\_ed = **3.3842857225**;

else If dnum = **541** and avg\_ed = **0** then avg\_ed = **2.5999999046**;

else If dnum = **689** and avg\_ed = **0** then avg\_ed = **2.6666666667**;

**Run**;

The idea behind this is:

1. To use mean value of the district if there are a few missing values and small (but considerable) number of total observations, then there may not be any outliers as such.
2. To use median value of the district if there are a few missing values and large number of total observations, since there may be outliers.
3. To use the mean/median value of all the observed values in the column if a particular district contains only the missing values and no observed values.

Now the data is ready to be used for modelling and testing.

Let’s split the data into two parts: train data and test data:

**Data** reg.train reg.test;

Set reg.data;

If ranuni(**100**) >= **0.85** then output reg.test;

Else output reg.train;

**Run**;

This means that the whole data set is randomly divided into two parts, where approximately 85% of data is in **reg.train**, and the rest of the data is in **reg.test**. So, we will build the model using the training data set and then, use the test data to check the accuracy of our prediction.

Now, all the operations will be done only on the training data.

1. **Preliminary Checks**

Now, some preliminary modelling checks.:

* To check if the continuous dependent variables have significant correlations with **Api00**:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Correlation\_with\_api00.xls';

ODS Trace On/listing;

**Proc** **Corr** Data = reg.train (drop = snum dnum growth yr\_rnd mealcat ed\_status);

With api00;

ODS Select PearsonCorr;

**Run**;

ODS Trace Off;

ODS HTML Close;

In the Excel file ‘Correlation\_with\_api00.xls’, it can be observed that all of the independent variables have significant correlations (p-value < 0.01).

* To check if the categorical variables have significant impact on **Api00**:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Impact of yr\_rnd on api00.xls';

ODS Trace On/Listing;

**Proc** **Anova** Data = reg.train;

Class yr\_rnd;

Model api00 = yr\_rnd;

ODS Select OverallANOVA;

**Quit**;

ODS Trace Off;

ODS HTML Close;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 1645551.463 | 1645551.463 | 105.48 | <.0001 |
| **Error** | 345 | 5382076.324 | 15600.221 |  |  |
| **Corrected Total** | 346 | 7027627.787 |  |  |  |

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Impact of mealcat on api00.xls';

ODS Trace On/Listing;

**Proc** **Anova** Data = reg.train;

Class mealcat;

Model api00 = mealcat;

ODS Select OverallANOVA;

**Quit**;

ODS Trace Off;

ODS HTML Close;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 5276496.486 | 2638248.243 | 518.27 | <.0001 |
| **Error** | 344 | 1751131.301 | 5090.498 |  |  |
| **Corrected Total** | 346 | 7027627.787 |  |  |  |

The above 2 tables with significant F values (low p-values) indicate that **yr\_rnd** and **mealcat** have significant impacts on **Api00** and hence must be taken into account.

1. **Model Building**

Now, we will build multiple models and run MLRM model diagnostic checks on each of the model at each step.

1. Let’s start with all the variables first:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\model1.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46 hsg not\_hsg some\_col

col\_grad grad\_sch avg\_ed full emer enroll mealcat/ VIF;

**Quit**;

ODS HTML CLOSE;

From Model1.xls, it can be seen that many variables have insignificant t-values. Also, all the parent-education related variables have oddly high VIF, which suggests rebundancy of one or more of these variables. This suggests Multicollinearity. Notice that **hsg** has the highest p-value (0.911).

1. Let’s try to remove **hsg** from model1:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\model1\_1.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46 not\_hsg some\_col

col\_grad grad\_sch avg\_ed full emer enroll mealcat

/VIF;

**Quit**;

ODS HTML CLOSE;

There still are many insignificant variables and some worrisome VIF values. We notice that with the exclusion of **hsg,** adjusted R-square has increased slightly, which is a good sign. Also, **Avg\_ed** is the most insignificant variable here.

1. Let’s try to remove **Avg\_ed** from model1\_1:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\model1\_2.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46 not\_hsg some\_col

col\_grad grad\_sch full emer enroll mealcat

/VIF;

**Quit**;

ODS HTML CLOSE;

There still are many insignificant variables and some worrisome VIF values. We notice that with the exclusion of **Avg\_ed,** adjusted R-square has increased slightly, which is a good sign. Also, **enroll** is the most insignificant variable here.

1. Let’s try to remove **enroll** from model1\_2:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\model1\_3.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46 not\_hsg some\_col

col\_grad grad\_sch full emer mealcat

/VIF;

**Quit**;

ODS HTML CLOSE;

There still are many insignificant variables and some worrisome VIF values. We notice that with the exclusion of **enroll,** adjusted R-square has increased slightly, which is a good sign. Also, **emer** is the most insignificant variable here.

