ROBUST REGRESSION AND APPLICATIONS

A THESIS PRESENTED BY ARZDAR KIRACI
TO
THE INSTITUTE OF
ECONOMICS AND SOCIAL SCIENCES
IN PARTIAL FULFILLMENT OF THE
REQUIRMENTS
FOR THE DEGREE OF MASTER OF ECONOMICS

BILKENT UNIVERSITY
SEPTEMBER, 1996

ROBUST REGRESSION AND APPLICATIONS

A THESIS PRESENTED BY ARZDAR KIRACI

TO

THE INSTITUTE OF

ECONOMICS AND SOCIAL SCIENCES

IN PARTIAL FULFILLMENT OF THE

REQUIRMENTS

FOR THE DEGREE OF MASTER OF ECONOMICS

BILKENT UNIVERSTY

SEPTEMBER, 1996

Arzdar Kiraci

tarafından değiştirilmiştir

I certify that I have read this thesis and in my opinion it is fully adequate, in scope and in quality, as a thesis for the degreee of Master of Economics.

Prof. Dr. Asad Zaman

I certify that I have read this thesis and in my opinion it is fully adequate, in scope and in quality, as a thesis for the degreee of Master of Economics.

Assoc. Prof. Dr. Syed Mahmud

I certify that I have read this thesis and in my opinion it is fully adequate, in scope and in quality, as a thesis for the degreee of Master of Economics.

Assis. Prof. Dr. Bahri Yılmaz

Approved by the Institude of Economics and Social Sciences

Director:

I certify that I have read this thesis and in my opinion it is fully adequate, in scope and in quality, as a thesis for the degreee of Master of Economics.

Prof. Dr. .Asad Zaman

I certify that I have read this thesis and in my opinion it is fully adequate, in scope and in quality, as a thesis for the degreee of Master of Economics.

Assoc. Prof. Dr. Syed Mahmud

I certify that I have read this thesis and in my opinion it is fully adequate, in scope and in quality, as a thesis for the degreee of Master of Economics.

Assis. Prof. Dr. Bahri Yılmaz

Approved by the Institude of Economics and Social Sciences

Director:

ABSTRACT

ROBUST REGRESSION AND APPLICATIONS

ARZDAR KIRACI
MA in Economics
Supervisor: Prof. Dr. Asad Zaman
September 1996

This study analyzes the effect of outliers in the regression analysis with the help of a written program in the programing language of GAUSS. The analysis relies on the subject of Robust Regression, which is explained and supported by experiments and applications. The applications contain examples to show the superiority of this technique.

Key Words: Robust Regression, Outlier, Leverage Point, Robust Distance, Least Median Squares, LMS, Least Trimmed Squares, LTS, Minimum Volume Ellipsoid, MVE, Minimum Covariance Determinant, MCD, Program for Robust Regression, Defect of Ordinary Least Squares Regression.

ÖZET

GÜÇLÜ REGRESYON VE UYGULAMALARI

ARZDAR KIRACI
Yüksek Lisans Tezi, İktisat Bölümü
Tez Yöneticisi: Prof. Dr. Asad Zaman
Eylül 1996

Bu çalışma GAUSS programlama dilinde yazılmış bir programla regresyonda yanlış etki gösteren noktaları incelemektedir. Güçlü Regresyon tekniği açıklanmakta, deneyler ve örneklerle desteklenmektedir. Örnekler Güçlü Regresyonun üstünlüklerini göstermektedir.

Anahtar Kelimeler: Güçlü Regresyon, Uzak Nokta, Etkili Nokta, Güçlü Mesafe, Medyan Kare Minimizasyonu, LMS, Belirli Toplam Kare Minimizasyonu, LTS, Elipsoid Hacmi Minimizasyonu, MVE, Kovaryans Determinantı Minimizasyonu, MCD, Güçlü Regresyon Programı, Regresyon Hataları

Acknowledgements

I would like to express my gratitude to Prof Dr. Asad Zaman for his valuable courses in Statistics and Econometrics. I would also thank to Assoc. Prof. Dr. Syed Mahmud and Assis. Prof. Dr. Bahri Yılmaz for their valuable comments..

Contents

1. Introduction

The Ordinary Least Squares Regression (OLS) is the oldest and very easily applicable type of regression. A person familiar with matrix algebra or a scientific calculator, can get the results in a very short time. The results are reliable and counted as admissible until the discovery of the Bayesian Estimators.

The results of the OLS are reliable, if the data are also reliable. If there is a possibility of corrupted data or wrong recordings, blind application of the OLS leads to very different results. This is due to the fact that OLS is a very equalitarian type of mathematic process. Every data has to be counted in the regression with its tendencies.

It is shown that even one corrupted data leads to conflicting and possible opposite results. For example it may show a variable as significant while it is not or makes a variable become dropped. This comes from the fact that OLS tries to minimize the residuals of all the points, as a result, a data far away from all the other data points tendency increases the residuals of all the other points.

If we could make a metaphor, if the residuals represent the desire of food of the persons in a society, then each would have different desire. It is sometimes the case that, although the biological organism does not need to be fed always, people eat much more than their need. Assuming that there is an outlier person in such a small society, and also assume that people are trying to apply the equalitarian OLS idea.

If for most of the people the desire for food is just their biological need, than OLS idea would make them loose weight while the outlier fad person would gain kilos. However, if they say that we believe in OLS but after a Robust reasoning, and if they apply this idea, then the outlier would be forced to loose the excess kilos, which is in fact more equalitarian, because it does not harm everybody.

As a result, blind application of OLS may lead to harmful results. Therefore, it is suitable to go through the concept of the effects of corrupted data on OLS.

Robust Regression can be explained as the a technique to identify the data which do not conform to the tendency of the majority of observations. It is a natural result that after identifying such data's it is logical to cancel their effects. In order t make the Robust Regression analysis easier, I have written a program in GAUSS using the tools that can be used in this kind of regression. It identifies the points, which change the results in their absence.

The thesis contains the explanation of the written program ROBUST, experimented data and results drawn out of these experiments and applications. The applications contain the illustration of the superiority of the Robust Technique used.

2. Robust Regression

2.1 Outliers and OLS

In the literature we have standard assumptions and we can state the robustness of an estimator to how well the estimator works under failures of the standard assumptions. The most important assumption is the normality assumption. Normality is important because of the central limit theorem and computational easiness. The central limit theorem suggest that many random variables may be reasonably well approximated by normal distribution. The second and more important fact is that computational procedures with normal distributions are easy to carry out[1].

Theoretical and technical developments have introduced new techniques, which doesn't require the normality. In addition, easiness of the normality is replaced by computer power. New computers have made possible to shorten the calculation time and even introduced techniques to simulate the computations[2] by regenerating some variables. The robust procedures are the result of these developments. Sometimes calculations require exact normality, the lack of this requirement leads to very different results, even when they are close to normal. Robust procedures do not require normality and therefore are superior to the classical procedures[3].

---

[1] We could have also used the absolute value instead of the square term, and then minimized the sum of absolute residuals. However, minimizing the sum of squared residuals is substantially easier.
[2] Bootstrapping or generating random numbers with help of computer.
[3] Central Limit Theorem also produces approximate normality but classical procedures require exact normality.

2.1.1 Sensitivity of OLS

In figure 1 we can see how sensitive the OLS (Ordinary Least squares) estimator is for a regression of the type $y=\beta_1 x+\beta_0$ in case of bad data or an outliers[4]. The bold line shows the actual trend of the data in the case that the first point is recorded correctly (case a). In the other cases, as the point deviates from the actual position the estimator gets worse. Therefore, the OLS is very sensitive to corrupted data.

If we define the i^{th} residual as $r_i^2(\beta) = (y_i - x_i\beta)^2$ then the LS regression would try to minimize the sum of squared residuals, which is : $\sum_i r_i^2$.

In figure[5] 1 we see the rotation of the LS line when the residual of the first point gets bigger and bigger. OLS tries to minimize the total sum of squares and in this case the first point adds a very large number to the sum. So it has to be minimized more than the others. As a result, the true tendency cannot be selected in this case.

In figures[6] 2 and 3 we have the same situation. In figure 2.a.) we have a negative relation between x and y, however, in 2.b.) it seems to be that there is no relation between x and y. In all of the figures 2,3 and 4 we see points which are very far apart from the point cloud and therefore they named as *leverage* point.

As explained above, the influence of the leverage points is obvious and it is not always the case that they have a negative effect. In the first three figures they add a large amount to the sum of residuals, increase the variance and show wrong results. However, in figure[7] 4 the leverage point is in accordance with the observations and the tendencies. This point is a justification to the tendency of the data and therefore may determine the value of R^2. We can see from figure 5, that in case a.) the points decreases the value of R^2 to a lower degree with

---

[4] Outliers is a special kind of data, which does not show the general trend of the other observations..

[5] Appendix page i

[6] Appendix page i

[7] Appendix page ii

wrong direction compared to the dashed tendency. In b.) part it increases R^2 to a higher degree and shows a relation where in fact there is none.

Given that OLS is extremely sensitive to outliers and abredant data, a natural way to continue is to identify this sensitivity. The main idea is to delete one or several observations and study the impact on various aspects of the regression. When one observation is deleted, all of the regression statistics change, but we should keep in mind that if there are wrong observations in the data set these statistics are in fact wrong. A number of methods have been proposed to asses the impact of dropping an observation on various regression statistics.

The intended objective of sensitivity analysis is to asses whether the OLS regression results are seriously affected by the presence of a small group of observations. While sensitivity measures taken in combination can, in hands of experts, achieve this objective, there now exist simple ways, which can routinely identify the extend to which OLS is affected by outlying subgroups of observations. These include the high breakdown regression estimators, including LMS[8], LTS[9],MVE[10] and MCD[11]. Application of these techniques immediately reveals the presence of subgroups of observations that differ substantially from the other points or exert undue influence on, the regression results. The written program named Robust, uses this tools to identify the outliers, which are explained in a while.

2.2 Least Median Squares, LMS

Least median squares technique has several properties, which makes it very attractive for the preliminary analysis. In particular, if results from an LMS analysis are similar to OLS, we can safely conclude that no small subgroup of the data is causing undue distortions of the OLS results.

