Paper 253-25

Detecting Curvilinear Relationships in PROC REG

Mel Widawski, UCLA, Los Angeles, CA

ABSTRACT

We all know the world is not flat, but many researchers continue to model their world as a linear system. Many systems encountered in research are not linear; the relationship between motivation and errors is one example. Luckily, many relationships that are not linear may be expressed as linear relationships. We can detect these relationships with the PARTIAL option on the MODEL statement in PROC REG. I will show you how to interpret the plots produced, and demonstrate that curvilinear relationships are revealed in partial plots that cannot be detected in simple scattergrams. Finally, I will show you how to create transformed variables to represent the nonlinear terms, and the value of centering in creating Quadratic terms.

INTRODUCTION

Various types of curvilinear relationships may be encountered in research. Some are routinely implied by transformations of the dependent variable. One of the most common is the log transformation used to alleviate a problem with heteroscedasticity. The need for this is routinely discovered in residual plots.

Another form of curvilinear relationship involves a term that is some power of one of your variables. The Quadratic or squared term is the most common. Since this usually involves only one of the predictor variables it is sometimes not detectable by either a residual plot, or a bivariate plot of the independent variables with your criterion variables. The partial residual plots are useful to detect this type of relationships. The second section will demonstrate this technique.

When adding a quadratic term to your model a technique called centering is useful for cleanly assessing the relative linear and quadratic contributions to your model. It is also useful to prevent the introduction of multicollinearity into the model with the addition of quadratic or higher terms.

The SAS® PROC REGRESSION features that will be used are the PLOT statement, the OUTPUT statement, and the PARTIAL option of the MODEL statement.

The PLOT statement is useful for generating residual plots as well as bivariate plots of the original variables.

The OUTPUT statement writes the residual, predicted value, and other statistics to an output data set along with all of the original variables.

The PARTIAL option creates a plot of the residual of the dependent variable with all of the other variables removed, and the residual of each predictor with all other variables removed. It is useful in detecting hidden curvilinear relationships.

EXPONENTIAL RELATIONSHIPS

The first type of curvilinear relationship you may encounter is the exponential relationship. It is usually discussed under a heading of Normality or Heteroscedasticity. Many people are familiar with skewed distributions on variables and using log transformations with these variables, especially if they are dependent variables. What seems to escape notice is the fact that using the log transformation implies that the underlying relationship is exponential.

DETECTING HETEROSCEDASTICITY

Heteroscedasticity is a violation of one of the assumptions of linear regression, which assumes a constant variability about the regression line. If the variability increases as the values of the predicted value increases then certain transformations are applied. Among the choices are the log, square root, and reciprocal transformations.

Usually the need for one of these transformations is determined by examining the residual plot. If the residual plot is fan shaped then heteroscedasticity is assumed.

The following example demonstrates use of the PLOT statement in PROC REG to produce residual plots:

```
PROC REG DATA=in.hetero;
  MODEL yb = x1 x5;
  PLOT  R.*P.;
  OUTPUT OUT=outres P=pred R=resid ;
RUN;
```

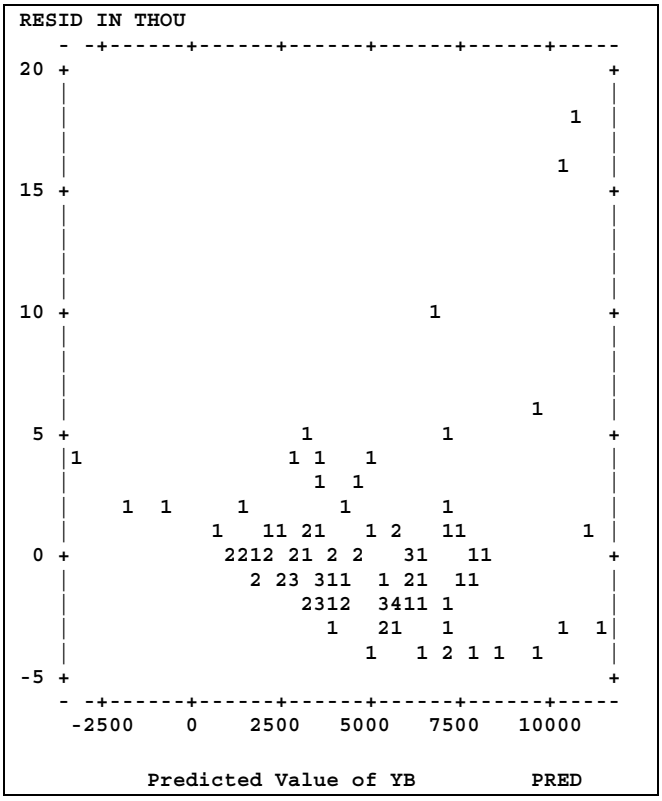
The OUTPUT statement allows you to add the predicted value and the residual value to the original variables in a new data set called OUTRES, which will be used below.

In our sample data set the variable YB is a variable generated by a DATA step to have a specific relationship with the independent variables, X1 and X5. I am using manufactured variables so that I can tell you without any possibility of error that a certain underlying relationship exists.

In the result of the PROC REG, notice that the model as a whole is significant with a p<.0001. The model accounts for almost 41% of the variance in the dependent variable YB.

```
Analysis of Variance

                Sum of          Mean
Source  DF     Squares        Square   F Val  Prob>F
Model    2 803868181.28 401934090.64  33.51  0.0001
Error   97   1163140896 11991143.258
C Total 99 1967009077.3

Root MSE    3462.82   R-square 0.4087
Dep Mean    4704.31   Adj R-sq 0.3965

Parameter Estimates

          Param   Stand  T for H0:
Var  DF   Estim   Error  Param=0   P>|T|
INTER 1 -34355  4888.97  -7.027   0.0001
X1    1  220.69   32.42   6.807   0.0001
X5    1 1668.27  345.57   4.828   0.0001
```

Look at the result of the residual plot below and notice the fan shaped plot of the residuals. This is a clear condition of heteroscedasticity.



