INTERACTIVE LEARNING TOOLS

David S. Zamar
Simon Fraser University, Canada
Ruben H. Zamar
University of British Columbia, Canada
ruben@stat.ubc.ca

Although the intuitive idea of robustness is simple and appealing some key robustness concepts and measures are mathematically involved and hence difficult to grasp by students. We show how interactive graphic tools can be used to motivate and demonstrate the use of robust methods. We also show how interactive graphical tools can also be used to help understand and "visualize" main robustness concepts such as influence function and breakdown point. We developed interactive tools for simple linear regression and multivariate location and covariance matrix. Due to space limitation we focus on regression and only briefly comment on the multivariate location and covariance matrix.

INTRODUCTION

It is our belief that in the same sense that one picture is worth one-thounsand words, one interactive graphical tool may be worth one thousand pictures. Many statistical concepts such as robustness, randomization and blocking in design of experiments, classification and clustering, sampling distributions, bootstrap, and hypothesis testing can be presented, demonstrated and more easily understood using interactive graphical tools.

The influence of a single point on the least squares estimator of regression may be measured by an index such as Cook's distance that takes into account the combined effect of leverage and outlierness of the point. As experience indicates, leverage and outlierness are subtle concepts difficult to grasp. For example, students usually struggle to understand the differences between good and bad leverage points. Unfortunately, an analysis of a formula or the presentation of a set of static pictures usually fails to convey a full understanding of the situation.

Regarding multivariate location and covariance matrices, it is usually difficult to convey to students the meaning of these summary statistics. We believe in the pedagogical convenience of exploiting their geometrical interpretation; that is, use the fact that the estimated covariance matrix gives the shape of the concentration ellipsoid and the multivariate location pinpoints its center. A simple approach is to choose a given percentage of data points to be enclosed inside the smallest ellipsoid, with the estimated center and shape, and superimpose it onto the data cloud. This strategy allows us to easily visualize and compare the properties of different multivariate location and covariance estimators, emphasizing robustness issues.

We developed interactive tools to aid the understanding of the properties of robust estimates of regression and multivariate location and covariance matrix. However, in this paper we concentrate on the simple linear regression setup to illustrate the main ideas underlying our project.

ROBUST SIMPLE LINEAR REGRESSION ESTIMATES

In this section we briefly describe the robust estimates of the parameters of the simple linear regression model used in our interactive tool.

Siegel's Repeated Medians

The breakdown point (BP) of an estimator is the smallest proportion of the observations that must be replaced by arbitrary values in order to force the estimator to produce values arbitrarily far from the parameter values that generated the original data (Donoho and Huber, 1983). Siegel (1982) defined the first regression estimate with maximal BP = 1/2, called repeated medians (RM), by performing a median-based operation over the ratios

$$r(i,j) = \frac{y_j - y_i}{x_j - x_i}.$$

In this case, the robust slope estimate is defined as

$$\hat{\beta} = Med_{1\le i\le n}\{Med_{j\in J_i}\{r(i,j)\}\},$$

where $I = \{(i,j) : x_i \neq x_j\}$ and $J_i = \{j : (i,j) \in I\}$, for $1 \le i \le n$. The robust intercept estimate is then calculated as

$$\hat{\alpha} = Med_{1\le i\le n}\{y_i - \hat{\beta}x_i\}.$$

Example

The computation RM is illustrated in this small example that includes only 6 data points.

x	*y*
1.20	5.01
1.50	18.00
2.10	6.74
3.40	10.77
4.10	10.47
5.00	12.44

The robust estimates of the slope, β, and intercept, α, are obtained as follows. For each point, we calculate the median of the slopes of the lines connecting this point to the others in the dataset:

$$\beta_1 = Med\left\{\frac{18.00-5.01}{1.5-1.2}, \frac{6.74-5.01}{2.1-1.2}, \frac{10.77-5.01}{3.4-1.2}, \frac{10.47-5.01}{4.1-1.2}, \frac{12.44-5.01}{5.0-1.2}\right\} = 1.955$$

$$\beta_2 = Med\left\{\frac{5.01-12.44}{1.2-1.5}, \frac{6.74-18.00}{2.1-1.5}, \frac{10.77-18.00}{3.4-1.5}, \frac{10.47-18.00}{4.1-1.5}, \frac{12.44-18.00}{5.0-1.5}\right\} = -0.588$$

$$\vdots$$

$$\beta_6 = Med\left\{\frac{5.01-12.44}{1.2-5.0}, \frac{18.00-12.44}{1.5-5.0}, \frac{6.74-12.44}{2.1-5.0}, \frac{10.77-12.44}{3.4-5.0}, \frac{10.47-12.44}{4.1-5.0}\right\} = 1.955$$

