Paper 1760-2014

# Scenarios Where Utilizing a Spline Model in Developing a Regression Model Is Appropriate

Ning Huang, University of Southern California

## ABSTRACT

Linear regression has been a widely used approach in social and medical sciences to model the association between a continuous outcome and the explanatory variables. Assessing the model assumptions, such as linearity, normality, and equal variance, is a critical step for choosing the best regression model. If any of the assumptions are violated, one can apply different strategies to improve the regression model, such as performing transformation of the variables or using a spline model. SAS® has been commonly used to assess and validate the postulated model and SAS® 9.3 provides many new features that increase the efficiency and flexibility in developing and analyzing the regression model, such as ODS STATISTICAL GRAPHICS. This paper aims to demonstrate necessary steps to find the best linear regression model in SAS 9.3® in different scenarios where variable transformation and the implementation of a spline model are both applicable. A simulated data set is used to demonstrate the model developing steps. Moreover, the critical parameters to consider when evaluating the model performance are also discussed to achieve accuracy and efficiency.

## INTRODUCTION

In this paper, statistical tests and terms in linear regression will be reviewed, but former knowledge and experience on applied regression analysis is recommended.

### LINEAR REGRESSION OVERVIEW

Linear regression can model the relationship between the continuous dependent variable and independent variable by fitting the observed data into a linear equation. The dependent variable is what we usually refer to as outcome, or Y. The independent variable is referred to as predictor, explanatory variable, or X, which we suspect contributes to the outcome. Epidemiological investigations employ linear regression to understand the association of continuous outcomes with its explanatory variables, such as “is higher blood pressure associated with larger body weight”.

Building a linear regression model is to find the curve (straight line, parabola, etc.) that best fits the data, closely approximating the true (but unknown) relationship between X and Y (Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Keith E. Muller, 2008). Simple linear regression take only one explanatory variable into the model, while multiple regression models deal with more than one X, where adjusted effect estimates can be obtained by taking the effect of potential confounders into account.

The general form of linear regression is $Y = \alpha + \beta_1 * X_1 (+ \beta 2 * X_2 \ldots + \beta\kappa * X_k) + E$, where $E$ denotes residual that is the distance of each observed Y from the corresponding mean value of estimate Y for the given X. We can take it as the error that cannot be explained by the model. Regression coefficients are $\alpha$ as the intercept and $\beta$ as the slope, respectively.

### MODEL ASSUMPTIONS

The diagnostics of linear regression focuses on the error, which is the residual that cannot be explained by the model. In order to get an accurate model, 4 major assumptions need to be satisfied: (1) Linearity, which requires residuals scatter evenly around zero so that the mean value of predicted Y for each specific value of X is a linear function of $\beta$(s); (2) Independence, which requires Y conditional on X are statistically independent. (3) Normality, which requires the residuals distribute normally on any fixed value of X; (4) Homoscedasticity, which requires the residuals have constant variance for any fixed value of X. If any of these are violated, the estimate yielded by the regression model may be biased or misleading.

### ALLOWING MORE MODEL FLEXIBILITY

Having more flexibility in the model can solve the problem of unsatisfied assumptions. It can be improved by applying appropriate transformation to the variables, using polynomial independent variable model or applying linear spline regression model. We may also categorize the independent variables, consider adding additional predictors that are related to the existing predictors and outcome, or perform data cleaning to exclude outliers.

In epidemiology and molecular biology, the residuals of linear relationships between the outcome and biological parameter do not normally distribute due to large outliers. A nonlinear transformation, such as log-transformation of the dependent or independent variable is commonly used to solve this problem. For example, insulin and triglycerides serum concentration was log-transformed because of asymmetrical distribution in a study of the associations of maternal pre-pregnancy body mass Index and gestational weight gain with adult offspring cardiometabolic risk factors (Hochner et al., 2012).

Piecewise linear regression splines model is also a commonly-used tool in biomedical investigations which allow more flexibility with knots joining several lines or curves together. The knots are introduced into the model when there are certain “cutoff” values in the predictor. For example, Wang et al. (2013) used piecewise splines with 5 knots set to fit sharply the nonlinear shapes and found the substantial differences between the association of U.S. background mortality rates with BMI and age.

