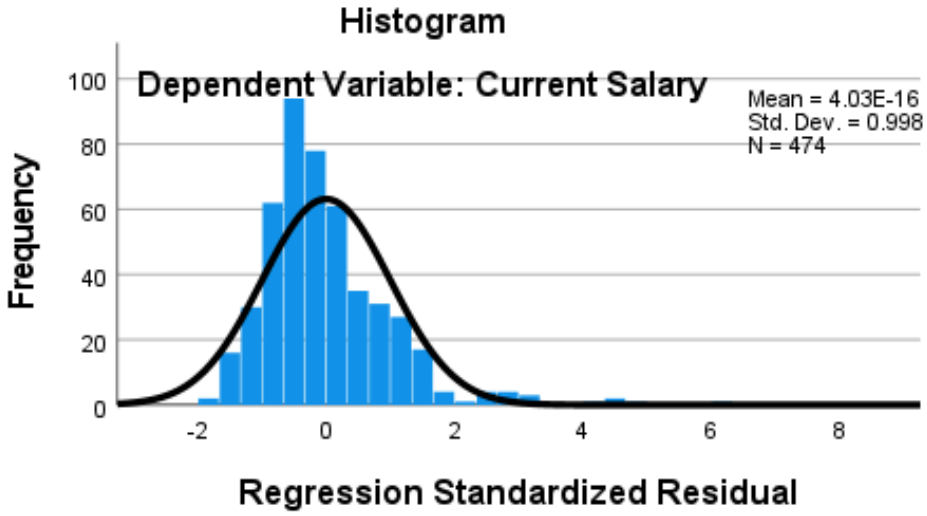
Checking Regression Assumptions Graphically

Jon Peck 3/15/2021
jkpeck@gmail.com

Linear regression has excellent statistical properties when the assumptions behind it are satisfied. The assumptions include a correct specification, independence of the regressors and the error term, independence of the error terms across the cases, and constant error variance (homoscedasticity). For some statistics, normality of the error term is needed. Violations of these assumptions have consequences of varying seriousness.

Statistical tests are available for some but not all these assumptions. Often, however, graphical displays are the best way to assess them. For example, a residual histogram gives a good idea of whether the normality assumption is satisfied and whether outliers are present. This chart, for example, shows some skew and large outliers to the right. Formal tests confirm this, but the plot tells the story better.



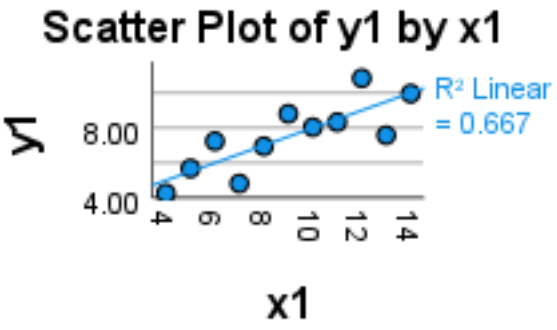
Tests of Normality

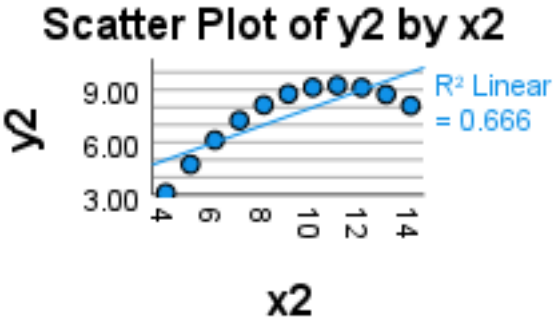
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Deleted Residual	0.113	474	0.000	0.884	474	0.000