We can now use PROC UNIVARIATE to further examine the residuals output to OUTRES above. The code that accomplishes this is presented below:

```
PROC UNIVARIATE DATA=outres PLOT NORMAL;
  VAR resid yb;
RUN;
```

Look at the results of PROC UNIVARIATE below and notice the pronounced skew and Kurtosis. Notice also that the test for normality, W fails with a p<.0001.

Moments

N	100	Sum Wgts	100
Mean	0	Sum	0
Std Dev	3427.667	Variance	11748898
Skewness	2.774061	Kurtosis	11.08419
USS	1.1631E9	CSS	1.1631E9
CV	.	Std Mean	342.7667
T:Mean=0	0	Pr>|T|	1.0000
Num ^= 0	100	Num > 0	38
M(Sign)	-12	Pr>=|M|	0.0210
Sgn Rank	-615	Pr>=|S|	0.0338
W:Normal	0.769157	Pr<W	0.0001

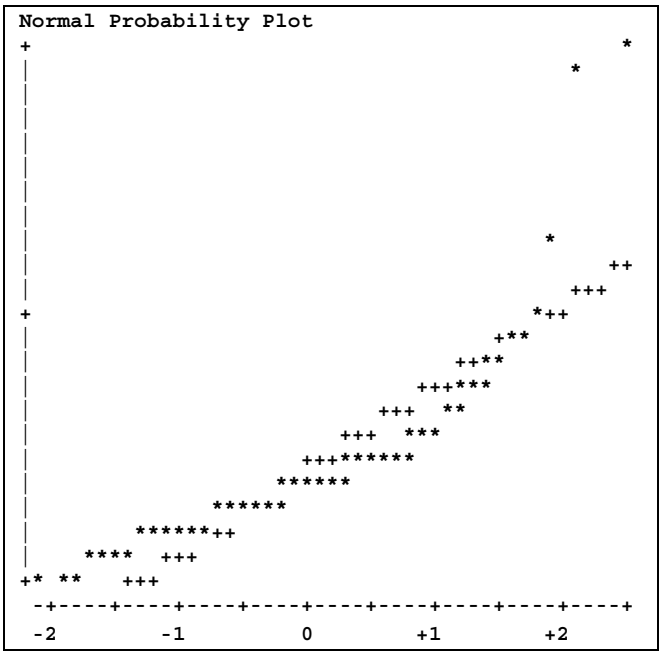
The stem and leaf and box plots to the right also reveal the degree of skew of the residual. It is produced because the PLOT option was specified on the PROC UNIVARIATE statement above. It is because skewed residuals are often the result of an underlying exponential relationship that it has become common to use the log transform when these values are encountered.

Variable=RESID Residual

Stem	Leaf	#	Boxplot
17	7	1	*
16	2	1	*
15			
14			
13			
12			
11			
10			
9	5	1	*
8			
7			
6	3	1	0
5	46	2	0
4	68	2	0
3	579	3	|
2	0356	4	|
1	0034577	7	|
0	0012233455677778	16	+--+--+
-0	9985555444311110	16	*-----*
-1	8877776554433211110000	22	+-----+
-2	97765555211100	14	|
-3	997210	6	|
-4	5211	4	|

Multiply Stem.Leaf by 10+3**

The following normal probability plot produced by PROC UNIVARIATE also demonstrates the lack of normality.



THE LOG TRANSFORM

When this problem is encountered the first remedy most people attempt is to transform the dependent variable using the log transform. This is accomplished in a data step.

The following source code creates YBLOG, which is the log transform of YB. And then the PROC REG is rerun with the transformed variable as the dependent variable in the model.

```
DATA fixed;
  SET in.hetero ;
   yblog = log(yb) ;
   ybsqrt= sqrt(yb) ;
RUN;

PROC REG DATA=fixed;
  MODEL yblog = x1 x5;
  PLOT  R.*P.;
  OUTPUT OUT=outres2 P=pred R=resid ;
RUN;
```

The residual and predicted values are output into variables named RESID and PRED in the data set OUTRES2. The residuals are also plotted.

If the transformation worked and the underlying relationship is exponential then the regression model should improve, and the residual plot should be more oval than fan shaped.

Dependent Variable: YBLOG

Source	DF	Sum of Square	Mean Squa	F Val	Prob>F
Model	2	41.486	20.74	77.02	0.0001
Error	97	26.121	0.26		
C Total	99	67.607			

Root MSE	0.518	R-square	0.6136
Dep Mean	8.130	Adj R-sq	0.6057
C.V.	6.382		

The model is still significant but now it accounts for over 61% of the variance of the dependent variable, and if you remember before we only accounted for less than 41% of the variance. Prediction has substantially improved. This is compelling evidence that we are on the right track in discovering the underlying relationship.

