REDUCED MAJOR AXIS REGRESSION: TEACHING ALTERNATIVES TO LEAST SQUARES

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

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. Both independent and dependent measurements may have multiple sources of error. Thus the underlying least squares regression assumptions can be violated. Reduced Major Axis (RMA) regression is specifically formulated to handle errors in both the x and y variables. It is an excellent topic to teach students the importance of understanding the assumptions underlying the statistical procedures commonly used in practice as well as showing them that alternatives may better satisfy the actual needs.

REGRESSION ISSUE BACKGROUND

Almost all linear regression analysis performed 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. This is an excellent real world area to illustrate to students the impact of often over-looked assumptions.

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 two methods of measuring the same variable as in the example data used in this paper. 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 text book *Biometry* (Sokal & Rohlf, 1995) in which it appears under the title of Model II regression. This paper demonstrates the potential improvements brought about by RMA regression.

The specific data used in this paper is what is called in-line inspection (ILI) measurements of corrosion depths in a pipeline by what is typically called a smart pigging device. Typically the ILI 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 are identified and sized. This data has multiple sources of measurement and identification errors. When parts of the underground pipeline are excavated, field measurements are made of 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 very small portion of all the potential anomalies reported by the ILI tool is excavated. The limited field measurements are paired with the associated ILI calls. 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. In this context the ILI calls are the independent variable and the field measurement the dependent variable as we want to estimate actual field depth at un-excavated locations.

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

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

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. This is common for many other applications encountered in the world.

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. 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, 2012).

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). It is a question as to whether this result of over and under prediction is a reasonable match for reality or whether it is an artifact of the methodology employed.

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

The example given consists of a 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 larger than most samples typically available. Table 1 shows the y-intercept and the slope for both approaches.

Table 1. Least Squares, RMA regression coefficients.

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

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.

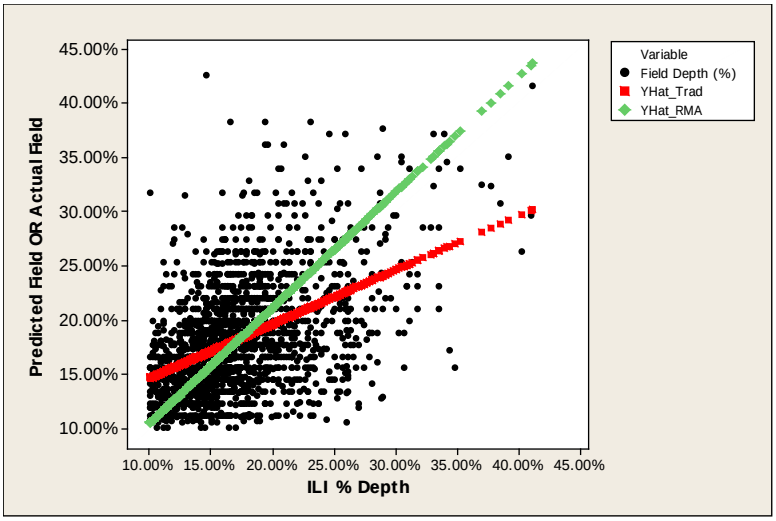

Figure 1. Predicted Least-squares and RMA Regression over the range of the ILI calls. YHat_RMA is the predicted RMA field depth while YHat_Trad is the traditional least-squares prediction for the field depth.

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

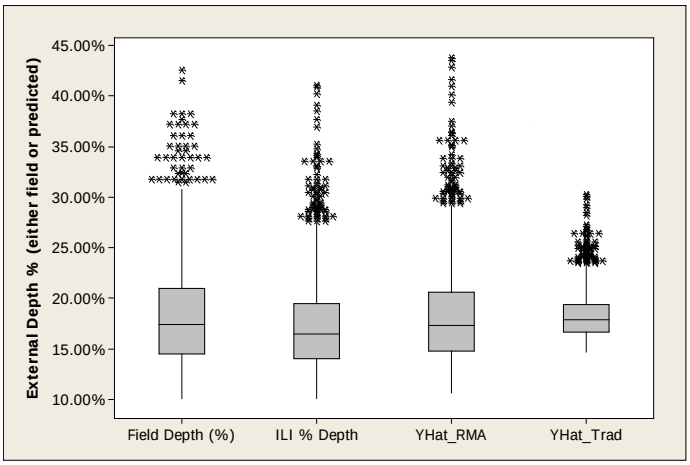

Figure 2. Box plots of dependent variable Field Depth (%), independent variable ILI % Depth, and the two predictions: YHat_RMA and the least squares YHat_Trad.

Figure 2 shows in a much clearer fashion a 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(%).

Blind application of even a well-known and sound methodology may lead to poor modeling results and violation of key assumptions generally over-looked in both teaching and in the application world-wide of the corresponding methodology. It is essential that the impact of the conceptual underpinnings be understood and taught in classes. Examples such as the differences between least-squares and RMA regression are simple to teach and to clearly point out the resulting impact. Making students aware that assumption matter and do impact the results is essential for the sustainable use of statistical modeling and believability of results.

CONCLUSIONS

It is important to use 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.

REFERENCES

Sokal, R. R., & Rohlf, F. J. (1995). Section 14.3 “Model II regression”. In *Biometry* (3rd ed) (pp. 541-549). New York: W. H. Freeman.

"Total Least Squares." *Wikipedia*. Wikimedia Foundation, Accessed 17 Nov. 2013.

Galton, F. (1886). Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland, 15,* 246-263.