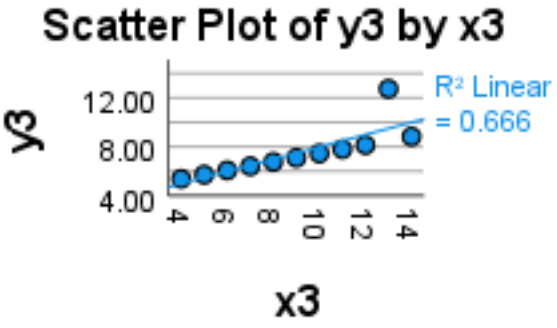
a. Lilliefors Significance Correction

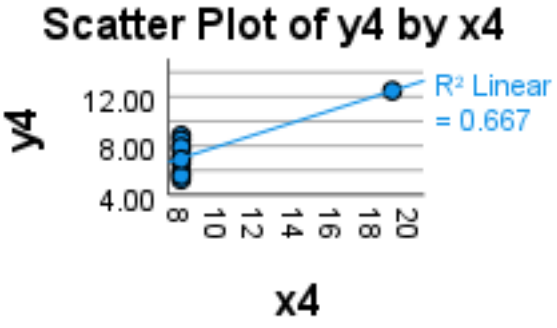
The Anscombe quartet

Perhaps the most famous example of what hides behind the regression line is the Anscombe quartet. These four plots have exactly the same regression line, but they differ wildly in how well the line describes the data. The plots tell the story.



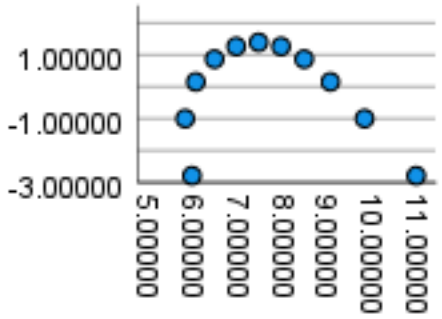






While the first plot shows a case where the regression line is a reasonable summary, the second shows that the functional form is wrong; the third shows a big outlier, and the fourth shows the impact of a high leverage point that seriously distorts the line.

Looking at the residuals vs predicted values plot for the second equation, the functional form problems are screamingly obvious.

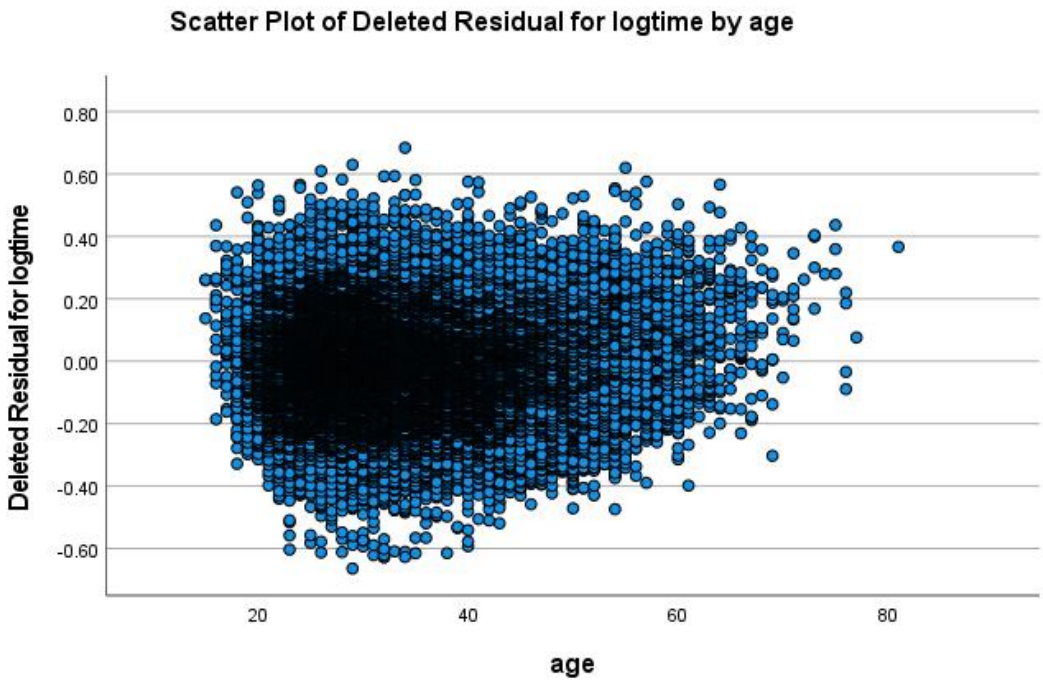


With larger datasets, the story may not be so obvious.

