Q^2 : A Measure of Linearity

Eric Marland
marlandes@appstate.edu
Gregory Rhoads
rhoadsgs@appstate.edu
Michael J. Bossé
bossemj@appstate.edu
José Almer Sanqui
sanquijat@appstate.edu
William Bauldry
bauldrymc@appstate.edu
Department of Mathematical Sciences
Appalachian State University
28608
USA

Abstract

In determining if a set of bivariate data can be accurately modeled by a linear function, one could use linear regression and the value of R^2. Unfortunately, some published resources have been found to incorrectly interpret a high value of R^2 as evidence of a linear relationship between the variables. Indeed, at times the values of the variables may be independent of each other and a linear regression may not be appropriate. Herein, using only mathematics from grades 10-14, we propose a novel measure, Q^2, to indicate the measure of linearity of a scatterplot of points. While Q^2 shares many of the properties of R^2, Q^2 is invariant under rotation, and so is a more appropriate tool to compare two independent data sets for linearity. Herein, rather than presenting either Q^2 or R^2 as superior to the other, we propose the complementary nature of the two measures and that, by investigating Q^2, students can gain deeper understanding of R^2. This paper provides a link to a dynamic applet and instructions to accompany the reading and assist the reader to further investigate this topic and glean additional insights.

1 Introduction

Let's say that we are given a scatter plot of bivariate data. We perform a vertical least squares linear regression on the data set, determine the line of best fit, and find the associated value for R^2. We now

rotate the entire scatter plot about the centroid of the data set, $(\bar{x}, \bar{y})$, by an angle θ. The vertical least squares regression line of the rotated data still passes through the centroid and so is a rotation of the regression line of the original data. Some immediate observations can be made [2]:

1. The "shape" of the scatter plot has not changed.
2. The rotation angle of the regression line has no direct relationship to θ.
3. The value of R^2 changes as the data are rotated.

While these findings may be interesting, one may query when or why one would ever rotate a set of data. The answer may not be in rotating the data as much as in orienting the data on a coordinate grid. Let us consider the following scenario:

> You've lost contact with an unmanned aircraft that has crashed over a flat expanse of uninhabited land. You need to find a particular artifact from the debris and so you send a reconnaissance aircraft to take pictures of the debris field of the crash site. However, the debris field has no natural coordinate system by which to contextualize the location of each visible piece of debris (data point). In order to determine a reasonable search area for the artifact, two interconnected ideas come to play: the linearity of the data (of the scatter plot of the debris field) and the shape of the debris field. If the unmanned aircraft had a relatively mild angle of descent, the debris field may indeed be quite linear. However, the selection of one coordinate system on which the data points will be mapped may produce a least squares regression line with a large R^2 value while a different coordinate system may produce a different regression line with a small R^2 value. One must now decide which coordinate system is best on which to map the debris field. The selection of the coordinate system and its associated regression line and value for R^2 can help to hone in on possible locations for the artifact.

In this scenario, rotating the coordinate system produces the same effect as keeping the coordinate system static and rotating the data. Thus, in modeling real world phenomena, rotating the data may be a valid heuristic. Altogether, these observations lead to some important additional understandings [2]:

4. R^2 is not rotationally invariant and
5. since the "shape" of the scatter plot has not changed - and thus its linearity has not been affected - R^2 cannot be a measure of the linearity of the data, as is improperly reported in some resources [2].

This immediately leads to the natural question: If R^2 is not a measure of the linearity of the data, then what is? Prior to considering this question, one may ask why we may want a measure of linearity on a data set. In respect to the debris field, determining the linearity of the data could have provided hints regarding how wide the crash site should be investigated in order to find the artifact sans concerns for selecting one particular coordinate system over another.

In an introductory statistics class, students are often asked to perform a linear regression on bivariate data. However, when data is nonlinear, a linear regression may be inappropriate. Preceding the

practice of performing a linear regression, it may be valuable to employ a test of linearity such as the Q^2 measure developed here. Additionally, understanding Q^2 may help students better conceptualize the meaning and application of R^2.

The remainder of this paper develops the measure Q^2 which will determine the linearity of a set of data. Later in this paper, the reader is provided a link to a dynamic applet to further investigate ideas associated with R^2 and Q^2. While beyond the scope of this paper, future applications of Q^2 may be determined to meet needs in the areas of pattern recognition and computer science.