1. Let’s try to remove **emer** from model1\_3:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\model1\_4.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46 not\_hsg some\_col

col\_grad grad\_sch full mealcat

/VIF;

**Quit**;

ODS HTML CLOSE;

There still are many insignificant variables and some worrisome VIF values. We notice that with the exclusion of **emer,** adjusted R-square has increased slightly, which is a good sign. Also, **some\_col** is the most insignificant variable here.

1. Let’s try to remove **some\_col** from model1\_4:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\model1\_5.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46 not\_hsg

col\_grad grad\_sch full mealcat

/VIF;

**Quit**;

ODS HTML CLOSE;

There still are many insignificant variables and some worrisome VIF values. We notice that with the exclusion of **some\_col,** adjusted R-square has increased slightly, which is a good sign. Also, **mealcat** is the most insignificant variable here.

1. Let’s try to remove **mealcat** from model 1\_5:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\model1\_6.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46 not\_hsg

col\_grad grad\_sch full

/VIF;

**Quit**;

ODS HTML CLOSE;

There still are many insignificant variables and some worrisome VIF values. We notice that with the exclusion of **mealcat,** adjusted R-square and R-square have decreased by a margin which is where we need to have a trade-off between the loss in R values and exclusion of insignificant variables. Since the decrease is very low, we may consider removing more variables. Also, **not\_hsg** is the most insignificant variable here.

1. Let’s try to remove not\_hsg from model1\_6:

ODS HTML FILE = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\model1\_7.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46

col\_grad grad\_sch full

/VIF;

**Quit**;

ODS HTML CLOSE;

With only a marginal decrease in R values, we now have a model with all the variables as significant.

Our regression equation is:

Est(Api00) = (566.51535) – (2.26654\***meals**) – (1.13465\***ell**) – (25.14417\***yr\_rnd**) - (1.81631\***mobility**) + (4.69048\***acs\_k3**) + (2.03543\***acs\_46**) + (0.63746\***col\_grad**) + (2.05598\***grad\_sch**) + (1.39318\***full**)

1. **MLRM Model Diagnostics**

Now, we have to run some diagnostics checks on our model to see if any of the fundamental assumptions of MLRM getting violated. It is a crucial part of a modeling since any such violation of assumptions may lead to severe errors and/or misleading relationships of independent variables with the dependent one.

* **Multicollinearity**: The VIF column in Model1\_7.xls is a measure of the severity of the problem. Since all the VIF values are less than 10, we can say that the model is free from any significant multicollinearity.
* **Normality of Residuals**: It is an important assumption to be validated by the model since the individual t-tests for significance would not be valid with this assumption to be true. We can test it using Shapiro-Wilk, Anderson-Darling tests of normality.

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46

col\_grad grad\_sch full

/VIF;

output out = reg.predict1\_7 predicted = pred residual = res;

**Quit**;

**Data** reg.predict1\_7 (Keep = snum api00 pred res);

Set reg.predict1\_7;

**Run**;

ODS HTML File = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Model1\_7\_residual\_normality\_test.xls';

ODS Trace On/Listing;

**Proc** **Univariate** Data = reg.predict1\_7 normal;

Var res;

ODS Select TestsForNormality;

**Run**;

ODS Trace Off;

ODS HTML Close;

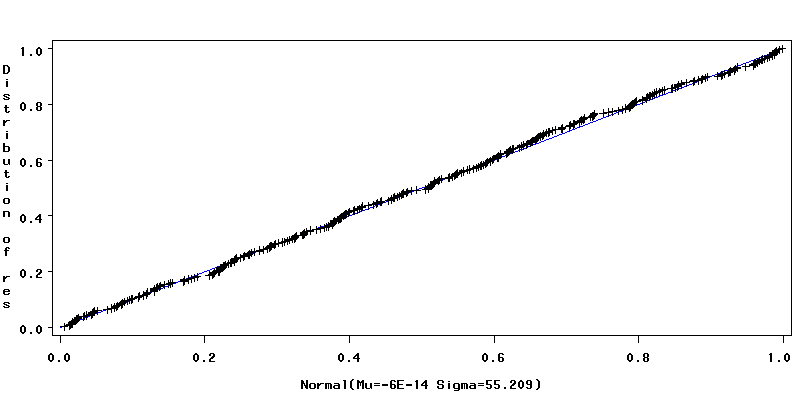
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tests for Normality** | | | | |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.995467 | **Pr < W** | 0.4115 |
| **Kolmogorov-Smirnov** | **D** | 0.027709 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.036877 | **Pr > W-Sq** | >0.2500 |
| **Anderson-Darling** | **A-Sq** | 0.304869 | **Pr > A-Sq** | >0.2500 |

Since p-values are high, we do not have enough evidence to reject the null hypothesis that the distribution is normal. Therefore, we conclude that the error distribution is normal. We may obtain a ppplot, solidifying our result:

**Proc** **Capability** Data = reg.predict1\_7;

ppplot res;

**Run**;



The more the curve is close to the diagonal line, the more is the distribution close to the normal distribution. This indicates the normality of our residuals.