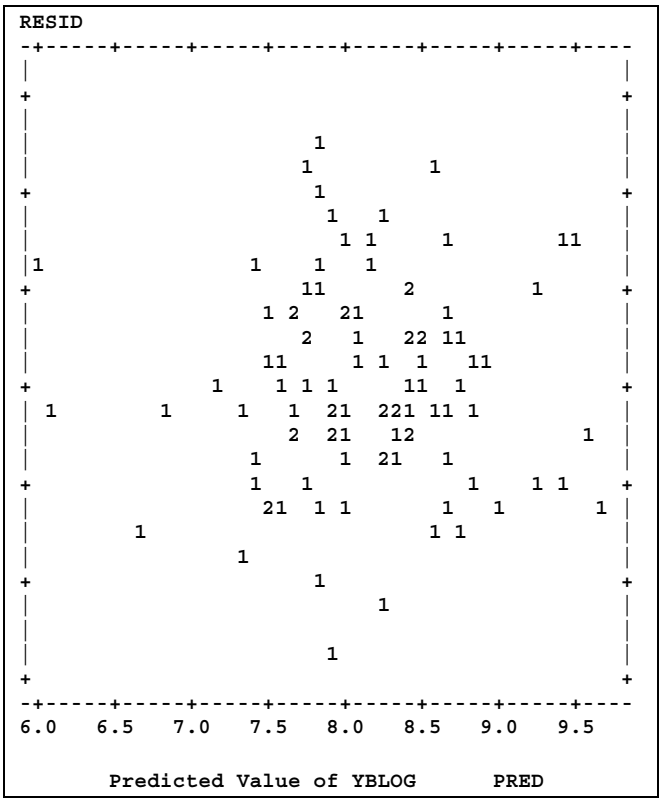
The following are the new parameter estimates. Notice that with the log transformed variable the intercept is no longer significantly less than zero. The parameters have decreased; this is to be expected since magnitude of the dependent variable has decreased. It does not mean that they are any less important.

Parameter Estimates

Var	DF	Param Estim	Stand Error	T for H0: Param=0	P>|T|
INT	1	-0.73	0.7326	-0.999	0.3201
X1	1	0.05	0.0048	10.358	0.0001
X5	1	0.37	0.0517	7.263	0.0001

The residual plot has also changed; it is more oval than fan shaped. The seeming outliers are simply chance variation with the small sample size. The values were generated using a normal random number function.

This residual plot follows :



While the above plot is not completely oval, it certainly is less fan shaped. It is about what you would expect with random, normally distributed error variance.

The results of another PROC UNIVARIATE shows a marked improvement in skew, kurtosis and tests for normality.

Moments

N	100	Sum Wgts	100
Mean	0	Sum	0
Std Dev	0.513665	Variance	0.263851
Skewness	0.073144	Kurtosis	-0.16873
USS	26.12129	CSS	26.12129
CV	.	Std Mean	0.051366
T:Mean=0	0	Pr>|T|	1.0000
Num ^= 0	100	Num > 0	46
M(Sign)	-4	Pr>=|M|	0.4841
Sgn Rank	-22	Pr>=|S|	0.9402
W:Normal	0.982736	Pr<W	0.6724

Notice that the skew is now .073144 and the kurtosis is now –0.16873, and both are marked improvements. The test for normality yields a W: .982, p=.6724, which shows no deviation from normality.

The stem and leaf and box plots also show marked improvement in the distribution of the residuals. Both look more nearly normal and far less skewed. They follow on the next page.

```
Variable=RESID          Residual

 Stem Leaf                      #  Boxplot
   12 0                         1      |
   11 1                         1      |
   10 39                        2      |
    9                                  |
    8 34                        2      |
    7 57789                     5      |
    6 138                       3      |
    5 8                         1      |
    4 11344559                  8      |
    3 13456                     5  +------+
    2 2445588                   7  |      |
    1 029                       3  |      |
    0 15677899                  8  |  +   |
   -0 988887775411             12  *------*
   -1 9988650                   7  |      |
   -2 655433100                 9  |      |
   -3 85222                     5  +------+
   -4 0                         1      |
   -5 766400                    6      |
   -6 9865411                   7      |
   -7 81                        2      |
   -8 40                        2      |
   -9                                  |
  -10 93                        2      |
  -11                                   |
  -12                                   |
  -13 3                         1      0
      ----+----+----+----+
  Multiply Stem.Leaf by 10**-1
```

In conclusion, remember that when you transform a dependent variable you are implying that there is an underlying relationship. A log transformation implies an exponential relationship.

Thus:

```
LOG(YB)= -0.73+ 0.05x1+ 0.37x5
```

Is the same as:

```
YB= EXP(-0.73+ 0.05x1+ 0.37x5)
```

This is the underlying relationship expressed in terms of the original variable.

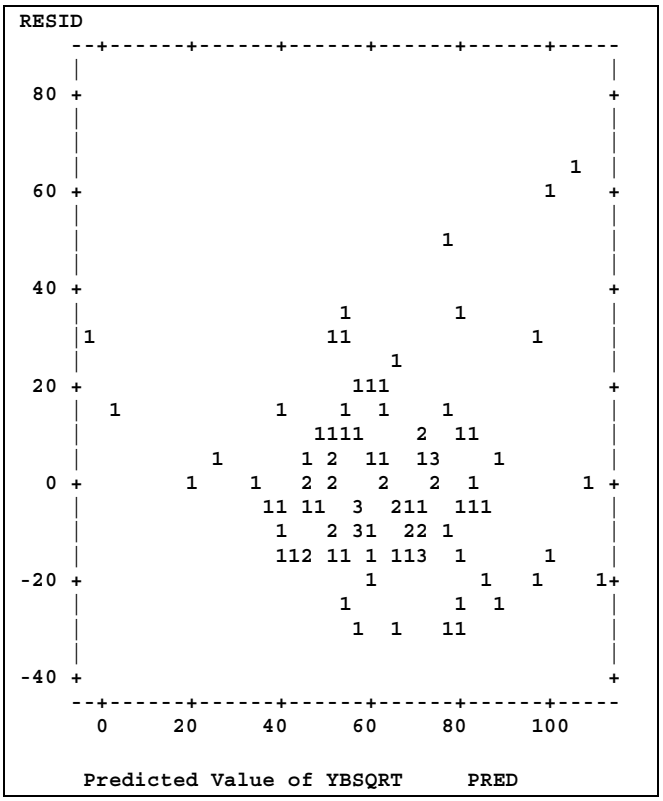
SQUARE ROOT TRANSFORMATIONS

The square root transformation is usually the next one tried if the log transformation does not work. To demonstrate that the log transformation was correct in the previous example we will use the variable YBSQRT created above.

