Reduced Major Axis Regression to Improve Oil & Gas Pipeline Integrity

William V. Harper, PhD, PE, Otterbein University, Mathematical Sciences, Westerville, Ohio, USA, wharper@otterbein.edu

Neil A. Bates, P. Eng., Det Norske Veritas (Canada) Ltd., Calgary, Alberta, Canada, neil.bates@dnv.com

ABSTRACT

The theoretical underpinnings of standard least squares regression analysis are based on the assumption that the independent variable (often thought of as x) is measured without error as a design variable. The dependent variable (often labeled y) is modeled as having uncertainty or error. Pipeline companies use the inline inspection (ILI) metal loss depth as the x variable and field measured excavation depth as the y variable. Both measurements have sources of error. Thus the underlying least squares regression assumptions are violated. Often one common result is a regression line that has a slope much less than the ideal 1-1 relationship.

Reduced Major Axis (RMA) Regression is specifically formulated to handle errors in both the x and y variables. It is not commonly found in the standard literature but has a long pedigree including the 1995 text book Biometry by Sokal and Rohlf in which it appears under the title of Model II regression. In this paper we demonstrate the potential improvements brought about by RMA regression.

BACKGROUND

ILI data is based on assessment made by what is commonly referred to as a smart pigging device. Typically these use some form of ultrasonic or magnetic flux signals through pipeline walls as the pigs migrate down the oil or gas pipeline. Based on the resulting signal response, potential anomalies of various types (e.g., corrosion) are identified and sized. This is not a static measurement as it comes from data that a pigging device collects while travelling down the line. This data is then analyzed by both a computer algorithm and human beings to predict the size (i.e. length, width, and depth) of the metal loss. It has multiple sources of measurement and identification errors.

When parts of the underground pipeline are excavated, field measurements are made of any metal loss due to corrosion uncovered. These are also subject to error both from the field measurement tools used as well as human error. Nonetheless the excavation data is the current “gold standard” in the pipeline industry.

Typically only a small portion of all the potential anomalies reported by the ILI tool can be excavated. The limited field measurements are paired with the associated ILI calls as best possible. Sometimes there are multiple ILI calls in an area and it is not always clear which ILI call should be paired with the field excavation measurement. From the matched ILI to field data a regression relationship is desired that can then be applied to the usually thousands of ILI calls that were not dug up and measured.

The importance of comparing in-line inspection calls to field measured excavation data should not be underestimated. Neither should it be undertaken without a solid

understanding of the methodologies being employed. Such a comparison is not only a key part of assessing how well the ILI tool performed, but also for any subsequent use of the ILI data. The use of the matched ILI and field data regression analysis are commonly used to provide the basis for assessing the likelihood of leaks or failures from unexcavated ILI calls. Additional methods such as statistically active corrosion [5] and probability of exceedance [6] help develop an integrity maintenance plan for the pipeline operator.

REGRESSION ISSUE BACKGROUND

Both field depth measurements and the corresponding ILI estimated depths can have error as discussed above. Excavations are expensive and such data should be given a proper treatment. Almost all linear regression analysis performed in the world is based on a least squares methodology that has much to offer. However the assumptions underlying such applications are commonly ignored. In doing so the resulting regression fits are sometimes disappointing and may not be appropriate for the intended modeling.

With both variables subject to error the underlying least squares regression assumptions are violated. Often one common result is a regression line that has a slope much less than the ideal 1-1 relationship between the ILI estimated depth of an anomaly (such as a corrosion pit) compared to its field measurement depth. Reduced Major Axis (RMA) Regression is one method specifically formulated to handle errors in both the x and y variables. It is not commonly found in the introductory regression literature but has a long pedigree including the 1995 text book *Biometry* by Sokal and Rohlf in which it appears under the title of Model II regression. This paper demonstrates the potential improvements brought about by RMA regression for this pipeline application.

Building on a solid comparison between ILI data and excavations provides the foundation for hopefully more accurate predictions and management plans that reliably provide longer range planning. This may also result in cost savings as the time between ILI runs might be lengthened due to a better analysis of such important data.