Heteroscedasticity

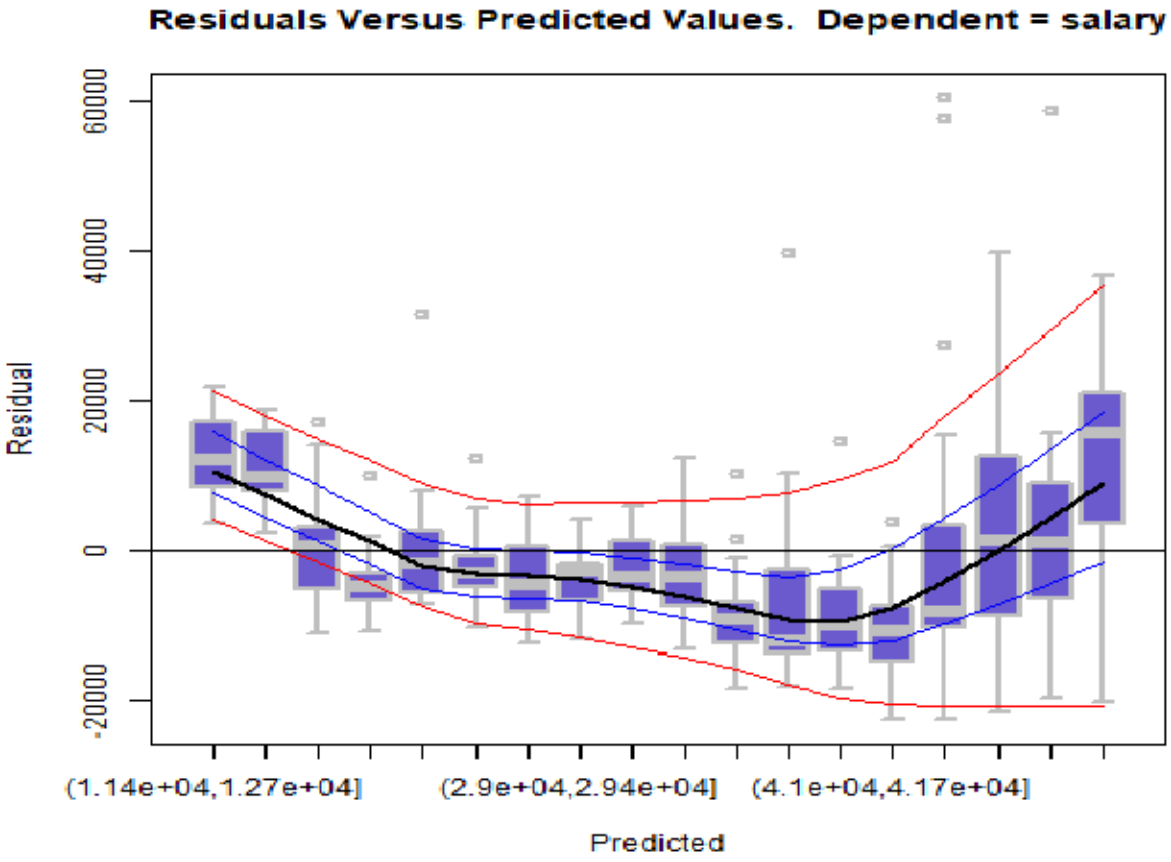
Not shown in any of these plots is a good graph for assessing heteroscedasticity. That is the situation where the error variance is not constant. Often this happens when the scale of the data varies substantially over the sample. Heteroscedasticity does not cause bias in the coefficient estimates, but their reported standard errors will be incorrect, and the estimates are not statistically efficient.

Several heteroscedasticity tests are available in Statistics in the UNIANOVA procedure – Breusch-Pagan, White, and an F test, but, again, graphical methods may be more illuminating. However, in large samples, the picture may be obscured. Here is a residual plot looking for varying error variance versus age, which is one of the independent variables. There are 28,628 cases in this dataset. The dependent variable is the log of times in a marathon.