The following PROC REG uses YBSQRT as the dependent variable:

```
DATA fixed;
  SET in.hetero ;
   yblog  = log(yb) ;
   ybsqrt = sqrt(yb) ;
RUN;

PROC REG DATA=fixed;
  MODEL ybsqrt = x1 x5;
  PLOT  R.*P.;
  OUTPUT OUT=resid P=pred R=resid ;
RUN;
```

If this were the proper transformation then the regression results should be improved. Remember that using the log transform the proportion of variance explained was over 61%, now using the square root transformation only 54% of the variance is explained. This is clearly not an improvement in performance.



The residual plot above still shows a fan shape. So the heteroscedasticity problem is not alleviated. When PROC UNIVARIATE is run on the residuals there is still a deviation from normality as shown by the excerpt of the results below.

```
W:Normal    0.923445   Pr<W    0.0001
```

The distribution is still skewed:

```
Stem Leaf                      #  Boxplot
   6 14                        2      0
   5
   5 0                         1      0
   4
   4
   3 6                         1      |
   3 00023                     5      |
   2 7                         1      |
   2 122                       3      |
   1 5557                      4      |
   1 0224                      4      |
   0 56666788889              11  +------+
   0 01122233344              11  |  +   |
  -0 3333222100               10  *------*
  -0 999887777666555          15  |      |
  -1 44443331100000           14  +------+
  -1 986666555                 9      |
  -2 4320                      4      |
  -2 885                       3      |
  -3 10                        2      |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```

In conclusion, it would be safe to assume that the exponential model is a better fit for the data. And in fact YB was created from the **e** raised to the power of 0.05*x1+ 0.37*x5+ a random error component. This is the correct underlying relationship. The converse is possible with other data sets, which might have an underlying quadratic relationship (Square Root Transform).

OTHER CURVILINEAR MODELS

There are times when the correct model involves additional variables that are transformations of the predictors. The classic case of this is the quadratic equation. Detecting these models can be tricky, and at times researchers will attempt to do a log transformation instead of the appropriate transformation. Many times residual plots will have a classic fan shape. Simple XY plots may not reveal the effect if there are other contributing variables to mask this relationship. The following analysis is done on a data set where the variables were created according to a specific underlying model

DETECTING A CURVILINEAR EFFECT

There is an option on the MODEL statement that is very useful for detecting curvilinear relationships. The PARTIAL option requests partial regression plots. In the following example we have three predictors X1, X2, and X3 all of which are related to the dependent variable Y2. A partial plot will be created for each of the predictors. The residual of the dependent variable after regressing all the other predictors is plotted against the residual of an independent variable after regressing the other predictors on it. When this is done hidden curvilinear effects can be revealed.

The program below requests a partial regression plot:

```
PROC REG DATA=in.nonlin;
  MODEL y2 = X1 X2 X3 /PARTIAL;
  PLOT R.*P.  y2*x1 y2*x2 y2*x3 ;
  OUTPUT OUT=resid P=pred R=resid ;
RUN;
```

Look at the following output. At first glance it looks like the model is pretty good, as the model explains almost 72% of the variance of Y2. All three of the predictors contribute to the model, and the model itself is significant. Most researchers would stop at this point and thank their lucky stars.

Analysis of Variance

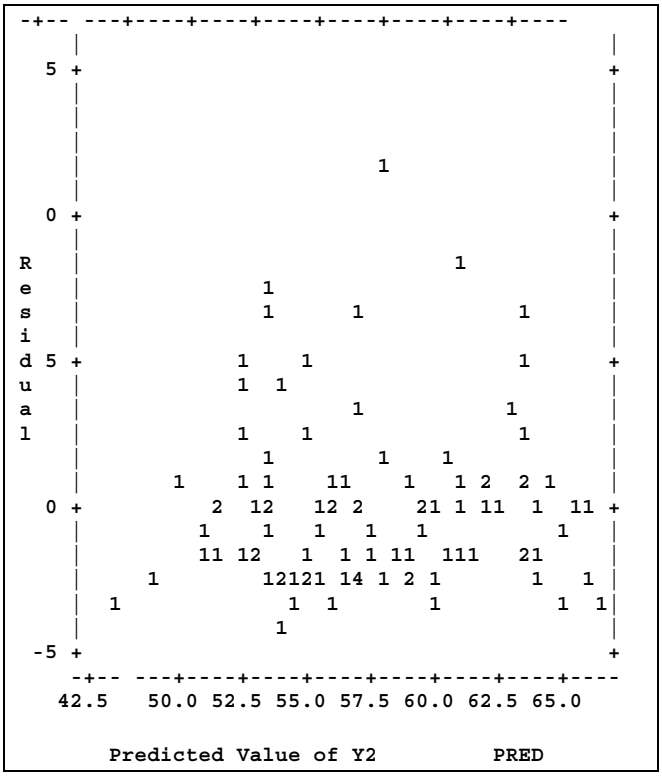
Source	DF	Sum of Squares	Mean Square	F Val	Prob>F
Model	3	2260.819	753.60	88.33	0.0001
Error	96	818.998	8.53		
C Total	99	3079.817			

Root MSE	2.92083	R-square	0.7341
Dep Mean	57.49389	Adj R-sq	0.7258
C.V.	5.08024		

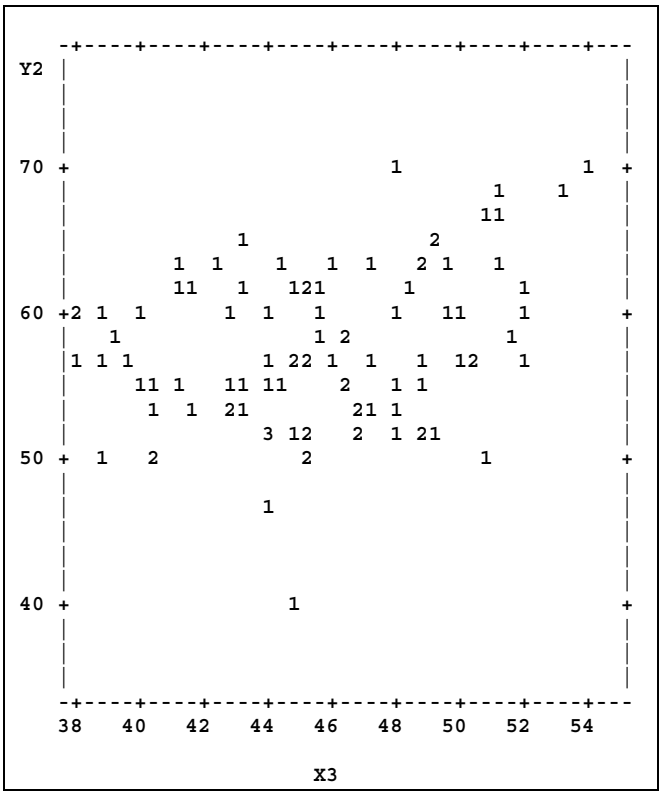
Parameter Estimates

Var	DF	Param Estim	Stand Err	T for H0: Param=0	P>|T|
INTER	1	-1.581	4.3484	-0.364	0.7169
X1	1	0.405	0.0275	14.728	0.0001
X2	1	0.096	0.0336	2.867	0.0051
X3	1	0.297	0.0730	4.071	0.0001

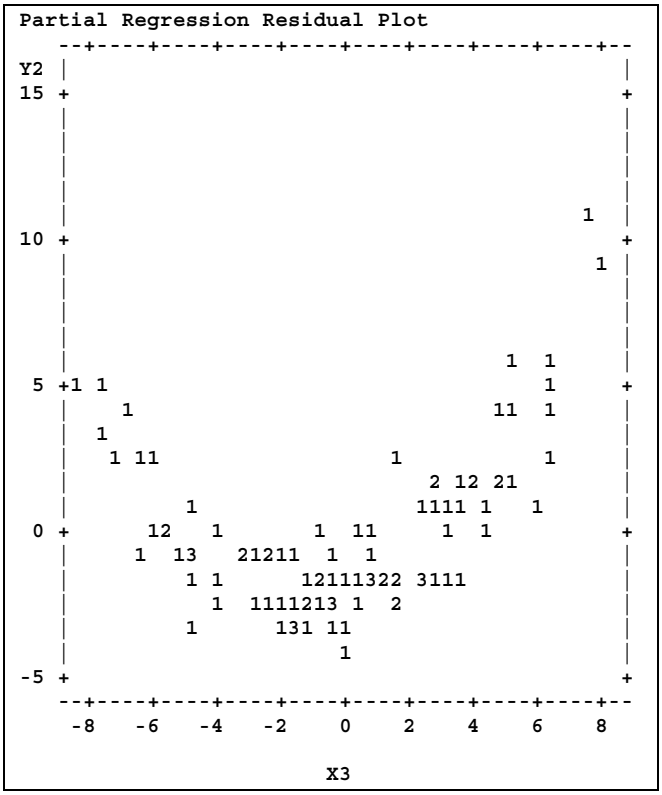
Notice that the Residual Plot following implies some transformation on the dependent variable. There is a fan shape, but it would be a mistake to jump to this conclusion and try a log or a square root transformation. In general if you see a fan shape in a residual plot you should look at the partial regression residual plots before proceeding to transform the dependent variable. (Plot has been slightly condensed to enable it to be shown here.)



Before looking at the partial regression plot we can look at the XY plot below. Notice that there is no indication of any relationship besides a weak linear relationship. We still have to keep looking.



The partial regression plot below tells another story. You don't have to search to find the answer, because a glance reveals the quadratic relationship between X3 and Y2. The result is a nicer parabola than I could ever draft in high school.



Notice, a PROC UNIVARIATE on the residuals from this analysis produced the stem and leaf, and box plots below. The test of normality from this procedure was significant (p<.0001), and this would lead many researchers to conclude that the dependent variable might need a transformation. But we can see from the partial plot above that there is another relationship in the data involving only one of the predictor variables.

