Should the Residuals be Normal?

How a grain of truth can become a mountain of misunderstanding

Donald J. Wheeler

The analysis of residuals is commonly recommended when fitting a regression equation to a data set. It has even been recommended for the analysis of experimental data where the independent variable is categorical (i.e. treatment levels). In both of these contexts it has been said that the residuals should be "normally distributed." This paper shall look at this idea and make suggestions about what does and does not make sense.

Back when I was teaching the graduate level regression theory class at the University of Tennessee the question of the normality of the residuals was not discussed in any of the textbooks. The analysis of residuals simply did not include any consideration of the histogram of residual values. However, it seems that the importance of having normally distributed data and normally distributed residuals has grown in direct proportion to the availability of software for performing lack-of-fit tests. "After all, if it is part of the software it must be important, mustn't it?" Unfortunately, as with other aspects of statistical analysis, the analysis of residuals involves much more than simply using all the options in your software.

Regression

In the regression problem you are looking for some function, or combination of functions, of the independent variables that will explain a substantial proportion of the variation in the dependent variable. The object is to extract all of the systematic variation and to leave nothing but noise behind. And the residuals are defined as the difference between the original data and the predicted values from the regression equation. In this context the analysis of residuals is a simple graphic technique to see if there are any obvious patterns left within the unexplained portion of the variation of the dependent variable. The emphasis is upon not missing patterns that might suggest a relationship between the independent and dependent variables. It is not about what is the shape of the histogram of the residuals.

Early in my career a couple of my graduate students who were in over their head got me involved in a civil engineering research project. As I scrambled to catch up I threw a bunch of data into the computer and performed a series of regressions. As I interpreted these regressions I told the principal investigator that the width of the local streets did not have an impact on the midblock accident rate. When I did this he told me, in no uncertain terms, that I was wrong, and that the width of the street did, indeed, affect the midblock accident rate. So I decided to do what

I had neglected to do before, I would look at the data. The residuals for the regression I had fit to these data are shown in figure 1.

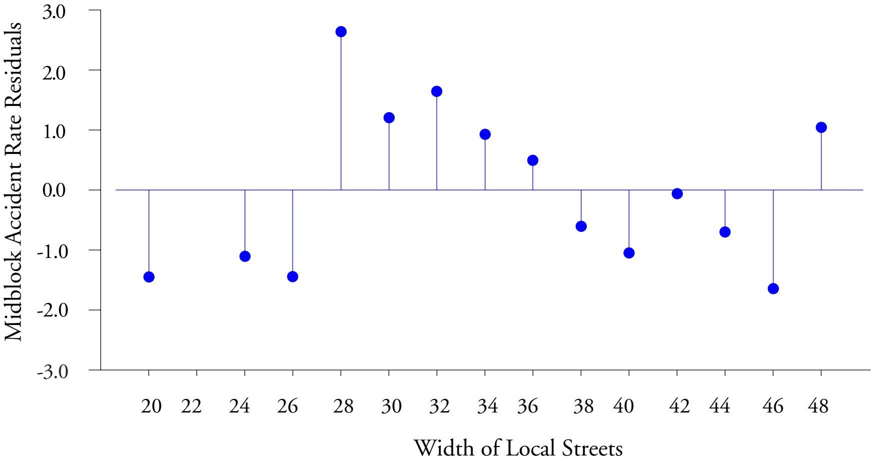

Figure 1: Residuals for Midblock Accident Rates

It is traditional to show the residuals plotted in order according to the values of the independent variable. In figure 1 the horizontal scale shows the width of the local streets while the vertical scale shows the values of the residuals about my regression line. Clearly there is a pattern here. Narrow streets are safe, wide streets are safe up to a point, and widths in between are dangerous. The principal investigator was right. I had fit the wrong regression model to these data, and the residual plot revealed my mistake.

Figure 2 shows the histogram of these 14 residuals. With only fourteen values you are never going to detect a lack of fit with any probability model you might choose. There is virtually nothing to be learned from the histogram of the residuals simply because it does not show the residuals in the context of the problem.

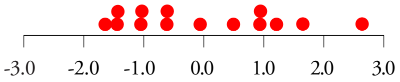

Figure 2: Histogram of Residuals for Midblock Accident Rates

As may be seen in this example, we analyze residuals to see if there are any discernible patterns in those residuals when they are arranged in order according to the corresponding values of any of the independent variables. It is relationships that are being sought. Now, admittedly, when there are no more relationships to be found, and when the residuals are nothing but noise, we might expect our residuals to have a mound-shaped histogram. With enough data we might even expect it to begin to look quasi-normal. However, while having a normal histogram is a *consequence* of having residuals that are pure noise, it cannot be taken to be an *indicator* that the residuals do not contain further relationships with the independent variables. This is a critical distinction that has deep roots in the history of statistics.

At the beginning of the Nineteenth Century Gauss and Laplace demonstrated that the normal distribution was the appropriate model for measurement error. This is why residuals that contain no further relationships with the independent variables will tend to have a mound-shaped histogram.

By 1840 Adolphe Quetelet had reversed the Gauss-Laplace theorem to conjecture that if the histogram is mound-shaped it must represent pure noise and no further investigation is necessary. Unfortunately, this conjecture of Quetelet's was, and still is, incorrect. This was conclusively demonstrated by Sir Frances Galton in 1875. This means that we cannot use the histogram of the residuals to infer that we have indeed found the right regression equation. This makes the placing of the residuals on a probability plot, or testing the residuals for a lack of fit simply a triumph of computation over common sense. The only strong results that can be obtained from such tests will always be inconclusive in terms of the regression analysis simply because they will lack context.