### ALWAYS PLOT THE DATA FIRST

Selecting a good model allows us to summarize the association and predict precisely based on the model. We usually begin modeling with plotting the data. By examining these initial plots, we can quickly assess whether the data have linear relationships or transformation is needed. A plot that resembles a straight line (Fig.1A) indicates the existence of linear association. When we see an imperfect straight line with a bent (Fig.1B), certain kind of transformation approach is needed in the linear regression model, which in this case can be logarithm transform of the predictor. If the data appear with multiple kinks (Fig.1C), linear spline model may be necessary to gain a better model fit.

In more cases, the approach to be taken may not seem obvious judging only from the data appearance. In this paper, we will see a data structure in this kind and the appropriate steps to better fit the linear model. More specific and reliable criteria are proposed in this paper to help measure the model performance. Selecting important statistics among the many types of linear regression helps compare models with accuracy and efficiency.

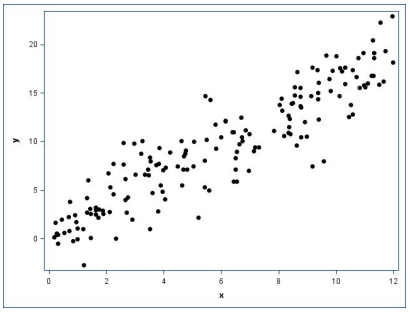


**Figure 1A**

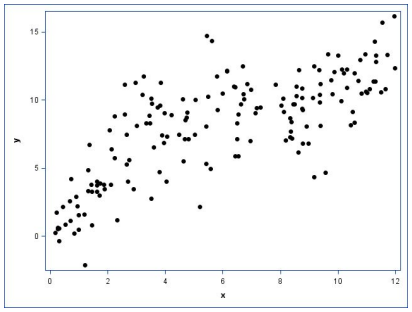


**Figure 2B**

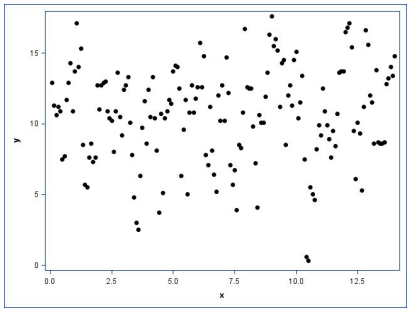


**Figure 3C**

## DATA SET

A simulated data set based on generic data is used to illustrate the model building. There are 500 data points in this data set. We use “Y” as the dependent variable and “X” as the independent variable. The X-Y plot can be generated with the SGPLOT Procedure (Fig. 2) with the following program:

```
libname SAS2014 "C:\SAS2014\codes and output";
%let N = 500;                                   /* size of the sample data*/
data SAS2014.lrdata(keep=x y);
call streaminit(387915);
do i = 1 to &N;
      X=10+8*rand("Uniform");                   /* explanatory variable     */
      eps = 750*rand("Normal", 0, 0.5);         /* error term               */
      if x<13 then Y=-1005.95+303.08*x +eps;
      else if x>=13 then y=-1005.95+303.08*x-202.63*(x-13)+eps;
      output;
end;
run;
PROC sgplot data=SAS2014.lrdata;
    scatter y=y x=x;
run;
```

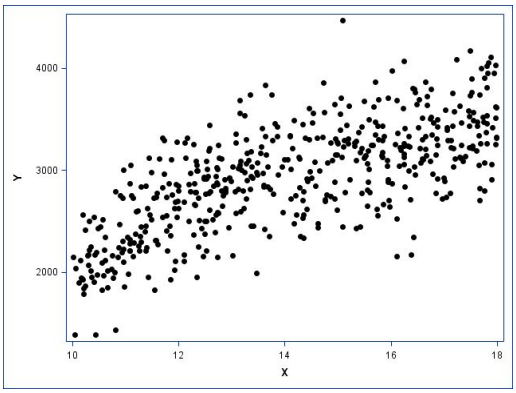



**Figure 2. X-Y Scatter Plot of the Simulated Data Set**

The Y can be modeled as a function of X based on simple linear regression (the SLR model) $Y = \alpha + \beta * X + E$. However, the curvature exists that from lower left to upper right so that data points with X of either very low or very high value lie below the straight line suggested by the data, while the data points with middling X values lie on or above that straight line. In this case, taking logarithms of the X values (the Log X model) or using a linear spline model may both promote linearity. The performance of these three models will be discussed and compared.

