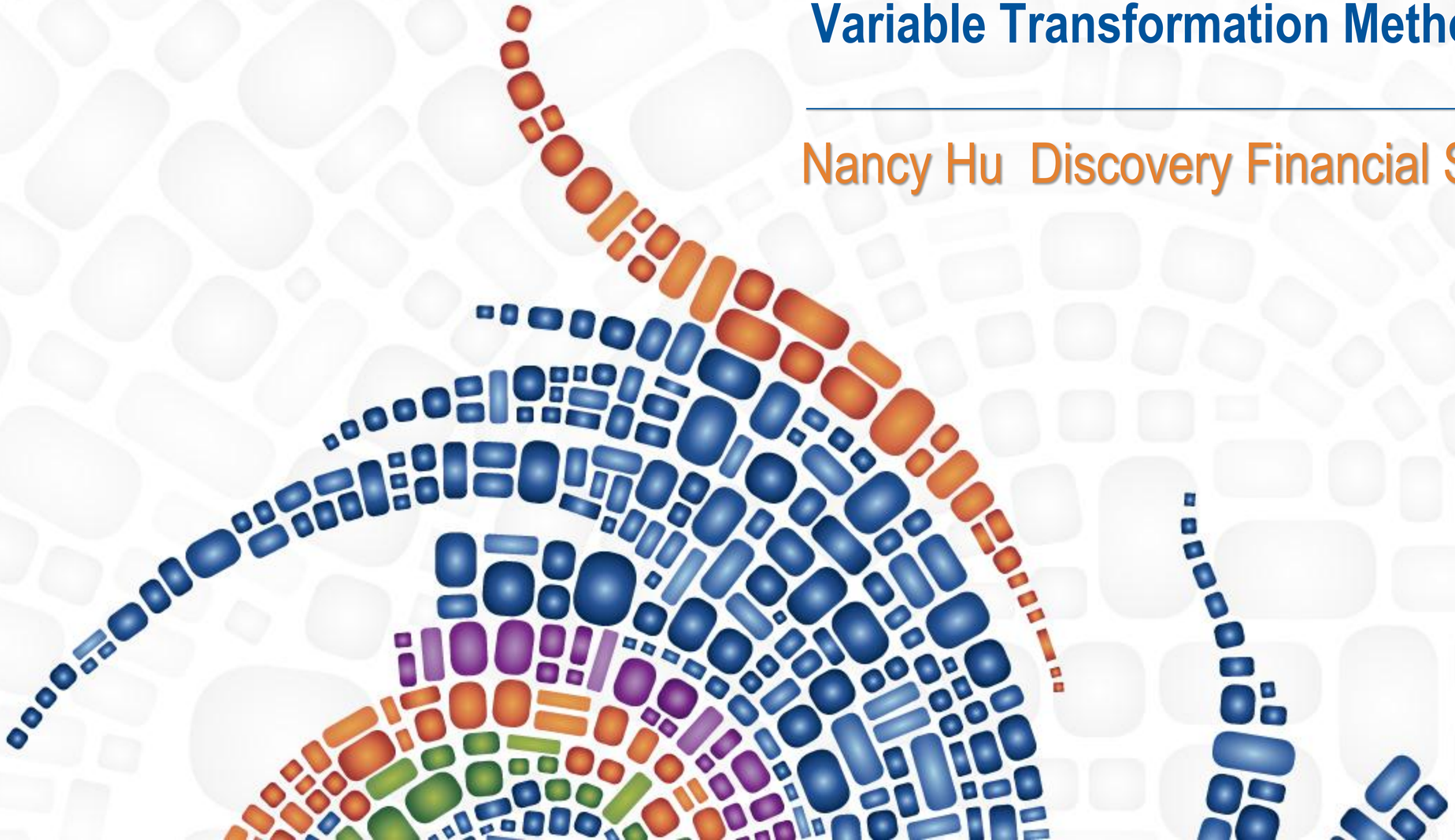
SAS® GLOBAL FORUM 2015

The Journey Is Yours

# Introduce a Linear Regression Model by Using the Variable Transformation Method

Nancy Hu  Discovery Financial Service



SAS GLOBAL FORUM

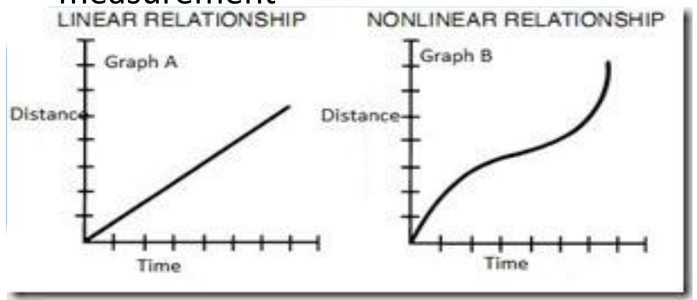
# Introduce a Linear Regression Model by Using Variable Transformation Method