REGRESSION APPROACHES

Least Squares Regression

In many commercial spreadsheet programs and major statistical packages, least-squares is the default method for performing a linear regression. Least squares regression minimizes the sum of squared deviations (errors) of the vertical distance between the actual y values and their corresponding predictions, typically termed *y-hat* where the *hat* implies an estimate. A key assumption in such a design is that the independent variable x is measured without error. Often in pipeline integrity the horizontal or x predictor variable is the ILI call and the vertical or y variable is the matched field measurement. Both the ILI call (x) and the field measurement (y) in this case are subject to error. Thus from a theoretical statistics perspective, there are problems using least squares regression for such modeling efforts.

Reduced Major Axis Regression

Reduced Major Axis (RMA) has its roots in various fields including biological applications. For example from fish capture data the biologist may want to develop a

predictive model to predict fish weight for a given breed based on its length or vice versa. However in this case both weight and length are subject to errors when trying to collect data from live fish. Similarly, metal loss data will have error in both the tool reported depth and field measured depth [1]. The RMA approach can be found in the text book by Sokal & Rohlf (1995) [2]. Other names that may appear in the literature for RMA are geometric mean regression, least products regression, diagonal regression, line of organic correlation, and the least areas line (Wikipedia, January 3, 2012) [3].

A common experience for fitting a least squares regression predicting field measurements as a function of ILI calls is that the resulting model under-predicts deeper calls and over-predicts shallow calls. This is reminiscent of the problem of regression to the mean (Galton, F., 1886) [4] although this paper will not pursue the impact of this interesting but different issue with respect to pipeline integrity. It is a question as to whether this result of over and under predicting is a reasonable match for reality or whether it is an artifact of the methodology employed.

While one desires an accurate predictive model so that ILI calls for corrosion pits that have not been excavated can be reasonably quantified, a common desire among many integrity engineers is to be conservative. For predicting pit depths, a model that generally under-predicts deeper pits is non-conservative.
RMA minimizes the sum of the areas (thus using both vertical and horizontal distances of the data points from the resulting line) rather than the least squares sum of squared vertical distances. One of the issues with standard least squares regression is the inability to treat the least squares regression equation $y = a + bx$ as an ordinary equation and back solve to obtain an equation that predicts x from y. With least squares, when one interchanges the x and y variables, the resulting regression equation is not the equivalent of $x = (y - a)/b$. Additionally, doing so with least squares results in the paradox of similar over and under prediction when the variables are interchanged. With an RMA equation, one can perform this simple algebraic feat as it will match the equation RMA one would obtain with the variables interchanged - i.e., the resulting RMA regression is the equivalent of $x = (y - a)/b$.

EXAMPLE - LEAST SQUARES VERSUS RMA

This paper presents illustrates the differences in various aspects of least squares versus RMA. The example given consists of a relatively large data set of matched excavation pit depths and the ILI calls. The data set has 1,812 ILI external pit depths from a single ILI run that have been matched with excavation field data and is much larger than most samples typically available. The authors' routine can be used to compute both least squares and RMA estimates. Table 1 shows the y-intercept and the slope for both approaches.

Coefficient	Traditional Least Squares	Reduced Major Axis
Intercept	0.096220	-0.00225
Slope	0.501700	1.070149

Table 1. Least Squares, RMA regression coefficients.

In this example, which is fairly typical of least squares for such data, results in the expected field measured *y* (called "Pit Depth (%)" for this application) regression equation y = 0.09622 + 0.5017 * x where *x* is the ILI %Depth (% of wall thickness). At this point, it is worthwhile to examine the issues associated with the regression equation. In many applications, pig calls are not reported (or filtered out) if they are less than some threshold such as 10%. Assuming this is a reasonable lower bound the least squares regression equation y = 0.09622 + 0.5017x will predict no values less than approximately 14.6% wall thickness. If there was a pig call of 100%, the least squares equation only predicts 59.8%. This is a concern that is too often overlooked. The following figures will better illustrate some aspects of this issue.

The RMA equation is y = -0.00225 + 1.070149 * x. For a 10% ILI call, the RMA predicted value is 10.5% and for a 100% ILI call the predicted value is 106.8%. While a wall thickness greater than 100% is not possible, one starts to see that, at least in this example, the RMA covers a predictive range of importance and is not limited to such a tight interval as will be shown more explicitly in plots that follow. Instead of showing both axes ranging from the possible full range from 0% to 100%, Figure 1 focuses on the actual range of the values in the data to provide a more detailed view in which the 1,812 pairings reside. *YHat_RMA* is the predicted RMA field depth while *YHat_Trad* is the traditional least-squares prediction for the field depth.