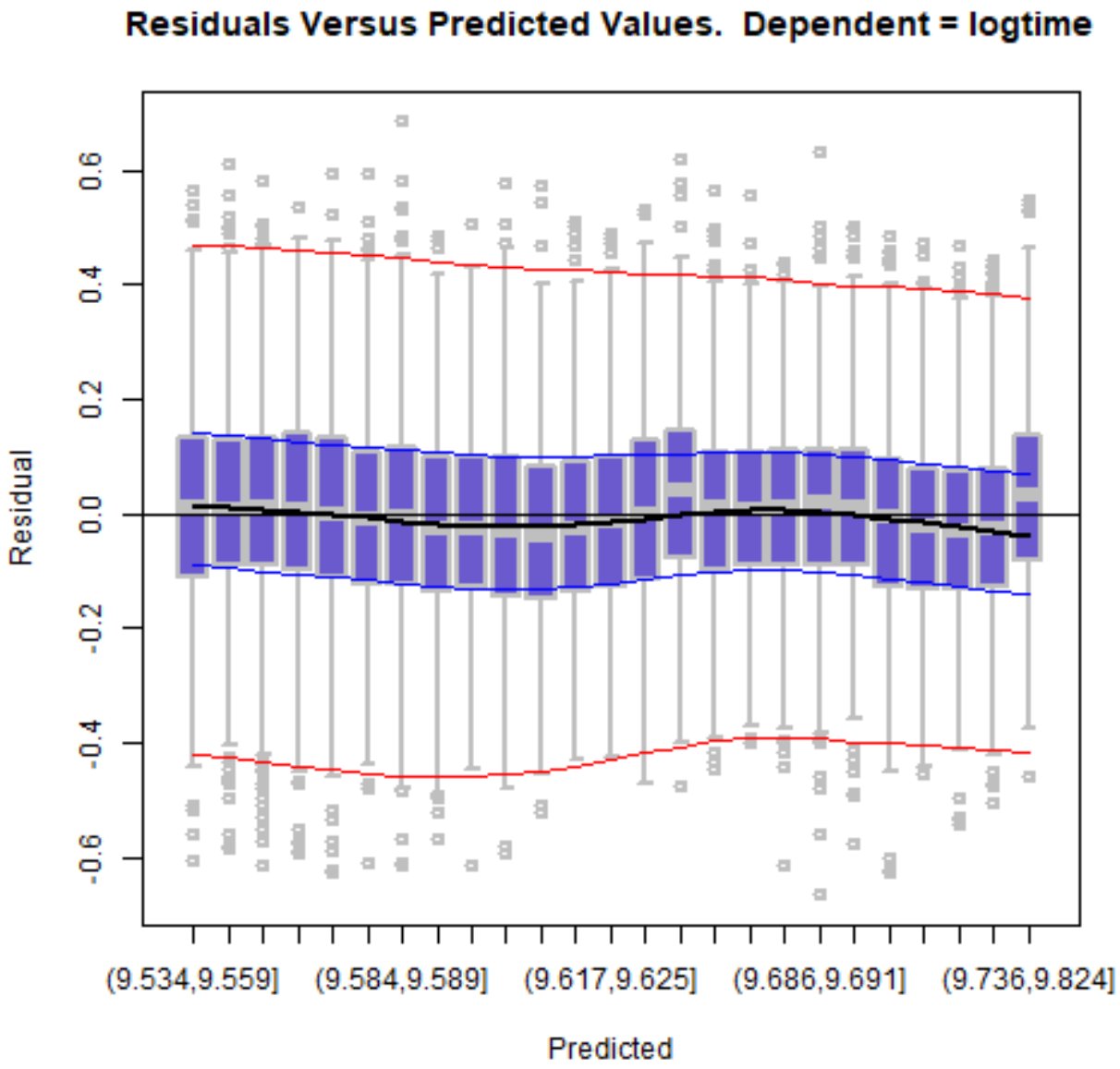


Maybe the error variance diminishes with age, but it is hard to tell.

We need to use some statistical aggregation in the graph to make the picture clearer. I am going to use a set of boxplots connected at the medians. Here is one from a different dataset. The pattern here is very clear. It is obvious that the residual variance is increasing with the predicted values, but you can also see that the residual mean varies with these values, which violates a very important assumption. That may indicate a functional form problem. You can also see that the outliers are all on the same side of zero. The lines curving horizontally across the graph are loess fits to the quantiles of the residual bins with the heavy one at the median.