The REG, UNIVARIATE, and SGPLOT procedures and ODS STATISTICAL GRAPHICS in SAS® 9.3 are used to demonstrate the basic steps in generating linear regression model, assessing assumptions and fitness.

## PERFORMING SIMPLE LINEAR REGRESSION ANALYSIS

Based on the X-Y plot from Figure 2, a simple linear regression may exist between Y and X. Performing the SLR model first is the simplest way to better understand their relationship. The REG Procedure is a general-purpose procedure for regression and the following program can be used:

```
ODS rtf file = "SLR Model";
PROC REG DATA=SAS2014.lrdata;
    MODEL y=x/clb r;
/*CLB requests the 100(1-α)% upper and lower confidence limits for the parameter
estimates and r requests an analysis of the residuals*/
    OUTPUT out=lrreg p=pred r=resid;
/*Output the regression results into lrreg data set which stores the predicted
value of Y as pred and residual of the model as resid;*/
RUN;
ODS rtf close;
```

```
                              Analysis of Variance

                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F

Model                       1       68520922       68520922     497.20    <.0001
Error                     498       68630552         137812
Corrected Total           499      137151475

              Root MSE            371.23086    R-Square     0.4996
              Dependent Mean     2926.04849    Adj R-Sq     0.4986
              Coeff Var            12.68711

                              Parameter Estimates

                                      Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|       95%
Confidence Limits

Intercept     1      684.79195      101.87528       6.72      <.0001      484.63363
884.95028
X             1      159.26196        7.14241      22.30      <.0001      145.22899
173.29492
-------------------------------- OMITTED RESULTS --------------------------------
                    Sum of Residuals                           0
                    Sum of Squared Residuals            68630552
```

```
                    Predicted Residual SS (PRESS)       69173725
```

**Output 1. Output from The REG procedure: The SLR model.**

The linear regression function can be written as Y = 684.79195 + 159.26196 * X. The regression coefficients are all statistically significant (P<0.0001).

## PERFORMING SIMPLE LINEAR REGRESSION WITH LOGARITHM OF THE PREDICTOR

We can use a DATA step to generate the natural logarithm of X (lnX variable):

```
DATA SAS2014.lrdata2;
set SAS2014.lrdata;
    lnx=log(x);
RUN;
```

The simple regression model of Y and lnX can be generated with the REG Procedure:

```
ODS rtf file = "log x model";
PROC REG DATA=SAS2014.lrdata2;
MODEL y=lnx/clb r;
OUTPUT out=lrreg2 p=pred r=resid;
RUN;
ODS rtf close;
```

```
                              Analysis of Variance

                                   Sum of          Mean
Source                     DF        Squares         Square    F Value    Pr > F

Model                       1       70713229       70713229     530.04    <.0001
Error                     498       66438246         133410
Corrected Total           499      137151475

                    Root MSE            365.25352    R-Square     0.5156
                    Dependent Mean     2926.04849    Adj R-Sq     0.5146
                    Coeff Var            12.48283

                              Parameter Estimates

                      Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|       95%
Confidence Limits

Intercept     1    -2963.09494      256.31846     -11.56      <.0001    -3466.69381
-2459.49607
lnx           1     2238.95800       97.25009      23.02      <.0001     2047.88695
2430.02905
-------------------------------- OMITTED RESULTS ---------------------------------
                    Sum of Residuals                           0
                    Sum of Squared Residuals            66438246
                    Predicted Residual SS (PRESS)       66957465
```

**Output 2. Output from The REG procedure: The Log X model.**

The linear regression function can be written as Y=-2963.09494+2238.95800*ln(X).The regression coefficients are all statistically significant (P<0.0001).