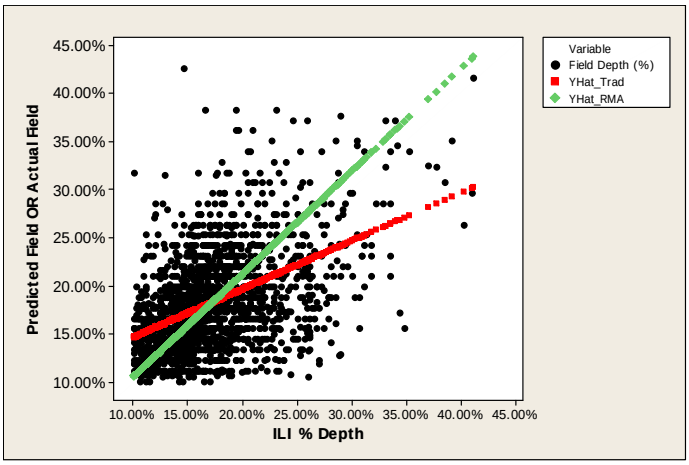

Figure 1. Predicted Least-squares and RMA Regression over the range of the ILI calls. *YHat_RMA* is the predicted RMA field depth while *YHat_Trad* is the traditional least-squares prediction for the field depth.

Figure 2 shows box plots for the following:

1. Y variable: Field Depth (%) which is the field measurement
2. X variable: ILI % Depth
3. YHat_RMA: predicted y using RMA
4. YHat_Trad: traditional predicted y using least squares

Figure 2 shows in a much clearer fashion the concern listed in the prior paragraph. The range of predictions for the least squares regression is much too narrow to adequately model the field measured pit depths. Each of the four items is shown with a box plot. The lower part of the box is the 25th percentile, the middle line is the median or 50th percentile, and the top line is the 75th percentile. The lines (known as whiskers) extending out of the box go to the most extreme values that are not potential outliers.

Any potential outliers (per the box plot methodology) are represented by asterisks (*). Note the unrealistic small range of distribution in the traditional least squares regression (*YHat_Trad*) versus the other three plots. Note also the predicted RMA regression distribution (*YHat_RMA*) is fairly similar to the distribution of Field Depth(%).

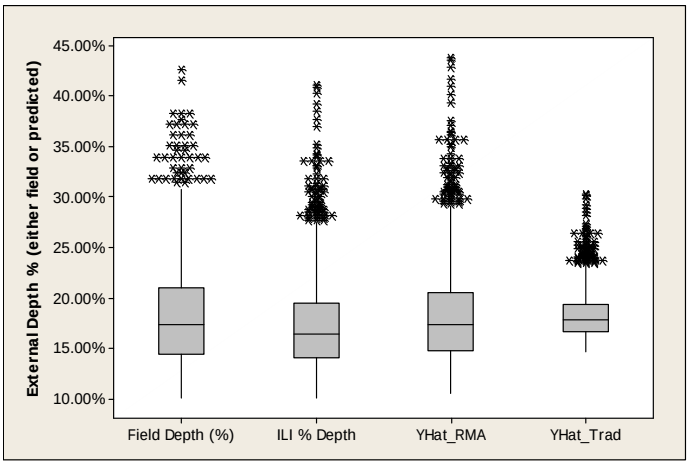

Figure 2. Box plots of dependent variable Field Depth (%), independent variable ILI % Depth, and the two predictions: *YHat_RMA* and the least squares *YHat_Trad*.

Figures 3 and 4 also illustrate the difference between the RMA and least squares regression predictions with a super-imposed 1-1 line. The primary observation from Figures 3 and 4 is the limited range on the predicted least squares *yhat* that is shown here on the horizontal axis with the dependent (*y*) variable Field Depth (%) on the vertical axis. The coverage shown in the plots illustrate a previous point - i.e., while the RMA predictions range roughly from about 10% to 45% wall thickness, matching the range of the field measurements, the least squares predictions range from approximately 15% to 30% wall thickness. Such comparisons should be considered before using a traditional least squares regression when comparing ILI to field measurements. Indeed this potential for concern may be generalized to cover a wider domain of pipeline integrity modeling issues.