The problem with the OLS is that it fails in the case of even one of the observation is bad, as we have seen in the previous section. Obviously OLS will fail again if the number of bad

---

[8] Least Median Squares, by Peter J. Rousseew 1984
[9] Least Trimmed Squares, by Peter J. Rousseew 1984
[10] Minimum Volume Elipsoid, by Peter J. Rousseew 1984.
[11] Minimum Covariance Determinant, by Peter J. Rousseew 1984.

recordings increases. We obviously need to deal with larger number of bad observations. The LMS has the property that it works even when half of the data is contamined.

If we want to explain the process of LMS we have to consider some specific residuals. For any β, let $r_i^2(\beta) = (y_i - x_i\beta)^2$ be the squared residual at the t-th observation. Rearrange these squared residuals in increasing order by defining $r_1(\beta)$ the smallest residual, $r_2(\beta)$ the second smallest and $r_k(\beta)$ the k^{th} in the order of residuals. The LMS is defined to be the value of β for which the median of the squares is the smallest possible. The median is most of the time the half of the number of observations.

If we take the median of the residuals, we allow some points to have very large residual numbers. This situation avoids cases as in figure 2-5 and it acts as if ignoring the presence of far apart points.

In the case of figure 1, LMS will give the first data the highest residual and the assigned coefficient b will be as if the first data is not present. In figures 2 and 3 the dashed lines show the result of the minimization of the median residual. They are similar to the original LS regression with the correct data. More explanation about this type of regression is in the program algorithm part. As a result we can summarize LMS as:

$$\underset{\beta}{Minimize\_} \underset{i}{med}(r_i^2)$$

In order to minimize the this special number we can consider to take subgroups of sum fixed number. If we assume that we have 10 observations and 3 variables. We can consider all of the subgroups of size 3 made up of these 10 observations, which makes 120 possibilities. You take one of these subgroups, draw the best-fit line going through the and consider this line as the regression line and look at the median residual. If this is the smallest among the 120 other median residuals then record the beta or coefficient as the solution of the LMS coefficient. This process can be used also in the following sections.

2.3 Least Trimmed Squares, LTS

In this regression type the outliers are identified by looking for cases where the sum of the smaller number of residuals is preferred. It is more alike to OLS than the LMS. In this case the first half of the smallest residuals are added together. Again defining $r_1(\beta)$ the smallest residual, $r_2(\beta)$ the second smallest and $r_k(\beta)$ the k^{th} in the ordered form from smallest to the largest, you add up the first h of them, where h is half of the data or any number that does not include the corrupted data.

As a result, LTS corresponds to minimize: $\sum_{i=1}^{h} r_i^2$

The difference between LMS and LTS is in the considered number of points. LMS concentrates only on one point while LTS concentrates on at least half number of points. So the concept of efficiency says that both are not as efficient as the OLS, which considers all of the points. The LTS has at least 50% of efficiency, while the LMS has 0% . However, if we use this type of regressions to identify the outliers and then make OLS after the dropped observations then the efficiency increases. In other words, if we delete these outliers and run OLS in the rest (Reweighted OLS), then we have the highest possible efficiency.

Both LMS and LTS are means to detect the minimum possible residuals and so detect the outliers. In order to identify the leverage points, which play the decisive role we need to introduce two more methods.

2.4 Minimum Volume Ellipsoid, MVE and Minimum Covariance Determinant, MCD

The idea in the MVE is to detect the points that are far away from the rest of the point crowd. In figure 5 the three points are the leverage points, which are far away from the center of the crowd.

The classical method to determine the leverage points is the Mahalanobis distance, which has the following form:

$$MD^2(x_i, X) = (x_i - T(X))C(X)^{-1}(x_i - T(X))^t$$

n: Number of observations

$$T(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad C(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - T(X))^t(x_i - T(X))$$

The arithmetic mean *T(X)* is a kind of measure for the center of the point cloud and the matrix *C(X)* a kind of measure for the variance or spreadness. If a point far away from the center with some factor larger than the spreadness is identified as leverage point by the classical method. The classical method of leverage point detection suffers from the same reason as the OLS does. It considers all of the points and in turn the influence of the outliers also. This may be beneficial for the bad leverage points.

With the increasing number of bad data the spreadness matrix *C(X)* increases and it becomes even harder to identify the points far apart. In figure[12] 6 we can see the 97.5 tolerance ellipsoid that is a special ellipsoid. After that ellipsoid the points are identified as leverage point and are rejected. The big ellipsoid, which is influenced by the three leverage points contains them. The robust distance ellipsoid, however, identifies them.

The way to identify the outlying points is in the same way as in the previous two topics, namely, finding a way to discriminate them in forming some subgroups. In MVE you form a subgroup of some predetermined number of elements and minimize the a function, which is proportional to the volume they form. This volume should contain at least half of the elements or the number of points wanted. At the end the subgroup that has the minimum volume and containing the desired number of elements is accepted as the base or central group. According to this subgroup *T(X)* and *C(X)* are determined. ,

Standartized residuals are formed by dividing the residuals by standard deviation, and usually standartized residuals with large number are accepted as outliers. In this case also we want to identify points far away from the cloud if they have a distance larger than the chi-

[12] Appendix page ii

square distribution 97.5% and variable times of freedom. These are the leverage points. Formally, they are far apart if:

$$(x_i - T(X))C(X)^{-1}(x_i - T(X))^t \geq \chi^2_{degree,0.975}$$

In the case of MCD, you take all possible combinations of half number of data points and try to minimize the covariance matrix *C(X)*. In this case also the minimizing subgroup determines the mean and covariance and similar to MVE the points exceeding some value are regarded as *leverage* points.

In all of the four robust regression procedures, the process is to look for some subgroups. The idea is that if there is a tendency than this subgroup should have the pure tendency and all others are an accumulation to it. The ones not suited will be seen as outlier or distant point (leverage point).

2.5 Famous Example

This is a famous example, because it reflects a real world situation where the facts are

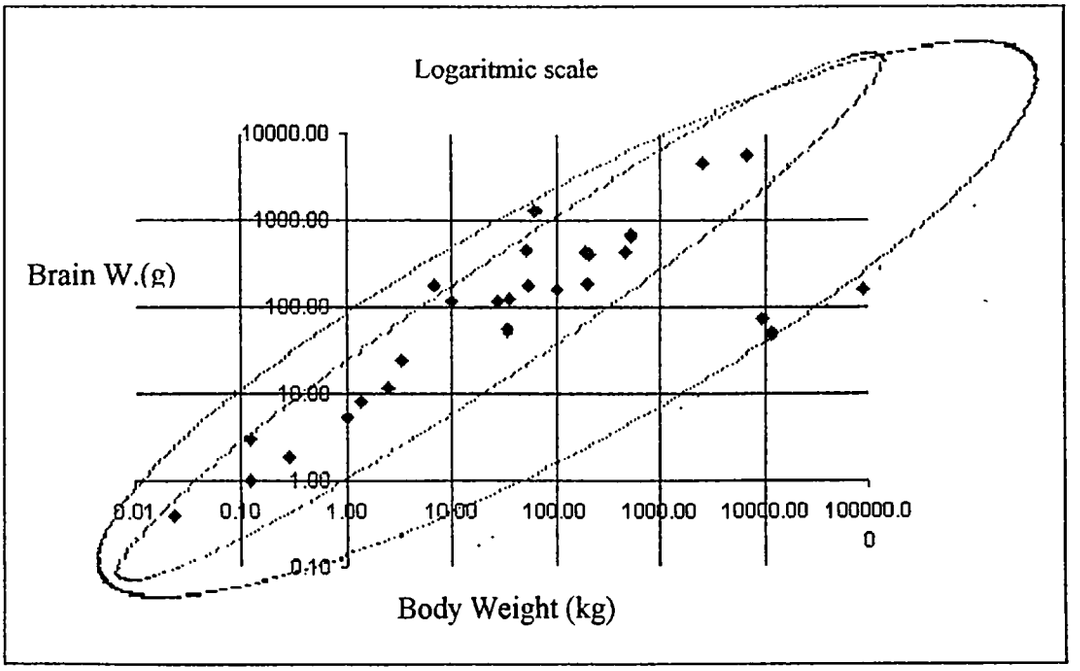


Figure 7. Scatter Plot of Table1

observable and used by most of the authors in this concept. Table[13] 1 summarizes the relation between brain-weight of a species and the body-weight of it with the previous figure.

In figure[14] 6 and 7 we can see tolerance ellipsoid for the MD and the robust distance. Obviously as in table 1, the classical methods are unable to detect the difference between the recent habitants and the extreme cases, which are the very oldest and very new ones. The MD identifies only one dinosaur as leverage point, while the robust distance identifies all of the five species that are different from the other points. The results of OLS are:

F-Value = 40.26061840	Res. SS.= 28.41083006	Std. err= 1.045334508	
R2-Value=0.6076100612	Est. SS.= 43.99375337	Case Num=	28
Adj. R2 =0.5925181405	Tot. SS.= 72.40458343	Var. No.=	2

Variable	Est.Coe.	Confidence Intrv.	Str. err	T-Value	P-Value
Body W.	1.225	(0.6880 to 1.762)	0.19307	6.3451	0.00000
Constant	-0.71659	(-1.8858 to 0.452)	0.42037	-1.7047	0.10018

If we exclude the variables which have a standartized residual larger than 2.5 deviations and change the direction of the direction against the direction of the trend of the data and repeat the regression we get the following results:

F-Value = 556.9207362	Res. SS.= 1.710742488	Std. err=0.2854188641	
R2-Value=0.9636628370	Est. SS.= 45.36895075	Case Num=	23
Adj. R2 =0.9619324959	Tot. SS.= 47.07969324	Var. No.=	2

Variable	Est.Coe.	Confidence Intrv.	Str. err	T-Value	P-Value
Body W.	1.2834	(1.1292 to 1.4376)	0.05438	23.599	1.3387E-16
Constant	-1.0677	(-1.4001 to -0.7352)	0.11721	-9.1089	9.6762E-09

The results suggests that leaving the bad residual points out we get an almost exact relation between these variables. The residual sum of squares decreases drastically and the standard error makes forecasting four times precise. The R^2-value and the F-value increases to show the increase of precision in the regression. Therefore, if OLS shows a relation between the variables we have to keep in mind the possibility of figure 2,3 and 5. An outlier may cause to make a positive relation look like negative or show no relation.

