

How To: Deal with Heteroscedasticity

Using STATGRAPHICS Centurion

by

Dr. Neil W. Polhemus

July 28, 2005

Introduction

When fitting statistical models, it is usually assumed that the error variance is the same for all cases. Whether performing an analysis of variance or fitting a regression model, the assumption of constant variance is important to insure that:

- (1) model parameters are estimated efficiently.
- (2) prediction limits and comparisons between different groups of cases are calculated properly.

The situation in which the error variance is different for different cases is referred to as *heteroscedasticity*.

In real life, it is not uncommon for the error variance to increase as the mean response increases. If the response follows a Poisson distribution, the variance will be proportional to the mean. In STATGRAPHICS Centurion, procedures such as *Poisson Regression* deal with non-constant variance as a matter of course. For other types of data, the manner in which the variance changes may not be known ahead of time and must therefore be determined from the data.

This case study discusses methods for identifying and dealing with heteroscedasticity. We will consider the use of both variance stabilizing transformations and weighted least squares.

Sample Data

Our example is from the excellent book by Neter, Kutner, Wasserman, and Nachtsheim titled *Applied Linear Statistical Models*, 4th edition (Irwin, 1996). They report on a study of 54 women aged 20 to 60 years old, where the goal was to determine the relationship between diastolic blood pressure and age. The data is contained in the file *howto6.sfb*, a small section of which is shown below:

<i>Subject</i>	<i>Age</i>	<i>Blood Pressure</i>
1	27	73
2	21	66
3	22	63
4	26	79
5	25	68
6	28	67

Figure 1: First 6 Rows of Sample Data

Step 1: Plot the Data

The first step in analyzing any new set of data is to plot it.

Procedure: X-Y Plot

The data in this case involve two quantitative variables. To display the relationship between them, a simple *X-Y Scatterplot* will suffice. This plot is used so heavily in data analysis that it may be invoked by pressing the *X-Y Scatterplot* button  on the main toolbar. On the data input dialog box, indicate the variables to be plotted on each axis:

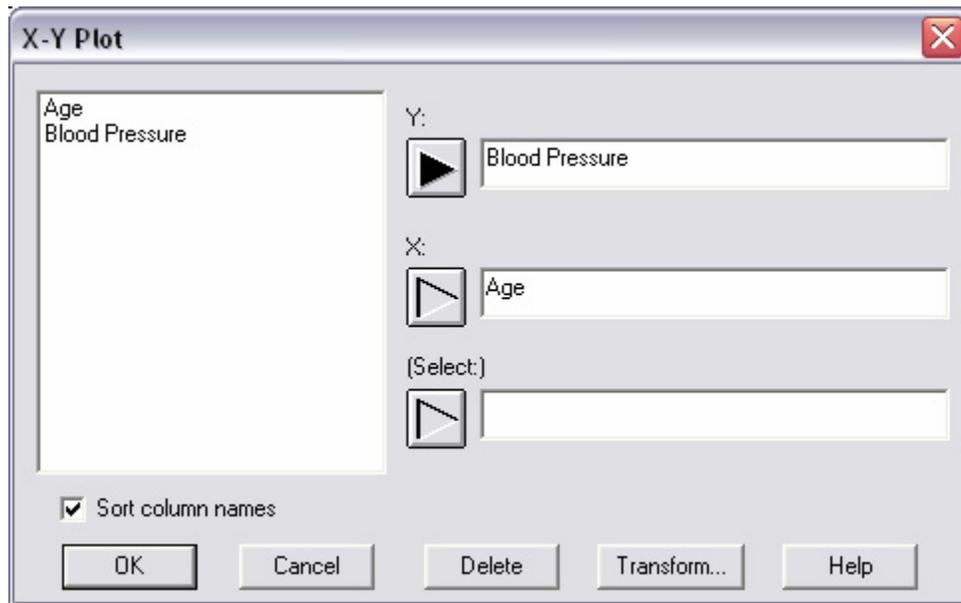


Figure 2: Data Input Dialog Box for X-Y Scatterplot

The resulting plot displays the 54 observations:

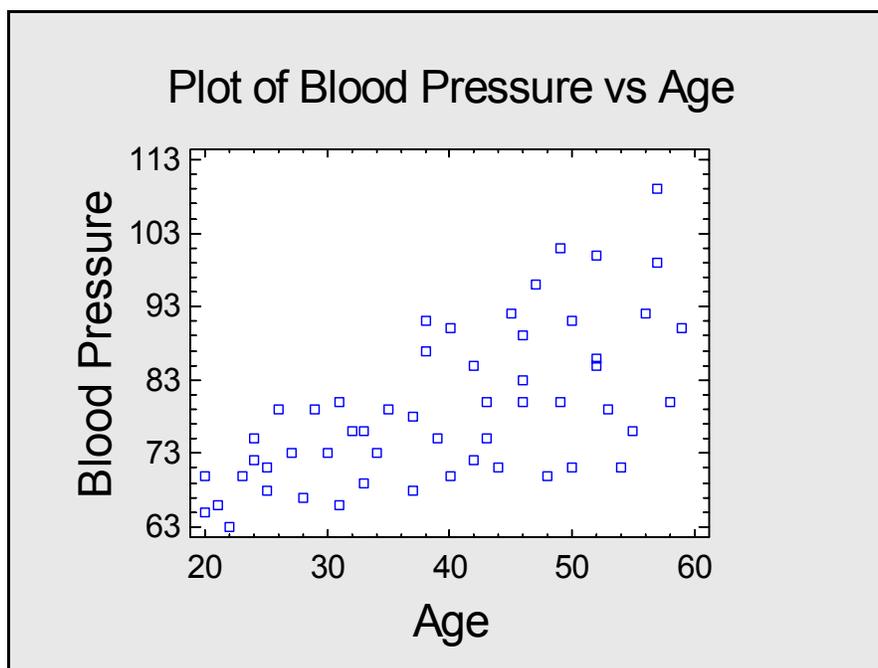


Figure 3: X-Y Scatterplot for 54 Subjects

There is a noticeable increase in both the mean blood pressure and the variability of blood pressure with increasing age.