2 Initial Statistical Background and the Comparison Line

In addition to measures of central tendency, fitting a line to a scatterplot of data (either by hand or using technology) is one of the earliest exposures students have to statistics. Connected to the regression line, students investigate and interpret the coefficient of determination (a measure of how much error one would expect when using the linear model to make predictions), R^2, and the Pearson correlation coefficient, r, which measures the linear dependence between the two variables x and y. While the value of R^2 has several valid interpretations, for this discussion it is useful to recognize that R^2 can be seen as the percentage reduction in the sum of squared distances by using the regression line over a comparison line (either a vertical or horizontal line). We now consider this idea in more detail, and how it leads to a definition of Q^2, through three cases:

1. In the case when all errors are in the y-values of a data set, it is appropriate to determine the vertical regression line and compare the sum of the squares of the vertical distances from the regression line to the sum of the squares of the vertical distances from the comparison line $y = \bar{y}$. (See Figure 1.)

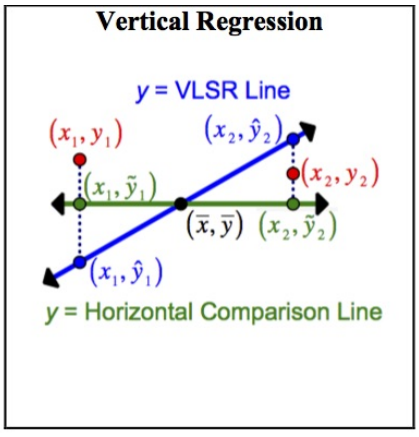

Figure 1: Vertical Regression

2. Similarly, in the case when all errors are in the x-values of the data, it is appropriate to determine the horizontal regression line and compare the sum of the squares of the horizontal distances from the regression line to the sum of the squares of the horizontal distances from the comparison line $x = \bar{x}$.

Prior to considering the other case, it is important to observe a few relationships between the data errors, the regression line, and the comparison line:

a. The comparison line is perpendicular to the direction of the error in the data.

b. The regression line and the comparison line intersect at the data's centroid.

In order to develop the measure Q^2, we will ensure that the appropriate regression line and the comparison lines: (a) are perpendicular; (b) intersect at the centroid; and (c) are invariant on the rotation of the data. These characteristics make the selection quite natural.

3. Some data has recognized possible errors in both x- and y-values. For instance, students are given a stop watch to measure the distance a ball rolls after descending a ramp. Let us assume that the x-component of the data is time and the y-component is distance. We readily accept that there may be some error in our observations of the distance the ball rolled at x seconds. However, due to the human imprecision of manipulating the stopwatch, there may also be some error in the actual versus recorded time at any value of x. Based on the authors' former experiences in laboratory work, this error could be quite large. Thus, error can be assumed in both variables.

When we can know that (A) error is in both directions with equal variance or (B) there is no assumable information regarding error in either direction, it is appropriate to use an orthogonal least squares regression line, which considers the sum of the squares of the orthogonal distances from the regression line to the sum of the squares of the orthogonal distances from the line perpendicular to the orthogonal regression line through the centroid (See Figure 2). It is appropriate to determine the orthogonal regression line and compare.

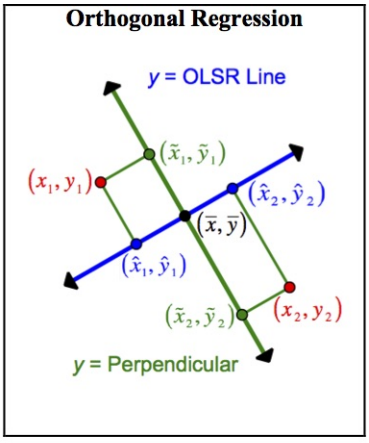

Figure 2: Orthogonal Regression

In addition, when there is no natural set of axes, the orthogonal least squares regression line may also be more appropriate than the traditional vertical linear regression model, since error in a particular variable may not hold definite meaning. Notably, rotation of the data by angle θ about the centroid will produce a rotation in the orthogonal regression line by the same angle [2]. However, the

orthogonal least squares regression line does not have a recognized associated R^2 value as a measure of the goodness of fit of the regression line.

