Building Statistically Significant Robust Regression Models in Empirical Software Engineering

Sandro Morasca
Dipartimento di Scienze Politiche, della Cultura e dell'Informazione
Università degli Studi dell'Insubria
Via Carloni 78, I-22100, Como, Italy
sandro.morasca@uninsubria.it

ABSTRACT

Outliers have been a constant source of problems in the analysis of Empirical Software Engineering data. In some cases, outliers are due to corrupted data, while they may be the result of highly unlikely circumstances in others. In either case, outliers may unduly greatly bias data analysis, as is the case with Ordinary Least Squares (OLS) regression. Robust data analysis techniques have been proposed to address this problem. In this paper, we describe an existing robust linear regression technique based on the Least Median of Squares (LMS) and provide a statistical significance test for the associations obtained with it. We also apply LMS and OLS regression to real-life, publicly available Empirical Software Engineering data sets, to compare the results obtained and investigate commonalities and differences between LMS and OLS from a practical point of view.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics—*complexity measures, product metrics*

General Terms

Measurement, Application

Keywords

Robust regression, Data analysis, Outliers, Statistical significance, Effort prediction, Defect prediction

1. INTRODUCTION

In a data set, data points are hardly ever located in a nice way that allows data analyzers to build a regression line in a straightforward and easy way. The presence of outliers is a common problem that arises when analyzing real data. An outlier is a data point that lies far from the bulk of the data points and which may overly and unduly influence the regression model. For instance, take the data set shown in Figure 1 (taken from [5]). It is clear that all points but one (point A) are located on a straight line. Point A is so far from the rest of the data points that is manages to irresistibly attract the Ordinary Least Square (OLS) regression line, which is shown in the figure. That singular data point may be corrupted, due to some specific conditions that may be highly unlikely to occur, or just due to a statistical fluctuation. This regression line has a lower explanatory value than the straight line that goes through all the other points. We are sacrificing understanding the linear relation among all data points but one because of only one data point.

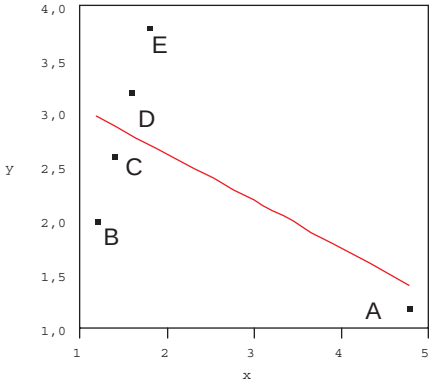

Figure 1: Overinfluential outlier.

In this paper, we use a specific kind of so-called robust regression, whose goal is to produce linear models that are not biased by few overinfluential outliers. This kind of robust regression, based on the Least Median of Squares (LMS), was introduced by Rousseeuw and Leroy [5] to this end. To the best of this author's knowledge, LMS regression has only been used in Empirical Software Engineering for the analysis of data about Web application designs [1]. So, the first contribution of this paper is to illustrate and discuss LMS regression (Sections 2, 3, and 4). By illustrating the characteristics of LMS, we would like to make it simpler for researchers and practitioners in Empirical Software Engineering to evaluate whether LMS regression may be useful when they carry out data analyses.

When it was introduced [5], LMS regression did not come with ways for assessing the statistical significance of LMS regression models. The second contribution of this paper is to introduce a statistical significance test for LMS regression models (Section 5). This test is based on less constraining

assumption than statistical significance tests used in Ordinary Least Squares (OLS) regression, so it can be used to assess the statistical significance of regression models when that may not be possible with OLS regression models.

To show the application of LMS regression to real-life data sets, Section 6 describes the results obtained by using LMS regression on two data sets belonging to the PROMISE repository [3], namely "desharnais_1_1" and "qqdefects." These results are also compared to the results obtained with OLS regression. More applications of LMS regression need to be carried out, and further studies are necessary at the theoretical level as well, as outlined in Section 7.

2. LMS REGRESSION

The goal of robust regression techniques is to allow data analyses not to be overly influenced by few data points. The robustness of a regression technique can be quantified via its so-called breakdown point, as follows. Suppose that an estimator $T(Z)$ is built based on a sample Z of n data points. Let us study the influence of k out of those n data points, so let Z' be any sample of n data points, k of which are different from those in Z and $n-k$ of which coincide with $n-k$ data points in Z.

The bias due to the modification of k data points is defined as the supremum of the deviations computed over all such Z', i.e., $bias(k,T,Z) = sup_{Z'}|T(Z') - T(Z)|$ [5]. The estimator breaks down if its bias is infinite, i.e., k outliers may cause an arbitrarily large difference between the estimator computed on the Z and the estimator computed on Z'. The breakdown point of the estimator is the smallest proportion k/n of modified data points that can make the estimator deviate to an arbitrarily large extent. For instance, the sample mean as an estimator of the expected value of a distribution and the OLS estimators of the coefficients of an OLS regression model have a breakdown point given by $1/n$. This is the worst behavior possible for an estimator, as the estimator may be led astray even by only one corrupted data point. To make matters worse, when n becomes arbitrarily large, this breakdown point becomes arbitrarily small, as it is 0 in the limiting case when n tends to infinity. Dealing with outliers requires a much higher breakdown point than $1/n$.

Robust estimators have higher breakdown points than traditional estimators. For instance, the sample median has a 50% breakdown point. At least 50% of the data must be corrupted for the sample median to become arbitrarily large. However, a higher breakdown point may be a problem. For instance, having a 60% breakdown point would imply that we can no longer tell the correct data points from the corrupted ones, since the majority of data points are corrupted.

Here, we focus on LMS regression, which is a specific type of robust regression. For notational convenience, in what follows, given a dependent variable y and an independent variable x, $\{y_i\}$ denotes the multiset of sample values obtained for y and $\{x_i\}$ the multiset of corresponding sample values obtained for x. Also, $est = est(x;par)$ denotes the estimation of variable y by means of function $est(x;par)$, which depends on x and is built by means of parameters here collectively denoted as par. Finally, $est_i = est(x_i;par)$ is the estimated value corresponding to x_i. The difference between an actual and an estimated value is called a residual $r_i = y_i - est_i$. Note that a multiset of residuals may very well exist even when $\{y_i\}$ is a set and not a multiset.

LMS regression is based on the minimization of the *median* of the squared residuals $r_i^2 = (y_i - est_i)^2$. Its basic idea actually finds its motivations in the outlier problems that are typical of OLS regression, which is based on the minimization of the *average* of the squared residuals $\sum_{i\ in 1..n}(y_i - est_i)^2/n$. As explained above, the sample mean and the sample median have very different breakdown points.