Step 2: Fit a Linear Model to the Data

The shape of the relationship in Figure 3 suggests that a linear model of the form

$$Y = a + bX + \varepsilon$$

would model the data well, where

Y = blood pressure

X = age

ε = random error

The random error is usually assumed to follow a normal (Gaussian) distribution with mean equal to 0 and variance equal to σ^2 . In this case, the standard deviation of the error terms (σ) appears to be a function of X .

Procedure: Simple Regression

To fit a straight line to this data using the usual least squares method, STATGRAPHICS Centurion provides a *Simple Regression* procedure:

- If using the Classic menu, select: *Relate – One Factor – Simple Regression*.
- If using the Six Sigma menu, select: *Improve – Relate – One Factor – Simple Regression*.

Complete the data input dialog box as shown below:

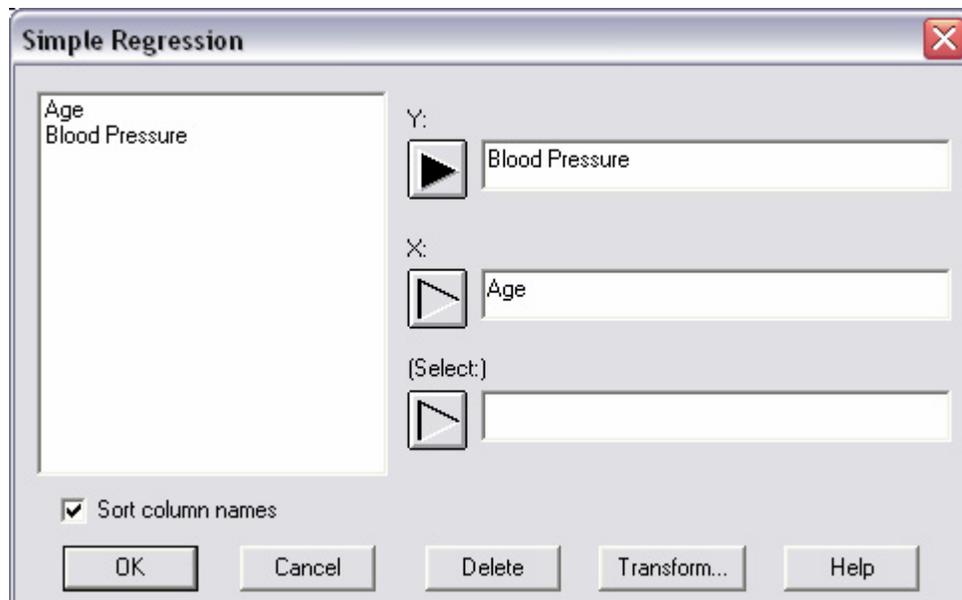


Figure 4: Data Input Dialog Box for Simple Regression

When OK is pressed, an analysis window will be generated containing a plot of the fitted model:

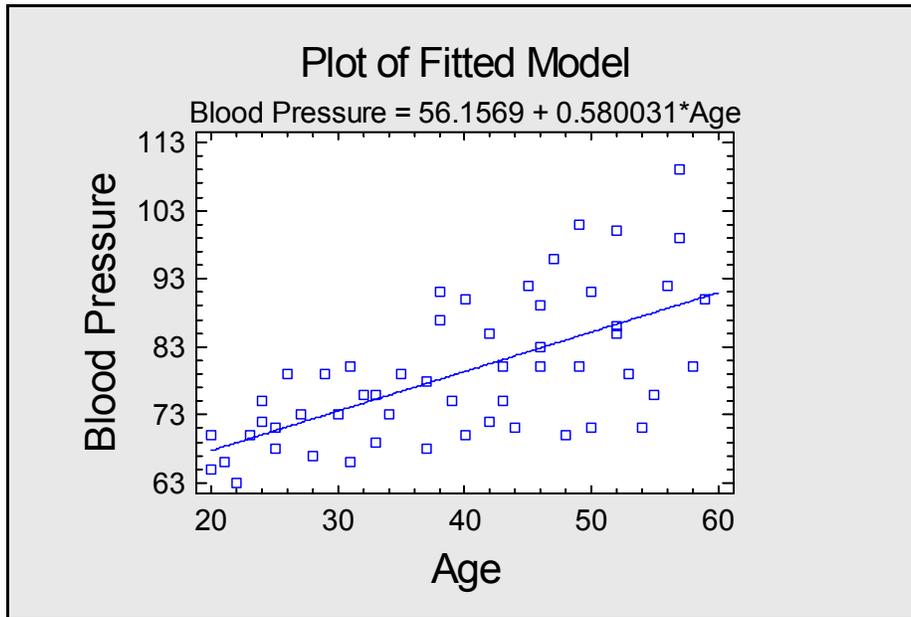


Figure 5: Fitted Linear Model Using Least Squares

Simple Regression generates the model

$$\hat{Y} = \hat{a} + \hat{b}X$$

using ordinary least squares, which means that it finds the line for which the residual sum of squares

$$SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}X_i)^2$$

is as small as possible. Unfortunately, this weights a residual $\hat{\varepsilon}_i$ of a given magnitude the same for all values of X, although a large residual is more significant when X is small (for this data), since the error variance at small X is less.