In respect to an orthogonal least squares regression, [1] state that the worst-fit line is perpendicular to the best-fit line (i.e., orthogonal least squares regression). This means the smallest sum of squared orthogonal distances from any line through the centroid is the orthogonal least squares line, and the largest sum of squared distances will be the perpendicular line.

Employing the power of the orthogonal regression line, our proposed measurement of linearity, Q^2, will compare the minimum and maximum sum of squared orthogonal distances and will ensure that Q^2 is rotationally invariant - a feat which R^2 could not perform.

3 Developing Q^2 as the Measure of Linearity

Before defining Q^2 as a measure of linearity, let's first consider the definition and meaning of R^2. Assume we are given data points (x_i, y_i), and denote the means of the x-values and y-values by $\bar{x}$ and $\bar{y}$ respectively. The sum of squares of deviations from the mean is $SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$. If we let $\hat{y}_i$ be the predicted values at x_i from the vertical linear regression line, the sum of squares of the errors from the regression line are $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. We then define $R^2 = \dfrac{SS_{yy} - SSE}{SS_{yy}} = 1 - \dfrac{SSE}{SS_{yy}}$. In calculating R^2, we use only the data at our disposal to make predictions and use the mean of the y-values, $\bar{y}$, as the predicted y-value for any given x-value. In other words, we use the line $y = \bar{y}$ to make predictions. Then SS_{yy} is the sum of squared vertical distances from our data points to the line $y = \bar{y}$ and SSE is the sum of squared vertical distances from our data points to the vertical regression line. Thus, $SS_{yy} - SSE$ is the reduction in the sum of squared vertical distances by using the regression line, and, therefore, R^2 is the percentage reduction in the sum of squared vertical distances by using the regression line from the sum of squared vertical distances using the line $y = \bar{y}$.

If we wish to develop a linearity measure, which we call Q^2, it should have properties paralleling those of R^2 for vertical regression.

1. The value of Q^2 should lie between 0 and 1.
2. If $Q^2 \approx 0$, there should be very little confidence that the data are linear.
3. If $Q^2 \approx 1$, the data values are very linear (close to the orthogonal regression line).
4. The value of Q^2 should be invariant under rotation of the data about the centroid, since the linearity of the data is independent of rotation.

Let's consider two examples below. In Data Set 1, the data points do not appear to be linear while in Data Set 2, the data points appear to be well represented by a line. Therefore, we would expect the value of Q^2 to be smaller for Data Set 1 than 2. Hence, Data Set 2 should have a larger percentage improvement of the sum of squares over the comparison line (defined below), Data Set 1 will have a negligible improvement in the sum of squares over the comparison line.

Figure 3: Data Sets 1 and 2

4 Defining Q^2

Based on the interpretation of R^2 and the four desired properties, we define Q^2 in the following manner:

> Q^2 is the percentage reduction in the sum of squared orthogonal distances using the orthogonal regression line as opposed to using the line perpendicular to the orthogonal regression line through the centroid of the data.

In order to generate a formula for Q^2, assume we have data points (x_i, y_i) and define the sum of the squared variances and the covariance as:

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ and } SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

In [3], we find that if $y = \beta_1 x + \beta_0$ is the orthogonal regression line, then

$$\beta_1 = \frac{SS_{yy} - SS_{xx} + \sqrt{(SS_{yy} - SS_{xx})^2 + 4SS_{xy}}}{2SS_{xy}} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

.

This demonstrates that the orthogonal regression line can be calculated in closed form from the data alone. This important point will later be discussed in more detail.

Now, let $(\hat{x}_i, \hat{y}_i)$ be the point on the orthogonal regression line closest to (x_i, y_i) and $(\tilde{x}_i, \tilde{y}_i)$ be the point on the line perpendicular to the orthogonal regression line through the centroid closest to (x_i, y_i) (See Figure 2). We can now define

$$SS_{orth} = \sum_{i=1}^{n}\left\{(\hat{y} - y_i)^2 + (\hat{x} - x_i)^2\right\} \text{ and } SS_{perp} = \sum_{i=1}^{n}\left\{(\tilde{y} - y_i)^2 + (\tilde{x} - x_i)^2\right\}.$$

Then $Q^2 = \dfrac{SS_{perp} - SS_{orth}}{SS_{perp}} = 1 - \dfrac{SS_{orth}}{SS_{perp}}$.

