**Checking Assumptions of Multiple Regression with SAS**

**1. Detecting Outlier**

**I. Box Plot Method**

If a value is higher than the 1.5\*IQR above the upper quartile (Q3), the value will be considered as outlier. Similarly, if a value is lower than the 1.5\*IQR below the lower quartile (Q1), the value will be considered as outlier. In SAS, **PLOTS** options in **PROC UNIVARIATE**tells SAS to generate Box Plot graph.

[**SAS Macro for Identifying Box Plot Method**](http://www.listendata.com/2014/10/identify-and-remove-outliers-with-sas.html)

**II. Studentized Residuals Method**

**Studentized Residuals : Meaning**

Before jumping into studentized residuals, we need to understand the meaning of residuals.

***Residuals****is the difference between the observed value and the predicted value.*

***Standardized Residuals****is the residuals divided by the standard error of estimate.*

***Studentized Residuals****is the residuals divided by the standard error of the residual with that case deleted.*

*If absolute value of studentized residual is greater than 3, the observation is considered as an outlier.*

**SAS Code**  
  
/\* Studentized residuals - Check Outliers\*/

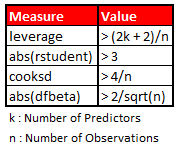
ods graphics on;

proc reg data=reg.crime;  
model crime=pctmetro poverty single / stb clb;  
output out=stdres p= predict r = resid rstudent=r h=lev cookd=cookd dffits=dffit;  
run;  
quit;  
ods graphics off;  
  
/\* Print only those observations having absolute value of studentized residual greater than 3\*/  
proc print data= stdres;  
var r crime pctmetro poverty single;  
where abs(r)>=3;  
run;

**III. Cook's D Method**

Higher the Cook's D is, the more influential the point is.

*If the Cook's D value is greater than 4/(number of observations), the value is considered as an outlier.*

**SAS Code**  
  
proc print data=stdres;  
where cookd > (4/51);  
var cd crime pctmetro poverty single;  
run;  
  
[](http://3.bp.blogspot.com/-yg8U-x8zEek/VPNGmNtkuuI/AAAAAAAADkA/iVBaEryEVFQ/s1600/outlier.png)  
  
**Consequences of Outliers**

*Outliers can affect the estimates of the independent variables*

**Treatment of Outlier**

**1. Percentile capping based on distribution of a variable**

It means replacing extreme values with the largest/smallest non-extreme observation.  
  
In layman's terms, capping at the 1st and 99th percentile means values that are less than the value at 1st percentile are replaced by the value at 1st percentile, and values that are greater than the value at 99th percentile are replaced by the value at 99th percentile.

[**SAS Macro for Percentile Capping**](http://www.listendata.com/2015/01/detecting-and-solving-problem-of-outlier.html)

**2. Compare Models with or without Outliers**

*Smaller the RMSE, Better the Model.*

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model’s predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit.  *Lower values of RMSE indicate better fit.*  
  
**2. Linear Relationship between Dependent and Independent Variables**  
  
**I. Scatter plot of independent variable vs. dependent variable**  
  
ods graphics on;  
proc reg data=reg.crime;  
model crime=pctmetro poverty single / **partial**;  
run;  
quit;  
ods graphics off;

**II . Run correlation between dependent variable and independent variables**  
  
There should be a moderate and **SIGNIFICANT** correlation score between dependent variable and independent variable.  
  
proc corr data=reg.crime;  
var pctmetro poverty single;  
with crime;  
run;

**Consequences of Non-Linear Relationship**  
  
If the assumption of linearity is violated, the linear regression model will return incorrect (biased) estimates. In short, the coefficients as well as R-square will be underestimated.

**Treatment of Non linear Relationship**

1. When the error variance appears to be constant (**Homoscedasticity**), only X needs be transformed to linearize the relationship. Transform independent variable to Log10(X), Inverse(X), Square root(X), Square(X), Exp(X), 1/X, Exp(-X).  
  