Also displayed in the analysis window is the *Analysis Summary*, which shows the estimated intercept and slope with their standard errors:

Simple Regression - Blood Pressure vs. Age					
Dependent variable: Blood Pressure (diastolic)					
Independent variable: Age (years)					
Linear model: Y = a + b*X					
Coefficients					
	<i>Least Squares</i>	<i>Standard</i>	<i>T</i>		
<i>Parameter</i>	<i>Estimate</i>	<i>Error</i>	<i>Statistic</i>	<i>P-Value</i>	
Intercept	56.1569	3.99367	14.0615	0.0000	
Slope	0.580031	0.0969512	5.98271	0.0000	
Analysis of Variance					
<i>Source</i>	<i>Sum of Squares</i>	<i>Df</i>	<i>Mean Square</i>	<i>F-Ratio</i>	<i>P-Value</i>
Model	2374.97	1	2374.97	35.79	0.0000
Residual	3450.37	52	66.3532		
Total (Corr.)	5825.33	53			

Correlation Coefficient = 0.638511
 R-squared = 40.7697 percent
 R-squared (adjusted for d.f.) = 39.6306 percent
 Standard Error of Est. = 8.14575
 Mean absolute error = 6.29301
 Durbin-Watson statistic = 2.34294 (P=0.8754)
 Lag 1 residual autocorrelation = -0.185587

The StatAdvisor

The output shows the results of fitting a linear model to describe the relationship between Blood Pressure and Age. The equation of the fitted model is

$$\text{Blood Pressure} = 56.1569 + 0.580031 * \text{Age}$$

Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between Blood Pressure and Age at the 95% confidence level.

The R-Squared statistic indicates that the model as fitted explains 40.7697% of the variability in Blood Pressure. The correlation coefficient equals 0.638511, indicating a moderately strong relationship between the variables. The standard error of the estimate shows the standard deviation of the residuals to be 8.14575. This value can be used to construct prediction limits for new observations by selecting the Forecasts option from the text menu.

The mean absolute error (MAE) of 6.29301 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is greater than 0.05, there is no indication of serial autocorrelation in the residuals at the 95% confidence level.

Figure 6: Simple Regression Analysis Summary

The estimated model coefficients and standard errors are:

$$\text{Intercept: } 56.1569 \pm 3.99367$$

$$\text{Slope: } 0.580031 \pm 0.0969512$$

The overall R-squared statistic, which measures the proportion of variability in *Blood Pressure* that has been explained by the model, is approximately 40.8%.

When fitting a line to this data, it would be better to give less weight to a residual of a given magnitude for values of X where the error variance is large. There are two primary approaches normally used to do this:

- (1) Fit the model after applying a variance stabilizing transformation to Y. In many cases, a transformation such as a square root, logarithm, or reciprocal transforms the problem into a metric where the error variance is approximately constant. Ordinary least squares can then be applied in the transformed metric.
- (2) Fit the model using weighted least squares. In this case, the residuals are weighted inversely to their variance, and estimates of the intercept and slope are obtained by minimizing:

$$SSE_w = \sum_{i=1}^n w_i \hat{\epsilon}_i^2 = \sum_{i=1}^n w_i (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n w_i (Y_i - \hat{a} - \hat{b}X_i)^2$$

The hard part, of course, is usually determining how the error variance changes so that an appropriate transformation or set of weights can be obtained.

Step 3: Examine the Error Variance

An important step when fitting any statistical model is plotting of the residuals. In the *Simple Regression* procedure, a plot of the residuals $\hat{\epsilon}$ versus X may be created by pressing the Graphs button  and selecting *Residuals versus X*:

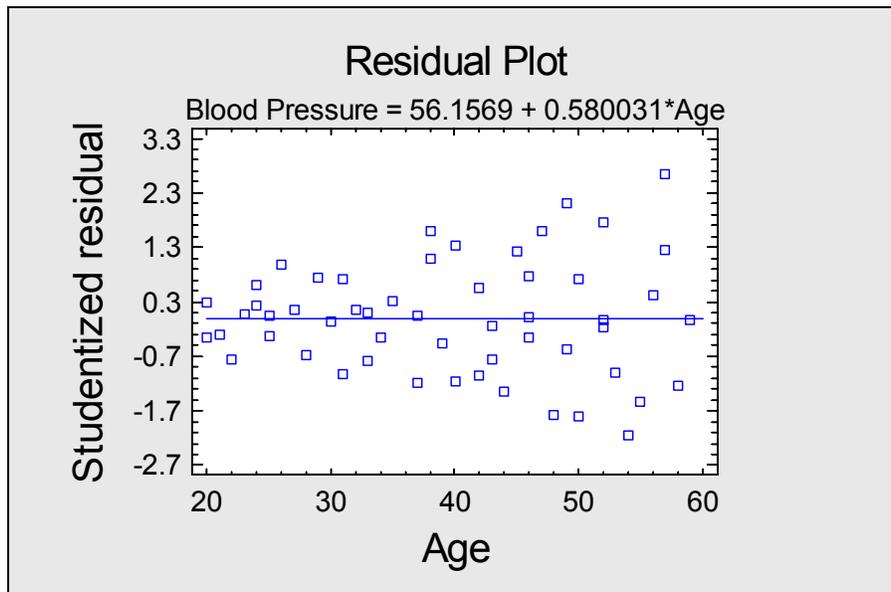


Figure 7: Plot of Residuals versus Age

Note the strong funnel shape, indicating that the error variance increases as X increases. To model the error variance:

1. Press the *Save results* button  on the analysis toolbar. Complete the dialog box as shown below to save the residuals to a column of datasheet A:

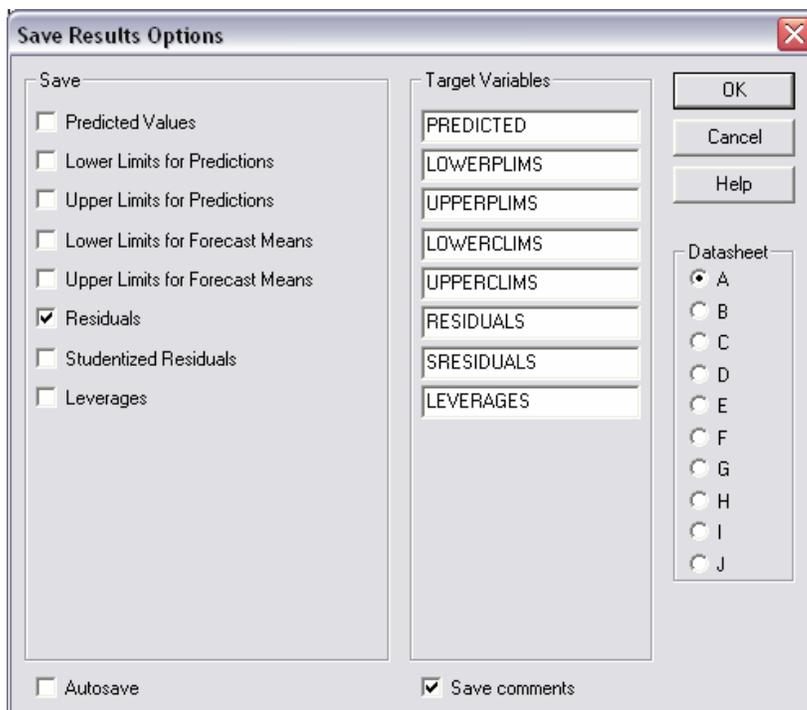


Figure 8: Dialog Box for Saving Results

2. Divide the data into several groups according the value of *Age* and estimate the standard deviation in each group. This is most easily done using the *Subset Analysis* procedure. You can run this procedure by:
 - If using the Classic menu, select: *Describe – Numeric Data – Subset Analysis*.
 - If using the Six Sigma menu, select: *Analyze – Variable Data – Multiple Sample Comparisons – Subset Analysis*.

Complete the data input dialog box as shown below:

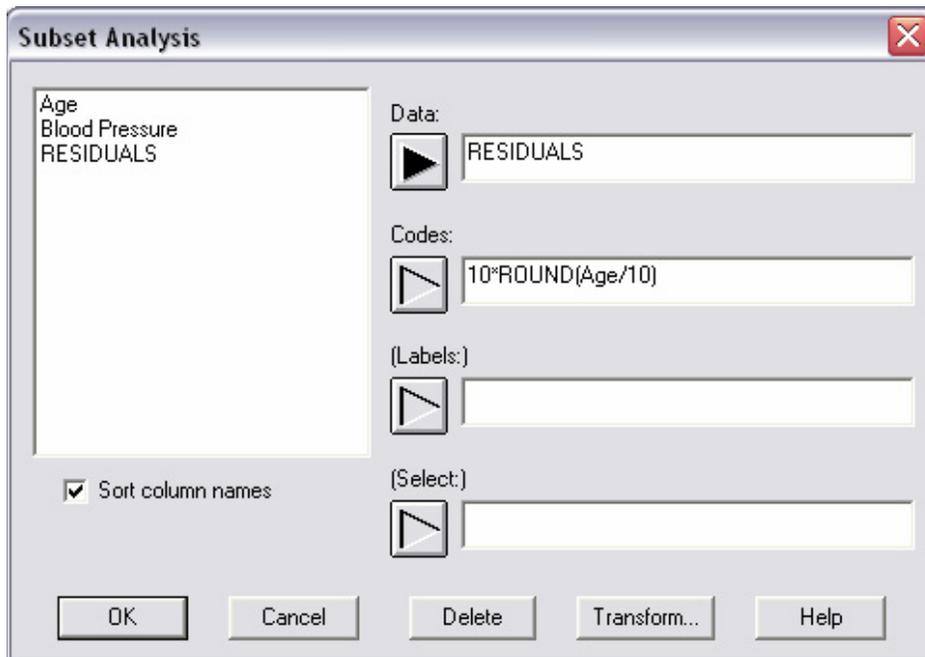


Figure 9: Data Input Dialog Box for Subset Analysis

The expression in the *Codes* field groups the *Residuals* by rounding *Age* to the nearest multiple of 10. The result is shown in the following *Scatterplot*, generated by default by the *Subset Analysis* procedure:

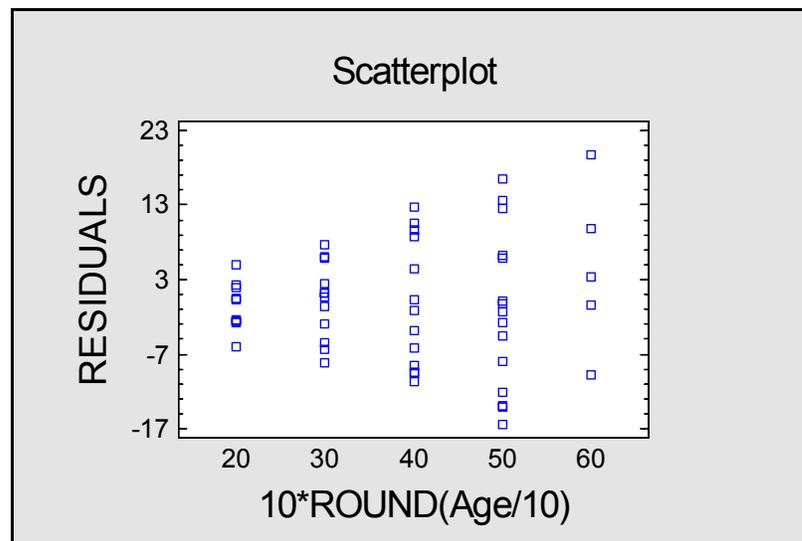


Figure 10: Scatterplot of Residuals Rounded into Groups

3. We can now press the *Graphs* button  on the analysis toolbar and select the *Sigma Plot*. This plot displays the standard deviation for each group of observations:

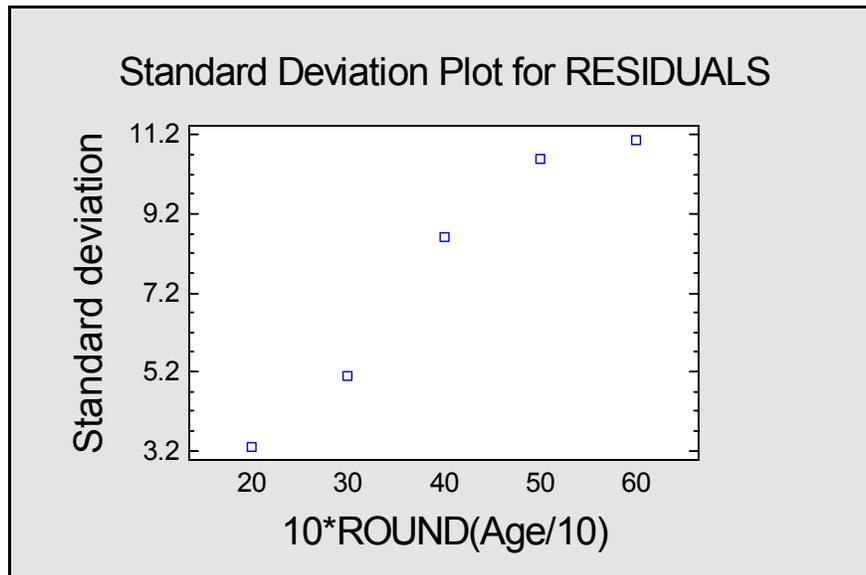


Figure 11: Scatterplot of Residuals Rounded into Groups

We can also use the *Tables* button  to create a table of the standard deviations:

Summary Statistics		
Data variable: RESIDUALS		
		<i>Standard</i>
<i>10*ROUND(Age/10)</i>	<i>Count</i>	<i>Deviation</i>
20	9	3.2962
30	12	5.08238
40	13	8.60559
50	15	10.5909
60	5	11.0787
Total	54	8.06853

Figure 12: Table of Residual Standard Deviations by Age

It will be noted that the relationship is nearly linear, suggesting that the error standard deviation increases proportionately to the value of *Age*. We may also save the group standard deviations and plot them with a fitted line as shown below:

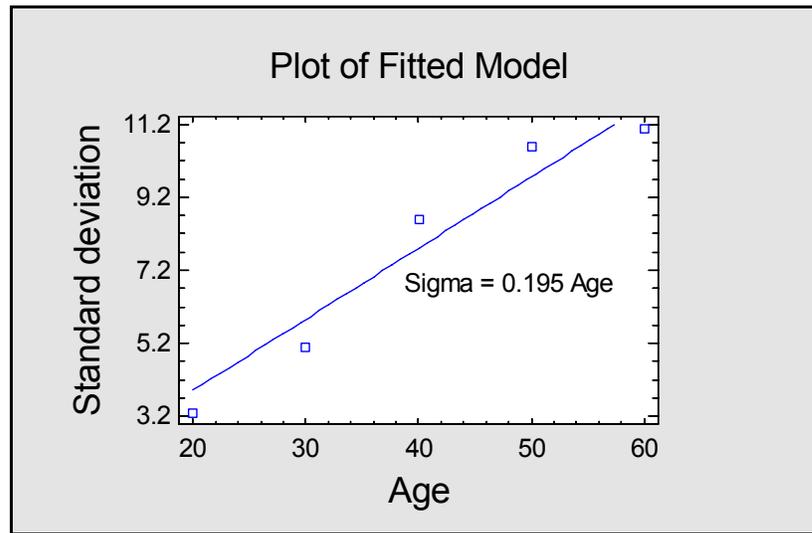


Figure 13: Plot of Residual Standard Deviation with Model for Sigma

A very plausible model for this data is that the error standard deviation is directly proportional to *Age*. This means that the coefficient of variation, defined by the standard deviation divided by the mean, is constant at approximately 19.5%.

Step 4: Apply a Variance Stabilizing Transformation

When the variance of the response variable increases as its mean increases, it is sometimes possible to stabilize the variance by applying a transformation to *Y*. The table below shows some common situations:

<i>Situation</i>	<i>Variance stabilizing transformation</i>
Error variance proportional to mean response: $\sigma^2 \propto \mu_Y$	$Y' = \sqrt{Y}$
Error standard deviation proportional to mean response: $\sigma \propto \mu_Y$	$Y' = \log Y$
Error standard deviation proportional to mean response squared: $\sigma \propto \mu_Y^2$	$Y' = \frac{1}{Y}$

Figure 14: Table of Variance Stabilizing Transformations

Our earlier analysis suggests that a logarithm may help stabilize the variance. If we begin with the linear model

$$Y = a + bX$$

then taking logarithms of both sides results in

$$\log Y = \log(a + bX)$$

Assuming that the errors are additive after taking logs, the problem is a nonlinear one which must be solved using nonlinear least squares.

Procedure: Nonlinear Regression

To fit a model using nonlinear least squares in STATGRAPHICS Centurion:

- If using the Classic menu, select: *Relate – Multiple Factors – Nonlinear Regression*.
- If using the Six Sigma menu, select: *Improve – Relate – Multiple Factors – Nonlinear Regression*.