For the sake of clarity, since the concept of median is central for the research described here, it is necessary to clarify that we use the so-called *low median* in this paper, as it is well known that the median value may not be unique for a multiset (or even a set) of ordinal values. In a multiset of ordinal values, we define the low median as the minimum value such that the total number of occurrences of the elements greater than the low median is not greater than $n/2$. The low median clearly coincides with the median if the median is unique. For brevity, we use the term "median" instead of "low median" throughout the paper. Also, we refer to the multiset of values that are not greater than the low median as the lower half of the multiset, even though, strictly speaking, this lower "half" may contain more than $n/2$ occurrences of the elements.

These are the general ideas underlying LMS regression. However, for simplicity's sake, we focus on the following two special cases for the estimation model in this paper

- constant LMS regression: $est = c$
- univariate LMS regression: $est = a \cdot x + b$.

So, we here focus on linear univariate LMS regression models, since the constant LMS regression case can be seen as a special case of the more general linear univariate LMS regression models. At any rate, most of our results are applicable to multivariate linear models with an arbitrary number of independent variables, as as we point out throughout the paper where appropriate. Also, we are interested in the constant LMS regression case to introduce a statistical significance test for LMS regression models (see Section 5).

It is worth noting that, when using the constant LMS regression model, a new, robust indicator (m_{LMS}) of central tendency in its own right is defined as the value of parameter c that minimizes $med\{(y_i - c)^2\}$. As a comparison, it is well known that the value that minimizes $\sum_{i\in 1..n}(y_i - c)^2/n$ is the average $m = \sum_{i\in 1..n} y_i/n$.

We now define and list a few properties for indicators of central tendency, which m_{LMS} shares.

1. If min and max denote, respectively, the minimum and maximum sample values, then $min \leq m_{LMS} \leq max$.
2. Indicator m_{LMS} is a Chisini mean [2]. A function $F(z_1,\ldots,z_n)$ of n variables leads to a Chisini mean

M if for every choice of the variables $< z_1, \ldots, z_n >$, with $M = F(z_1, \ldots, z_n)$, M is the only value such that $M = F(M, ..., M)$.

3. Indicator m_{LMS} does not depend on the specific indexing chosen for the data points in $\{y_i\}$, i.e., it is truly a function of the multiset $\{y_i\}$.

4. Indicator m_{LMS} is a homogeneous function of degree 1, i.e., if m_{LMS} minimizes the median of squares with data set $\{y_i\}$, then $\lambda \cdot m_{LMS}$ minimizes the median of squares with data set $\{\lambda \cdot y_i\}$ for any given value of λ.

5. Suppose that m_{LMS} minimizes the median of squares with data set $\{y_i\}$ and that one data point y_j is replaced by another data point y'_j, with $y'_j \geq y_j$, to obtain a new data set. The value m'_{LMS} that minimizes the median of squares with this new data set is such that $m'_{LMS} \geq m_{LMS}$. Symmetrically, if $y'_j \leq y_j$, we have $m'_{LMS} \leq m_{LMS}$.

As for the univariate LMS regression model, we minimize the median of the squared residuals $r_i^2 = (y_i - ax_i - b)^2$, i.e., finding the values of parameters a and b that minimize $med\{(y_i - ax_i - b)^2\}$. As a comparison, OLS finds the values of parameters a and b that minimize the average of the squared residuals $\sum_{i\ in1..n}(y_i - ax_i - b)^2/n$ (note that this is perfectly equivalent to minimizing the sum of squared residuals $\sum_{i\ in1..n}(y_i - ax_i - b)^2$, as is usually said in OLS).

For illustration convenience, we first investigate constant LMS regression (Section 3) and we then proceed to univariate LMS regression (Section 4).

3. CONSTANT LMS REGRESSION

Here, we show how m_{LMS} can be computed. The final result is that m_{LMS} is the midpoint of the narrowest interval that contains at least $\lceil n/2 \rceil$ occurrences of data points in $\{y_i\}$. For instance, in the data set $\{y_i\}$ of Figure 2, we have $m_{LMS} = 8.5$, which is the midpoint of interval $[6, 11]$, which is the narrowest interval that contains at least $\lceil 7/2 = 4 \rceil$ occurrences of data points. Just for comparison's sake, the sample mean is $m = 8.14$ and the sample median is $med = 7$.

First, m_{LMS} is comprised between the absolute minimum $yMin$ and the absolute maximum $yMax$ values in $\{y_i\}$. By contradiction, suppose that $m_{LMS} < yMin$ and let $\tilde{r}_i^2 = (y_i - m_{LMS})^2$ be the i-th square residual. We can always find another value $v = m_{LMS} + \delta$ for some $\delta > 0$ such that, for all i, the square residual $r_i^2 = (y_i - v)^2$ of the i-th data point computed from v is smaller than the square residual of the i-th data point computed from m_{LMS}. To show this, $(y_i - v)^2 < (y_i - m_{LMS})^2$ can be rewritten as $(y_i - m_{LMS} - \delta)^2 < (y_i - m_{LMS})^2$, and, through computations, as $2(y_i - m_{LMS}) > \delta$. Since $y_i - m_{LMS} > 0$ for all i, a suitable positive value of δ always exist. As $r_i^2 < \tilde{r}_i^2$ for all i, the median of multiset $\{r_i^2\}$ is lower than the median of multiset $\{\tilde{r}^2\}$, so the value we chose for m_{LMS} does not actually minimize the median square residual.

We now show how m_{LMS} is computed, by using the data sample in Figure 2 as an example when needed. Our discussion is based on the following two observations.

1. The median of any multiset can be computed based on the knowledge of the lower half of the data points (or, equivalently, the upper half).

2. The median is simply the maximum of the lower half.

To find m_{LMS}, we compute the median of multiset $\{\tilde{r}_i^2\}$ of the square residuals of the y_i's from m_{LMS} (i.e., $\tilde{r}_i^2 = (y_i - m_{LMS})^2$). Because of observation 2 above, we need to find the lower half of multiset $\{\tilde{r}_i^2\}$. To build this lower half, we need to find the appropriate half of multiset $\{y_i\}$, composed of those y_i's with the lower half of distances to m_{LMS} (for clarity only, note that this half of multiset $\{y_i\}$ by no means coincides with the lower half of multiset $\{y_i\}$).

