Demonstrating Least Squares Fitting Properties Using a Plot Equivalent to an F test

Frank J. Matejcik
Industrial Engineering
South Dakota School of Mines and Technology
501 E. St. Joseph Street, Rapid City, SD 57701 -3901
Phone (605) 394-6066 e-mail: Frank.Matejcik@sdsmt.edu

Roger W. Johnson
Mathematics and Computer Science
South Dakota School of Mines and Technology
501 E. St. Joseph Street, Rapid City, SD 57701 -3901

Keywords: graphics; plots; regression; outliers; residual plots; ancient tools

ABSTRACT
This paper describes a plot (Katya’s Triangle) equivalent to an F-test. The plot demonstrates some commonly taught properties of regression. Datasets, which are selected for their teaching value, are used as our regression examples. Additionally, the plot may be used to show the influence of particular observations and may be helpful for visual learners. The plot is derived using the Ancient Tools scheme.

INTRODUCTION
F tests for regression are generally presented with difficult mathematical notation. Multiple summations, multiple subscripts, and matrices are often avoided in college texts, and are challenging to undergraduates. The more traditional presentations of these tests can be supplemented by graphical calculations. These graphs may allow visual learners to gain fundamental insights more readily. Alternately, students comfortable with the notation will test their understanding with the graphs. Both groups of students should benefit from a parallel presentation of both graphs and traditional notation. Moreover, these graphs are new regression diagnostic plots.

This paper includes six additional sections. The Ancient Tools section describes background information about related plots and teaching methods. The next section describes the notation of regression and includes comments on the difficulty in teaching the subject. The core of the explanatory material is included in the section titled Katya’s triangle for regression. The following two sections show the use of Katya’s triangle used on demonstration data, and real data simple linear regression.

ANCIENT TOOLS
Matejcik (1997) and Matejcik & Heiberger (1997) have presented alternative presentations of statistics, which may have industrial applications and are appealing to educators. This approach relies on four geometric principles, and constructions using triangles and rulers. Both the geometry and the physical tools are thousands of years old, the approach is called Ancient Tools. Animated demonstrations of these geometric principles constructions are at http://silver.sdsmt.edu/~fmatejci/.

Ancient Tools methods may be used for many of the statistics taught in college courses including Statistics for Quality Control, First Statistics classes, and Design of Experiments. Additionally, Ancient Tools may be used to present elementary concepts: Addition, Subtraction, Multiplication, Division, Ratios, and Square Roots. Matejcik lectured on these topics while serving as a Fulbright lecturer in the Philippines. The site http://silver.sdsmt.edu/~fmatejci/ATPS.htm is a support document for Matejcik’s lecturers to high school and elementary school teachers in the Philippines.

An uncommon graphical tool that is frequently used in the Ancient Tools system is a repeated use of the Pythagorean theorem for the square root of the sum of squares. Figure 1 illustrates this property for the roots of 2, 3, and 4.

Figure 1: Repeated use of the Pythagorean Theorem
This paper makes extensive use of this property.

**NOTATION**

[Diagram of regression terms]

In our discussion we use these sums:

\[ SST = \sum (\bar{Y} - Y)^2, \quad SSR = \sum (\hat{Y} - \bar{Y})^2 \]

\[ SSE = \sum (\hat{Y} - Y)^2 \]

Mathematical notation in figure 2 and the preceding formulas have accents and the repeated use of the terms \( Y \) and \( SS \). A plot describing these mathematical relationships may aid some learners. Such a plot is described in the next section.

**KATYA’S TRIANGLE**

By recalling the Pythagorean theorem and \( SST = SSR + SSE \), we may form the right triangle in figure 3. Box, Hunter, and Hunter (1978) use a triangle like figure 3 to introduce ANOVA.

[Diagram of inner triangle]

Recall that \( SSR/SST = R^2 \), and that \( R^2 \) is equivalent to F for testing in regression. Observe now that the cosine of the left angle is \( \sqrt{R^2} \). By noting appropriate properties of the Arccosine function and the square root function, we argue that the left angle in the triangle in figure 3 is equivalent to F for testing in the ANOVA table for regression. Figure 4 adds lines to indicate the 95% and 99% significance test levels. Figure 4 has four line segments coming from its leftmost point and resembles the letter “K”, which motivates the name “Katya’s Triangle”. “Katya” is Frank Matejcik’s daughter.

[Diagram of Katya’s Triangle with decision lines]

Figure 4 is expanded using the property shown in figure 1 to display terms of each group \( \sqrt{SST} \), \( \sqrt{SSR} \), and \( \sqrt{SSE} \). Completed expanded plots are shown in figures 6, 11, 15, 17, 19, 23, and 27.