---

[13] Appendix page vii

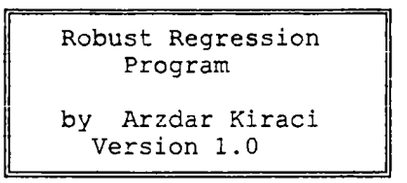
[14] Appendix page ii

3. The Program Robust

3.1 A Guide to The Program Robust

I have met no program yet that includes four of the techniques LMS, LTS, MVE and MCD. Therefore there exists also no program that decides on the *good* or *bad leverage* points. I hope to be useful to the ones who use it.

The program starts with:



at the top and a help line at the bottom indicating the possible operation or the limits of the input during the choice. For example we see at the bottom:

```
Please use the Arrow keys and also Spacebar or ENTER for option change
Press any key to continue
```

During the menu choices you have to use the cursor keys to move the cursor to the choices and press space or ENTER (which do the same job and both are used to accept the choice where the cursor is on). If a choice is made the squared brackets disappear the printing changes to large caps, for example, before choice:

```
Type of Robust Regression:     [Lms]     [Lts]     [Mve]     [Mcd]
```

and after pressing space:

```
Type of Robust Regression:     [Lms]     [Lts]      MVE      [Mcd]
```

At any time you can move the blanking cursor key to the desired level and change the selection.

There are two main menus, one to make general choices and second to make regression specific choices. The first menu looks like the following:

```
Type of Robust Regression:      [Lms]      LTS       [Mve]      [Mcd]
Combinations:                   [Randomx .....]       EXACT

Input file name :brainlog.dat
Output name     :brainlog.lms

Case number     :28
Variable number :_


Variable position   :

How much output     :           [Small]          [Medium]         [Large]

Data plot on        : [Non]     [Estimated]      [Index]        [Both]

Outer Diagnostics :                  [Yes]      [No]

                                    [Execute]


═══════════════════════════════════════════════════════════════════════
Enter variable number in your data set
```

If we go through the explanation of each step we can say the followings: For the type of the regression you jump to the position of the regression you want to perform and press space, for example if you choose MCD it will appear to be:

```
Type of Robust Regression:      [Lms]      [Lts]       [Mve]       MCD
```

For the choice of the replications, you can choose to have all of the possible combinations or randomly generate some subgroups. If the data set is very large with many variables it may take a very long time to arrive at an exact solution. With a large number of random generations you will end high probability in an exact solution.

If you are in the position "[Exact]" press space and make your choice. If you want random generations, you have to enter the number of replications you want. In this case if your cursor is blanking in the place of [Randomx .....], begin to write the number. If you press a non-number element a beep signal will be heard, requiring you to give the correct number. This will look like as follows:

```
Combinations:                    Randomx 1250        [Exact]
```

In the third position you have to choose the name of the input file. The file has to be in the same directory as where you run this program. If you write the file name wrong or if it does not exist an error message will appear. As long as the cursor is in the same position you can enter the name, however you have to press Enter or space in order to be accepted. If it is accepted, the name will be written immediatelly. It will look like the following while you are writing:

```
Input file name :brainl_
```

After that you have to enter the output file name, with the care that if there exists a file with the same name that it will be erased. Same rules as input file name applies also here. As an example we can consider the main menu above.

In the option "Case number:" you have to enter the observation number. Your data set have to be in row major ordered, that is every row in your data set have to contain one observation. If your data set is not compatible with the information you entered the program will not execute. The main menu is an example for the appearance.

In the option "Variable number:", you have to enter the number of columns in the data set. It should include all of the variables, even if there exists an index colomn. In the next stage you can exclude the variables you don't want to have in your analysis.

In the next stage where you enter the position of the variables include to give the names of the variables and their positions. If you press space or enter when you are in this option a screen will appear as follows:

Name	Position
Res.Var.	0

ok

You write the name of the response variable in this box and give its position. By moving the cursor key to the bottom you can finish this section. Again, if you make a wrong input an error message will appear and a new input is desired. After that the following menu will appear:

Variables to exclude			
Name	Exc- Inc-	Name	Exc- Inc-
Res.Var.			
Var 1	[ ] [x]		
Var 2	[ ] [x]		

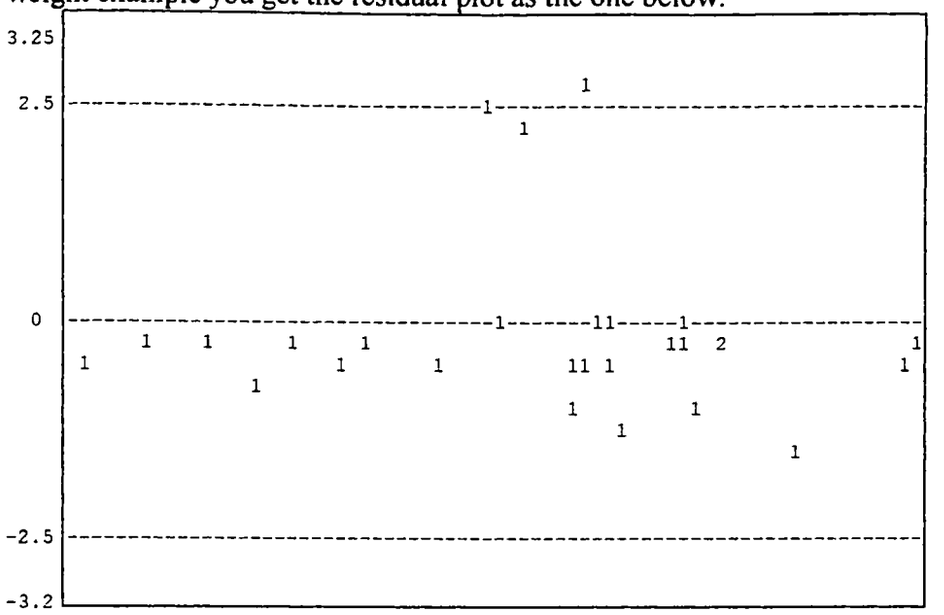
ok

In this menu you can change at any time the variable name and if you want it to be excluded from the regression you can go to the "Exc-" part and exclude it. At the end you go the "ok" part and press space.

Then you are asked how much the lenght of your output should be. If you choose small output then your results will be limited with the basic results. The observations are not printed on the output file. If you choose medium output the OLS results are extended. In large the maximum possible output is produced and if you want outlier diagnostics you have to make that choice to see the list of good and bad leverage points, as in example below:

How much output : [Small] [Medium] LARGE

It also possible to see the distance and residual plots. They use the DOS prompt as output so they are not supported by any graphic option possibility. For the brain weight and body weight example you get the residual plot as the one below:



The numbers indicate the number of points in this range. The horizontal axis is either the index, which is the order of the data, or it is the order of the estimated response variable from smaller to the larger one. The same is true for the distance plots. The residual plots are for LMS and LTS, the distance plots are for MVE and MCD. As in the example below if you select "Non" no plot is drawn. If you select "Estimated" the estimated values vs. the residuals are drawn, this case changes to index vs. residual plot if you select the mode "Index". If you select both, both are drawn. The same output is produced with different locations on the horizontal axis. In example below you can make the index choice.

```
Data plot on        : [Non]     [Estimated]     INDEX       [Both]
```

In example below you make the decision to outlier diagnostics. Outlier diagnostics runs as follows: First you run LMS or LTS to provide the necessary residuals for the outlier diagnostics and the run MVE or MCD to get the distances necessary for the last part.

```
Outer Diagnostics  :          [Yes]          [No]
```

At the end you move the cursor to the bottom to execute the program. If there are missing or wrong values, the explanation will appear and the correct values are desired. Only after that you can go to the stage below.

```
[Execute]
```

In the next stage the following menu appears for LMS,LTS and MVE. The MCD does not the menu because it automatically takes half of the data to the minimization process. For LMS and LTS we have:

```
Do you want a constant to be added : YES          [No]
Give the size of the elementary set: ALL          [Size:]
Which element should be minimized  : _
                                [Execute]
```

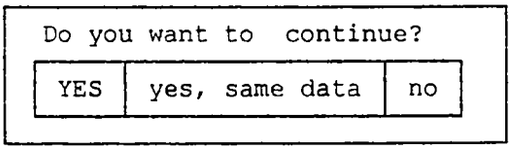
For MVE we have:

```
Do you want a constant to be added : YES          [No]
Give the size of the elementary set: ALL          [Size:]
Ellipsoid contains how many points : _
                                [Execute]
```

If you don't want to add constant you move the cursor to the right side and press space, by default a constant is added. By moving the cursor to the size part you can select the size of the elementary subgroup, but a more time consuming choice is the choice of all elementary subsets starting with the minimum possible and ending with the regression on all the elements At the end selecting the best subgroup that minimizes the target element (LMS) or target sum (LTS). At the last part the size of the target element is given. While executing the program the following screen will appear:

```
Cycles left: 1000
Time passed:    1 sec
```

After going through the calculations the last choice must be made to continue with same data or not. If you select Yes all information must be again entered, but if you choose the old date you just have to enter the new regression type, but we don't have to forget to change the output file name if necessary. At the end the following screen will appear:



3.2 Algoritms used in Robust

3.2.1 Exact Algoritm

The idea in robust regression was to generate subgroups and find the best subgroup that minimizes objective function. The objective function was the median residual in the regression LMS, it was the partial sum of the smalllest residuals in LTS, it was the volume funciton in MVE and it was the covariance matrix in the MCD.