Let us take any value $v \in [yMin, yMax]$ and let us compute $med\{(y_i - v)^2\}$. To this end, we need to

- compute the square distances of all data points in $\{y_i\}$ to v;
- order these square distances;
- find a data point $yMed(v)$ such that $(yMed(v) - v)^2$ is the median of $\{(y_i - v)^2\}$. In general, given a value v, there might exist more than one data point $yMed(v)$ such that $(yMed(v) - v)^2$ is the median of $\{(y_i - v)^2\}$. For illustration purposes only, we assume for the time being that there exists only one such $yMed(v)$. We show that this temporary assumption does not really affect our reasoning in Remark 1 below.

Thus, we need to find all those data points

- that are closer to v than the remaining data points, since the ordering of square distances is the same as the ordering of distances;
- whose occurrences are strictly necessary to account for at least $\lceil n/2 \rceil$ occurrences in $\{y_i\}$.

Given a value v, there always exists an interval $I(v) = [yLow(v), yHigh(v)]$

- to which these $\lceil n/2 \rceil$ occurrences belong;
- in which no other data points fall;
- such that $yLow(v)$ and $yHigh(v)$ are data points in $\{y_i\}$, since, if either $yLow(v)$ or $yHigh(v)$ was removed from $I(v)$ the resulting interval would contain less than $\lceil n/2 \rceil$ occurrences of data points in $\{y_i\}$.

For instance, in Figure 2, $I(3) = [1, 7]$.

So, $yMed(v) = yLow(v)$ if $|v - yLow(v)| > |v - yHigh(v)|$; $yMed(v) = yHigh(v)$ if $|v - yLow(v)| < |v - yHigh(v)|$; either $yMed(v) = yLow(v)$ or $yMed(v) = yHigh(v)$ if $|v - yLow(v)| = |v - yHigh(v)|$, i.e., v is the midpoint of $I(v)$.

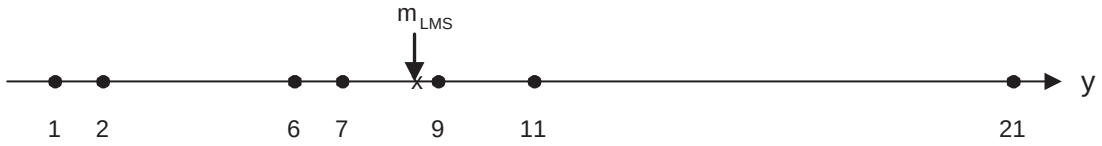

Figure 2: Computation of m_{LMS}.

Two cases are possible: 1) $v \notin I(v)$; 2) $v \in I(v)$. For example, point 13 in Figure 2, for which the median distance is 7, computed as the difference between 13 itself and the point with value 6, does not belong to $I(13)$. Let us now focus on case 1), to show that such a v cannot be m_{LMS}, and let us suppose that $v > yHigh(v)$, so $yMed(v) = yLow(v)$ and the distance of v from $yMed(v)$ is obviously $v - yLow(v)$. Let us take any value $u \in I(v)$. Its maximum distance from data points in $I(v)$ is $yHigh(v) - yLow(v)$, which happens when either $u = yLow(v)$ or $u = yHigh(v)$. It is clear that $yHigh(v) - yLow(v) < v - yLow(v)$, so this shows that all points in $I(v)$ have a lover maximum distance from $\lceil n/2 \rceil$ occurrences of data points in $\{y_i\}$ than v. So, if $I(v)$ is an interval for which the chosen value of v is actually m_{LMS}, then $v \in I(v)$. Now, given one such interval $I(v)$, it is immediate to show that the midpoint $(yLow(v)+yHigh(v))/2$ is actually the value that minimizes the distance from the farthest data point in $I(v)$, since it is equidistant from the extremes of the interval. Thus, m_{LMS} must be one of these midpoints. There is only a finite number of those intervals, since their left extreme and their right extreme must be values in $\{y_i\}$. So, we can actually find m_{LMS} by brute force, by computing the maximum distance of the points in each interval $I(v)$ from the midpoint of the interval itself, and pick the midpoint that minimizes this maximum distance. For the example of Figure 2, these are the possible intervals: $[1, 7]$; $[2, 9]$; $[6, 11]$; $[7, 21]$. Now, the maximum distance from the midpoint of an interval $I(v)$ is obviously $(yHigh(v) - yLow(v))/2$, so m_{LMS} is the midpoint of the narrowest interval $I(v)$ that contains at least $\lceil n/2 \rceil$ occurrences of data points in $\{y_i\}$. In Figure 2, this is interval $[6, 11]$, so $m_{LMS} = 8.5$.

Remark 1. Given a value v, there may be more than one interval that contains just enough data points to account for at least $\lceil n/2 \rceil$ occurrences of data points in $\{y_i\}$. For instance, this is the case of $v = 13.5$ for the data set of Figure 2, which is associated with intervals $[6, 11]$ and $[7, 21]$, since there may be two possible values for $yMed(v)$, i.e., $yMed(v) = 6$ and $yMed(v) = 21$. However, this does not influence our results, since m_{LMS} is the midpoint of the narrowest interval $I(v)$ that contains at least $\lceil n/2 \rceil$ occurrences of data points.

Remark 2. "Degenerate" cases may occur, For instance, if all data points are equidistant, $\lceil n/2 \rceil$ intervals exist that contain at least $\lceil n/2 \rceil$ occurrences of data points in $\{y_i\}$. In this case, all intervals $I(v)$ have the same width, so, it is not possible to identify the narrowest one. However, this should not be a surprise, because "degenerate" cases may exist even for medians, as more than one median may exist between the "low" and the "high" median too. Like with medians, averaging mechanisms may be used for m_{LMS}, but this is beyond the scope of the paper.

4. UNIVARIATE LMS REGRESSION

We here concisely sketch how to build the univariate LMS regression line based on the discussion on constant LMS regression of Section 3. The univariate LMS regression line lies halfway between the two parallel straight lines that are closest to each other if their distance is measured along the y axis and are such that at least $\lceil n/2 \rceil$ occurrences of data points lie between or on them.

To show this, suppose that $est = a \cdot x + b$ is the univariate LMS regression line. Thus, it is the straight line such that $med\{(y_i - ax_i - b)^2\}$ is minimal. Each single $(y_i - ax_i - b)^2$ may be viewed as the squared distance, measured along the y-axis, between the actual value y_i and the LMS regression line. We call this type of distance y-distance. Like in Section 3, the value of $med\{(y_i - ax_i - b)^2\}$ is obtained as the largest value of the squared y-distance of the set of data points that contain at least $\lceil n/2 \rceil$ occurrences of data points and also contain the occurrences of the data points with the smallest y-distances to the LMS regression line. These occurrences of data points will therefore lie in a strip across the LMS regression line, which is a part of the (x, y) plane delimited by two parallel lines.