For Katya’s Triangle, \( \sqrt{\sum W_i^2} \) is graphically constructed as follows. Here the \( W_i \) values could be any real numbers. Draw a segment \( W_i \) in length from an origin. The direction of the segment need not be specified. At the terminating end of the segment draw a segment \( W_i \) in length perpendicular to the first segment. This second segment in the previous sentence is ambiguous because the segment may be drawn in two directions. Draw the \( W_i \) segment counterclockwise to the origin if \( W_i \) is positive. Conversely, draw the \( W_i \) segment clockwise to the origin if \( W_i \) is negative. Continue by drawing segments \( W_i \) in length for \( i = \{3, 4, \ldots, n\} \) perpendicular to the line formed by the origin and the previously terminated segment. Segments with positive \( W_i \) are drawn counterclockwise with respect to the origin, and segments with negative \( W_i \) are drawn clockwise with respect to the origin. Complete the plot by drawing line segment from the terminating point of the \( W_i \) segment to the origin. The length of this final segment is \( \sqrt{\sum W_i^2} \).

For Katya’s Triangle, first plot \( \sqrt{SSR} \) by letting \( Y_i - \bar{Y} \) and rotating the plot so that the resultant segment is horizontal with the origin at the left. Next, plot \( \sqrt{SSE} \) by letting \( Y_i - \bar{Y} \) and rotating the plot so that the resultant segment is vertical with the origin attached to the right end of the \( \sqrt{SSR} \) plot. Complete Katya’s Triangle by plotting \( \sqrt{SST} \) by letting \( Y_i - \bar{Y} \) and rotating and translating the plot so that the origin is attached to the \( \sqrt{SSR} \) origin and the triangle is completed as in Figures 3, 4, and 5. This structure creates many nearly symmetric relations.

Next, we discuss how Katya’s triangle looks in common introductory examples.

**DEMONSTRATION**

The following examples are used in Wilson (2002) to demonstrate difficulties in regression. The original data is from Anscombe(1973). The cases include a case with correct
assumptions, a case with perfect line except for one outlier, a parabolic case, and an unbalanced case. The conditions are visible in Katya’s triangle.

The figures 6 through 9 describe a good case, which includes three residual plots, and a scatter plot to verify that the application is appropriate. Figure 6 is a Katya’s triangle plot with colors added. If colors are not rendered correctly as you read this document check [http://www.hpcnet.org/AncientTools](http://www.hpcnet.org/AncientTools) for a corrected version.

The color scheme in Katya’s triangle is shown in figure 7 for the first time. The polygon associated with $\sqrt{SST}$ is in black and on top to imply “total”. The polygon associated with $\sqrt{SSR}$ is in blue and on the bottom. The polygon associated with $\sqrt{SSE}$ is in red and on the right to imply “residuals”. The 99% and 95% significance line are shown in green and yellow, respectively. Additionally, the shape of Katya’s triangle may be checked.

When the fit is good, Katya’s triangle often looks like a bird, because there is only one sign change in the black and blue polygons and they tend to mirror each other. The residual polygon generally has segments that change direction (see figure 6), and a large portion of a residual plot with segments in the same direction shows a problem in the assumptions.

Figures 11 through 12 show data that are perfectly linear except for one outlier. In both figures 11 and 12 there are noticeably long segments in the black and red polygons. Also, large arcs are seen in the red polygons of figures 11 and 12.

Figures 13 through 15 show data on a parabola rather than a line. Figure 15 shows a large arc in the red polygon showing problem with the assumptions.
The whiskey data shows a subtle higher order function in figure 18. The red polygon in figure 19 shows a large arc. However, in figure 19 the red polygon is small. Figure 20 is an expanded view of the red polygon in figure 19, in figure 20 the polygon is rotated one quarter turn clockwise from that of figure 19.

Figures 16 and 17 show unbalanced data, which also can be observed Katya’s triangle. Figure 17 is shows a clear pattern with two large arcs in the red polygon. Additionally, figure 18 shows an unusual pattern with a large blue arc and one long line segment in the blue polygon.

Figures 21 through 28 use data from the record times in the 1500 meter track event. The year 1900 is plotted at year zero. Figure 21 shows an outlier in the scatter diagram. The residual polygon in figures 23 and 24 show the outlier, and also large arcs indicating additional problems with the assumptions. The large arcs indicate that many residuals of the same sign occur in series. This characteristic of the residual is much more subtle in figure 22 the usual residual plot.

**EXAMPLES**

The following examples demonstrate the use of Katya’s triangle with some real data. The first example shows alcohol content in as a function of age of whiskey. The second data set shows record times in the 1500 meter track event as a function the year the records were set.
SUMMARY

We have demonstrated that plots describing the calculations for the F tests for regression may be prepared using Ancient Tools techniques. Additionally, we have demonstrated their use with introductory and real data simple linear regression examples. We have observed that new plots make some problems more visible than they appear in traditional plots.

REFERENCES


Frank J. Matejcik and Clare A. Heiberger
Results of teaching SPC with ancient tools to jewelry workers Proceeding of the Fourth International Applied Statistics in Industry Conference June 3-5, 1997, Kansas City, Missouri.

Frank J. Matejcik, Teaching SPC with ancient tools, 44th annual Minnesota Quality Conference March 11-12, 1997 Minneapolis, Minnesota