In the exact algoritm all possible combinations of the subsamples had to be generated. In order to achieve this goal, two matrices of the elementary subgroup size are formed. Assuming that there are 15 observationsand the elementary subgroup is of the size 4, then the two matrices look like the following when initialized.

$$indxit = \begin{bmatrix} 1 & 2 & 3 & \underset{\uparrow}{4} \end{bmatrix} \qquad icr = \begin{bmatrix} 12 & 13 & 14 & \underset{\uparrow}{15} \end{bmatrix}$$

The last colomn of the matrix indxit acts as a counter. After getting and ordering the residuals, the objective function is calculated and if it has the minimum it is recorded as best index. In order to get the other subgroup the last colomn is incremented by one. It is incremented till it is equal to the last colomn or the corresponding colomn in the icr matrix. After incrementing the matrix indxit gets the following form.

$$indxit = \begin{bmatrix} 1 & 2 & 4 & \underset{\uparrow}{5} \end{bmatrix} \qquad icr = \begin{bmatrix} 12 & 13 & 14 & \underset{\uparrow}{15} \end{bmatrix}$$

Every time the last colomn of indxit equals to the last colomn of the matrix icr, the colomn before is incremented. This is true also for the other colomns, if they are equal to the corresponding colomn in icr the previous colomn is incremented by one and the following colomns have the numbers following it. For example in the following case we see before incrementing and after incrementing.

Before: $indxit = \begin{bmatrix} 2 & 13 & 14 & \underset{\uparrow}{15} \end{bmatrix} \qquad icr = \begin{bmatrix} 12 & 13 & 14 & \underset{\uparrow}{15} \end{bmatrix}$

After: $indxit = \begin{bmatrix} 3 & 4 & 5 & \underset{\uparrow}{6} \end{bmatrix} \qquad icr = \begin{bmatrix} 12 & 13 & 14 & \underset{\uparrow}{15} \end{bmatrix}$

This goes on until both matrices are equal to each other. The number of trials equals:

$$\frac{(number\ of\ observations)!}{(number\ of\ observations - variable\ number)!\ *\ (variable\ number)!}$$

This value increases exponentially with increasing number of observations, as in the figure[15] 8 in the appendix part. Therefore in large number of observations, even with the fastest PC a data set with 75 observations may take days to finish the program. Therefore, the alternative solution which takes a shorter time is the random generations method.

3.2.1 Random Algoritm

The idea in the random algoritm is to generate a random index set and select the subsample according to that index. The trade-off between exactness and randomness is that in random samples if you select a small number of random generation you may miss the global solution. In the following sections we will try to show that the results of random generation also produces satisfactory results.

In order to explain theorethically the possibility that also random generations produces good results we have to define ε , which is the ratio of bad data in the data set. We have the following formula for the probability that the global solution is reached.

$$P = 1 - (1 - (1 - \varepsilon)^{Variable\ number})^{Random\ Generations}$$

We want this number P to be 95% or 99%, as large as possible. Table 2 below summarizes that how many numbers of generations needed for the percent number of contamined data to the dimension in hand if the observation number is very large compared to the variable number.

[15] Appendix page iii

Table 2. Number of random generations needed in % contamined data

Variable Number	5%	10%	20%	25%	30%	40%	50%
1	1	2	2	3	3	4	5
2	2	2	3	4	5	7	11
3	2	3	5	6	8	13	23
4	2	3	6	8	11	22	47
5	3	4	8	12	17	38	95
6	3	4	10	16	24	63	191
7	3	5	13	21	35	106	382
8	3	6	17	29	51	177	766
9	4	7	21	36	73	296	1533
10	4	7	27	52	105	494	3067

If we compate this table with the case that how many combinations is needed by exact solution we get Table 3 below.

Table 3. Number generation for exact solution for the given data number

	2	3	4	5	10	20	40
Variable Number							
1	2	3	4	5	10	20	40
2	1	3	6	10	45	190	780
3		1	4	10	120	1140	9880
4			1	5	210	4845	91390
5				1	252	15504	658008
6					210	38760	3838380
7					120	77520	18643560
8					45	125970	76904685
9					10	167960	273438880
10					1	184756	847660528

We see that for a dataset of size 20 and 4 variables, containing half of bad observations the random generation requires 47 number of generations while the exact solution requires 4845 number of computations. The exact solition requires 10 times more computation. This however, does not imply that the random generations solution in computer is 10 times faster.

In most of the computer programs there is no built in random number generaor that generates different numbers of data for the given size. If you want to have a subsample with different indices than you have to spend some time on getting new indices for the dublicated ones. This decreases the ratio of the exact time solution to the random from 10 to lower degrees.

The random generations in the program Robust work in the same principle as explained above. If we would explain it by a case assume that we have 15 observations and our subsample is of the size 6. We first fill the matrix "indxit" with random index numbers,as below.

$$indxit = [15 \quad 3 \quad 7 \quad 3 \quad 12 \quad 8]$$

If the variable number is not negligible in comparison with the observation number, then the dublicated index numbers increases. The program, therefore, checks if the matrix has dublications in it. If the dublications are small in number, for example, 20% of the total observations then it again tries six times to fill the dublicated ones with only one new generated index and checks if it is just before choosen.

If the duplications are in small number or if the observation number is large compared to the subgroup chosen then the above process is suitable. If however the subgroup element size is not small compared to the case number, generating an index and looking if it found before, and if it exists generate a new one may cause an infinite large trial and error loop. In this case it is more suitable to switch to a more guaranteed method of filling the duplications. For this case we generate a new rest matrix "rester", which is made up of indices that do not exist in the first generated matrix "indxit". For the matrix above we get the following matrix.

$$rester = [1 \quad 2 \quad 4 \quad 5 \quad 6 \quad 9 \quad 10 \quad 11 \quad 13 \quad 14]$$

In order not to lose the randomness property, from the rest matrix we randomly select one and add to the index matrix. For the example above we randomly generate a number of 1 to the rest size, which is 10 in this case and select the index number there and add it to the index matrix.until the index matrix is filled. During this process the rest matrix shrinks in every trial by one. After the choice above the rest matrix becomes as follows.

$$rester = [1 \quad 2 \quad 4 \quad 5 \quad 6 \quad 9 \quad 10 \quad 13 \quad 14]$$

3.3 Effectiveness

In this we will ask the question that if the random process is effective, because it has a shorter calculation time. The experiments performed are summarized in the tables in the appendix[16] and below. I have taken two extreme cases one is where observation number is very large compared to the variable number and one for which it is not negligible.

The first case is for Wood data set for which an observation number of 20 against a variable number of 4 exists. The results are summarized below. For the first table below we see that for replications below 8000 the estimated coefficients for the reweighted OLS differ but the significance level has only an error for replication below 1000.

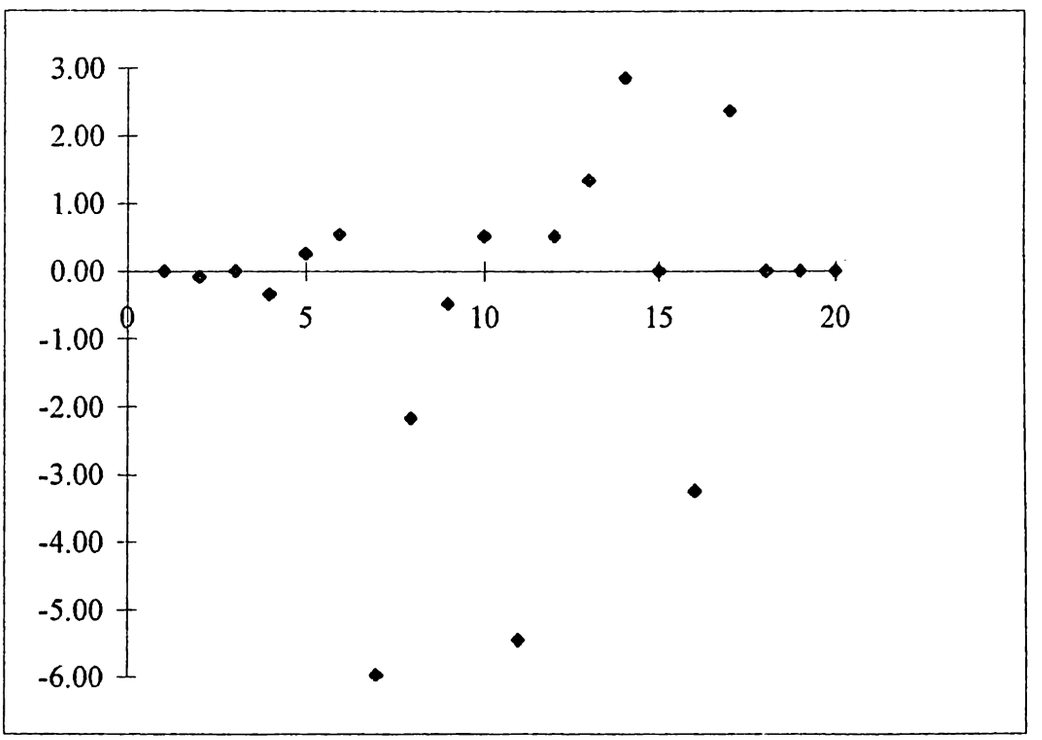
To get the correct answer for the reweighted OLS we have to make at least 8000 replications because only after that we find the correct outliers. For the replication number below 8000 we get for the worst case a 20% deviation of one of the coefficients. For example, for 500 replications we have 9 results out of 20, which have different coefficients than the exact solution. Two of them have a deviation of the coefficient from the true value of 27%, more to that 18 results have identified the confidence interval correctly. Therefore, if you want to see the significance of some coefficients you have to select a replication number larger than 500.

[16] Appendix page 9,10

Replication	*Elementary set found*	*Excl.Data found*	*Excl. Set found*	*Reweigted dev*	*Coef*	*Conf. int. identified*
38760						
500	5	11	9	2x%27		18
1000	14	14	6	6x%20		20
4000	17	17	3	1x%20		20
8000	20	20				20
20000	20	20				20
40000	20	20				20

Table of experimenting on Wood.dat for LMS

The data set wood.dat is made up of 20 observations and 4 variables and below we have the normalized data plot after LMS



Previously we have given numbers in Table 2, which was for the case that variable number was negligible with respect to observation number. We have to note that by looking at the figure we can see that there are no clear cut outliers, which makes difficult to find good subsamples.

As a result, if the number of variables is relative to observation number not negligible, if the variables are spread or have high variance and if the number of outliers is large we have to increase replication number up to degree of 8000 replications, which is one fifth of the replication number needed by the exact solution but takes two times less duration to be executed.