Analysis of Variance

With a set of experimental data where *n* observations on some response variable have been obtained for each of *k* treatments, the regression model degenerates into a step function model where a separate step is fitted to each treatment average. Since this step function will "fit" each treatment average perfectly, the "residuals" will simply be nothing more or less than the variation about the average within each treatment.

Treatment One	*4*	*5*	*5*	*4*	*8*	*4*	*3*	*7*	*Average = 5*
Treatment Two	*0*	*2*	*1*	*5*	*3*	*2*	*0*	*3*	*Average = 2*
Treatment Three	*6*	*9*	*9*	*7*	*8*	*7*	*9*	*9*	*Average = 8*

Figure 3: Data Set Two

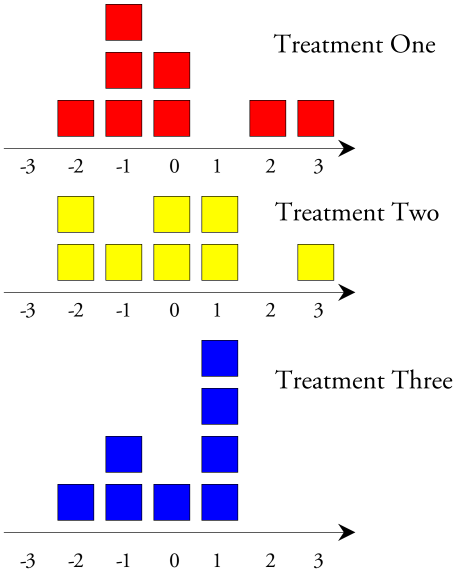

Figure 4: Residuals for ANOVA for Data Set Two

Here the treatment levels define a categorical independent variable, and so when we plot the residuals we are simply plotting the histograms for each subgroup shifted so that each histogram is centered on zero.

Since the step-function model has already fitted each treatment average perfectly there are no more degrees of freedom for lack of fit. No regression model can ever fit these data more precisely than the ANOVA model. So why analyze the residuals? About the only reason to analyze the residuals from an ANOVA model is to check for the presence of extraneous variables not included in the experiment (also known as assignable causes of exceptional variation.) The histogram in Figure 5 does not help in doing this. Testing for a lack of fit will not help find extraneous variables. Neither will plotting these values on a probability plot.

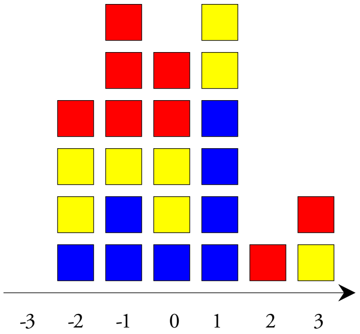

Figure 5: Histogram of Residuals for ANOVA for Data Set Two

The most effective way to check for the presence of extraneous variables is to plot these residuals on an *XmR* Chart and look for points outside the limits on the *X* Chart. This plot is shown in Figure 6.

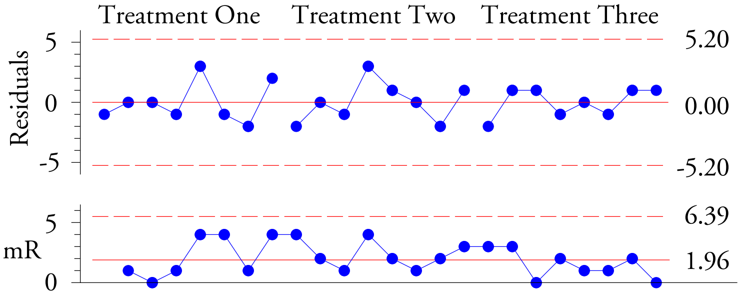

Figure 6: XmR Chart of Residuals for ANOVA for Data Set Two

Any points outside the limits on the *X* chart should be taken as evidence that one or more extraneous factors not considered in the experiment have a dominant effect upon the response variable. Such a finding should modify and inform your interpretation of the experimental results since you would know that you do not know the whole story. When there are no points outside the limits of the *X* chart for residuals, then you have no evidence of any extraneous factors, and the simple interpretation of your experimental results will be the appropriate interpretation.

So, as before, the analysis of the histogram of the residuals from an ANOVA model does not provide any useful results in terms of the analysis of the experimental results. It cannot tell you that your model is inadequate since you already have the best model possible for your data. Analyzing the histogram of the residuals is like rummaging around in the dust instead of looking at the relationships found by your ANOVA model. The important thing in both regression and ANOVA is to understand what the model tells you about the relationships between the inputs and the responses. The analysis of residuals is simply a secondary check used to see if you may have missed something. While patterns in the running record, or signals on the X chart may be helpful, there is little point in knowing that the residuals show a lack of fit with some probability model.

Summary

The whole point in the analysis of residuals is the discovery of patterns that fit in with the context of the data. Since you are a much more sophisticated pattern recognition device than any algorithm, we plot the residuals in context. As illustrated here, this will usually require running records organized according to the independent variables in the study or the levels of various treatment factors in the study. Whenever we create a histogram of the residuals we are ignoring this context and organizing the residuals according to their own magnitudes. This organization may result in a bell-shaped histogram, but this will not be indicative of anything regarding our analysis of the data.

So, should the residuals be normally distributed? Possibly.

Is it important if they are not normally distributed? Not really.

Is it important that the residuals are reasonably homogeneous? Absolutely.

Be careful that you do not confuse that which is unimportant with that which is important simply because the software lets you test the unimportant.