## Abstract

This study is intended to assist Analysts to generate the best of variables using simple arithmetic operators (square root, log, loglog, exp and rcp). The advantage of this methodology is that the new variables may be more significant than the original ones. This paper gives a new way to compute all the possible variables using a set of math transformation.

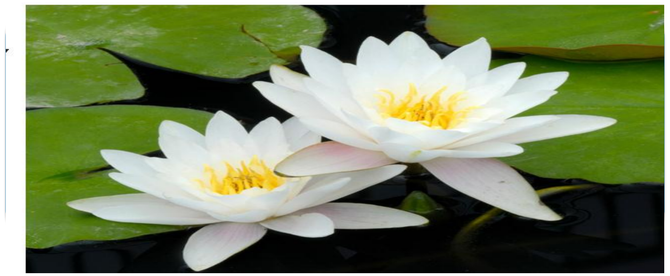
## Objectives

- Objective 1 assumption IID
- Objective 2 Variable Transformation
- Objective 3 Model procedure and measurement



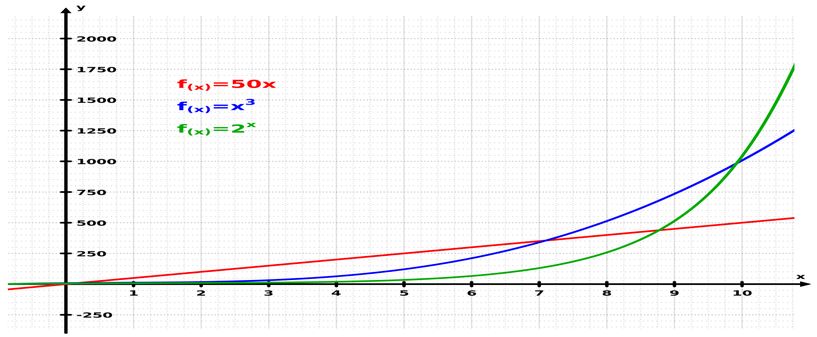

## Methods

- Method 1 – Sas SGplot
- Method 2 – Variable Transformation
- Method 3 – Error Plot



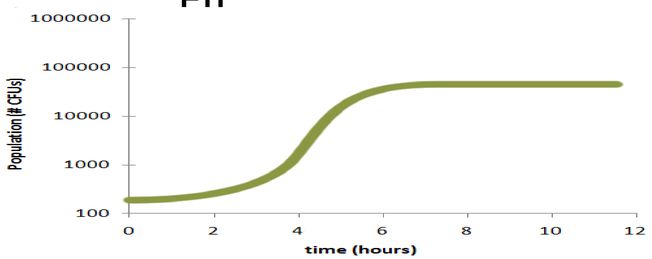
## Results

1. Assumptions regarding linear regression
2. Derived variable and variable transformation to fit better model
3. Variable selection multicollinearity and CORR.
4. creating the model using different statistical procedure
5. testing for assumption validation, , auto correlation and effects of outliers

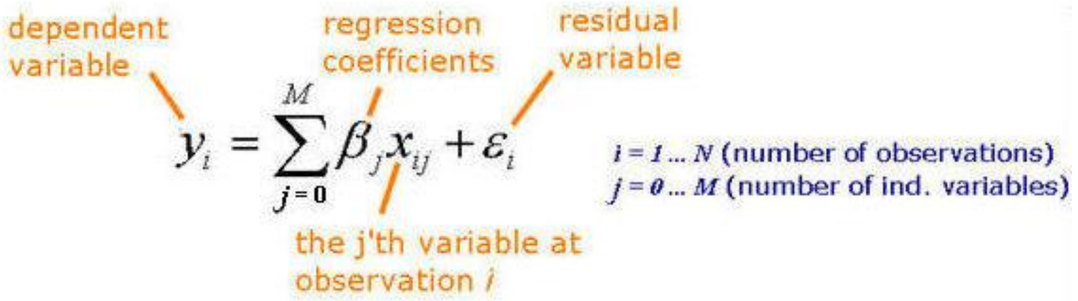



## Conclusions

- Conclusion 1 – S-plot can help you
- Conclusion 2 non-linear variable transform can be best choice
- Conclusion 3 Better Model Fit