5 Applying Q^2 to Data Sets

Figure 4 demonstrates the orthogonal least squares regression line (OLSR Line) and comparison line (Comp Line) for Data Sets 3, 4, and 5 with the associated values of Q^2 and R^2. Notably, as might be expected, the low values for Q^2 demonstrate that Data Sets 3 and 4 are quite nonlinear and the high value for Q^2 for Data Set 5 reveals that Data Set 5 is very linear, even though the value of R^2, based on the vertical least squares regression line (VLSR line), is very low. Notice that Data Set 3 is the same as Data Set 1, and Data Set 5 was generated by rotating Data Set 2 around the centroid until the orthogonal regression line is almost vertical. The value of Q^2 is the same for Data Sets 2 and 5 even though the value of R^2 is lower for Data Set 5 ($R^2 = 0.00$) than Data Set 2 ($R^2 = 0.94$). So, we verify that Data Set 2 is more linear than Data Set 1 as reflected by the higher Q^2 value.

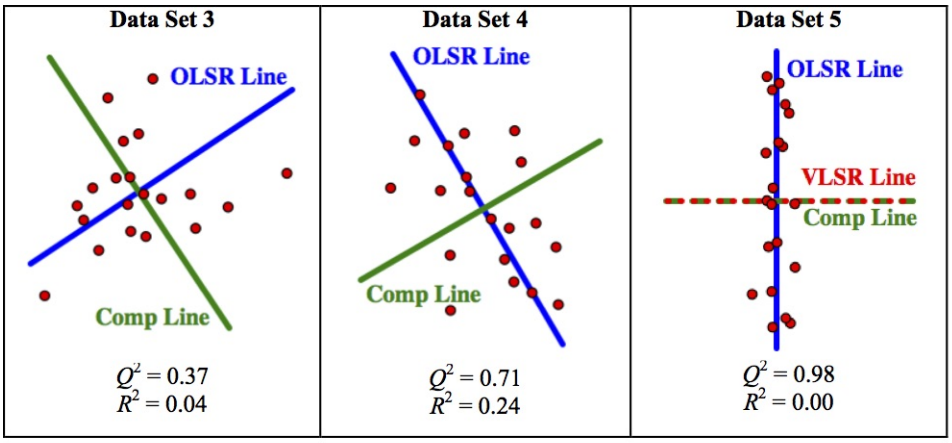

Figure 4: Data Sets 1 and 2 with regression and comparison lines

At this point, the reader might wonder if there is any significant value to Q^2 over the well-known R^2. Let us again consider Data Set 5 in Figure 4. We immediately recognize that the value of R^2 is quite low, despite the data being quite linear. (We later discuss why $Q^2 = 0.71$ is considered "low".) Furthermore, the vertical least squares regression line notably departs from the trend of the data. In fact, the steeper a data trend becomes, the worse the vertical least squares regression line represents the trend in the data [2]. However, the orthogonal least squares regression line better demonstrates the trend of the data and the value of Q^2 does not change regardless of the rotation of the data.

6 Experimenting with an Applet

To enhance the reader's experience, we have provided a dynamic applet to accompany the reading and assist the reader to further investigate this topic and glean additional insights. In order to open the applet, you may need the latest version of the free software Maple Player.

To use the applet, first click on "Click to Initialize". Second, use the slider to adjust the dispersion of the random points. Third, from the drop-down menu, select the number of points desired in the data plot. Fourth, select if you wish your points to be randomly generated or randomly perturbed about a randomly selected line. Fifth, using the check boxes, select if you wish to see the vertical or orthogonal regression lines or both. Select radians or degrees for an angle by which to rotate the data.

You can now drag the needle on the rotation tool to rotate the data about their centroid and observe the effects on the values of R^2 and Q^2. If the data are too compressed or land outside of the viewing screen, you can use the slider to zoom in or out of the graph to better see the data.