In the opposite extreme we have the data table of the data set brainlog, which we often mentioned before. This case has three dinosaurs as clear outliers and 2 species that are just on the line. The next table summarizes that in all cases at least in one second you'll get the correct result. This is so because all of them correctly identify the outliers and get the expected result for reweighted OLS. This result is compatible with the suggestions of table 2.

Replication	*Ellementary set found*	*Excl.Data found*	*Excl. Set found*	*Reweigted Coef dev*	*Conf. int. identified*
378					
15	9	20	20		20
30	6	20	20		20
60	9	20	20		20
150	16	20	20		20
300	17	20	20		20

The same fact is also true for that case also. As the MVE advices to take the elementary subset size as the variable number plus one(, which is one element more than the LMS or LTS). This in fact makes the execution time larger. It becomes also harder to find the correct elementary set. For the tables below we have the similar comments as above.

Type	*Replication*	*Execution time*	*Ellementary set found*	*Leverage pt found*	*Worst possible*
MVE					
Exact	77520	1370			
Random	250	5	0	13	5
	500	11	0	17	2
	2500	115	1	20	
	5000	229	1	20	
	10000	458	0	20	

As in the LMS part for the Brainlog data, we find in very few replications the target elementary set and leverage points. The execution time of MVE is larger because it needs more calculation and has one more variable to account. However this is not the case for the Wood data, it happens for the few replications worst possible happens,i.e. it identifies no leverage point.

Type	*Replicatio n*	*Execution time*	*Ellementary set found*	*Leverage pt found*	*Worst possible*
MVE					
Exact	3276	31			
Random	200	4	10	14	
	400	9	18	19	
	800	17	18	20	
	2000	35	20	20	
	4000	71	20	20	

Finally, the results obtained in random replications depend on the data set used. A data set with clear and small number of outliers will give the result in replications few than the number of fingers in a hand. In addition the variable number has to be negligible regarding the observation number. In contrast, if we have the opposite case it is preferable to use the exact solution.

4. Applications

In this section the idea that the outliera affect the results very much is investigated. Robust regression is applied on various data sets and the results are discussed. All data's are in the appendix part.

4.1 Urban Unincorporated Places[17]

4.1.1 Extreme Case

These data are taken form Country and City Data Book, 1962, Table 5, pages 468-475. They are examples for unincorporated places (with populations at least 25000), including the median family income for each. The last ten of these are representatives for the 10 highest median income and the other ten for the lower ones. I have taken these as an example because of the bias in selection and because the source I have taken used this also as an example.

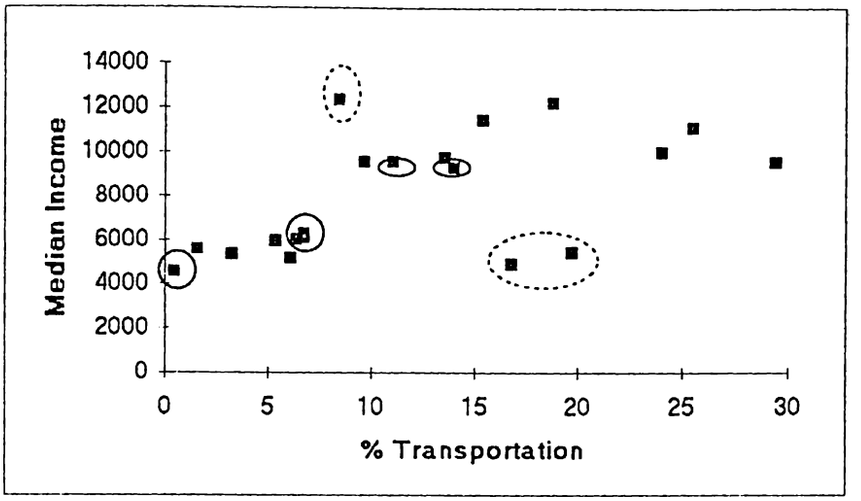
We get the following results:

F-Value = 29.03660301	Res. SS.= 6219615.259	Std. err= 751.9438603
R2-Value=0.9547869937	Est. SS.= 131342908.5	Case Num= 20
Adj. R2 =0.9219048072	Tot. SS.= 137562523.8	Var. No.= 9

Variable	Est.Coe.	Confidence Intrv.			Str. err	T-Value	P-Value
Varl 1	-58.7	(-349	to	231)	93.27	-0.62899	0.5422
Varl 2	-220	(-448	to	7.39)	73.199	-3.0084	0.011899
Varl 3	41.7	(-22.4	to	106)	20.622	2.0244	0.067904
Varl 4	33.5	(-51.5	to	119)	27.343	1.2266	0.24558
Varl 5	908	(-290	to	2110)	385.38	2.3572	0.038
Varl 6	-9.48	(-92.4	to	73.4)	26.656	-0.35551	0.72893
Varl 7	-251	(-515	to	12.7)	84.809	-2.9597	0.012982
Varl 8	23.4	(-51.9	to	98.8)	24.226	0.9677	0.35399
constant	2.13E+04	(-9080	to	5160)	9756.2	2.1783	0.052018

[17] Example taken from, Data Analysis and Regression, Frederick Mosteller & John W. Tukey

The author of the given example claimed that Varl 4, which is % using public transportation, showing a relation in this multi-dimensional space. He comes to the strange conclusion that the more affluent are using more public transport by making the following plot.



As the evidence shows there seems to be a suprising relation between these variables. One has to ask the question that "are the richer rich because they are more stingy". We see that the OLS could not identify any significance relation. After making the LMS and identifying the outliers we get the following results for OLS.

F-Value = 221.8849714	Res. SS.= 481333.6093	Std. err= 262.2249877
R2-Value=0.9960720052	Est. SS.= 122057936.1	Case Num= 16
Adj. R2 =0.9915828684	Tot. SS.= 122539269.7	Var. No.= 9

Variable	Est.Coe.	Confidence Intrv.	Str. err	T-Value	P-Value
Varl 1	-67.7	(-211 to 75.7)	35.847	-1.8892	0.10079
Varl 2	-371	(-496 to -247)	31.158	-11.917	6.6631E-06
Varl 3	-4.61	(-57.1 to 47.8)	13.115	-0.35176	0.73537
Varl 4	37.5	(-2.93 to 77.9)	10.107	3.7104	0.007551
Varl 5	1760	(782 to 2730)	243.78	7.2065	0.00017645
Varl 6	-22.4	(-64.5 to 19.7)	10.533	-2.1265	0.071037
Varl 7	-394	(-529 to -259)	33.788	-11.656	7.7253E-06
Varl 8	22.5	(-12.1 to 57)	8.628	2.6023	0.035304
Constant	3.2E+04	(1.72E+04to4.68E+04)	3692.2	8.6662	5.4508E-05

The results of the reweighted OLS show that there is no relationship between % transportation and the income of family. So people are not rich because the are stringy but because they have not moved to this city in recent years (Varl 8) and because they live in

suitable homes (Varl 5). In addition, they have changed their homes when they have changed their financial status (Varl2). OLS did not identify any of these variables as significant and at the end there seem to be a wrong relationship, by intution. However, LMS picked up the most logical ones as significant out.

We can see from the figure that the outliers are from the linear trend (ellipses) and not from the possible outliers (dashed ellipse). So classical OLS and the author failed in this case.

4.1.2 More Data

If we add more intermediate data to the previous example in order to see what happens to the results we get the followings.

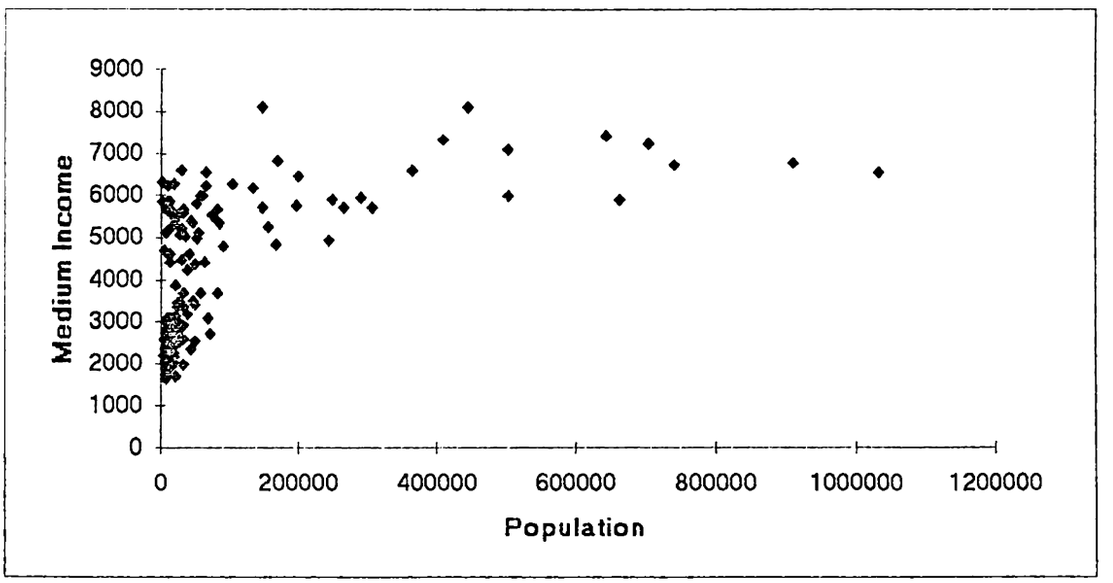
```
-------                OLS  Results                              --------
F-Value = 16.37391311     Res. SS.= 17571075.44     Std. err= 961.6618137
R2-Value=0.8733259904     Est. SS.= 121139900.0     Case Num=          28
Adj. R2 =0.8199895653     Tot. SS.= 138710975.4     Var. No.=           9

Variable   Est.Coe.   Confidence Intrv.     Str. err   T-Value    P-Value
--------   --------   ------------------    --------   --------   --------
Varl 2          155  (-74.6    to     384)     80.07     1.9314    0.06849
Varl 3        -16.9  (-135     to     102)    41.417    -0.40764   0.68809
Varl 4         15.7  (-29.5    to    60.9)    15.777    0.99446    0.3325
Varl 5         38.7  (-51.1    to     128)    31.339     1.2336    0.23241
Varl 6      1.4E+03  ( 206     to2.59E+03)    416.11     3.3582    0.00330
Varl 7         32.5  (-41.7    to     107)    25.908     1.2548    0.22478
Varl 8        -3.82  (-146     to     139)    49.724    -0.076796  0.93959
Varl 9         37.1  (-28.8    to     103)    22.996     1.6125    0.12334
Constant      -8260  (-2.69E+04to1.04E+04)     6502      -1.27     0.2194

---------          Reweighted  OLS  Results                         ------
F-Value = 196.1956855     Res. SS.= 938049.6419     Std. err= 279.5904210
R2-Value=0.9924125810     Est. SS.= 122694195.0     Case Num=          21
Adj. R2 =0.9873543017     Tot. SS.= 123632244.7     Var. No.=           9

Variable   Est.Coe.   Confidence Intrv.     Str. err   T-Value    P-Value
--------   --------   ------------------    --------   --------   --------
Varl 2        -48.6  (-142     to    44.7)    30.483    -1.5953    0.13663
Varl 3         -377  (-467     to    -287)    29.329    -12.849    2.2497E-08
Varl 4         -3.3  (-20.5    to    13.8)    5.6001    -0.59008   0.56608
```