# Model Assumption




## IID **independent, identically and Normal Distribution**

**Assumptions**

A **statistical model** is an expression that attempts to explain patterns in the observed values of a response variable by relating the response variable to a set of predictor variables and parameters.
Consider the following familiar statistical model:

$y = mx+c+e$

This simple statistical model relates a response variable (*y*) to a single predictor variable (*x*) as a straight line according to the values of two constant parameters:

*m* – the degree to which *y* changes per unit of change in *x* (gradient of line)

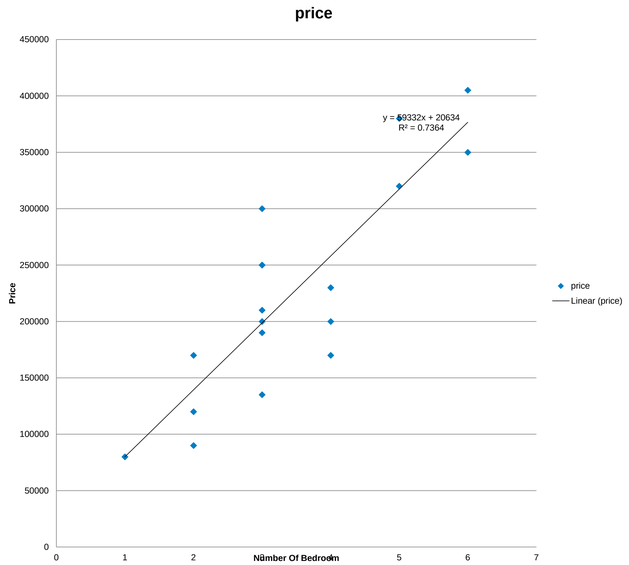
*c* – the value of *y* when *x* = 0 (intercept).

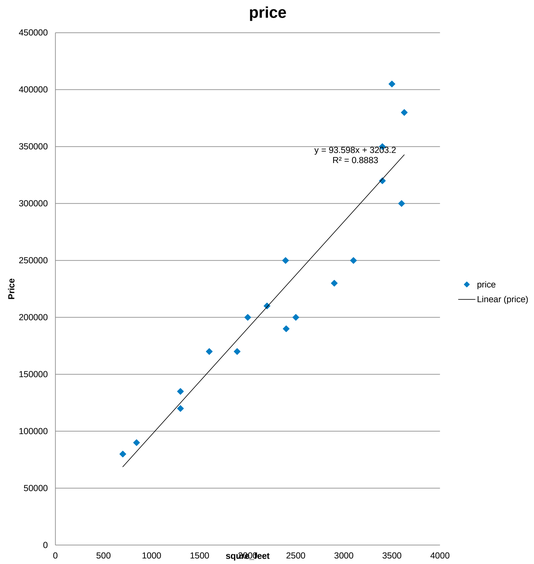
e is the residual.

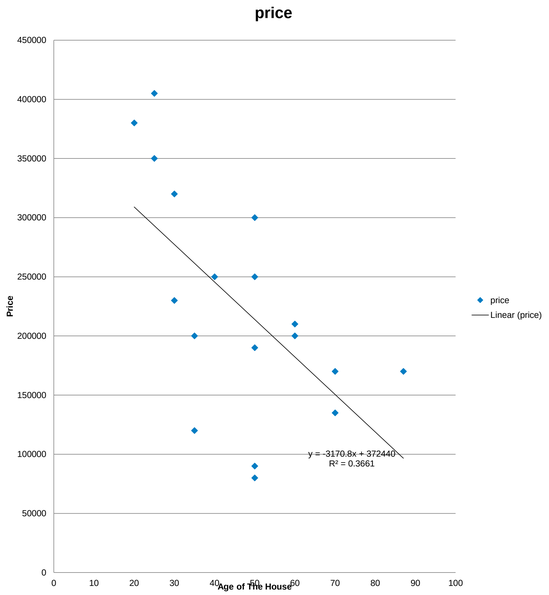
e can not be explained by the variables in the model. Most of the assumptions and diagnostics of linear regression focus on the assumptions of e. The following assumptions must hold when building a linear regression model. The error term is normally distributed with a mean of zero and a standard deviation of s2, N(0,s2). Although not an actual assumption of linear regression, it is good practice to ensure the data you are modeling came from a random sample or some other sampling frame that will be valid for the conclusions you wish to make based on your model.