## PERFORMING X SPLINE IN LINEAR REGRESSION

By observing the original X-Y plot, we can see that the relationship between Y and the X appears to change at approximately X of 13. We can use the following code to allow the regression line with a curvature at X of 13:

```
DATA SAS2014.lrdata3;
set SAS2014.lrdata;
    if x > 13 then xspl = x-13;
    else xspl = 0;
RUN;
```

The linear spline regression model can be generated with the REG Procedure:

```
ODS rtf file = "spline model";
```

```
PROC REG DATA=SAS2014.lrdata3;
MODEL y=x xspl/clb r;
OUTPUT out=lrreg3 p=pred r=resid;
RUN;
ODS rtf close;
```

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 73689912 | 36844956 | 288.55 | <.0001 |
| Error | 497 | 63461563 | 127689 | | |
| Corrected Total | 499 | 137151475 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 357.33634 | R-Square | 0.5373 |
| Dependent Mean | 2926.04849 | Adj R-Sq | 0.5354 |
| Coeff Var | 12.21225 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | -946.27118 | 274.47240 | -3.45 | 0.0006 | -1485.54045 | -407.00191 |
| X | 1 | 297.52955 | 22.79333 | 13.05 | <.0001 | 252.74639 | 342.31271 |
| xspl | 1 | -196.08685 | 30.81930 | -6.36 | <.0001 | -256.63903 | -135.53467 |

OMITTED RESULTS

| | |
|---|---|
| Sum of Residuals | 0 |
| Sum of Squared Residuals | 63461563 |
| Predicted Residual SS (PRESS) | 64175538 |

**Output 3. Output from The REG procedure: The spline model.**

The linear regression function can be written as Y = -946.27118 + 297.52955 * X - 196.08685 * $(X-13)^+$, where the spline term $(X-13)^+$ means only taken when X-13 is positive. The regression coefficients are all statistically significant (P<0.05).

## LINEAR REGRESSION MODEL PERFORMANCE

| | SLR MODEL | LOG X MODEL | SPLINE MODEL |
|---|---|---|---|
| **$R^2$** | 0.4996 | 0.5156 | **0.5373** |
| **MSE** | 137812 | 133410 | **127689** |
| **Press** | 69173725 | 66957465 | **64175538** |

**Table 1. Summary of Linear Regression Models Results (The Best Performance in Each Category is Shown in Bold)**

Coefficient of determination, or $R^2$ (ranges from 0~1) provides a measure of the percentage that the true association explained by the model. When $R^2$ is low, the cause can be either the linear relationship of the outcome to the predictors is poor, the model does not fit well or usually a combination of both. Based on the $R^2$ value (Table 1), the Log X model and the spline model both improve the strength of the linear relationship while the spline model fits the best.

The mean squared prediction error (MSE) can be used to measure the average of the squares of the residuals. MSE=0 indicates a perfect positive or negative straight-line relationship exists. It is a very good parameter to reflect the strength of models and we want to choose the model with smaller MSE. In this case (Table 1), MSEs of the Log X model and the spline model are both smaller than the one of the SLR model, while the spline model has the smallest.

The predicted residual sums of squares (PRESS) statistic provides a summary measure of the difference between the observed value and the predicted value, when the model was fit without that point. The model with the lowest

values of PRESS indicates the best structures. In this case (Table 1), the Log X model and the spline model both have smaller PRESS than the SLR model, while the spline model has the smallest.

## ASSESSING MODEL ASSUMPTIONS

$R^2$ is not a measure of the appropriateness of the straight-line model. Thus, the model assumptions must be met to assure the model estimate linear relationship well. Also, the random error component E has a normal distribution with mean 0 and variance $\sigma^2$ (Kleinbaum, Kupper, Nizam, Muller, 2008). During the model generating process with PROC REG we compute the residuals and output to corresponding data sets for further analysis. The assumptions can be evaluated with the following SAS code (only presenting the part of the SLR model as an example):

```
ODS rtf file = "SLR Model";
* -- Assumption 1: Does the linearity assumption hold?;
PROC SGPLOT data=lrreg;
    scatter x=pred y=resid;          *usual scatterplot of resid vs. pred;
    loess x=pred y=resid ;           *loess smooth of this rel'n;
    refline 0;                       *add a reference line at y=0;
    title 'Evaluate linearity assumption with Loess smooth';
run;
* -- Assumption 2: Are residuals normally distributed?;
PROC UNIVARIATE plot normal data=lrreg;
    var resid;
    title 'Evaluate the distribution of residuals for normality';
run;
* -- Assumption 3: Does the homoscedasticity assumption hold?;
PROC GPLOT data=lrreg;
    symbol1 v=star;
    plot resid*pred;                 * -- plot residuals predicted value;
    title 'Evaluate homoscedasticity assumption';
run;
ODS rtf close;
```