The robust slope estimate, $\hat{\beta}$, is obtained by taking the median of these medians (hence the name repeated medians). The robust intercept is obtained by calculating the median of the partial residuals $y_i - \hat{\beta}x_i$ for $1 \le i \le 6$.

$$\hat{\beta} = Med\{1.955, -0.588, ..., 1.955\} = 1.894$$

$$\hat{\alpha} = Med\{5.01 - 1.894\times 1.2, 18.00 - 1.894\times 1.5, ..., 12.44 - 1.894\times 5.0\} = 2.868$$

Figure 1 displays the 6 points together with the least squares and repeated median regression lines.

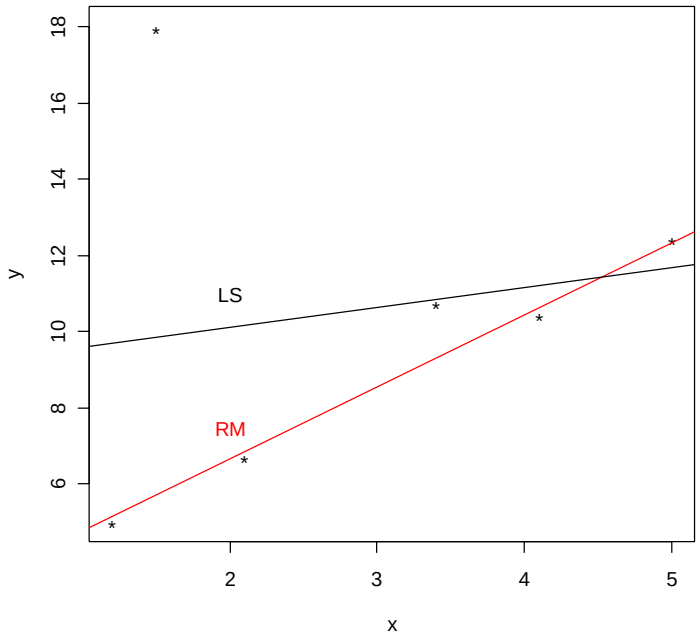


Figure 1: Least Squares (LS) and Repeated Medians (RM) Regression Lines

INTERACTIVE TOOLS

We have developed several interactive tools to help visualize statistical concepts. In this paper we chose to present our applet on robust simple linear regression, which is available online at http://hajek.stat.ubc.ca/~ruben/data/robustRegressionApplet. The applet is currently undergoing modifications and the final version will be completed this summer.

Applet Goals

The goals of this applet are to demonstrate the effects of outliers on the LS regression line and to demonstrate that outliers have a much lesser effect on the robust regression line. The tool may be used to visualize the influence of a point (measured by Cook's distance in the current implementation), as well as the concepts of good leverage points, bad leverage points, and regression outliers. Moreover the applet may also be used to visualize the concepts influence function, of contamination bias and breakdown point.

Using the Applet

There are two methods in which regression data points may be obtained for use by this applet:

- Generate a random dataset of a given size.
- Read data from a built in library. *This feature is not yet available and will be included in the final version*.

By clicking on the reset button, found in the top right hand corner of the applet, a new random set of data points are generated and displayed. The number of data points generated is given by the selected value of the drop down menu found below the reset button. The randomly generated datasets share the same slope, intercept and random error model. In future versions, these quantities will be parameters with user-defined values.