Var1 5	40.5	(10.3 to 70.7)	9.8628	4.11	0.0014464
Var1 6	1.61E+03	(1.13E+03to2.09E+03)	157.69	10.20	2.8814E-07
Var1 7	-18.4	(-46.3 to 9.45)	9.1051	-2.0242	0.065793
Var1 8	-396	(-497 to -295)	32.918	-12.035	4.6804E-08
Var1 9	28.5	(5.81 to 51.3)	7.4211	3.845	0.0023314
Constant	3.2E+04	(2.08E+04to4.32E+04)	3655.9	8.750	1.4835E-06

Again, LMS has improved some variables but still leaving Var 4 insignificant. This example took 8 hours to get the exact solution. Additional data decreased precision of the R^2 and standard error of the OLS, but leaving the reweighted in almost same level of precision.

4.2 Median Income and Population in US cities[18]

In this example we will use a large sample of 146 cases. We have two cases, one with the ordinary data and one with the logarithm of the data. The following figure is for the ordinary data with the results below.



-------	OLS Results	--------
F-Value = 15.01678080	Res. SS.=35835142273406	Std. err= 498853.8410
R2-Value=0.09443519563	Est. SS.=3737003309758	Case Num= 146
Adj. R2 =0.08814655116	Tot. SS.=39572145583164	Var. No.= 2

[18] Example taken from, Data Analysis and Regression, Frederick Mosteller & John W. Tukey

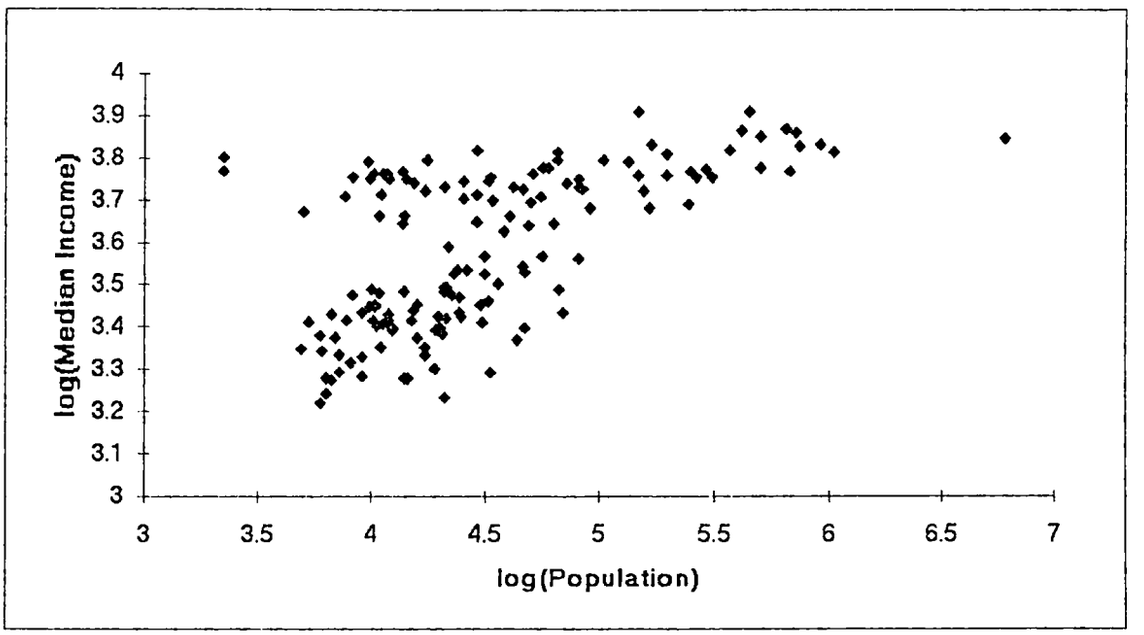
Variable	Est.Coe.	Confidence Intrv.	Str. err	T-Value	P-Value
Population	93.8	(30.6 to 157)	24.217	3.8751	0.00016133
Constant	-2.66E+05	(-5.53E+05to 2.1E+04)	109940	-2.4201	0.01676

Reweighted OLS Results

F-Value = 4.332517245	Res. SS.= 15100337314	Std. err= 12108.06412
R2-Value=0.04036537441	Est. SS.= 635169629.3	Case Num= 105
Adj. R2 =0.03104853339	Tot. SS.= 15735506943	Var. No.= 2

Variable	Est.Coe.	Confidence Intrv.	Str. err	T-Value	P-Value
Population	1.74	(-0.559 to 4.04)	0.83669	2.0815	0.03987
Constant	1.3E+04	(4190 to 2180)	3204.4	4.0564	9.709E-05

The case for the logarithmic plot follows with the results.



The results of the logarithmic case are below.

OLS Results

F-Value = 80.04771624	Res. SS.= 33.82940551	Std. err=0.4846920952
R2-Value=0.3572797687	Est. SS.= 18.80532398	Case Num= 146
Adj. R2 =0.3528164338	Tot. SS.= 52.63472948	Var. No.= 2

Variable	Est.Coe.	Confidence Intrv.	Str. err	T-Value	P-Value
log(pop)	1.93	(1.36 to 2.49)	0.21526	8.9469	1.6515E-15
Constant	-2.41	(-4.42 to -0.387)	0.77287	-3.112	0.0022412

```
---------          Reweighted  OLS  Results                          ------
F-Value = 131.6675077     Res. SS.= 20.56615765     Std. err=0.3903100159
R2-Value=0.4937515966     Est. SS.= 20.05847941     Case Num=          137
Adj. R2 =0.4900016084     Tot. SS.= 40.62463706     Var. No.=            2

Variable   Est.Coe.   Confidence Intrv.     Str. err   T-Value    P-Value
--------   --------   -----------------     --------   --------   --------
log(pop)       2.06  ( 1.59   to    2.53)    0.17963     11.475  1.0729E-21
Constant      -2.85  (-4.54   to   -1.17)    0.64277    -4.4407  1.849E-05
```

I have selected these two cases for the following reason. The ordinary data with linear regression was a wrong choice. Actually, there is a tendency in the data showing that the larger the population size in the city the richer is the city. Possibly due to that fact the richer is the median family. The rule is not general but it is true for the most of the cities.

Applying robust procedure for such a wrong linear model it throwed 41 data out of 146, from the sample. This comes from the fact that the model is not linear. Looking back at the figure we see that there are a lot of rich medium income families living in small cities. This is a bit unusual, these cities may have an extra advantage than only depending on the population.

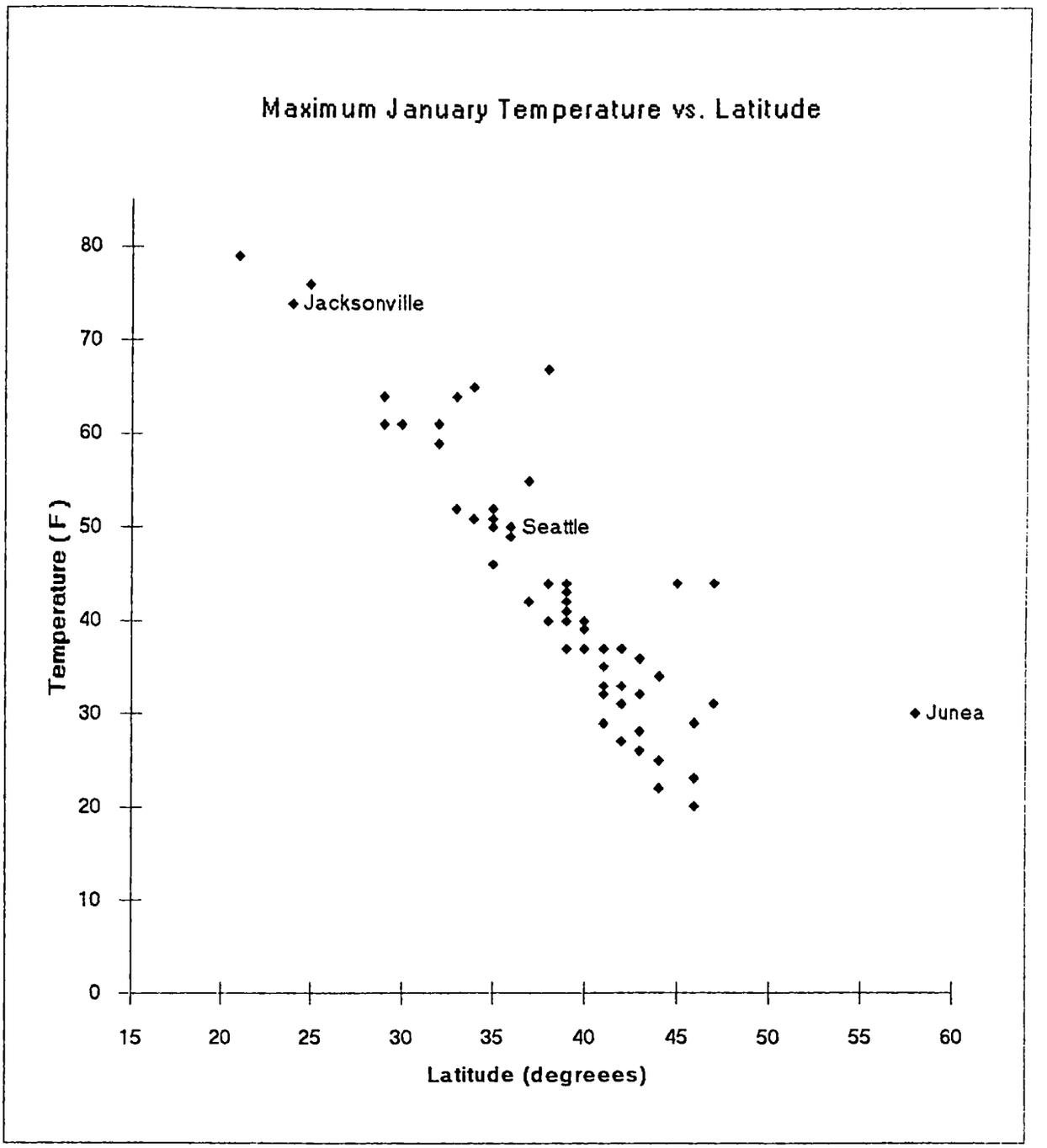
The choice of the log model is more logical, because as the population grows, growing wealth has to bedistributed. However the population grows faster. Wealth is attractive and therefore population size increases faster, because of birth or emigration.