```
Variable=RESID          Residual

 Stem Leaf                            #  Boxplot
   11 8                               1     *
   10
    9
    8 2                               1     0
    7 14                              2     0
    6 57                              2     0
    5 2                               1     0
    4 12379                           5     |
    3 00                              2     |
    2 1123                            4     |
    1 0679                            4     |
    0 1111223445555558889            19  +--+--+
   -0 7543332222100                   13  *-----*
   -1 9877766554332110                16  |     |
   -2 988776665544433321110000        24  +-----+
   -3 22210                           5     |
   -4 2                               1     |
      ----+----+----+----+----
```

Notice the skew in the plot above; the original dependent variable would also appear to have a skewed distribution. I present this information as a warning against assuming that a certain transformation, especially of a dependent variable needs to be made without checking partial plots as well as residual plots.

ADDING A QUADRATIC TERM

The next step is to test the model including the quadratic term. A quadratic term is simply the square of one of your predictor variables. It is also possible to center the predictor before squaring it. Centering consists of subtracting the mean of the variable from its value for each case.

The following program may be used to create both centered and uncentered quadratic terms. PROC MEANS may be used to determine the mean of the variable for centering. X3SQ is the centered quadratic term for the variable X3, and X3SQO is the uncentered version. I have created both types of quadratic terms so that we may examine the differences. The value **45.624102** is the mean of variable X3 and is used for centering.

```
DATA fixed;
  SET in.hetero ;
   x3sq =(x3-45.624102)**2; /*centered*/
   x3sqo=(x3)**2;
RUN;

PROC REG DATA=fixed;
 MODEL y2= x1 x2 x3 x3sq / PARTIAL STB;
 PLOT R.*P.;
 OUTPUT OUT=resid P=pred R=resid;
RUN;
```

There is an additional option used on the model statement in this procedure, STB. It is used to obtain a standardized regression coefficient, which is useful in assessing the relative contribution of a predictor.

The results of the regression with the centered quadratic term are presented following. Notice that the proportion of variance accounted for is almost 96%, which is a considerable improvement over the original of 72% with only linear terms.

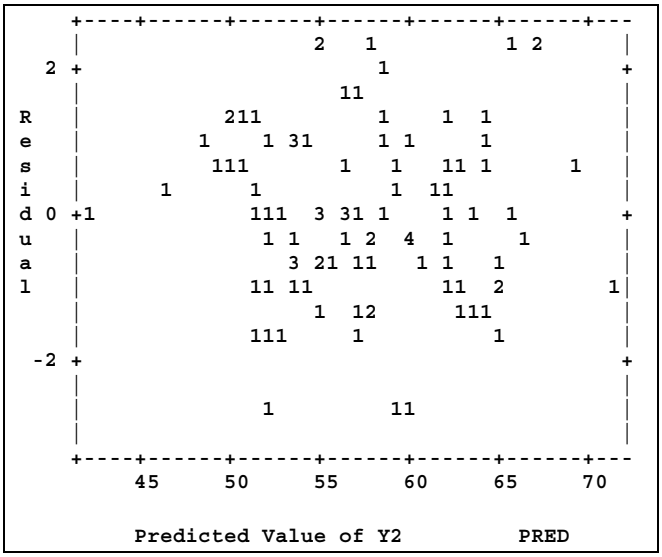
```
Analysis of Variance

              Sum of    Mean
Source   DF   Square    Square F Val  P>F

Model     4  2951.28  737.82 545.3 0.0001
Error    95   128.53    1.35
C Total  99  3079.81

Root MSE  1.16317     R-square  0.9583
Dep Mean 57.49389     Adj R-sq  0.9565
C.V.      2.02312


Parameter Estimates

          Param  Stand  T for H0:          Stand
Var   DF  Estim  Error  Param=0   P>|T|    Estim