Given two parallel lines, the line that minimizes the maximum y-distance to any point in the strip between the two parallel lines is located halfway between the two parallel lines, and we use Figures 3 and 4 to exemplify our reasoning. We use geometrical arguments to characterize these two parallel lines. Specifically, we show that the narrowest strip is delimited by two straight lines, one of which goes through at least one data point and the other goes through at least two data points. Let sup and inf denote the two parallel lines that delimit the narrowest strip, and let sup lie above inf, so, for instance let sup be defined by equation $y = a \cdot x + b_{sup}$ and inf by equation $y = a \cdot x + b_{inf}$, with $b_{sup} > b_{inf}$. We begin by showing that sup necessarily goes through at least one data point A and inf through at least one data point B. By contradiction, let us suppose that sup does not go through any data points. This is the case shown in Figure 3, where the strip delimited by sup and inf (the two dotted lines in Figure 3) encloses $\lceil 15/2 \rceil = 8$ occurrences of data points. For graphical convenience, the observations lying outside the strip are represented by '+'. The observations lying inside the strip are represented by dots, except for points A, B, and C, which take some "special" importance in what follows: point A is represented with an 'o' and points B and C are represented with 'x' for graphical convenience. Then, there exists another line sup' with equation $y = a \cdot x + b_{sup} - \delta$ with $\delta > 0$ such that at least $\lceil n/2 \rceil$ occurrences of data points remain in the strip delimited by sup' and inf. For instance, this is the thick dashed line in Figure 3. In other words, it is always possible to make sup

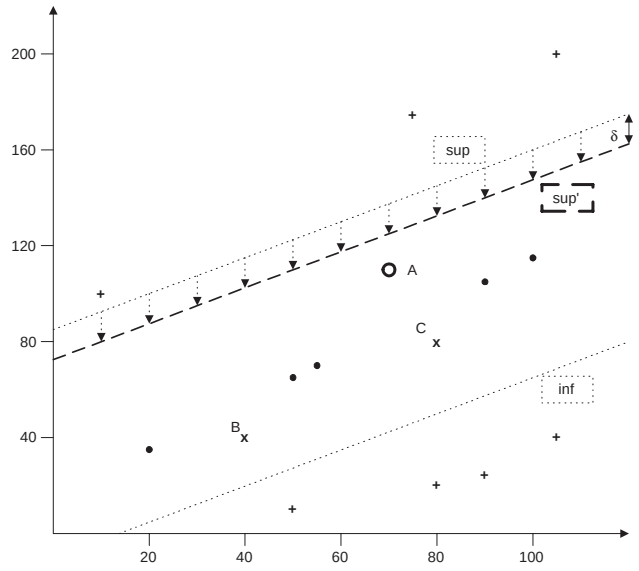

Figure 3: Shift of parallel lines.

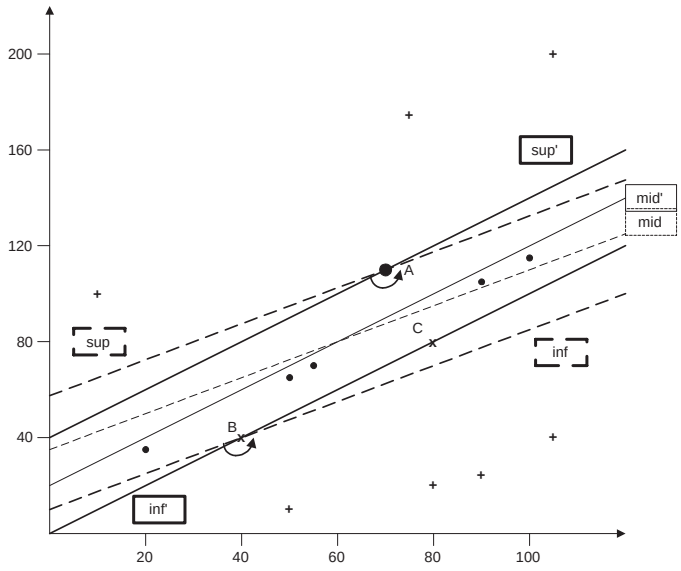

Figure 4: Rotation of parallel lines.

shift downwards without changing its slope and still leaving at least $\lceil n/2 \rceil$ occurrences of data points in the strip. Line *sup* may shift downwards until it "hits" one of the data points. In Figure 3 line *sup* may shift downwards until it hits point A, but it cannot shift any lower, because the strip delimited by sup' and inf would enclose $\lceil 14/2 \rceil = 7$ occurrences of data points. Line sup' is closer to inf, so the strip delimited by sup' is closer to inf is narrower than the strip delimited by *sup* is closer to inf, in terms of y-distance. The same kind of reasoning symmetrically applies to inf, which can shift upwards until it hits point B in Figure 3.

Now, let us show that either *sup* or inf actually go through at least two data points, with the aid of Figure 4. Again, by contradiction, suppose that *sup* and inf do not go through two data points. Since *sup* goes through point A with coordinates (x_A, y_A) and inf goes through point B with coordinates (x_B, y_B), as shown by the thick dashed lines in Figure 4, we can write their equations as $y = a(x - x_A) + y_A$ and $y = a(x - x_B) + y_B$, respectively. The line that lies halfway between them has equation $y = a(x - (x_A + x_B)/2) + (y_A + y_B)/2$, shown by line *mid* in Figure 4. Since point A is on the border of the strip, A has the greatest y-distance from the halfway line, so the square of its y-distance from the halfway line is also the value of $med\{(y_i - ax_i - b)^2\}$. The y-distance between A and the halfway line is $(y_A - y_B)/2 - a(x_A - x_B)/2$. That is obviously the same value as the y-distance between B and the halfway line. Let us assume for now that $x_A \neq x_B$, as is the case in Figure 4. This y-distance is a function of the slope a, so we can always find a value a' such that the new y-distance $(y_A - y_B)/2 - a'(x_A - x_B)/2$ of points A and B from the halfway line is lower, i.e., $(y_A - y_B)/2 - a(x_A - x_B)/2 > (y_A - y_B)/2 - a'(x_A - x_B)/2$. This happens when $a(x_A - x_B) < a'(x_A - x_B)$. In other words, we can make *sup* rotate around point A and inf around point B in such a way as to "rotate" the halfway line closer (as measured by the y-distance) to A and B. For instance, the new halfway line is represented by line mid' in Figure 4. Thus, we have a lower value for $med\{(y_i - a'x_i - b')^2\}$ than we had for $med\{(y_i - ax_i - b)^2\}$. We can make both delimiting lines rotate until either line hits a data point in the strip. In the example in Figure 4, we can make both lines rotate counterclockwise until line inf' hits point C. The lines cannot rotate any further because otherwise we would end up with a strip that contains $\lceil 14/2 \rceil = 7$ occurrences of data points.