* **Homoscedasticity**: It is yet another fundamental assumption of a Linear Regression Model which requires the variance of error terms to be constant over the range of predictor variables. It can be tested using White Test, that we can obtain using SPEC option in Proc Reg:

ODS HTML File = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Model1\_7\_Homosc\_WhiteTest.xls';

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46

col\_grad grad\_sch full

/spec;

ODS Select SpecTest;

**Quit**;

ODS Trace OFf;

ODS HTML Close;

|  |  |  |
| --- | --- | --- |
| **Test of First and Second** | | |
| **Moment Specification** | | |
| **DF** | **Chi-Square** | **Pr > ChiSq** |
| 53 | 65.73 | 0.1126 |

Since, p-value is high, we do not have enough evidence to reject the null hypothesis that the variance/spread of error terms is constant. Therefore, we may conclude that the errors are homoscedasticity. Further, we may obtain a plot of residuals vs predictions:

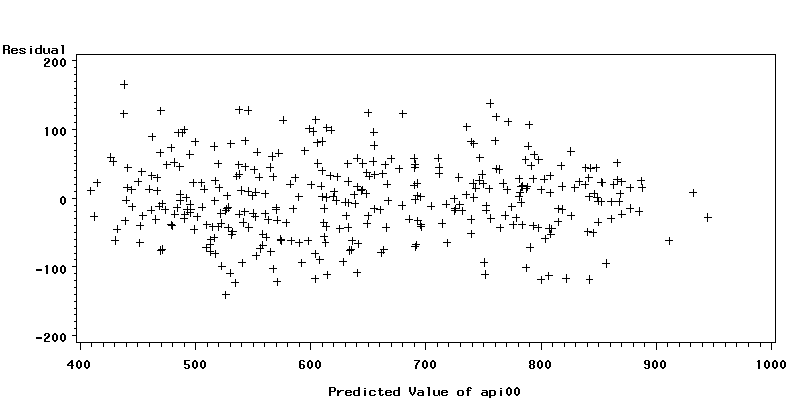
ODS HTML File = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Residual Homoscedastic Plot.jpeg';

**Proc** **Gplot** data = reg.predict1\_7;

Plot res\*pred;

**Quit**;

ODS HTML Close;



The graph seems to be fairly random which further indicates that the errors are homoscedastic.

* **Autocorrelation**: It is another fundamental assumption of a linear regression model that the error terms should not be correlated. We test for the autocorrelation using Durbin-Watson Test Statistic:

ODS HTML File = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Model1\_7\_AutocorrTest\_DWTest.xls';

ODS Trace On/Listing;

**PROC** **REG** DATA = reg.train;

MODEL API00 = meals ell yr\_rnd mobility acs\_k3 acs\_46

col\_grad grad\_sch full

/dw;

ODS Select DWStatistic;

**Quit**;

ODS Trace Off;

ODS HTML Close;

|  |  |
| --- | --- |
| **Durbin-Watson D** | 1.422 |
| **Number of Observations** | 347 |
| **1st Order Autocorrelation** | 0.286 |

DW Statistic value must be close to 2, ideally, to be free from any autocorrelation. But in practical problems, some level of autocorrelation is usually present and so, we can have some tolerance towards DW statistic. For the given number of observations, d = 1.422 may be considered to be somewhat satisfactory for the modeling purposes.

Therefore, we will choose this model as our model, based on the given sample.

1. **Model Accuracy**

Now, let’s run the model on the test data and check the accuracy of our model:

**Data** reg.test;

Set reg.test;

est\_api00 = **566.51535**-**2.26654**\*meals-**1.13465**\*ell-**25.14417**\*yr\_rnd-**1.81631**\*mobility+**4.69048**\*acs\_k3+**2.03543**\*acs\_46+**0.63746**\*col\_grad+**2.05598**\*grad\_sch+**1.39318**\*full;

Keep snum api00 est\_api00;

**Run**;

ODS HTML FIle = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Predictions1\_7.xls';

**Proc** **Print** Data = reg.test;

**Run**;

ODS HTML Close;

**Proc** **SQL**;

Create Table reg.mape as

Select mean((abs(api00 - est\_api00)/api00)\***100**) as Mape

From reg.test;

**Quit**;

ODS HTML FIle = 'C:\SASCOURSE\SASCOURSE\Linear Regression Case Study Materials\Linear Regression Case Study Materials\Reg\Final\_Mape.xls';

**Proc** **Print** Data = reg.mape;

**Run**;

ODS HTML Close;

Mape = 7.8 (approximately)