On the data input dialog box, enter an expression for the transformed values of Y and for the form of model to be fit:

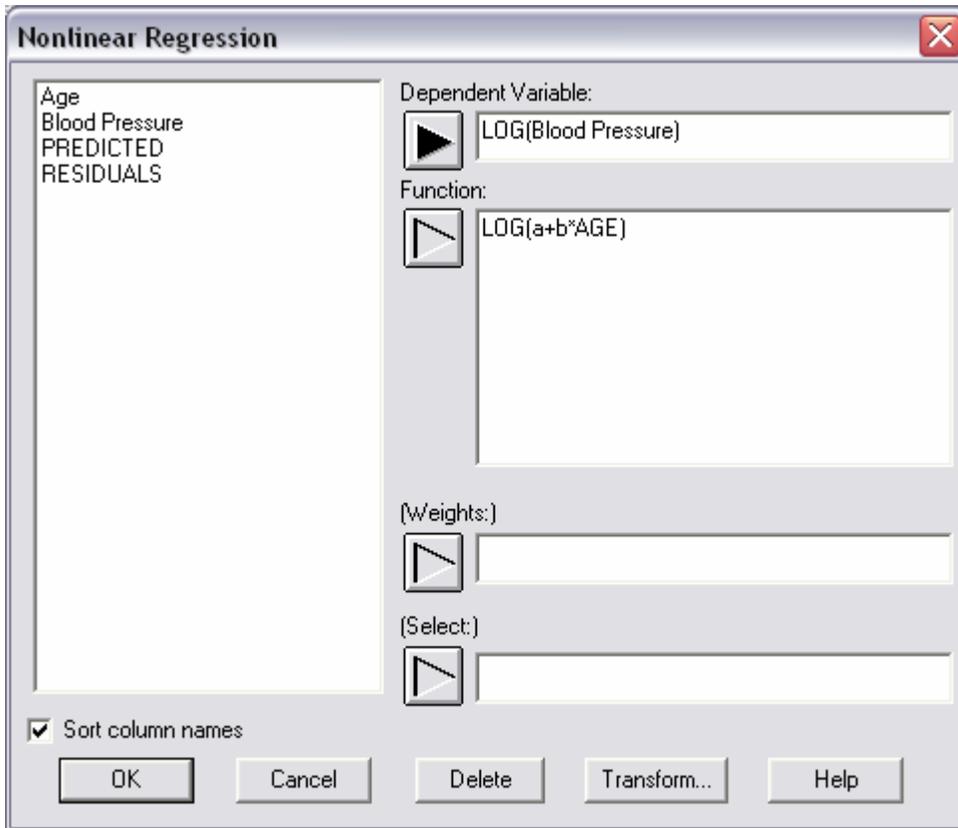
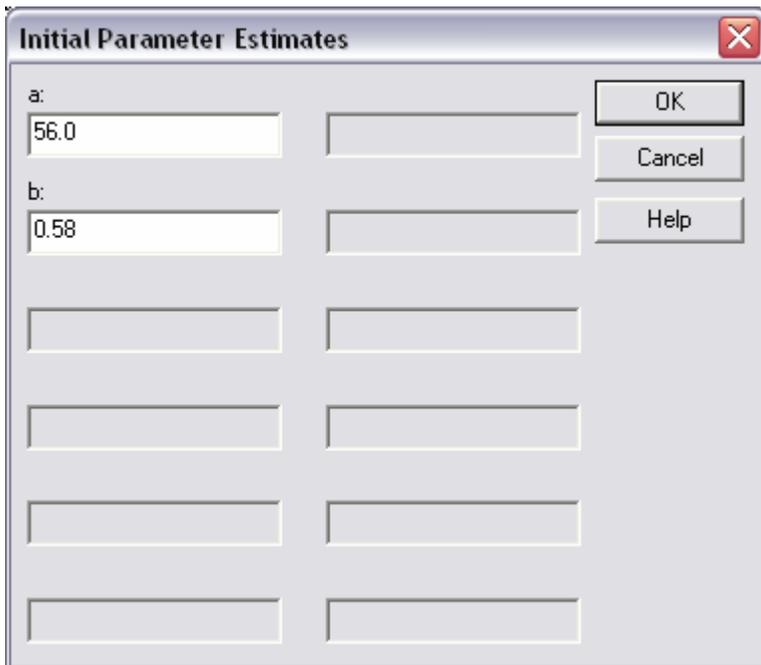


Figure 15: Data Input Dialog Box for Nonlinear Regression

A second dialog box will be displayed asking for initial estimates of the unknown parameters:



The dialog box titled "Initial Parameter Estimates" contains several input fields and control buttons. The first row is labeled "a:" and has a text box containing "56.0". The second row is labeled "b:" and has a text box containing "0.58". There are four empty text boxes arranged in two columns below the "b:" field. On the right side, there are three buttons: "OK", "Cancel", and "Help".

Figure 16: Initial Parameter Estimates for Nonlinear Regression

Initial estimates are needed because the nonlinear regression procedure performs a numerical search for the best fitting model. In the figure above, we have supplied the parameter estimates from the *Simple Regression* procedure as initial values.

The optimization will then be performed and the fitted model displayed:



Figure 17: Fitted Nonlinear Regression Model

The estimated coefficients are displayed in the *Analysis Summary*:

Nonlinear Regression - LOG(Blood Pressure)

Dependent variable: LOG(Blood Pressure)

Independent variables:

AGE (years)

Function to be estimated: LOG(a+b*AGE)

Initial parameter estimates:

a = 56.0

b = 0.58

Estimation method: Marquardt

Estimation stopped due to convergence of parameter estimates.

Number of iterations: 3

Number of function calls: 10

Estimation Results

			Asymptotic	95.0%
			Confidence	Interval
Parameter	Estimate	Standard Error	Lower	Upper
a	56.509	3.6115	49.2619	63.756
b	0.561454	0.0915119	0.377822	0.745087

Figure 18: Analysis Summary for Fitted Nonlinear Regression Model

A comparison of the estimated coefficients between the linear and nonlinear regressions is shown below:

	Simple Regression Using Ordinary Least Squares	Nonlinear Regression with Variance Stabilizing Transformation
Intercept	56.1569 ± 3.99367	56.509 ± 3.6115
Slope	0.580031 ± 0.0969512	0.561454 ± 0.0915119

Figure 19: Comparison of Model Coefficients

The *Nonlinear Regression* has resulted in a model that is not quite as steep as the original fit. The standard errors of the coefficients are also somewhat smaller.