For completeness only, even when $x_A = x_B$, we can make both delimiting lines rotate anyway until either one "hits" a data point even though this will not change the value of the y-distance.

Thus, in the univariate LMS regression case, the LMS regression line lies halfway between the two parallel straight lines that are closest to each other if their distance is measured along the y axis and are such that at least $\lceil n/2 \rceil$ occurrences of data points lie between or on them. We can extend this result to multivariate cases as follows. Given a set of data points with k independent variables and one dependent variable, the k-dimensional LMS regression linear variety (generalization of LMS regression line to the $k+1$-dimensional space) lies halfway between a k-dimensional linear variety that goes through at least $k+1$ data points and a parallel k-dimensional linear variety that goes through at least 1 data point. At any rate, we only take into account constant and univariate LMS regression here, so the multidimensional case will no longer be investigated in this paper.

5. STATISTICAL SIGNIFICANCE

Statistical regression techniques are usually used in conjunction with tests for evaluating the statistical significance of various properties of the regression curve. For instance, in OLS, tests are available for the statistical significance of the estimated values for the coefficients of the regression straight line and the estimated degree of correlation R^2. When it was proposed, LMS regression did not come with any statistical significance tests, so its results could be used as an indication of a central tendency of the regression line, but no further assessment could be made.

Here, we introduce a test for the statistical significance of an indicator of the degree of correlation R^2_{LMS}, which is the counterpart of R^2 used in OLS (which we denote R^2_{OLS} in what follows). We explain the rationale behind R^2_{LMS} by comparing it to the way R^2_{OLS} is built. So, we first provide a discussion of the way R^2_{OLS} is built (Section 5.1) and then we introduce R^2_{LMS} (Section 5.2). We introduce statistical inference for LMS regression in Section 5.3.

5.1 A Discussion on R^2_{OLS}

R^2_{OLS} is the instantiation for OLS of the more general indicator R^2_{GL}, where "GL" stands for Generalized Linear

$$R^2_{GL} = 1 - \frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{\sum_{i\in 1..n}(y_i - \bar{y})^2} = 1 - \frac{\frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{n}}{\frac{\sum_{i\in 1..n}(y_i - \bar{y})^2}{n}} = \quad (1)$$

$$1 - \frac{\frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{n-1}}{\frac{\sum_{i\in 1..n}(y_i - \bar{y})^2}{n-1}} \quad (2)$$

where $\bar{y}$ is the sample mean of the $\{y_i\}$ data set and a and b are estimated via some estimation procedure, not necessarily OLS. Thus, R^2_{GL} provides an indication of the change in the average square residual that the linear univariate model $est = ax + b$ provides over the simpler constant regression model $y = \bar{y}$. For instance, $R^2_{GL} = 0$ if and only if $\frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{n} = \frac{\sum_{i\in 1..n}(y_i - \bar{y})^2}{n}$, i.e., the two-parameter model $est = ax - b$ gives the same average squared residual as the single parameter model $y = \bar{y}$. Note that there exist values for a and b such that $R^2_{GL} < 0$. For instance, take $a = 0$ and $b \neq \bar{y}$. Since $\bar{y}$ actually minimizes $\frac{\sum_{i\in 1..n}(y_i - b)^2}{n}$, we have $\frac{\sum_{i\in 1..n}(y_i - b)^2}{n} > \frac{\sum_{i\in 1..n}(y_i - \bar{y})^2}{n}$, so $R^2_{GL} < 0$. Because of the continuity of R^2_{GL} with respect to a, we can even find an infinity of values $a \neq 0$ such that $R^2_{GL} < 0$. Since one of the desirable properties of R^2_{GL} is to be nonnegative, this shows that R^2_{GL} cannot really be used with all kinds of linear regression. Specifically, R^2_{GL} cannot be used with a and b estimated via LMS regression.

However, R^2_{GL} can be safely used in conjunction with OLS regression, since it can be shown that $\frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{n} \leq \frac{\sum_{i\in 1..n}(y_i - \bar{y})^2}{n}$ when a and b are estimated with OLS. Note that, in Formula (1), we purposely wrote $\frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{n}$ as the numerator and $\frac{\sum_{i\in 1..n}(y_i - \bar{y})^2}{n}$ as the denominator, instead of the more commonly used sums $\sum_{i\in 1..n}(y_i - ax_i - b)^2$ and $\sum_{i\in 1..n}(y_i - \bar{y})^2$. The reason is that Formula (1) shows that R^2_{GL} is based on the ratio between two averages, and these averages are the expressions that OLS minimizes. In addition, Formula (2) shows that R^2_{GL} can also be computed based on the ratio between two unbiased estimators: $\frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{n-1}$ is the estimator of the variance of the estimation error $y_i - ax_i - b$ and $\frac{\sum_{i\in 1..n}(y_i - \bar{y})^2}{n-1}$ is the estimator of the variance of $\{y_i\}$. The value of R^2_{OLS} can be interpreted as the fraction of the variance of the dependent variable of the sample data that cannot be explained by the OLS regression model.

Now, to make inferences with OLS, the following two assumptions need to be satisfied

- the true regression line is linear;
- the values of y for any given x are independent and identically normally distributed with mean $\alpha x - \beta$ and variance σ^2_e.

Under these conditions, it can be shown that R^2_{OLS} is an unbiased estimator of the true value $\rho^2_{OLS} = \frac{\sigma^2_{xy}}{\sigma^2_x \sigma^2_y} = 1 - \frac{\sigma^2_e}{\sigma^2_x}$, where σ^2_{xy} is the true value of the covariance of x and y, and σ^2_x and σ^2_y are the true values of the variances of x and y, respectively. The value of ρ^2_{OLS} can be interpreted as the fraction of the true variance of the dependent variable that cannot be explained by the OLS regression model.

Also, it can be shown that the true value of coefficient α is $\frac{\sigma_{xy}}{\sigma^2_x}$, of which the OLS estimator $a = \frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{\sum_{i\in 1..n}(y_i - \bar{y})^2}$ is an unbiased estimator.