Here are some points to observe:

1. The orthogonal regression line rotates precisely with the data and the value of Q^2 remains invariant in respect to the angle of rotation.
2. The vertical regression line does not rotate precisely with the data and avoids being too steep and the value of R^2 changes in respect to the rotation of the data. At times, the vertical regression line quite poorly represents the trend in the data.
3. Under some conditions, the orthogonal and vertical regression lines are quite close. This seems to occurs when the trend of the data is nearly horizontal.
4. It seems that R^2 is maximized when the rotated data has a trend with a slope relatively close to ± 1.

The reader can use the applet to investigate other ideas posed in this paper. Through experimenting with this applet, it is anticipated that introductory statistics students will come to better understand Q^2 as well as the R^2 value that they had previously encountered. Instructors can use the applet to have students investigate the ideas posed above along with others. Altogether, the applet will convince the user of both the distinction between Q^2 and R^2 and the value of Q^2 as a measure of the linearity of the data.

7 Justification of the Properties of Q^2

1. Since SS_{orth} and SS_{perp} are the minimum and maximum sum of squared distances respectively, $0 \leq \dfrac{SS_{orth}}{SS_{perp}} \leq 1$, so $0 \leq Q^2 \leq 1$.
2. If $Q^2 \approx 0$, then $SS_{orth} \approx SS_{perp}$, so all lines through the centroid have effectively the same sum of squared distances so the orthogonal regression is no better fit than any other line through the centroid. This means the data points are approximately circular.
3. If $Q^2 \approx 1$, then $SS_{orth} << SS_{perp}$, which means the orthogonal regression line is a significantly better fit than the perpendicular line suggesting the data points are very close to the orthogonal line and, hence, are highly linear.
4. Since the calculation of Q^2 uses the orthogonal regression line and that line is invariant under rotation, the value of Q^2 is also invariant under rotation.

8 Is $Q^2 = 0.71$ low?

When $R^2 = 0.71$, one would often say that the vertical linear regression quite well represents the data. However, we must remember that R^2 is the measure of the quality of the vertical regression line

in comparison to $y = \bar{y}$ (the horizontal line through the centroid of the data) and that $y = \bar{y}$ would rarely be a line of worst fit. Q^2, however, is the comparison of the line of best fit to the line of worst fit. Thus, Q^2 will generally be greater than R^2 (as shown in Figure 4).

Herein lies a problem with both Q^2 and R^2: How do we interpret the specific value? There is no universal agreement on which values of R^2 imply a strong, medium or weak relationship, as many different factors are involved. There is a similar danger to Q^2, to say that a value above, say, 0.9 would indicate a strongly linear data set or that it is twice as linear as a data set with a Q^2 value of 0.45. Nonetheless, we can recognize that as Q^2 approaches 1, the data are more linear and that a higher value represents data that is more linear than a lower value. It can be argued that the ranges of low, medium, and high values for Q^2 will, in the future, be determined more by the application or relationship being investigated and the desire of the investigator than by some prescribed definition for these ranges.

9 Discussion, Implications, and Conclusion

Determining how a given curve or a given set of points resemble a curve of a particular type is a problem that has been actively studied. Facial recognition is an application that requires one to find the best match of an attribute (ear shape, facial profile, etc ...) from among a set of candidate shapes. In [5], the authors describe a measure of linearity for a 2-d curve of unit length by computing the sum of the Euclidean distances between the endpoints and the centroid of the curve.

Our paper describes a measure to specifically determine how close a set of points in two dimensions is to a line. In [6], the authors describe other measures of linearity for a two-dimensional data set that are both translationally and rotationally invariant. The *average orientations* method computes the normal direction to the line through two random points and compares the average of these directions with the normal direction to the angle of orientation of the data set. The *triangle heights algorithm* takes three random points and computes a normalized height of the triangle formed by the three points and averages over many triangles. Both of these methods rely on samples of points rather than looking at all data values, so their value will depend on the samples chosen. The *rotation/correlation method* rotates the data so the angle of orientation is 45^o, and computes the Pearson correlation coefficient of the rotated data. Then it computes the correlation coefficient if the data is rotated another 90^o and chooses the larger of the two values. A number close to 1 indicates linearity. Our method utilizes fewer calculations. Altogether, our measure of linearity seems to have some advantages over other extant measures.