Step 5: Use Weighted Least Squares

The alternative to applying a variance stabilizing transformation is to use weighted least squares. By using weights w_i that are proportional to the error variance, the effects of heteroscedasticity can be mitigated. Some common cases are:

Situation	Least squares weights
Error variance proportional to X: $\sigma^2 \propto X$	$w_i = \frac{1}{X_i}$
Error standard deviation proportional to X: $\sigma \propto X$	$w_i = \frac{1}{X_i^2}$
Error variance proportional to \sqrt{X} : $\sigma^2 \propto \sqrt{X}$	$w_i = \frac{1}{\sqrt{X_i}}$

Figure 20: Table of Least Squares Weights

Procedure: Multiple Regression

To fit a regression model using weighted least squares in STATGRAPHICS Centurion:

- If using the Classic menu, select: *Relate – Multiple Factors – Multiple Regression*.
- If using the Six Sigma menu, select: *Improve – Relate – Multiple Factors – Multiple Regression*.

The dialog box should be completed as shown below:

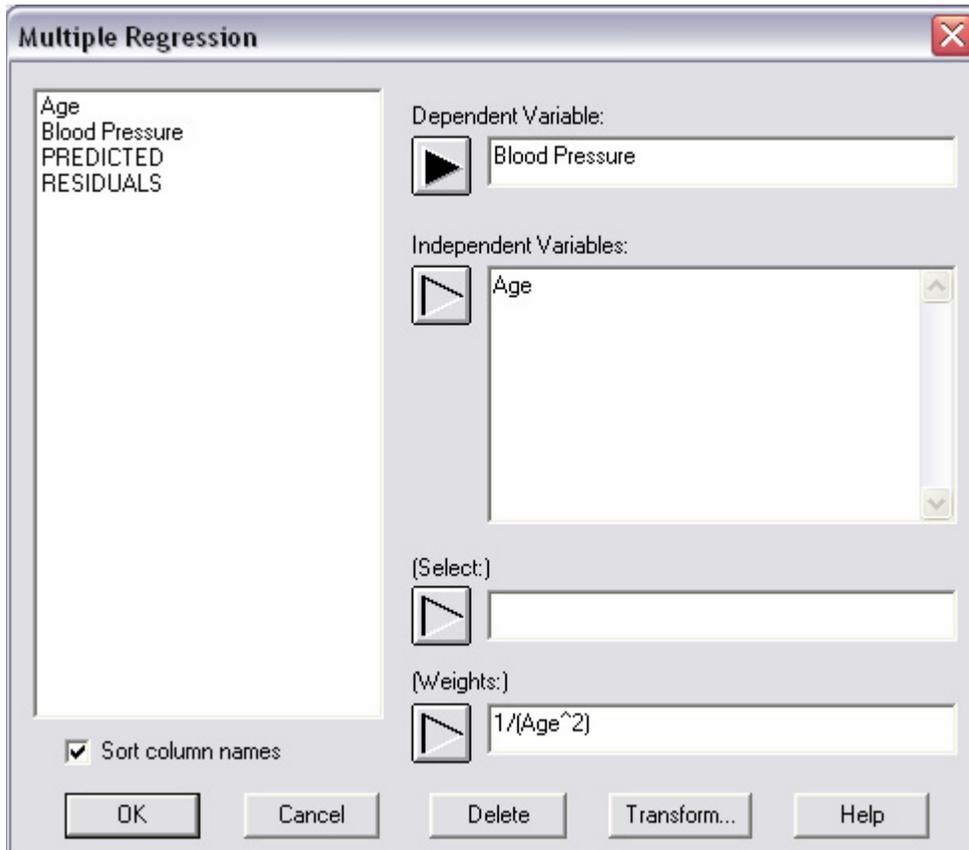


Figure 21: Data Input Dialog Box for Weighted Least Squares

In the *Weights* field, the expression $1 / \text{Age}^2$ has been entered to generate the weights w . This is based on our analysis in Step 3, which showed a proportional relationship between the standard deviation of the residuals from the ordinary least squares fit and the value of *Age*.

The *Analysis Summary* shows the results of the fit:

Multiple Regression - Blood Pressure

Dependent variable: Blood Pressure (diastolic)

Independent variables:

Age (years)

Weight variable: $1/(\text{Age}^2)$

		<i>Standard</i>	<i>T</i>	
<i>Parameter</i>	<i>Estimate</i>	<i>Error</i>	<i>Statistic</i>	<i>P-Value</i>
CONSTANT	55.831	2.78093	20.0764	0.0000
Age	0.588828	0.0815822	7.21761	0.0000

Analysis of Variance

<i>Source</i>	<i>Sum of Squares</i>	<i>Df</i>	<i>Mean Square</i>	<i>F-Ratio</i>	<i>P-Value</i>
Model	1.86441	1	1.86441	52.09	0.0000
Residual	1.86105	52	0.0357894		
Total (Corr.)	3.72546	53			

R-squared = 50.0451 percent

R-squared (adjusted for d.f.) = 49.0844 percent

Standard Error of Est. = 0.189181

Mean absolute error = 4.8556

Durbin-Watson statistic = 2.34038

Lag 1 residual autocorrelation = -0.184864

Figure 22: Analysis Summary for Weighted Least Squares

Summarizing all three models:

	<i>Simple Regression Using Ordinary Least Squares</i>	<i>Nonlinear Regression with Variance Stabilizing Transformation</i>	<i>Multiple Regression Using Weighted Least Squares</i>
<i>Intercept</i>	56.1569 ± 3.99367	56.509 ± 3.6115	55.831 ± 2.78093
<i>Slope</i>	0.580031 ± 0.0969512	0.561454 ± 0.0915119	0.588828 ± 0.0815822

Figure 23: Comparison of Model Coefficients

Using this approach, the fitted line is somewhat steeper than the original fit. The standard errors of the coefficients have also fallen noticeably.