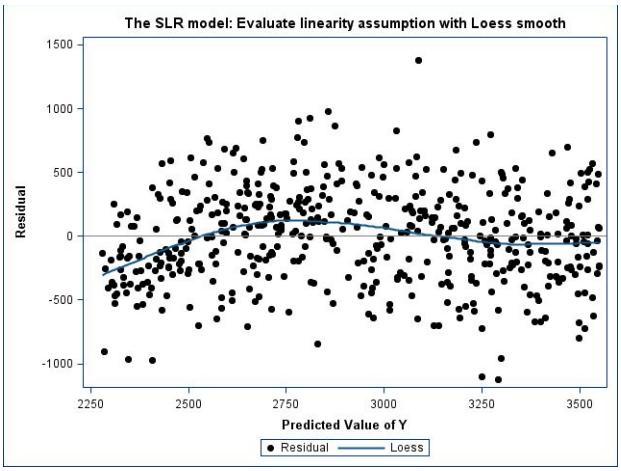
Output from The UNIVARIATE procedure is summarized into the following table for normality assumption analysis:

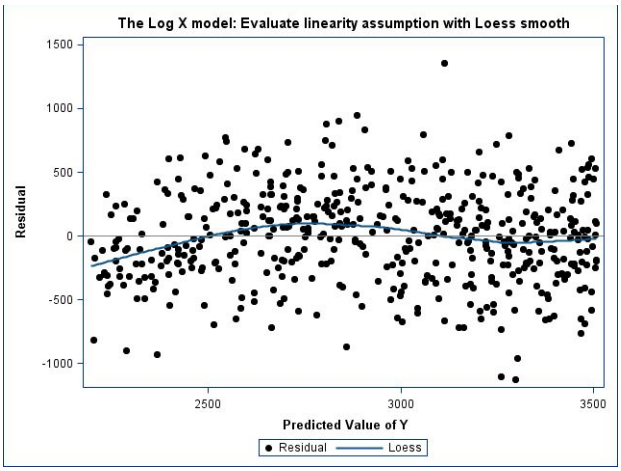
| | SLR MODEL | LOG X MODEL | SPLINE MODEL |
|---|---|---|---|
| **Residual Mean** | **0** | **0** | **0** |
| **Residual Median** | 3.350739 | **1.654937** | 2.416008 |
| **Skewness** | 0.0094062 | **-0.0032534** | -0.0269401 |
| **Kurtosis** | **0.10016755** | 0.13794118 | 0.20453514 |
| **Shapiro-Wilk Test for Normality** | 0.9457 | **0.9709** | 0.3235 |
| **Kolmogorov-Smirnov Test for Normality** | **>0.1500** | **>0.1500** | **>0.1500** |

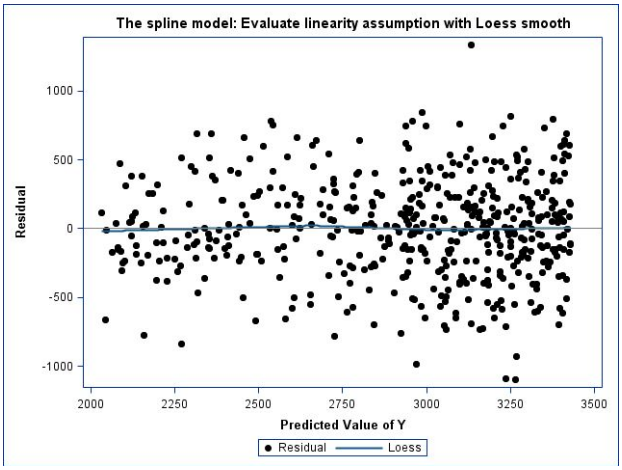
**Table 2. Normality Assumption Analysis with UNIVARIATE Procedure Output (The Best Performance in Each Category is Shown in Bold)**

All three models show satisfaction of normality based on the tests. Mind that sometimes the tests can be too stringent to apply on real-life data and models can be compared based on other statistics: The skewness and kurtosis show the level and orientation of skewness; Mean should be approximately equal to median in a normal distribution; The histogram of residuals should be bell-shaped and the probability plot is a straight line; In this data set, all three models has quite well bell-shaped histogram and straight line in probability plot (data not shown) with similarly slight skewness (Table 2). The SLR and Log X model satisfies normality assumption slightly better than the spline model.