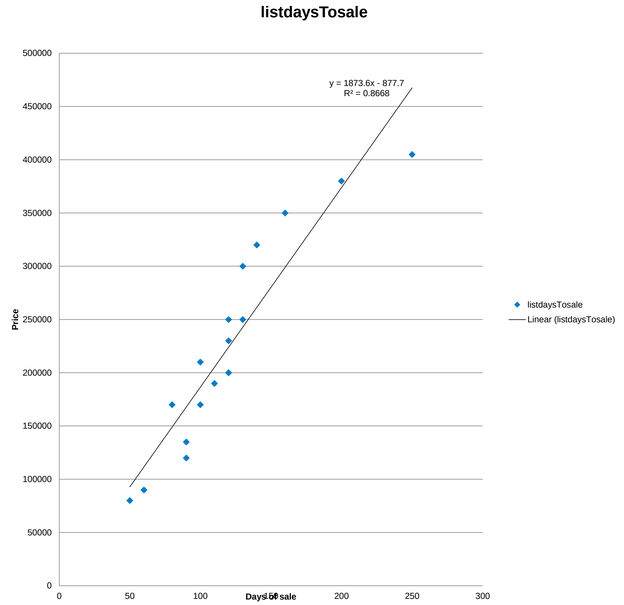
# Derived Variable and Variable Transformation.

- **SGplot, Log, Loglog, Sqrt, EXP and RCP Variable Transformation**
- Before you begin modeling, it is recommended that **you plot your data** when the number of the variable <=20. By examining these initial plots, you can quickly assess whether the data have linear relationships or interactions are present.  The code below will produce three plots.
- PROC SGPLOT DATA=HOUSES;
- PLOT PRICE*(AGE BEDROOMS SQFEET LISTDAYS);RUN;