Step 6: Compare the Models

As a final comparison, the plot below (created using the *Multiple X-Y Plot* procedure and then overlaying two graphs in the *StatGallery*) shows the data with the three fitted models:

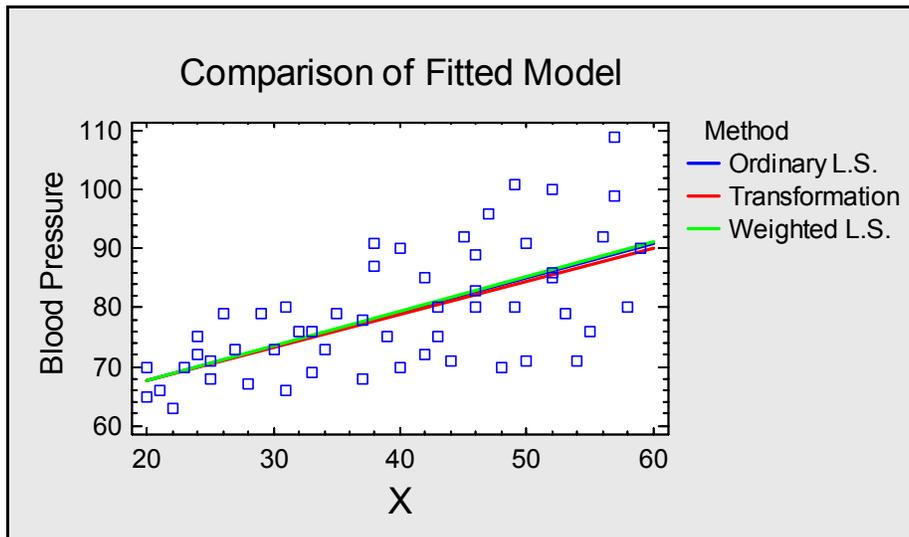


Figure 24: Plot of Three Fitted Models

Admittedly, the differences between the fitted models are not large in this case, but the data values have been treated more uniformly.

Part of the improvement gained by stabilizing the variance may be seen by examining the Studentized residuals. Returning to the *Simple Regression* window, we can press the *Tables* button  on the analysis toolbar and generate a table of unusual residuals:

Unusual Residuals					
Row	X	Y	Predicted Y	Residual	Studentized Residual
28	49.0	101.0	84.5784	16.4216	2.12
43	54.0	71.0	87.4786	-16.4786	-2.14
49	57.0	109.0	89.2187	19.7813	2.65

The StatAdvisor
 The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Blood Pressure deviates from a model fitted using all of the data except that observation. In this case, there are 3 Studentized residuals greater than 2.0, but none greater than 3.0.

Figure 25: Table of Unusual Residuals from Ordinary Least Squares Fit

Included in the table are any observations for which the Studentized residual exceeds 2.0 in absolute value. The Studentized residual equals the number of estimated standard errors that each observation lies from the fitted model when that observation is *not* used to fit the model (sometimes called *deleted residuals*). Observation #49 is 2.65 standard errors above the fitted model. When the weighted least squares fit is run, *no* Studentized residuals appear on this table. The Studentized residual for observation #49 falls to 1.90.

An even more important difference is seen when comparing the prediction limits around the fitted model. The plot below shows 95% prediction limits for the ordinary least squares fit and for the weighted least squares fit:

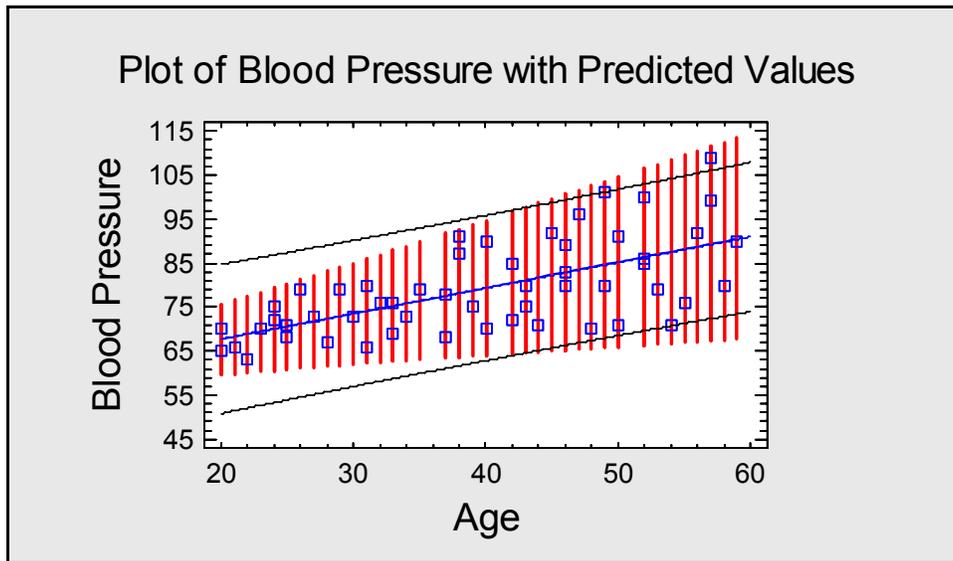


Figure 26: 95% Prediction Limits Using Unweighted and Weighted Least Squares

The solid bounds correspond to the limits from the unweighted least squares fit, which assumes constant variance (added to the *Plot of Fitted Model* in *Simple Regression* using *Pane Options*). The vertical lines show the bounds for the weighted least squares fit (created using the *Interval Plots* option in *Multiple Regression*). If we wished to identify limits within which 95% of all additional samples might fall, the weighted least squares approach gives a much better result.

Conclusion

When fitting regression models or performing an analysis of variance, the usual assumption is that the error variance is constant everywhere. In many cases, this is not true and can lead to results that are at best inefficient and at worst misleading. Two primary approaches are available to remedy such heteroscedasticity: variance stabilizing transformations and weighted least squares. As illustrated in this guide, both methods are easy to apply and can lead to results that are much more reliable.

Note: The author welcomes comments about this guide. Please address your responses to neil@statgraphics.com.