This problem of R^2 not being rotationally invariant is another reason why it is important to always look at the scatter plot of the data before interpreting its value, aside from potential outlier effect and the implications derived from an understanding of the source and nature of the data. In the future, it will be interesting to investigate the sensitivity of Q^2 to outliers compared to R^2.

If we return to our original question, where we were investigating the debris field of a crash of an unmanned aircraft and were searching for a particular artifact amongst the debris, we can see we have made some progress on a solution. The orthogonal regression line is the best approximation to the direction of impact to center the search, and the value of Q^2 is a measure of the spread of the debris from the impact direction. A high value of Q^2 would indicate a debris field close to the impact direction and the search should remain close to the impact line. A low value of Q^2 would indicate a

debris field with a larger spread from the impact direction, and the search should expand further from the orthogonal regression line.

It is important to clearly recognize one more aspect of Q^2. We have used Q^2 in respect to an orthogonal regression, because the orthogonal least squares regression line rotates appropriately with the data and Q^2 is rotationally invariant. However, much more significantly, since Q^2 was calculated in closed form based on the data alone, Q^2 can be recognized as generalizable to determine the linearity of any data set irrespective of the assumption of error in the measurement of the data in any direction. Thus, whether the recognized possible errors in measurements in a data set are in the y-direction (befitting a vertical regression), in the x-direction (befitting a horizontal regression), in both directions with equal variances (befitting an orthogonal regression), or in both directions with variances in a constant ratio not equal to 1 (befitting a Deming least squares linear regression, see [4]), Q^2 provides an applicable measure of the linearity of the data set.

Upon being introduced to the notion of Q^2, some student may naturally ask "Is Q^2 always better than R^2?" This question demonstrates some lingering misconceptions regarding R^2. These measures are of different characteristics: Q^2 is a measure of the linearity of the data set, and R^2 is a measure of the quality of the linear model representing the data. Rather than seeing one measure as superior to the other, we hope that students can recognize the complementary nature of the two measures. While we have previously stated that R^2 is neither a measure of the linearity of the data nor rotationally invariant, the greatest weakness in R^2 lies in when it is inappropriately employed or interpreted. Similarly, Q^2 can suffer from similar weaknesses when it is misinterpreted and misapplied.

Statistics instructors can provide data sets and have students use Q^2 to determine the linearity of each set. Then, as an open ended discussion, students can debate which data sets, based on their respective Q^2 value, warrant performing a linear regression and determining the value of R^2. As part of this discussion, they can debate ranges for which Q^2 could be considered low, medium, and high, and at what value is it justified that the data warrants a linear regression.

Measuring the linearity of a set of data points taken from a curve has been an important part of image processing. Various measures for this were compared in [6]. As in this paper, it seems that one of the greatest benefits of Q^2 is for comparing two data sets to see which is more linear rather than trying to quantify how linear a particular data set may be. An interesting extension to this investigation may be to modify this calculation to determine how linear a set of points in n-dimensional space may be. While this paper proposes a measure for the linearity of a data set, the future may produce rich and valuable applications of this measure to the fields of pattern recognition and computer science.

10 Bibliography

References

[1] Alciatore, D., and Miranda, R. (1995). The best least-squares line fit. Graphics Gems V, (91-97). DOI:10.1016/b978-0-12-543457-7.50022-x

[2] Bossé, M.J., Marland, E.S., Rhoads, G.S., and Rudziewicz, M., (2016). Searching for the Black Box: Misconceptions of Linearity. CHANCE. 29:4, 14-23, DOI: 10.1080/09332480.2016.1263094.

[3] Carroll, R. J., and Ruppert, D. (1996). The Use and Misuse of Orthogonal Regression in Linear Errors-in-Variables Models. The American Statistician, 50(1), 1?6. doi:10.1080/00031305.1996.10473533

[4] Fuller, W.A. (1987), Measurement Error Models. New York: John Wiley.

[5] Rosin, P., Pantović, J., Žunić, J., (2016). Measuring linearity of curves in 2D and 3D. Pattern Recognition, 49, (65-78). DOI:10.1016/j.patcog.2015.07.011

[6] Stojmenović, M., Nayak, A., and Žunić, J. (2006). Measuring Linearity of a Finite Set of Points. 2006 IEEE Conference on Cybernetics and Intelligent Systems. DOI:10.1109/iccis.2006.252284