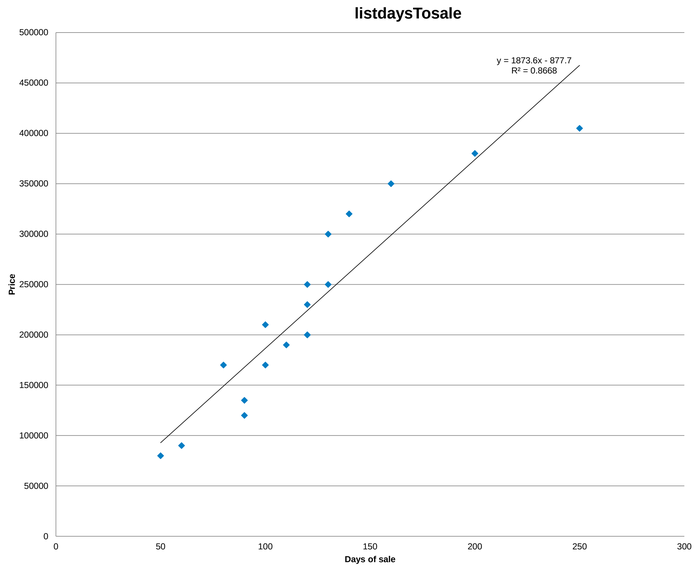



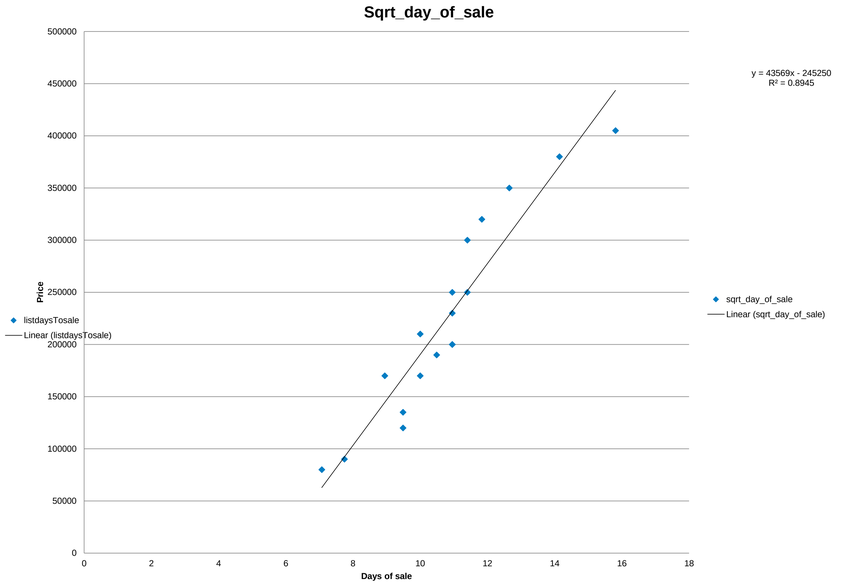



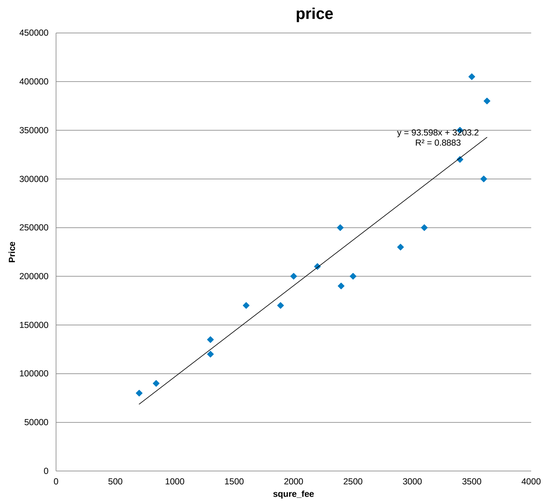



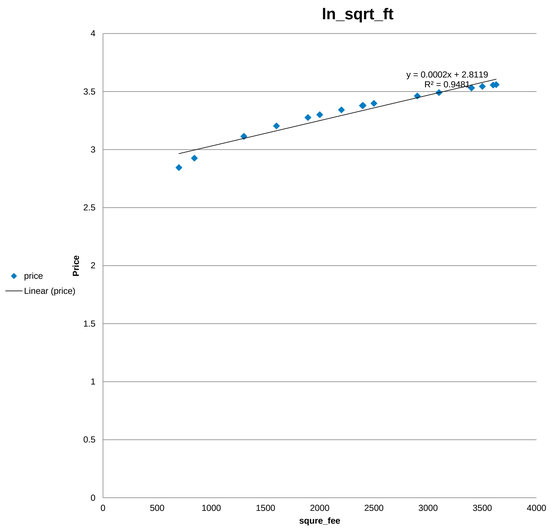
# Derived Variable SQRT, RCP and Log Transformation.

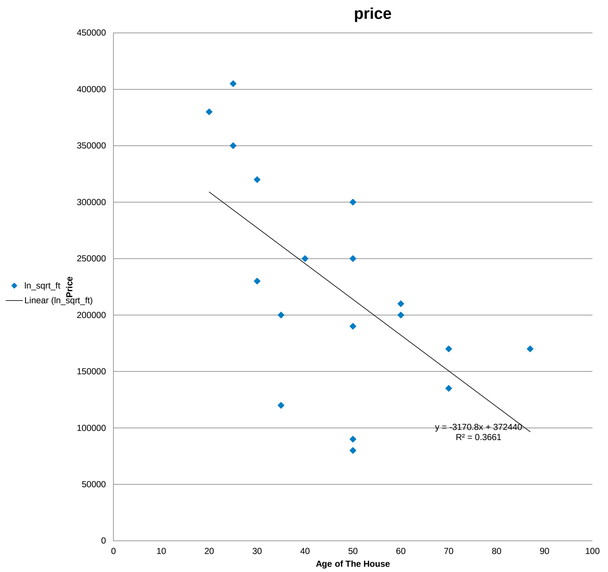
Variable Transformation can help to make better fit of the non-linear into linear, Based on the shape of the fit x with y, SQRT, Log, Exp or Rcp can be used.