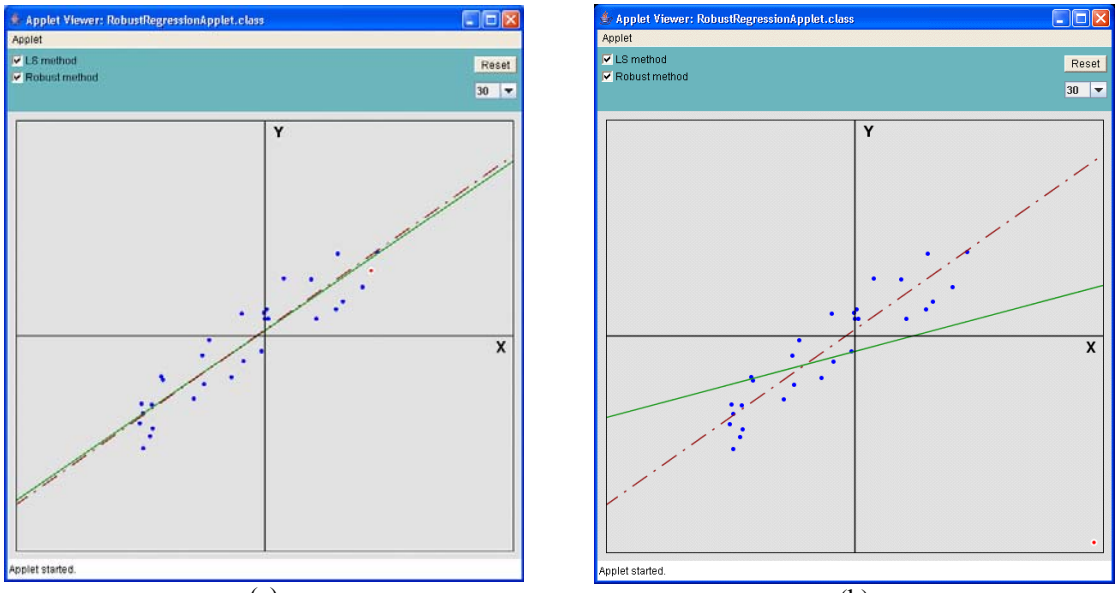
Points can be added by clicking the right mouse button. Each new point will appear directly under the mouse cursor. A point is selected (its color turns from blue to red) by placing the mouse cursor over top of it. A selected point may be deleted by holding the SHIFT key down while clicking the right mouse button. Multiple points may be selected by holding the SHIFT key down and sequentially left clicking on the chosen points. A selected point or group of points may

be dragged (within the default range of the *x* and *y* axes) by holding down the left mouse button and moving the cursor. In future versions, the user will be able to specify the range of the *x* and *y* axes in order to provide a zoom in/out feature.

The user can select to view the least squares regression line and/or the robust regression line by clicking on the corresponding selection box in the top left corner of the applet.

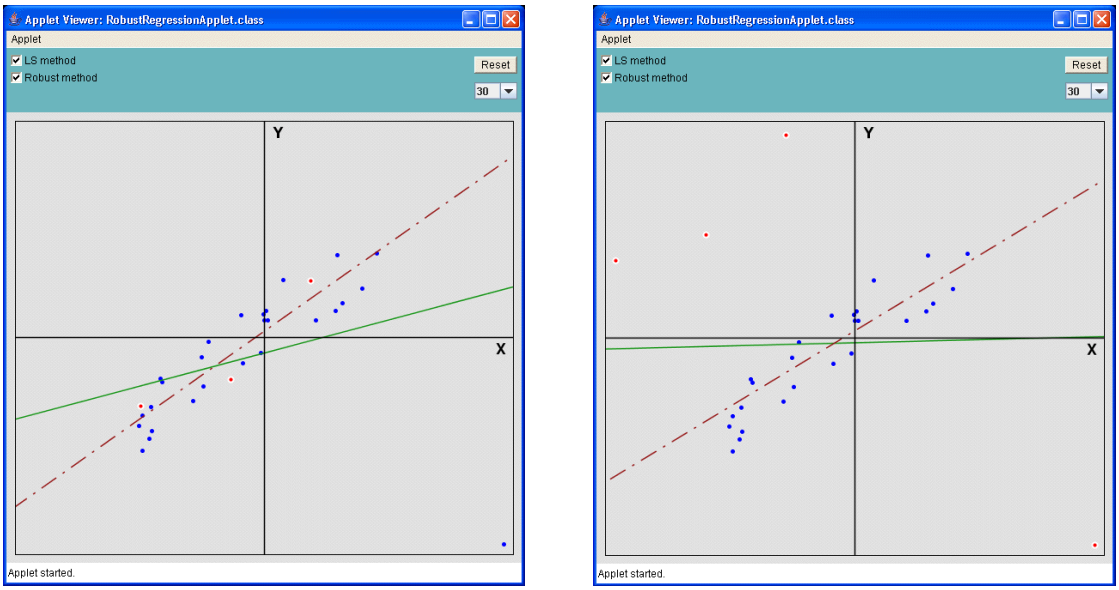
Figures 2 to 5 contain various snapshots of the regression applet in action with 30 randomly generated points. The robust (dashed-line) and least squares (solid-line) regression lines have been chosen to be displayed in Figure 2a. The red point has been selected (although the mouse cursor is not seen). Since the displayed data does not include any outliers both lines almost overlay each other. In Figure 2b the selected point has been dragged to the bottom right corner causing a noticeable change in the least squares line while the robust line remains undisturbed. The regression lines are updated in real time as the point is dragged. In Figure 3a the three red points have been selected. In Figure 3b the three selected points have been dragged to the top left corner of the plot. Again, the robust line remains almost undisturbed while the least squares line is dramatically affected. In Figure 4a seven additional points have been selected. In Figure 4b the four selected points are clustered together in the top left corner of the plot along with the previously acted on points. In this case, the eleven outliers cause a severe change on both regression lines. Finally, Figure 5 shows the least squares regression line alone with a single outlier. A display showing the coefficient of determination as a color gradient appears because the LS method has been selected alone. Moreover, because the outlier in the top left corner has been selected, its corresponding Cook's distance is also displayed.