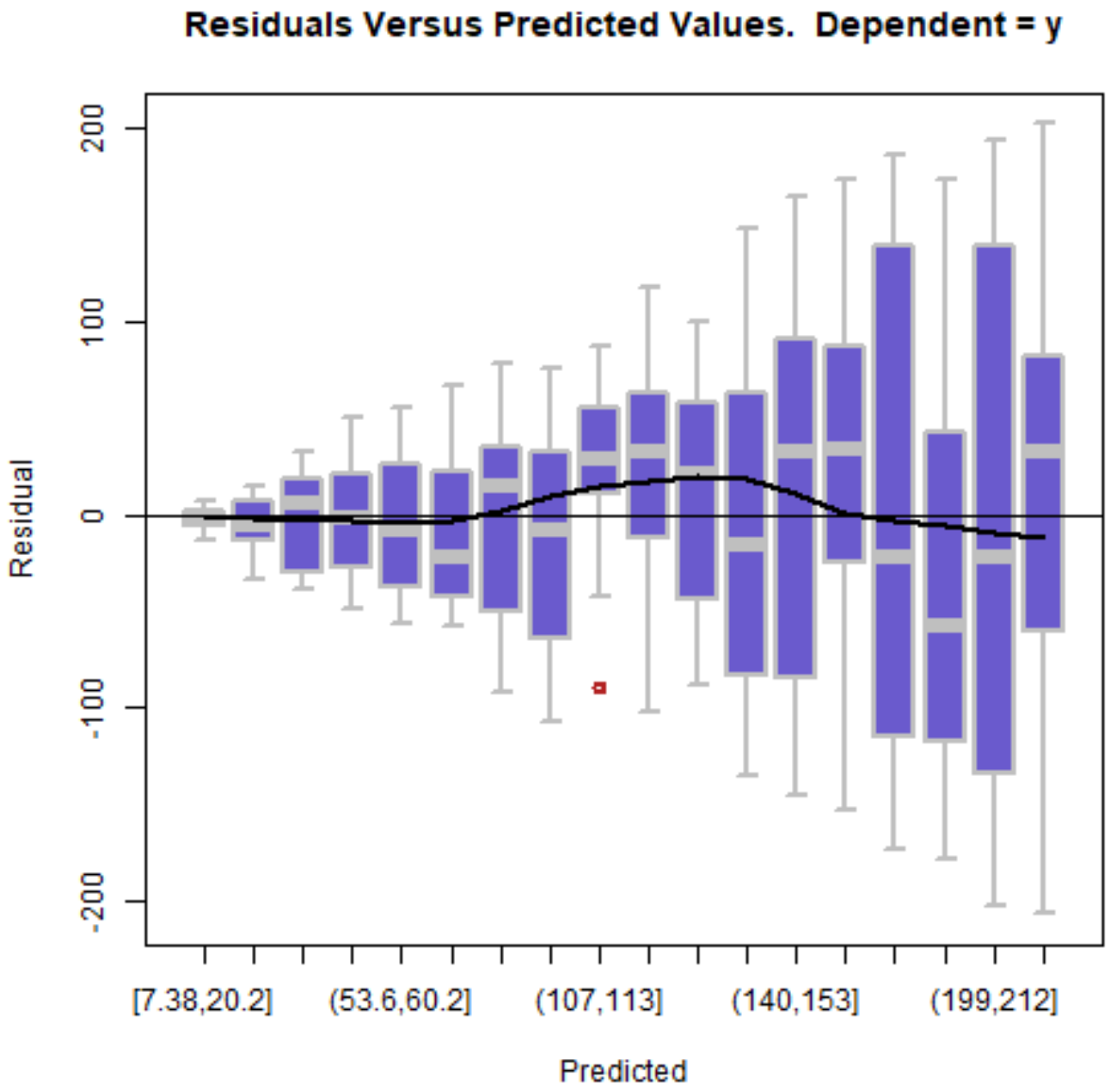


Now let's go back to the large dataset and do this plot. Here's what it looks like.



We can see that the boxes are nearly equal in size, but the outlier distribution becomes a little tighter as the predicted values increase, and the median line waves a bit in one interior region and sags at the extreme right end.

Here is a plot for some artificial data with a real heteroscedasticity problem.