INTER 1 -6.6366  1.746  -3.801   0.0003   0.000
X1    1  0.4345  0.011  39.350   0.0001   0.837
X2    1  0.0983  0.013   7.335   0.0001   0.154
X3    1  0.2945  0.029  10.117   0.0001   0.212
X3SQ  1  0.1370  0.006  22.591   0.0001   0.491
```

It might be useful to examine the parameter estimates from models with and without the quadratic term. These are presented following.

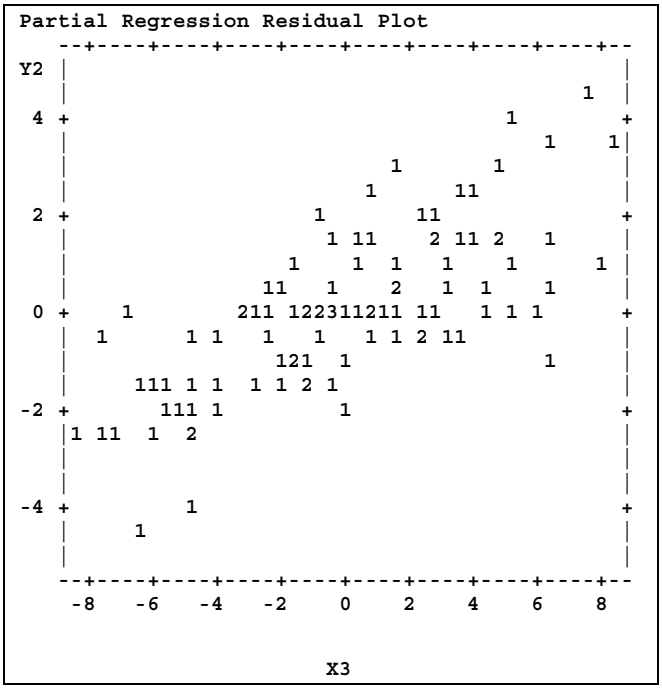
With Quad			Without Quad		
Var	DF	Param Estim	Var	DF	Param Estim
INTER	1	-6.637	INTER	1	-1.581
X1	1	0.434	X1	1	0.405
X2	1	0.098	X2	1	0.096
X3	1	0.295	X3	1	0.297
X3SQ	1	0.1370			

Notice that the parameter estimates for the linear terms are nearly identical in both models. The additional quadratic term is essentially orthogonal to the original model if it is centered.

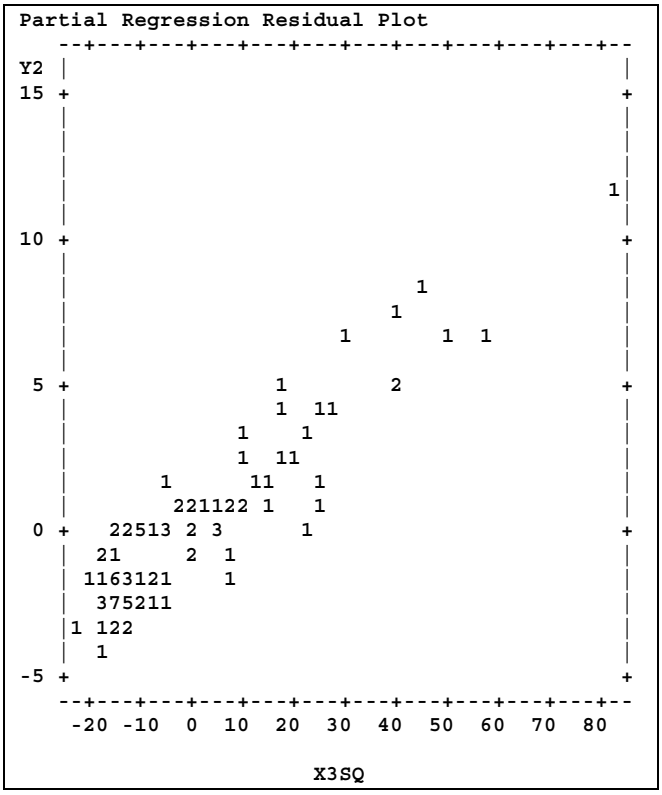
The resulting residual plot is no longer fan shaped.



The following partial regression residual plot shows only the linear relationship remaining between Y2 and X3, as the quadratic effect has been removed by the inclusion of X3SQ in the model.



The partial regression residual plot for Y2 with all of the other variables partialed out, and X3SQ with all of the other variables partialed out follows. Notice that this relationship is linear, that is what is meant by a linearizable curvilinear function. The skew of the independent variable is reflected in the skew of the dependent variable, and the resulting residual is no longer skewed.



The PROC UNIVARITATE test for normality run on the residuals of this model (including the centered quadratic for X3) yields a W statistic of .973 with a p>.o5, and thus there is no significant deviation of the distribution of the residuals from normal.

The stem and leaf and box plots show much less skew, and is more nearly normally distributed.

Stem	Leaf	#	Boxplo
2	222234	6	
1	55689	5	
1	00001123344	11	
0	5556667788899	13	+-----+
0	1111123334	10	| + |
-0	44333222222221111000	20	*-----*
-0	999988877766555	15	+-----+
-1	43332221100	11	
-1	866655	6	
-2			
-2	877	3	

This emphasizes the danger of transforming a dependent variable simply due to the shape of the distribution of that variable. A natural skew in a predictor may yield a skewed distribution in the criterion variable, but the residuals may be normal.

THE VALUE OF CENTERING

In order to fully appreciate the value of centering we need to run a model using the uncentered quadratic term for X3, X3SQO. The following program segment uses the simple square of X3.

```
PROC REG DATA=fixed;
 MODEL y2= x1 x2 x3 x3sqo
          / PARTIAL STB TOL COLLIN VIF;
 OUTPUT OUT=resid P=pred R=resid ;
RUN;
```

We also include options to furnish statistics for tolerance, collinearity, and variance inflation. This is done to demonstrate the effect of including an uncentered quadratic term in the model.

The overall model Analysis of Variance table is presented for comparison. Notice that the R-square is identical for both models. The overall regression model statistics are not affected by use of either form of quadratic term.

```
Centered Quadratic

Root MSE  1.16317  R-square  0.9583
Dep Mean 57.49389  Adj R-sq  0.9565
C.V.      2.02312



Un-Centered Quadratic

Root MSE  1.16317  R-square  0.9583
Dep Mean 57.49389  Adj R-sq  0.9565
C.V.      2.02312
```

The individual parameter estimates are a different matter. The parameter for the other variables (X1 and X2) is unchanged. And the estimate for the quadratic term is the same whether centering is used or not. But the parameter for the linear X3 is vastly different, and the sign is changed (.2945 becomes –12.2045).

Centered			Un-Centered		
Var	DF	Param Estim	Var	DF	Param Estim
INTER	1	-6.6366	INTER	1	278.5488
X1	1	0.4345	X1	1	0.4345
X2	1	0.0983	X2	1	0.0983
X3	1	0.2945	X3	1	-12.2045
X3SQ	1	0.1370	X3SQO	1	0.1370

Take a look at the tolerance and variance inflation factors presented below.

```
Parameter Estimates

          Param   Stand           Varian
Var   DF  Estim   Estim   Toler   Inflat

INTER  1 278.54   0.000    .      0.0000
X1     1   0.43   0.837   0.970   1.0302
X2     1   0.09   0.154   0.990   1.0093
X3     1 -12.20  -8.826   0.002 365.6366
X3SQO  1   0.13   9.048   0.002 365.2358
```

The tolerance has become very small and the VIF has become extremely large. To assess the variance inflation factor for each variable you can use the R^2 for the regression. The formula follows:

Comparison for VIF = 1/(1-R^2)

Using the R^2 of .9583 yields a comparison value of 23.98, and comparing this to the VIF statistics for each of the variables shows that at over 365 both X3 and X3SQO have problems with collinearity.

In the collinearity diagnostics below notice the condition index of 594.17. Generally condition indexes of greater than 30 is indicative of some collinearity, and indexes greater than 1000 indicate severe collinearity. Thus collinearity could be a problem when the uncentered quadratic term is used.