It can also be shown that the sample random variable

$$t = \sqrt{\frac{n-2}{1 - R^2_{OLS}}} \frac{R_{OLS}}{a}(a - \alpha)$$

has a Student's t distribution with $n-2$ degrees of freedom. This sample random variable can be used to test null hypotheses of the kind $H_0 : \alpha = \alpha_0$, where α_0 is a specified value for α. Now, suppose we want to test the following null hypothesis $H_0 : \rho_{OLS} = 0$. If $\rho_{OLS} = 0$, then $\alpha = 0$, so we can use the sample random variable $t = \sqrt{\frac{n-2}{1-R^2_{OLS}}} R_{OLS}$ to test this null hypothesis. Thus, the same statistic can be used to test this specific hypothesis on ρ_{OLS}.

5.2 Introducing R^2_{LMS}

Much in the same way as LMS regression is introduced in comparison to OLS regression, we here define R^2_{LMS} as follows

$$R^2_{LMS} = 1 - \frac{med_{i\in 1..n}\{|y_i - ax_i - b|\}}{med_{i\in 1..n}\{|y_i - \bar{y}_{LMS}|\}}$$

where $\bar{y}_{LMS}$ is m_{LMS} for the $\{y_i\}$ data set. Thus, we have replaced the average square residuals $\frac{\sum_{i\in 1..n}(y_i - ax_i - b)^2}{n}$ and $\frac{\sum_{i\in 1..n}(y_i - \bar{y})^2}{n}$ of Formula (1) $med_{i\in 1..n}\{|y_i - ax_i - b|\}$ and $med_{i\in 1..n}\{|y_i - \bar{y}_{LMS}|\}$, which are with their counterparts minimized in LMS regression. Note that using the median absolute residuals instead of the median square residuals $med_{i\in 1..n}\{(y_i - ax_i - b)^2\}$ and $med_{i\in 1..n}\{(y_i - \bar{y}_{LMS})^2\}$ is mostly a matter of taste, as it does not change the meaning of R^2_{LMS}, nor the statistical significance tests that can be used, as we show later. For completeness, we also use

$$S^2_{LMS} = 1 - \frac{med_{i\in 1..n}\{(y_i - ax_i - b)^2\}}{med_{i\in 1..n}\{(y_i - \bar{y}_{LMS})^2\}} \quad (3)$$

which we introduced in [1].

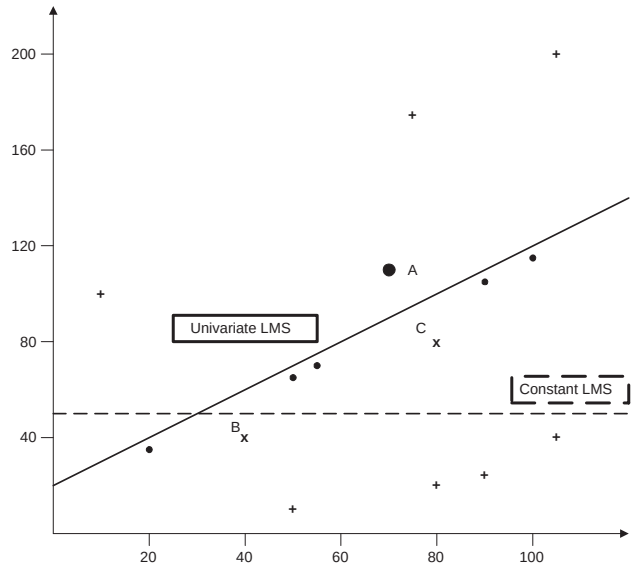

Figure 5: Constant vs. univariate LMS regression.

Note that a correlation indicator, R^2_{RR} is defined by Rousseeuw and Leroy [5] as

$$R^2_{RR} = 1 - (\frac{med_{i\in 1..N}\{|r_i|\}}{mad(y)})^2 \quad (4)$$

where $mad(y) = med_{i\in 1..N}\{|y_i - med_{j\in 1..N}\{|y_j|\}|\}$ is the median absolute deviation from the median. R^2_{RR} shows the improvement that a LMS regression model provides over predicting the value of y for each data point with a constant value: the median of y. R^2_{RR} ranges between 0 (there is no improvement) to 1 (the LMS regression model explains all uncertainty in the data set). However, R^2_{LMS} shows the improvement that an LMS regression model provides over predicting the value of y for each data point simply as the m_{LMS} of $\{y_i\}$, i.e., a simpler linear LMS regression model that has only the intercept but no independent variable. This is more similar to OLS analysis, in which R^2_{OLS} actually quantifies the improvement that an OLS model provides in explaining the uncertainty (i.e., the variance) of the dependent variable over a simpler model that predicts that the value of the dependent variable in each data point is given by the *mean* of y. It can be shown that $R^2_{RR} \geq S^2_{LMS} \geq R^2_{LMS}$, i.e., R^2_{RR} shows a greater improvement than R^2_{LMS}, because R^2_{RR} quantifies the improvement of using a linear LMS regression model over using the median (which does not minimize the median of the residuals), while R^2_{LMS} quantifies the improvement of using an LMS regression linear model over using m_{LMS} (which does minimize the median of the residuals).

5.3 Statistical Inference for LMS Regression

To make inferences with LMS regression, we introduce the following two assumptions that need to be satisfied

- the true regression line is linear;
- the values of y for any given x are independent and identically distributed.

So, we remove one condition—normality of the distribution of the residuals—that is an assumption used in OLS.

Statistical inference on R^2_{LMS} and a can be carried out by using distribution-free, nonparametric tests, as follows. First, like in the OLS case, $\rho^2_{LMS} = 0$ (where ρ^2_{LMS} is the true value of R^2_{LMS} implies $\alpha = 0$ (where α is the true value of a) and *vice versa*. Also, if $\alpha = 0$, then the true LMS regression line is $est = \beta$, i.e., the constant LMS regression line. So, the null hypothesis $H_0 : \alpha = 0$ for LMS regression implies that the distribution of absolute residuals is the one obtained with constant LMS regression. This also implies that the median of the distribution of absolute residuals obtained with constant LMS regression is the same as the median the distribution of absolute residuals obtained with univariate LMS regression. Therefore, we can use distribution-free, nonparametric statistical tests to this end. Specifically, we here use Fisher's sign test [4] with the variable $Z_i = |r_{i,ULMS}| - |r_{i,CLMS}|$ where $|r_{i,CLMS}|$ is the absolute value of the residual of the i-th observation with constant LMS regression and $|r_{i,ULMS}|$ is the absolute value of the residual of the i-th observation with univariate LMS regression. For instance, in Figure 5, the horizontal, dashed line represents the constant LMS regression line for the example used in Section 4 and the other straight like represents the univariate LMS regression line. The two sets of sample absolute residuals are obtained by taking the absolute values of the y-distances of the observations to the univariate LMS line and to the constant LMS line. These two sets are used in the statistical test of hypotheses.