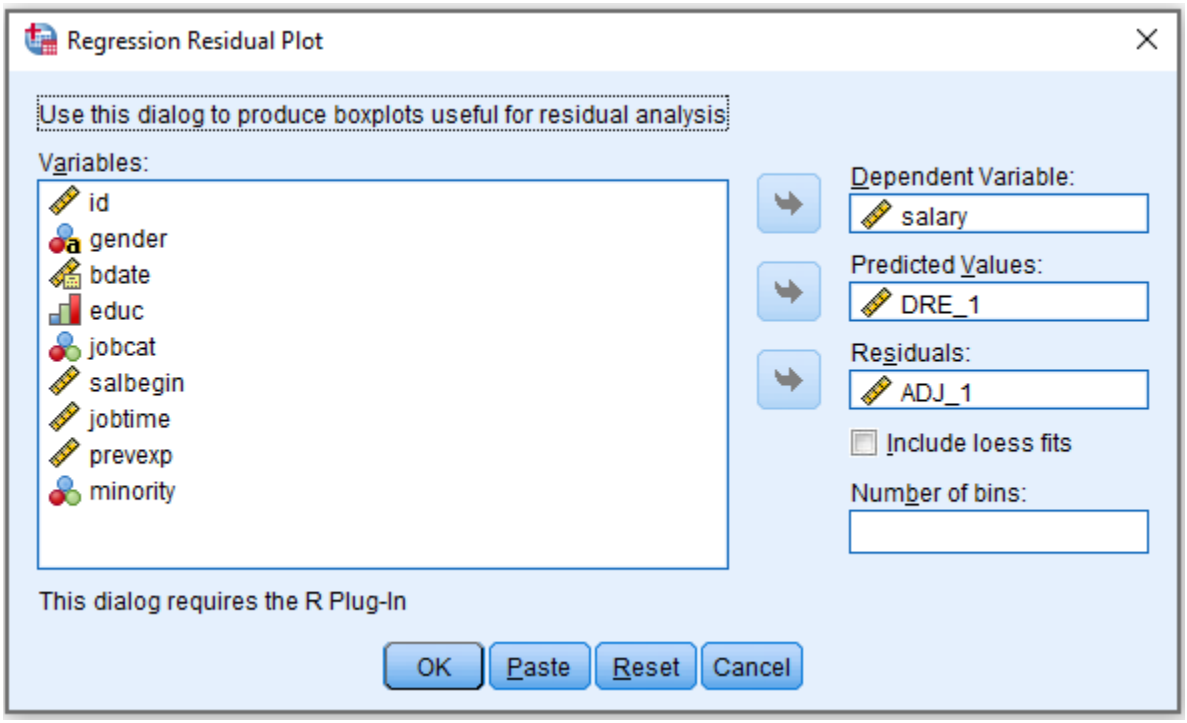
What Residual and Predicted Values to Use

The various regression procedures in Statistics can produce a variety of types of residuals and predicted values. For heteroscedasticity checking, it is not likely to matter much which type is used except that studentized residuals and other similar types might make heteroscedasticity more difficult to detect.

High leverage points, however, as in the fourth Anscombe plot above can cause what would be a large residual to appear smaller and make a problem with the functional form harder to detect. Therefore, when available, deleted residuals and deleted predicted values, where a data point does not contribute to its own residual or predicted value would be preferable where available.

The STATS RESIDUAL BOXPLOTS Extension Command

This command, which is used to create the boxplots above, can be installed via the *Extensions* > *Extension Hub* menu in Statistics. It requires the appropriate version of R and the R Essentials. It has a custom dialog box, which appears on the *Graphs* menu, and standard-style Statistics syntax. Here is the dialog box.



First, run the desired regression procedure and save the residuals and predicted values. Enter those in the appropriate slots. Check the loess box if you want those lines to appear.

This is the generated syntax.

```
STATS RESIDUAL BOXPLOTS DEPENDENT=salary
PREDVALS=DRE_1 RESIDUALS=ADJ_1
LOESS=NO.
```

In this example, the number of bins is determined by default based on the sample size.

Syntax help, as usual, is available by pressing F1 on an instance of the command in the Syntax Editor.

Acknowledgments

The idea for this plot came from a web post from several years ago for which I no longer have a reference.