What does that mean? Does that mean that the LMS has failed, a more advanced technology produced wrong results. No, the LMS behaved correctly but the model was wrong. Therefore, LMS is able to detect in extreme cases wrong regression models. By using that fact, it is possible to write a program to find a better fit equation by using robust regression.

As a result, in regressions with missing variables or wrong model the robust procedures detects many points as outliers. This is another superiority of the robust procedures. With this fact we can adjust our model but the classical OLS does not say anything.

4.3 Latitude versus Temperature[19]

This example is interesting if we look at the figure below.



[19] Example taken from, Data Analysis and Regression, Frederick Mosteller & John W. Tukey

As you go south the temperature increases as expected. However, there are some outliers because of their special position, to identify them we first look at the results.

```
F-Value = 86.97334953     Res. SS.= 2121.169555     Std. err= 6.100286957
R2-Value=0.8207096399     Est. SS.= 9709.748478     Case Num=          61
Adj. R2 =0.8112733051     Tot. SS.= 11830.91803     Var. No.=           4

Variable   Est.Coe.   Confidence Intrv.     Str. err   T-Value    P-Value
--------   --------   -----------------     --------   --------   --------
Latitude      -1.92  (-2.26   to   -1.58)     0.1282    -14.958  2.1636E-21
Longtitute    0.205  ( 0.079  to   0.332)  0.047442     4.3314  6.0722E-05
Altitude   -0.00176  (-0.00332to-0.000188) 0.000587     -2.986  0.0041604
Constant        100  ( 82.8   to    117)     6.495     15.408  5.5599E-22
```

All variables came out to be significant. The only outlier the OLS could identify was the Jacksonville. The results for reweighted OLS are.

```
F-Value = 665.1295723     Res. SS.= 182.0415429     Std. err= 2.011309927
R2-Value=0.9779453789     Est. SS.= 8072.080906     Case Num=          49
Adj. R2 =0.9764750709     Tot. SS.= 8254.122449     Var. No.=           4

Variable   Est.Coe.   Confidence Intrv.     Str. err   T-Value    P-Value
--------   --------   -----------------     --------   --------   --------
Latitude      -2.6  (-2.76   to   -2.44)    0.05971    -43.554  1.9483E-38
Longtitute  -0.275  (-0.394  to  -0.157)  0.043911    -6.2733  1.2254E-07
Altitude   0.00136  ( 0.000511to 0.00221) 0.0003157     4.3096  8.7848E-05
Constant       163  ( 150    to    176)     4.7579     34.309  6.5764E-34
```

After throwing out 12 cases a better approximation is reached. This is a good example for the fact that the LMS is able to increase precision. The question remains to ask here is that if it is feasible to throw away 20% of the data to get a better suited approximation.

We can say then that if the estimation we are going to make is for ordinary cities with no extraordinary position on the map then we can use the reweighted OLS, but if we are not sure that if the data we are using is corrupted or not, we can use both of them for comparison. For example, if I randomly select a data, which comes out to be Oklahoma City just to forecast. It has Latitude=35, Longtitude=97 and Altitude=1195. The OLS will give us a range of 35.33 to 65.83 and for reweighted OLS 41.93 to 51.97, where the actual data was 46. In addition, while putting the data I forgot to put a possible outlier candidate, which is San Juan with Latitude=18,

Longtitude=66 and Altitude=35. Repeating the calculations we get for OLS the range 63.65 to 94.15 and for the reweighted OLS 93.07 to 103.34. The measured unit was 81 and there is a deviation in robust case.

We have to discuss that the results that why a robust regression produced an unexpected result. In fact it is not an unexpected result. Robust procedure has identified 20% of the data as outlier, but most probably the data's used are not corrupted or biased. Therefore, in one out of five cases it is possible that robust procedure will fail in estimating the results. The results of the OLS were also acceptable so we should not always use the robust estimation blindly just to increase the precision.

4.4 Education and Income[20]

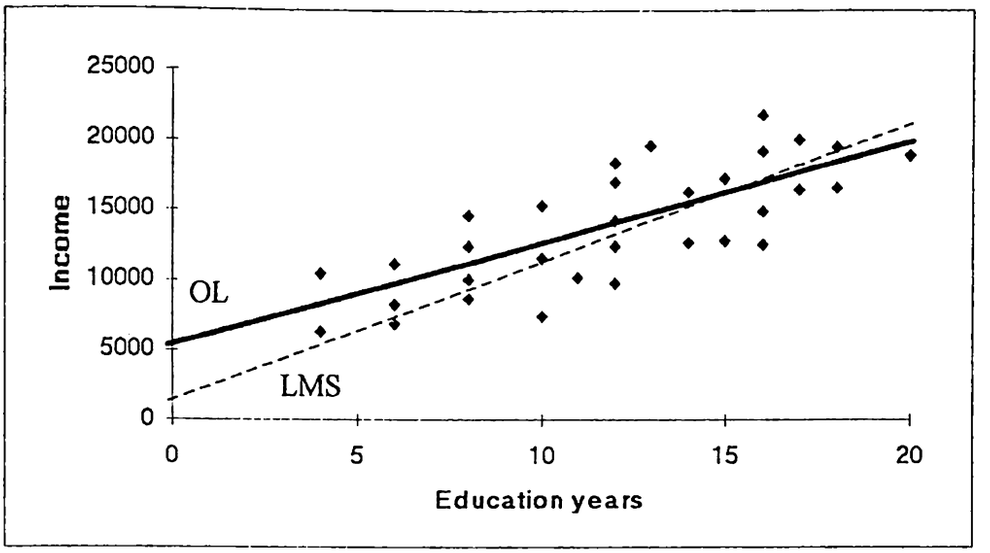
The data for this case comes from 306 interviewed employees on city payroll. A random sample of 32 people is selected out of it and the following results are obtained.

```
-------                  OLS  Results                                --------
F-Value = 39.87937474     Res. SS.= 242062725.4    Std. err= 2840.555846
R2-Value=0.5706887746     Est. SS.= 321777004.5    Case Num=          32
Adj. R2 =0.5563784004     Tot. SS.= 563839729.9    Var. No.=           2

Variable   Est.Coe.   Confidence Intrv.     Str. err   T-Value    P-Value
--------   --------   ------------------    --------   --------   --------
Educatio        739  ( 388     to1.09E+03)     116.94      6.315  5.8002E-07
Constant   4.99E+03  ( 514     to9.46E+03)     1490.5      3.345  0.0022224


-------                     LMS  Results                              --------
Variable  Est.Coe.
--------  --------
Educatio      956.5
Constant  966.0000000

     R2=0.7128379090
Std err= 3332.371805
```

---

[20] Applied Regression An Introduction, Micheal S Lewis-Beck

The following figure illustrates the difference between the OLS and the LMS line.



This illustrative example is for the case that when there is no outlier. If there is no outlier the reweighted OLS is the same as the classical OLS, however, the coefficients proposed by the LMS are different. This does not imply that again OLS or LMS have a defect. We don't have to expect in such cases that the coefficients should be the same, because of the different used techniques. If there are no outliers we can safely use the reweighted OLS.

In this case it is suitable to use both as approximator, bacause they have a different tendency. So in this case robustness becomes an alternative as there are no outlier.

5. Conclusions

As the name implies Robust Regression is an important tool in many respects. Direct or indirect use of it identifies bad data and improves results. With these facts in mind we applied it to many data sets. The results of these applications were impressive and promise succes in further research.

The identification of the outlier said in the applications that we may have missed some variables or using wrong model. In addition, blind application of robust technique causes problems, as shown previously. It has to have a reason while using the robust procedure to increase the precision or it may be damaging.

We also showed that Robust procedures find the significant variables after throwing outliers out. By using both OLS and LMS it was possible to comment on the outcomes. We used both in cases where the data was not clear. In data series where the precision is not exact with many possible outliers both of them will yied good results.

We should not forget that robust procedures are very powerful in identifiying the outliers in the corrapted data, but weak in high variance actual data. After all these, as a result, robust procedures are multi-purpose regressions, which are very useful.