If $\alpha = 0$, then the distributions of $|r_{i,CLMS}|$ and $|r_{i,ULMS}|$ coincide, and the median of their difference is null. Fisher's sign test's statistic B is the number of times $Z_i > 0$ in the data set. For simplicity, and because of the sufficiently large number of observations of the data sets we analyze in Section 6, we can use the normal approximation to B's exact distribution. Specifically, it can be shown that B asymptotically tends to a normal distribution with $n/2$ expected value and $n/4$ variance, so variable $(2B - n)/\sqrt{n}$ has an asymptotic standard normal distribution.

6. EXPERIMENTAL RESULTS

We have applied OLS and LMS regression to two data sets that belong to the PROMISE set of data sets. We used the statistical tool JMP to carry out OLS regression, while software was developed for LMS regression. Here, we summarize the results we have obtained. We selected these two data sets because they have different numbers of data points and they also allow us to deal with the estimation of *Defects* and *Effort*. At any rate, regression models are more often used for effort estimation than defect estimation. We used both dependent variables to show that the LMS approach may be used for the prediction of the number of defects too.

In OLS, we dealt with outliers by using an iterative approach, by removing one outlier at a time from the data set and then checking for further outliers on the new data set, until no more outliers could be found. We used Mahalanobis jackknife distances to identify outliers. The Mahalanobis distance of a point P in a set of data points is a measure of how far P is from the so-called "centroid" of the set of data points, which provides a concise idea of the location of the

data points. The Mahalanobis jackknife distance of P in a set of data points is a measure of how far P is from the set of data point after P has been removed from the set of data points. The idea is that P attracts the centroid, so it should not be taken into account when assessing the distance of P from all the other points. JMP uses a threshold value that is based on Fisher's F-distribution. A data point was classified as an outlier if its Mahalanobis jackknife distance was "too large." Iteratively, the outlier with the largest Mahalanobis jackknife distance was excluded from the data set. We used JMP to compute Mahalanobis jackknife distances. Specifically, we used JMP's predefined threshold for identifying Mahalanobis jackknife distances that are "too large," which, therefore, characterize outliers.

However, this iterative procedure may have drawbacks, especially in Empirical Software Engineering, in which large data sets are not all that common. The risk is removing too many observations and leaving too few to have a sufficient statistical basis to obtain results. On the other hand, leaving too many outliers in the data set may lead to unreliable results and problems from a statistical point of view (e.g., heteroscedasticity or the fact that residuals do not follow a normal distribution).

6.1 Analysis of Data Set desharnais_1_1

The first data set we analyzed is "desharnais_1_1" in the PROMISE repository [3], which has data about 81 projects. We used *Effort* (the development effort) as the dependent variable and *Transactions*, *Entities*, and *PointsNonAdjust* (obtained as the sum of *Transactions* and *Entities*) as the independent variables. Table 1 contains the following descriptive statistics for all of these variables (all tables are at the end of the paper):

- $Variable$ is the name of the variable
- n is the number of observations
- min is the minimum observed value
- max is the maximum observed value
- med is the sample median
- m_{OLS} is the ordinary mean (we have added the subscript "OLS" for additional clarity)
- σ_{OLS} is the standard deviation
- m_{LMS} is the value that minimizes the Least Median of Squares, as described in Section 3
- $mar_C = med\{|y_i - m_{LMS}|\}$ median absolute residual computed from m_{LMS}, i.e., by using the constant LMS regression model.

Table 2 contains the results for all the OLS univariate analyses we carried out. These are the data reported for each analysis: $Variable$ is the independent variable used; n is the number of data points used in the analysis: if $n = 81$, the analysis was carried out by using all of the data points, otherwise $n = 81 - \#outliers$; a is the estimate of the coefficient of the independent variable; b is the estimate of the intercept; p is the statistical significance of the OLS model: like in most of the literature, we use a 0.05 statistical significance threshold, i.e., the model is statistically significant if $p < 0.05$; R^2_{OLS} is the square of the correlation coefficient; w is the statistical significance of the Shapiro-Wilk W test for normality of the residuals: again, we use a 0.05 statistical significance threshold, i.e., the hypothesis that the residuals are normally distributed is rejected if $w < 0.05$.

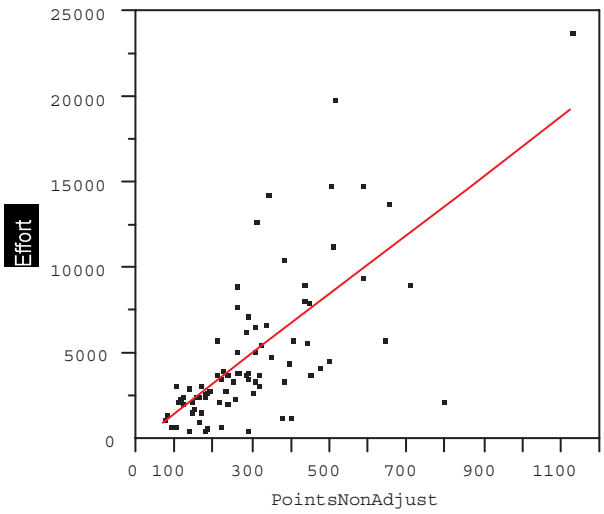

Figure 6: OLS and outliers.

Let us take the first OLS analysis, carried out with the entire data set and $PointsNonAdjust$ as the independent variable). For illustration purposes, Figure 6 also shows the OLS line obtained on the entire data set.

The scatterplot visually shows the existence of both outliers and heteroscedasticity. At any rate, in the remainder of the paper, we do not take into account heteroscedasticity, but we do investigate whether the residual distribution is normal. The value of $w = 0.0006$ shows that there is enough evidence to reject the hypothesis that the residuals are normally distributed. So, even though the model appears to be statistically significant, one cannot really accept these result as reliable. Let us now take the second OLS analysis, carried out with 57 observations (i.e., 24 outliers removed) and $PointsNonAdjust$ as the independent variable). In this case, the value of $w = 0.3276$ does not allow us to reject the hypothesis that the residuals are normally distributed. The other rows of Table 2 show the results for the other analyses. Notice that, in the last analysis of the table, the value of $w = 0.0130$ shows that there is enough evidence to reject the hypothesis that the residuals are normally distributed, even though 23 outliers have been removed.

Let us now carry out LMS regression analyses for the same data set. Table 3 summarizes the results, where $mar_U = med\{|y_i - ax_i - b|\}$ is the median absolute residual computed from the univariate LMS regression model.