```
Collinearity Diagnostics

                      Var   Var   Var   Var   Var
           Condit     Prop  Prop  Prop  Prop  Prop
#  Eigenv  Index      INT   X1    X2    X3    X3SQO

1 4.94455    1.00    0.00  0.00  0.00  0.00  0.00
2 0.03250   12.33    0.00  0.00  0.51  0.00  0.00
3 0.01760   16.76    0.00  0.27  0.41  0.00  0.00
4 0.00534   30.43    0.00  0.71  0.06  0.00  0.00
5 0.000014 594.17    0.99  0.01  0.00  0.99  0.99
```

This problem with collinearity means that X3SQO is highly linearly related to X3. This can be demonstrated by the simple correlation between X3 and X3SQO, and it can also be demonstrated that X3SQ, the centered quadratic, is unrelated or orthogonal to X3.

```
PROC CORR DATA=fixed;
   VAR x3 x3sq x3sqo;
RUN;
```

The correlation matrix follows:

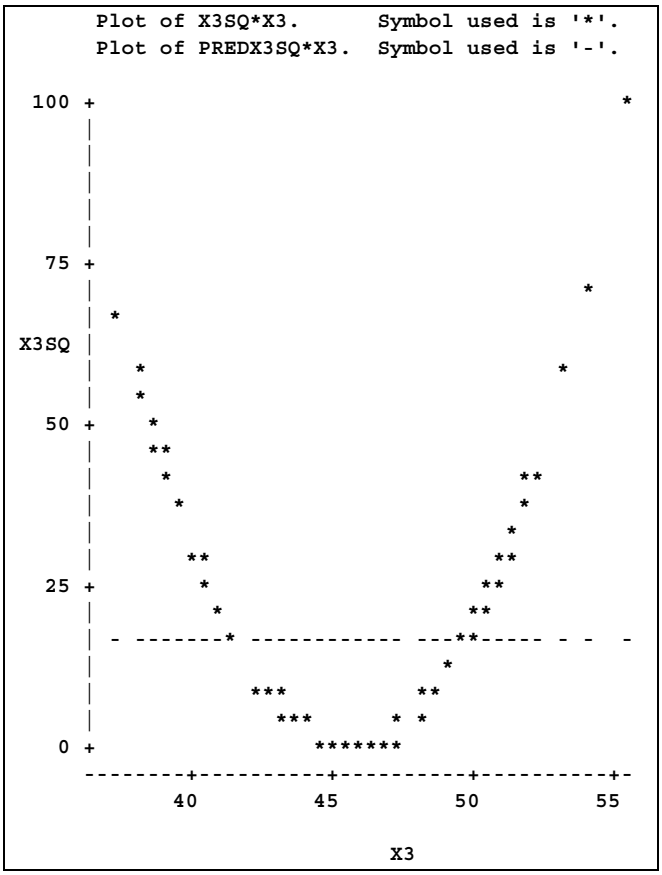
```
Pearson Corr Coeff
/ Prob > |R| under Ho: Rho=0
/ N = 100

         X3      X3SQ     X3SQO

X3     1.0000 -0.0058   0.9986
       0.0      0.9538   0.0001

X3SQ -0.0058   1.0000   0.04682
       0.9538    0.037   0.6437

X3SQO 0.9986   0.0468   1.0000
       0.0001    0.643   0.0
```

The correlation between X3SQ and X3 is nearly zero, but the correlation between X3SQO and X3 is nearly 1. Centering makes the quadratic term orthogonal to the linear. This orthogonal relationship is sometimes hard to grasp because there is clearly a relationship, it is just that the relationship is not linear.

A simple linear regression between X3 and X3SQ can demonstrate this by saving the predicted value and plotting the predicted value and the observed.

```
PROC REG  DATA=in.random;
   MODEL x3sq=x3;
   OUTPUT OUT=rand2 p=predx3sq;

PROC PLOT DATA=rand2;
   PLOT x3sq*x3='*' predx3sq*x3='-'/overlay;
RUN;
```

This produces the following plot.



Notice that the predicted value of X3SQ from X3 is a horizontal line. Of course partial plots would reveal the quadratic relationship.

CONCLUSION

Examining residuals is useful in detecting curvilinear relationships that can be linearized by a transformation on the dependent variable. But it is possible to be misled when there are quadratic or higher effects involving one or more of the predictors.

Beware also of transforming variables simply after examining the individual variables for lack of normality including skew. A dependent variable can show a skewed distribution if one of the predictor variables has a quadratic relationship with it.

The PARTIAL option on the MODEL statement in PROC REG is useful for detecting these effects even when they are hidden from examination of the simple XY plots. This is true because the effects of other variables in the model are partialed out of both the dependent variable and the independent variable in the plot.

Centering is useful when adding quadratic terms so that the quadratic and linear relationships can be examined in the same model. It also ensures against collinearity in your model.

REFERENCES

Freund, RJ and Littell, RC, *SAS® System for Regression, Second Edition,* Cary, NC: SAS Institute Inc., 1991. 329 pp.

Chatterjee, S and Price, B, *Regression Analysis by Example,* New York, NY: John Wiley & Sons, 1977. 228 pp.

SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2,* Cary, NC: SAS Institute Inc., 1989. 1351-1194.

ACKNOWLEDGMENTS

I would like to thank Sun Hwang for useful discussions regarding the subject matter and for reading a draft of this paper. I would like to thank Barbara Widawski, without whose editing this manuscript would be illegible.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mel Widawski
Principal Statistician
NPIStat
UCLA
Los Angeles, CA

Please contact through EMAIL at: mel@ucla.edu