2. When the error variance does not appear constant it may be necessary to transform Y or both X and Y. [**Run Box-Cox Transformations for Dependent Variable**](http://www.listendata.com/2015/08/checking-homoscedasticity-with-sas.html)

**3. Errors (Residuals) should be normally distributed**  
  
The Shapiro-Wilk W test can be used to check normality assumption. In this case, we set null hypothesis that residual is normally distributed.

*If the p-value is greater than .05, it means we cannot reject the null hypothesis that residual is normally distributed.*

**SAS Code**

proc reg data=reg.crime;  
model crime=pctmetro poverty single / stb clb;  
output out=stdres p= predict r = resid;  
run;  
  
proc univariate data=stdres normal;  
var resid;  
run;

**Consequences of Non-Normality of Errors**  
  
Many common tests of null hypotheses on regression results require normality. So if the residuals are not normal, then you cannot perform these hypothesis tests.  
  
**Treatment of Non Normality**

Transform the **DEPENDENT** variable. Try log, square root and reciprocal transformations.  
[**Run Box-Cox Transformations for Dependent Variable**](http://www.listendata.com/2015/08/checking-homoscedasticity-with-sas.html)

**4. Homoscedasticity**

There should be homogeneity of variance of the residuals. In other words, the variance of residuals are approximately equal for all predicted dependent variable values.

Another way of thinking of this is that the variability in values for your independent variables is the same at all values of the dependent variable.  
  
**I. Plot Residuals by Predicted values**  
  
proc reg data= reg.crime;  
model crime = poverty single;  
plot r.\*p.;  
run;  
quit;  
  
  
**II. White, Pagan and Lagrange multiplier (LM) Test**  
  
The White test tests the null hypothesis that the variance of the residuals is homogenous (equal). We use the / spec option on the model statement to obtain the White test.

*If the p-value of white test is greater than .05, the homogenity of variance of residual has been met.*

**With PROC REG ( No CLASS statement , No Pagan Test)**

*proc reg data= reg.crime;  
model crime = poverty single****/ SPEC****;  
run;*

***Note :****P-value greater than .05 indicates homoscedasticity.*

**With PROC AUTOREG (LM Test, CLASS statement for categorical variables)**

*proc autoreg data=reg.crime;  
model crime = pctmetro poverty single / archtest;  
output out=r r=yresid;  
run;*

***Note :****Check P-value of Q statistics and LM tests. P-value greater than .05 indicates homoscedasticity.*

**With PROC MODEL (White and Pagan Test , No CLASS statement for categorical variables)**

*proc model data= reg.crime;  
parms a1 b1 b2;  
crime = a1 + b1\*poverty + b2\*single;  
fit crime / white pagan=(1 poverty single)  
out=resid1 outresid;  
run;  
quit;*

*If the p-value of white test and Breusch-Pagan test is greater than .05, the homogenity of variance of residual has been met.*

**Consequences of Heteroscedasticity**

1. The regression prediction remains unbiased and consistent but inefficient. It is inefficient because the estimators are no longer the Best Linear Unbiased Estimators (BLUE).
2. The hypothesis tests (t-test and F-test) are no longer valid.

**Treatment of Heteroscedasticity**

[**Testing and Correcting Heteroscedasticity**](http://www.listendata.com/2015/08/checking-homoscedasticity-with-sas.html)  
 **5. Multicollinearity**  
Mutlicollinearity  means there is a high correlation between independent variables. The linear regression model MUST NOT be faced with problem of multicollinearity.  
  
**VIF (Variance Inflation Factor)**  
  
It measures how much the variance of an estimated regression coefficient is increased because of collinearity.  
**Interpretation :**If the variance inflation factor of a predictor variable were 9 (Sqrt(9) = 3) this means that the standard error for the coefficient of that predictor variable is 3 times as large as it would be if that predictor variable were uncorrelated with the other predictor variables.