The locally weighted scatter plot smoothing (LOESS) helps visualizing pattern in the relationship between the residuals and the predicted values of Y (Output 4). The SLR model has an obvious pattern of residuals and they don't evenly scatter around Y=0. The Log X model reduces the pattern while the spline model almost eliminates it.



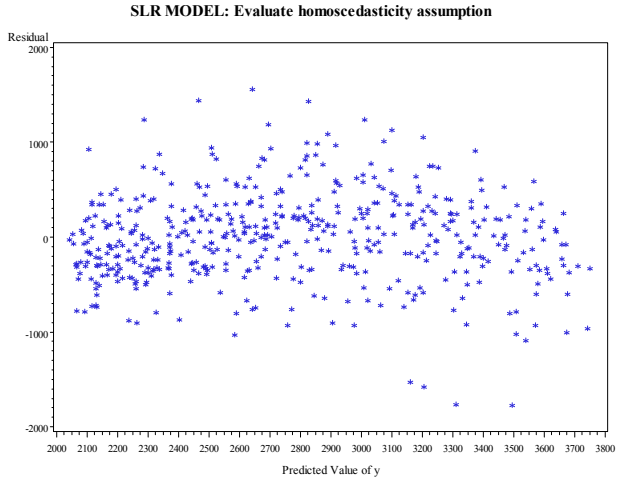



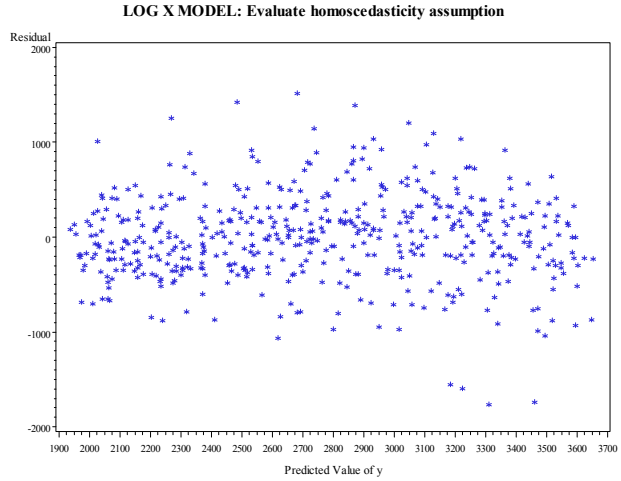



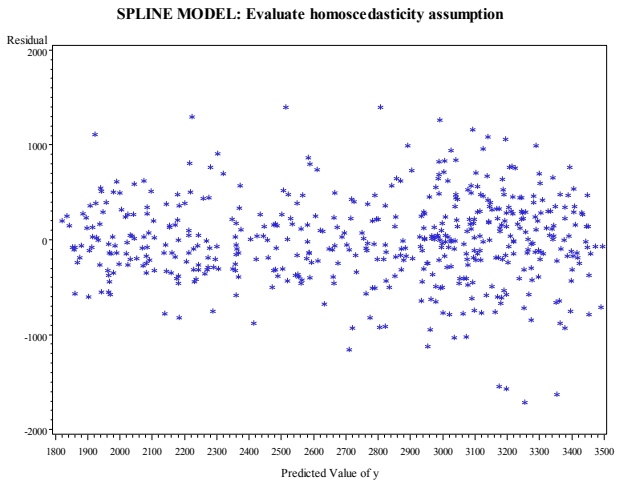

**Output 4. Output from The SGPLOT procedure for linearity assumption analysis**

To satisfy the homoscedasticity assumption, the residuals should have common variance across the range of X values (Output 5). The implementation of log X model and spline model improve the residuals' equal variance requirement at a certain level.