The reader is invited to try our regression and covariance applets by visiting the afore mentioned link. In the future additional applets will become available. Please send feedback to ruben@stat.ubc.ca or dzamar@sfu.ca.



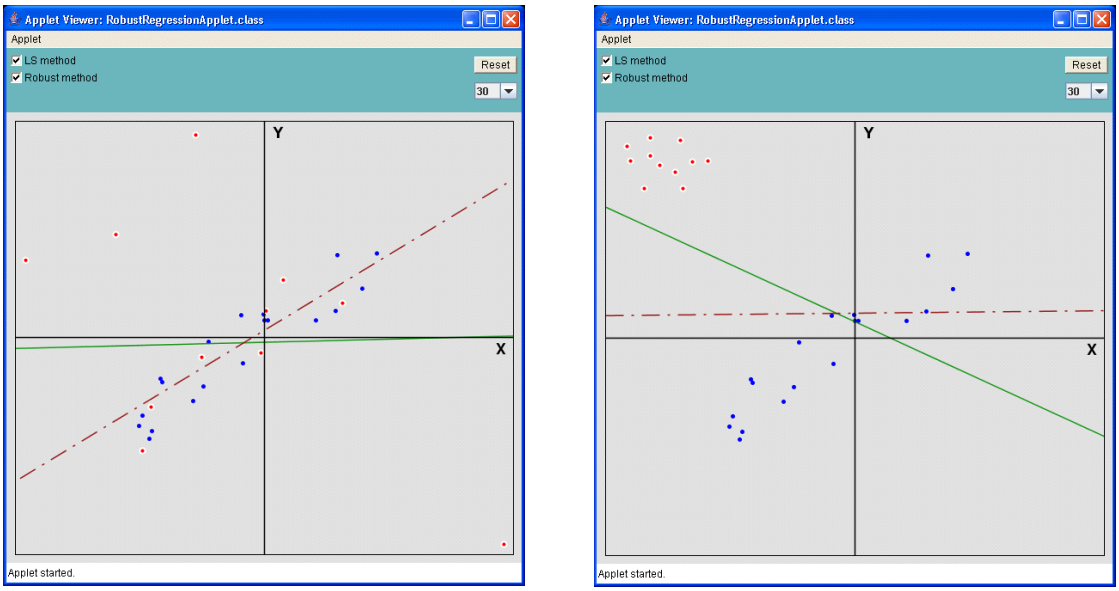
(a) (b)

Figure 2: Snapshots of Regression Applet



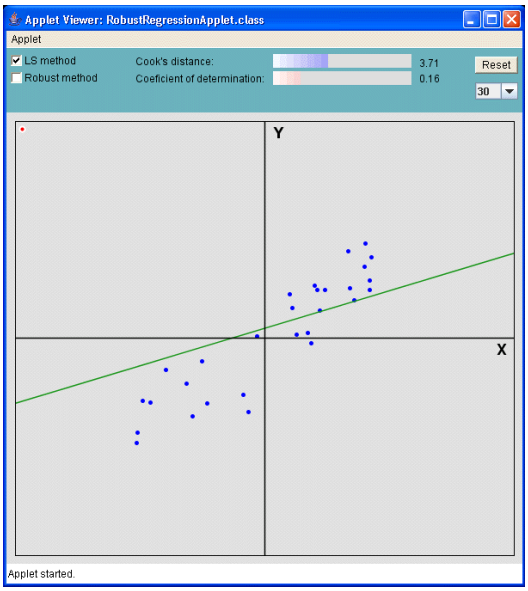
(a) (b)

Figure 3: Snapshots of Regression Applet



(a) (b)

Figure 4: Snapshots of Regression Applet



(g)

Figure 5: Snapshot of Regression Applet Displaying Only the LS Line (One Outlier Selected)

REFERENCES

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr. (Eds.), *A Festschrift for Z. Erich L. Lehman*, (pp. 157-184). Belmont, CA: Wadsworth.

Siegel, A. (1982). Robust regression using repeated medians. *Biometrika*, 69, 242-244.