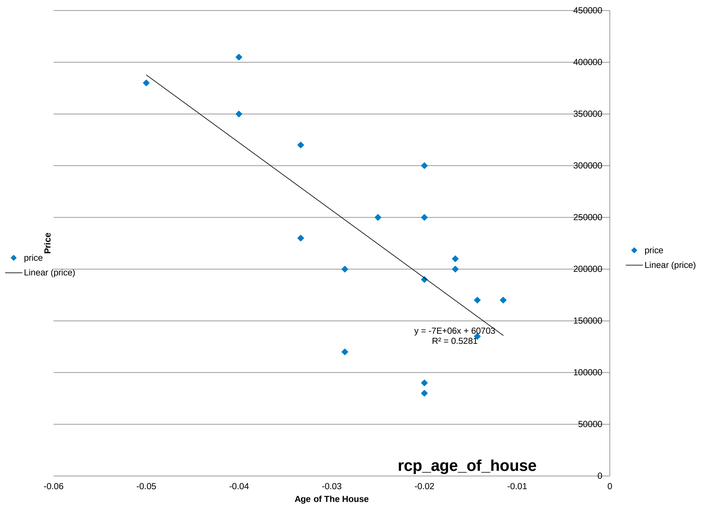


# Multicollinearity

Multicollinearity existed to change sign of model coefficent when VIF>10. Square root of vif can give you the error between two correlated Variable.  when your independent,X variables are correlated. A statistic called the Variance Inflation Factor, VIF, can be used to test for multicollinearity. **changing the model**, eg. drop a term; transform a variable; or use Ridg Regrssion.

Ridge Regression.  To check the VIF statistic for each variable you can use REG with the VIF option in the model statement.

PROC REG DATA=HOUSES; MODEL PRICE = BEDROOMS LOG_SQFT S1 S2 S3 BEDSQFT / VIF;

**Correlation** can be easily used to detect correlation between dependent variable and independent variables.

PROC CORR DATA=HOUSES;

VAR BATHS BEDROOMS SQFEET;RUN;

# GLM / REG

**The GLM procedure** can also be used to create a linear regression model. The GLM procedure is the safer procedure to use for your final modeling because it does not assume your data are balanced.  That is with respect to categorical variables, it does not assume you have equal sample sizes for each level of each category. GLM also allows you to write interaction terms and categorical variables with more than two levels directly into the MODEL statement.

```
PROC GLM DATA=HOUSES; CLASS STYLE;

MODEL PRICE = BEDROOMS LGT_SQFT STYLE
BEDROOMS*SQFEET;RUN;

PROC REG DATA=HOUSES; MODEL PRICE = BEDROOMS
LGT_SQFT STYLE BEDROOMS_SQFEET STYLE1 STYLE2
STYLE3;RUN;
```

# Test of Assumptions - Heteroscedasticity(SPEC), Time Series (Durbin-Watson) :Normal

We will validate the **"iid" assumption** of linear regression by examining the residuals of final model. Specifically, we will use diagnostic statistics from REG as well as create an output dataset of

Residual values for PROC UNIVARIATE to test. The following SAS code will do this for us.

PROC REG DATA=HOUSES;MODEL PRICE = BEDROOMS S1 S2 S3 /**DW SPEC** ;OUTPUT OUT=RESIDS R=RES;

The Durbin-Watson statistic test for first order correlation of error terms. The Durbin Watson statistic ranges from 0 to 4.0. **Generally a D-W statistic of 2.0 indicates the data are independent**. A small (less then 1.60) D-W indicates positive first order correlation and a large D-W indicates negative first order correlation.

PROC UNIVARIATE DATA=RESIDS **NORMAL PLOT**;VAR RES;RUN;

# Testing for Outliers

Outliers are observations that exert a large influence on the overall outcome of a model or a parameter's estimate. When examining outlier diagnostics, the size of the dataset is important in determining

The REG procedure can be used to view various outlier diagnostics. The Influence option requests a host of outlier diagnostic tests. The R option is used to print out Cook's D. The output prints statistics for each observation in the dataset.

```
PROC REG DATA=HOUSES;

MODEL PRICE = BEDROOMS S1 S2 S3 /INFLUENCE R;

RUN;
```

# Testing the Fit of the Model

The overall fit of the model can be checked by looking at the F-Value and its corresponding p-value
(Prob >F) for the total model under the Analysis of Variance portion of the REG or GLM print out. Generally, you want a Prob>F value less than 0.05. If your dataset has "replicates", you can perform a
formal Lack of Fit test. This test can be run using PROC RSEG with option LACKFIT in the model
statement.

```
PROC RSREG DATA=HOUSES
MODEL PRICE = BEDROOMS STYPE/LACKFIT;
RUN;
```

If the p-value for the Lack of Fit test is greater than 0.05 then your model is a good fit and no additional terms are needed.

**Session Title:** Introduce a Linear Regression Model by Using the Variable Transformation Method
**Session Type:** E-Poster
**Day Scheduled:** Monday, 4/27/15
**Start Time/End Time:** 14:45/ 15:15
**Location:** The Quad E-Poster Station 4

# Session ID 3052

SAS® GLOBAL FORUM

April 26-29
Dallas, TX