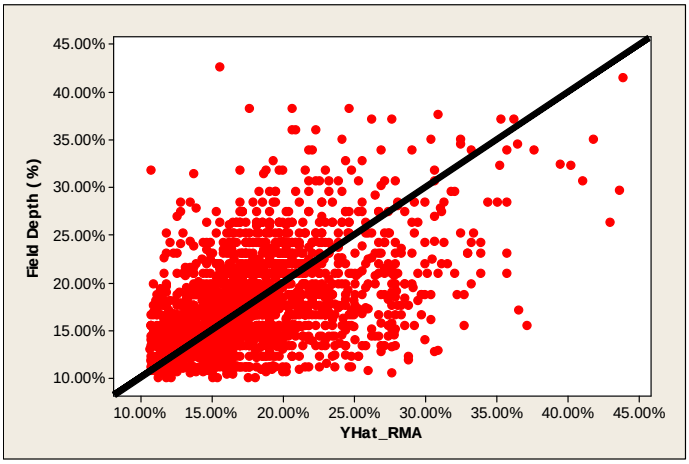

Figure 3. One to one plot of RMA regression predictions to actual Field Depth (%)

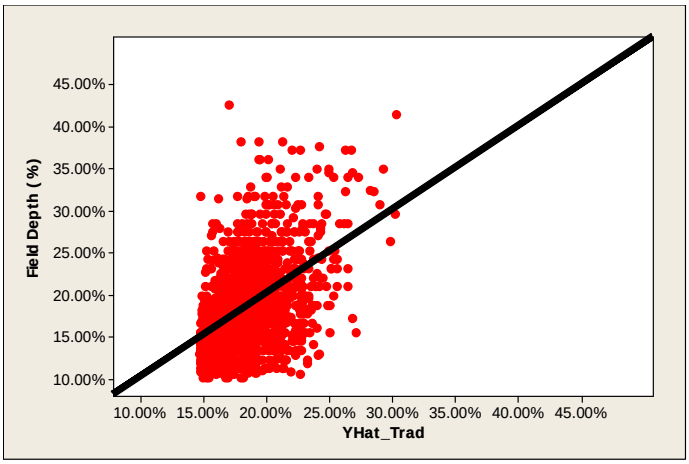

Figure 4. One to one plot of Least Squares regression predictions to actual Field Depth (%)

CONCLUSIONS

Typical analytical work such as modeling pipeline data generally costs little in comparison to the associated field investigation. The authors suggest the use of modeling methods that are theoretically sound and provide a reasonable approximation of reality. For regression oriented tasks, reduced major axis regression is worthy of consideration.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Martin Phillips for pointing us to RMA.

REFERENCES

1) Haines, H., McNealy, R., and Rosenfeld, M.J., “IPC2010-31483: Is the 80% Leak Criterion Always Appropriate?”, International Pipeline Conference 2010, ASME, Calgary, Alberta.
2) Sokal, Robert R., and F. James Rohlf, (1995), Biometry, 3rd edition, section 14.3 titled “Model II Regression”, pp. 541-549, W. H. Freeman, New York
3) http://en.wikipedia.org/wiki/Total_least_squares, January 3, 2012
4) Galton, F., 1886, “Regression toward Mediocrity in Hereditary Stature”, Journal of the Anthropological Institute of Great Britain and Ireland, 15, pp. 246-263
5) Maier, Clifford J., Pamela J. Moreno, William V. Harper, David J. Stucki, Steven J. Polasik, Thomas A. Bubenik, David A.R. Shanks, and Neil A. Bates, "Application and Validation of Statistically Based Corrosion Growth Rates", Proceedings of the 2012 International Pipeline Conference, IPC2012-90424, September 24-28, Calgary, Alberta, Canada, 2012, pp 1-7.
6) Vieth, Patrick H., “Corrosion Assessment and Probability of Exceedance”, Nov 6, 2008, http://www.slideshare.net/DNV_Columbus/pat-poe-slides-rev-1-presentation