**Output 5. Output from GPLOT procedure for homoscedasticity assumption analysis**






## INTERPRETING THE LINEAR REGRESSION

In epidemiology the linear regression model needs to be interpreted in a logical way in order to understand the association between the outcome and the predictors. The intercept is the estimated mean value of Y at X=0. The Slope indicates the estimated change in mean Y per increase of 1 unit in X.

In the SLR model, the mean Y is 684.79195 (95% CI = (484.63363, 884.95028)) at X=0 while the estimated mean Y increases by 159.26196 (95% CI = (145.22899, 173.29492)) per 1 unit increase of X. In the Log X model, the mean Y is -2963.09494 (95% CI = (-3466.69381, -2459.49607)) at ln(X) = 0 (original X = 1) while estimated mean Y increases by 2238.95800 (95% CI = (2047.88695, 2430.02905)) per 1 unit increase of ln(X) (original X increases by *e*).

The interpretation at log scale is too obscure for epidemiological application thus becoming a disadvantage of logarithm transformation model. The increase of one natural logarithm of X is difficult to translate to every 1 or standard deviation of X since the relationship is not linear. Moreover, the change of outcome may differ across the range of X due to biological facts, which is very common in epidemiology and appearing at a piecewise shape. This explains why the SLR model has the lowest model fit. The spline model provides more clearly-stated interpretation by localizing the influence of each data point to its particular segment and is more reasonable when interpreting. In this data set, when X<13, the estimated mean Y increases by 297.52955 (95% CI = (252.74639, 342.31271)) per 1 unit increase of X. When X>13, the estimated mean Y increases by 101.4427 (= 297.52955-196.08685) per 1 unit increase of X. The change of slope at X=13 (slope of the xspl variable) is statistically significant (P<0.0001).

## CONCLUSION

There exists no best model, but the best-fit model. Certain parameters can be considered when determining the model performance. The results indicate that while the simple linear regression fails to satisfy the model assumptions and has a poor fit, the logarithm transform of X and the linear spline model can both improve model performance by better satisfying the normality and homoscedasticity assumption, increasing the $R^2$, decreasing the MSE and PRESS statistic. In addition, the linear spline model has advantage in interpretation of the association, which makes it more appropriate than the other two models when applying to epidemiological data set.

The study is limited in that the data structure is based on one simulated data derived from generic epidemiological research data and that only one independent variable is examined in the model. It is not impossible, however, to conceptualize this data structure as a representative and perform more complicated regression analysis with the general results found in this paper.

## REFERENCES

Hochner H, Friedlander Y, Calderon-Margalit R, Meiner V, Sagy Y, Avgil-Tsadok M, Burger A, Savitsky B, Siscovick DS, Manor O., Associations of maternal prepregnancy body mass index and gestational weight gain with adult offspring cardiometabolic risk factors: the Jerusalem Perinatal Family Follow-up Study. Circulation, 2012;125:1381–1389

Kleinbaum, David G., Kupper, Lawrence L., Nizam, Azhar, Muller, Keith E., 2008, Applied Regression Analysis and Other Multivariable Methods, Thomson

Li, Arthur, 2013, Handbook of SAS DATA Step Programming, CRC press

PROPHET StatGuide, Possible alternatives if your data violate regression assumptions, Available at http://www.basic.northwestern.edu/statguidefiles/linreg_alts.html

Simpson, Pippa, Hamer, Robert, Jo, ChanHee, Huang, B. Emma, Goel, Rajiv, Siegel, Eric, Dennis, Richard, Bogle, Margaret, Assessing Model Fit and Finding a Fit Model, Proceedings of the SUGI 29

Wang, Y. Claire, Graubard, Barry I., Rosenberg, Marjorie A., Kuntz, Karen M., Zauber, Ann G., Kahle, Lisa, Schechter, Clyde B, Feuer, Eric J., Med Decis Making 2013;33: 176-197

## ACKNOWLEDGEMENT

I am thankful to Arthur Li and Kechen Zhao, for taking time out of their busy schedule to read my paper and provide guidance and suggestion. I also would like to thank Dr. Sandy Eckel for educating me on linear regression.

## CONTACT INFORMATION

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

Ning Huang

University of Southern California

E-mail: nienghuang@gmail.com

www.linkedin.com/pub/ning-huang/26/356/403/