*If VIF is greater than 5, there is a multicollinearity problem in the model.*

proc reg data = reg.crime;  
model crime = poverty single / **vif**;  
run;  
  
**Consequences of Multicollinearity**  
  
Multicollinearity inflates the standard errors, making it impossible to determine the relative  
importance of the predictors. In other words, the coefficients will be unreliable. Note that multicollinearity does not affect the efficiency of the estimators – they remain BLUE (Best Linear Unbiased Estimators).   
  
**Treatment of Multicollinearity**  
  
1. Run **PROC VARCLUS** with **HI** option (Principal Component Analysis). A variable that has the lowest 1-R2 ratio is likely to be a good representative for the cluster.  
  
2. Use centering: which is subtracting the mean from the predictor values before generating the square term. The resulting centered data may well display considerably lower multicollinearity.  
  
For example : Weight and Weight2 are faced with problem of multicollinearity.  
  
First Step : Center\_Weight = Weight - mean(Weight)  
Second Step : Center\_Weight2 = Center\_Weight ^2

**6. Independence of error terms - No Autocorrelation**  
  
It states that the errors associated with one observation are not correlated with the errors of any other observation. ***It is a problem when you use time series data.***Suppose you have collected data from labors in eight different districts. It is likely that the labors within each district will tend to be more like one another that labors from different districts, that is, their errors are not independent.  
  
proc reg data = reg.crime;  
model crime = poverty single / **dw**;  
run;  
  
PROC REG tests for **first-order autocorrelations** using the **Durbin-Watson**coefficient (DW). The null hypothesis is no autocorrelation.

*A DW value between 1.5 and 2.5 confirms the absence of  first-order autocorrelation. If DW value less than 1.5, it indicates positive autocorrelation. If DW value greater than 2.5, it indicates negative autocorrelation*

Autocorrelation inflates significance results of coefficients by underestimating the standard errors of the coefficients. Hypothesis testing will therefore lead to incorrect conclusions.   
  
**Another alternative test : Lagrange Multiplier Test**  
  
**It can be used for more than one order of auto correlation.** It consists of several steps. First, regress Y on Xs to get residuals. Compute lag value of residuals up to pth order. Replace missing values for lagged residuals with zeros. Rerun regression model including lagged residual variable as an independent variable.

*proc autoreg data = reg.crime;  
model crime = poverty single / dwprob godfrey;  
run;*

**Consequences of Autocorrelation**

*Autocorrelation inflates t-statistics by underestimating the standard errors of the coefficients. Hypothesis testing will therefore lead to incorrect conclusions. Estimators no longer have minimum variance but they will remain unbiased.*

**Treatment of Autocorrelation**

1. Add lagged transforms (lag value) of the dependent variable

2. Use PROC AUTOREG

It is advisable to build auto-regressive model with PROC AUTOREG for time series data.

**Important Point 1 :**Box Cox Transformation of Dependent Variable can solve problem of non-linearity, non-normality of error and heteroscedasticity.  
[**Run Box-Cox Transformations for Dependent Variable**](http://www.listendata.com/2015/08/checking-homoscedasticity-with-sas.html)

**Important Point 2 : RMSE for Training vs Test Sample**

The RMSE for your training and your test sets should be very similar if you have built a good model. If the RMSE for the test set is much higher than that of the training set, it is likely that you've badly over fit the data, i.e. you've created a model that tests well in sample, but has little predictive value when tested out of sample.  
  
**Important Point 3 : Transformation Rules**  
  
The specific transformation used depends on the extent of the deviation from normality.  
  
1. If the distribution differs moderately from normality, a square root transformation is often the best.  
2. A log transformation is usually best if the data are more substantially non-normal.  
3. An inverse transformation should be tried for severely non-normal data.

4. **If nothing can be done to "normalize" the variable, then you might want to dichotomize (2 categories) the variable.**