All the results are statistically significant. In general, it can be noticed that, for each of these three LMS regression lines,

the slope is lower than the slope of the corresponding OLS analyses, with or without outliers. This shows that high leverage points do not have high influence on LMS regression lines.

6.2 Analysis of Data Set qqdefects

We also used data set "qqdefects" in the PROMISE repository [3], which contains data about 31 projects. Table 4 contains the descriptive statistics for the three variables we used: *Defects* (Testing Defects), *Effort* (Total Effort), and *KLOC*.

We first used *Defects* as the dependent variable and *KLOC* as the independent variable. We also used *Defect* as the dependent variable and *Effort* as the independent variable. Table 5 summarizes the OLS results we obtained.

The results do not provide evidence to reject the null hypothesis that the distribution of residuals is normal for both OLS models where KLOC is the independent variable. However, the sample is small, so rejection of this null hypothesis could be due to insufficient power. On the contrary, there is enough evidence to reject the null hypothesis according to which the distribution of residuals is normal for both OLS models where Effort is the independent variable. So, inferences about the OLS models may not be reliable.

Table 6 shows the results we obtained for the corresponding analyses with LMS regression. Again, the slopes are lower than for the corresponding OLS analyses. The LMS regression models are statistically significant.

Next, we used Effort as the dependent variable and KLOC as the independent variable. Table 7 summarizes the OLS results we obtained. The results provide evidence to reject the hypothesis that the distribution of residuals is normal for both OLS analyses.

Table 8 summarizes the LMS results we obtained. However, in this case, the LMS regression model does not appear to be statistically significant.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have illustrated a robust regression technique that can be used to deal with outliers and that can help remove some assumptions that are used in other data analysis techniques, like OLS, which we have used here as a comparison technique. Outliers and lack of compliance with assumptions of existing data analysis techniques are often a problem in Empirical Software Engineering research.

We have introduced a statistical significance test for the robust regression technique. We have used both LMS and OLS regression on real-life data sets and the results for robust regression appear to be somewhat promising.

A number of activities still remain to be done, though. We here outline a few:

- check other robust data analysis techniques and compare them with LMS and OLS;
- extend the statistical significance test to multivariate LMS regression models and assess the statistical significance of the entire model and the coefficient of each independent variable;
- assess the power of the statistical significance test;
- check other possible statistical significance tests.

8. ACKNOWLEDGMENTS

The research presented in this article was partially funded by the IST project QualiPSo, sponsored by the EU in the 6th FP (IST-034763); the FIRB project ARTDECO, sponsored by the Italian Ministry of Education and University; and the project "La qualità nello sviluppo software," sponsored by the Università degli Studi dell'Insubria.

9. REFERENCES

[1] L. Baresi and S. Morasca. Three empirical studies on estimating the design effort of web applications. *ACM Trans. Softw. Eng. Methodol.*, 16(4), 2007.

[2] O. Chisini. Sul concetto di media. *Periodico di Matematiche*, pages 106–116, 1929.

[3] T. M. Gary Boetticher and T. Ostrand. PROMISE Repository of empirical software engineering data. West Virginia University, Department of Computer Science, 2007.

[4] M. Hollander and D. A. Wolfe. *Nonparametric statistical inference, second ed.* John Wiley & Sons, New York, 1999.

[5] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection.* John Wiley & Sons, Inc., New York, NY, USA, 1987.

Table 1: Descriptive statistics for data set desharnais_1_1

$Variable$	n	min	max	med	m_{OLS}	σ_{OLS}	m_{LMS}	mar_C
Effort	81	546	23940	3647	5046.31	4419.77	2786	1386
Transactions	81	9	886	140	182.12	144.04	93	53
Entities	81	7	387	99	122.33	84.88	65	34
PointsNonAdjust	81	73	1127	266	304.46	180.21	239	82

Table 2: OLS regression results for data set desharnais_1_1

$Variable$	n	a	b	p	R^2_{OLS}	w
PointsNonAdjust	81	17.30	-220.08	<0.0001	0.50	0.0006
PointsNonAdjust	57	13.40	291.27	<0.0001	0.50	0.3276
Transactions	81	17.85	1795.19	<0.0001	0.34	<0.0001
Transactions	58	11.39	1727.13	0.0004	0.20	0.2920
Entities	81	26.57	1796.34	<0.0001	0.26	<0.0001
Entities	58	30.43	755.80	<0.0001	0.43	0.0130

Table 3: LMS regression results for data set desharnais_1_1

$Variable$	a	b	p	mar_U	R^2_{LMS}	S^2_{LMS}
PointsNonAdjust	9.27	803.27	<0.0001	925.04	0.33	0.556
Transactions	7.29	2239.56	0.0027	1111.40	0.20	0.36
Entities	9.21	1677.13	0.0174	1150.24	0.17	0.31

Table 4: Descriptive statistics for data set qqdefects

$Variable$	n	min	max	med	m_{OLS}	σ_{OLS}	m_{LMS}	mar_C
Defects	31	5	1906	209	445.22	510.92	107	102
Effort	31	1308	53995	13388	16141.88	13771.05	9183.0	5419.0
KLOC	29	0.9	155.2	26.67	39.36	40.73	13.78	12.89

Table 5: OLS regression results for data set qqdefects with *Defects* as the dependent variable

$Variable$	n	a	b	p	R^2_{OLS}	w
KLOC	29	11.41	10.09	<0.0001	0.78	0.1978
KLOC	25	9.54	55.99	0.0004	0.43	0.1830
Effort	31	0.0215	98.25	0.0006	0.34	0.0009
Effort	22	0.0167	73.57	0.04	0.19	0.0243

Table 6: LMS regression results for data set qqdefects with *Defects* as the dependent variable

$Variable$	a	b	p	mar_U	R^2_{LMS}	S^2_{LMS}
KLOC	4.05	62.28	<0.0001	61.33	0.36	0.59
Effort	0.0053	62.60	0.0098	68.04	0.33	0.56

Table 7: OLS regression results for data set qqdefects with *Effort* as the dependent variable

$Variable$	n	a	b	p	R^2_{OLS}	w
KLOC	29	210.27	8680.12	0.0004	0.3819	<0.0001
KLOC	23	201.80	6096.7	<0.0001	0.7359	0.0375

Table 8: LMS regression results for data set qqdefects with *Effort* as the dependent variable

$Variable$	a	b	p	mar_U	R^2_{OLS}	S^2_{OLS}
KLOC	0.22	5.91	0.2887	4.11	0.41	0.66