6. References

1. Peter J. Rousseeuw and Annick M. Leroy, "Robust Regression and Outlier Detection", John Wiley & Sons, 1987
2. Douglas C. Montgomery and Elizabeth A. Peck, "Introduction to Linear Regression Analysis", Wiley & Sons, second edition,1991
3. Peter J. Rousseeuw and Bert C. Van Zomeren, "Unmasking Multivariate Outliers and Leverage Points", Journal of the American Statistical Assosciation, September 1990, Vol 85, No 441
4. Peter J. Rousseeuw, "Tutorial to Robust Statistics", Journal of Chemometrics, Vol 5, page 1-20.
5. Asad Zaman, "Lecture Notes on Statistics and Econometrics"
6. David C. Hoaglin, "Using Leverage and Influence to Introduce Regression Diagnostics", The College Mathemetics Journal
7. Frederick Mosteller and John W. Tukey, "Data Analysis and Regression", Addison-Wesley Publishing Company,1977
8. Micheal S Lewis-Beck, "Applied Regression An Introduction", a Sage University Paper,1976
9. The Gauss System Version 3.1, Aptech Systems, Inc. Maple Valley WA 1984-1993

Appendix

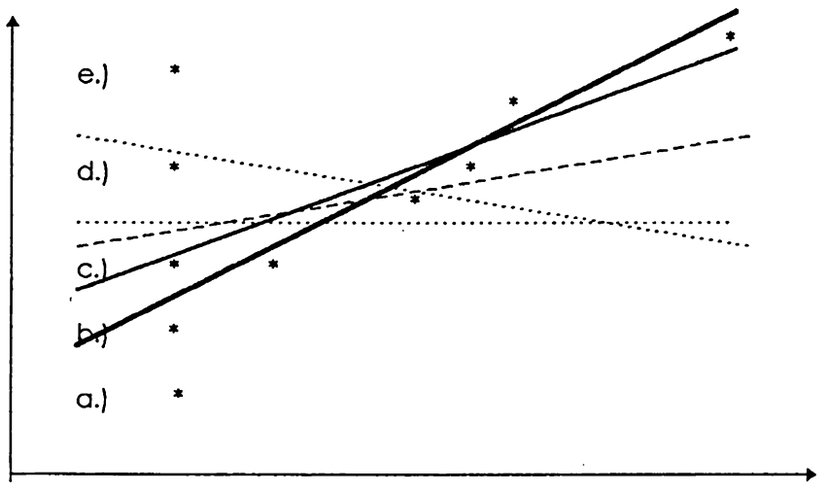


Figure 1. Outlier in y direction.

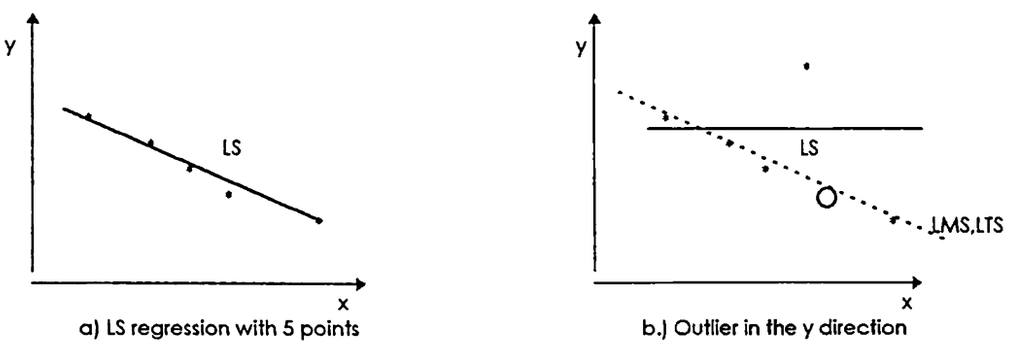


Figure 2. Outlier in the y direction.

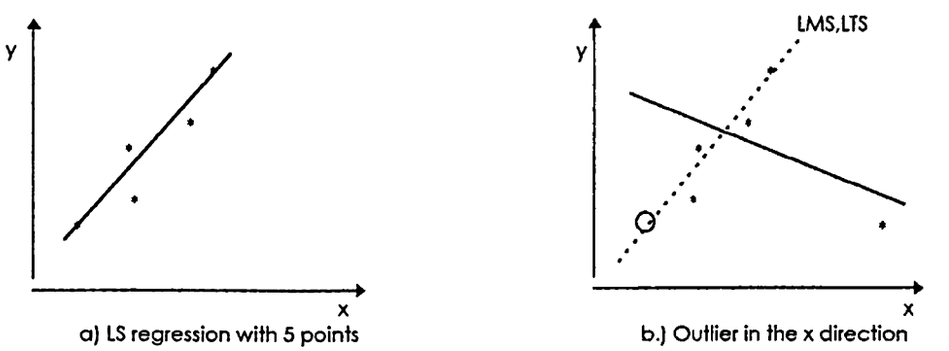


Figure 3. Outlier in the x-direction.

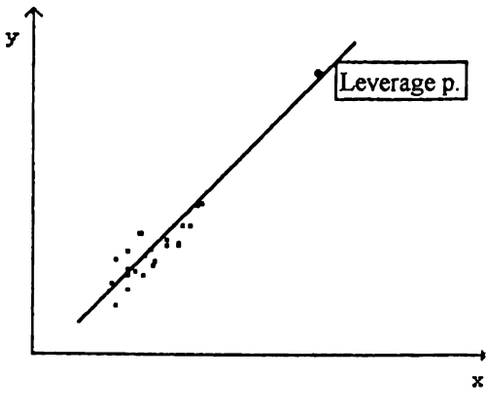


Figure 4.

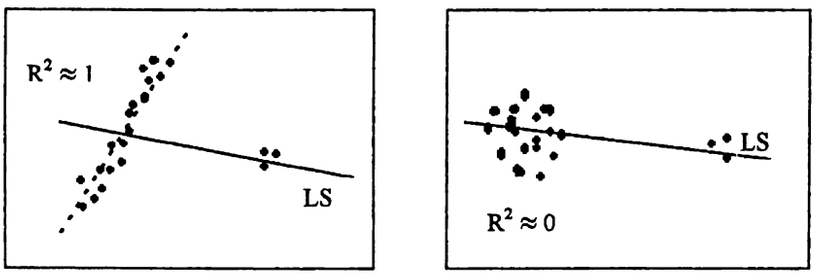


Figure 5. a.) b.)

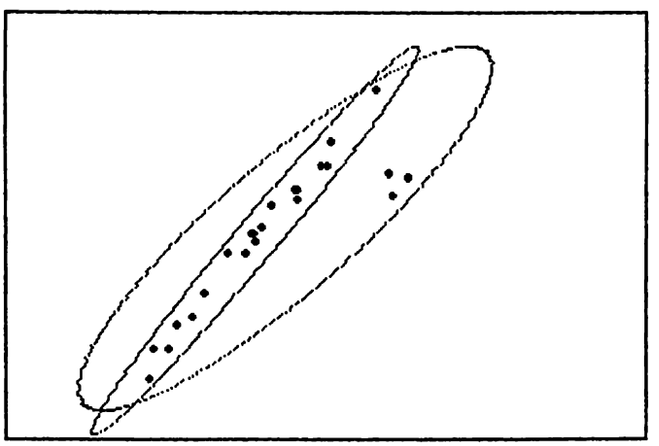


Figure 6 97.5% ellipsoids for classical and robust distances

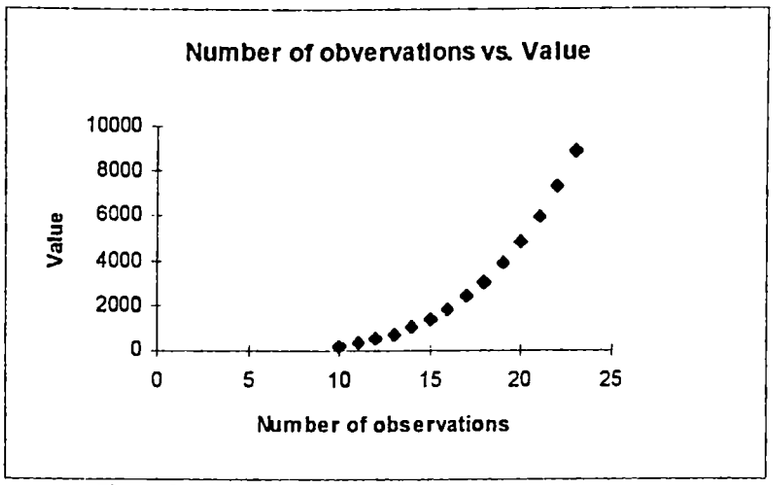


Figure 8

Table 1. Body weight Brain weight data

	Species	Body W.(kg)	Brain W.(g)	OLS Res.	LMS Res.	MD	Robust D.
1	Mountain beaver	1.35	8.10	-0.25	-0.07	1.01	0.58
2	Cow	465.00	423.00	0.16	1.16	0.70	0.59
3	Gray wolf	36.33	119.50	-0.26	-0.16	0.30	0.45
4	Goat	27.66	115.00	-0.35	-0.46	0.38	0.62
5	Guinea pig	1.04	5.50	-0.17	0.23	1.15	0.74
6	Diplodocus	11700.00	50.00	2.59	9.12	2.64	5.49
7	Asian elephant	2547.00	4603.00	-0.35	-0.57	1.71	1.68
8	Donkey	187.10	419.00	-0.21	-0.06	0.71	0.75
9	Horse	521.00	655.00	-0.02	0.58	0.86	0.69
10	Potar monkey	10.00	115.00	-0.77	-1.84	0.80	1.46
11	Cat	3.30	25.60	-0.47	-0.80	0.69	0.66
12	Giraffe	529.00	680.00	-0.03	0.53	0.87	0.71
13	Gorilla	207.00	406.00	-0.16	0.13	0.68	0.66
14	Human	62.00	1320.00	-1.26	-3.48	1.72	2.94
15	African elephant	6654.00	5712.00	-0.06	0.37	1.76	1.42
16	Triceratops	9400.00	70.00	2.32	8.25	2.37	4.90
17	Rhesus monkey	6.80	179.00	-1.16	-3.10	1.22	2.32
18	Kangaroo	35.00	56.00	0.11	1.07	0.20	0.46
19	Hamster	0.12	1.00	-0.20	0.18	1.86	1.26
20	Mouse	0.02	0.40	-0.42	-0.50	2.27	1.45
21	Rabbit	2.50	12.10	-0.20	0.09	0.83	0.47
22	Sheep	55.50	175.00	-0.27	-0.23	0.42	0.58
23	Jaguar	100.00	157.00	0.03	0.75	0.26	0.28
24	Chimpanze	52.16	440.00	-0.77	-1.86	1.05	1.74
25	Brachiosaurus	87000.00	154.50	2.85	9.93	2.91	5.84
26	Rat	0.28	1.90	-0.17	0.25	1.59	1.07
27	Mole	0.12	3.00	-0.75	-1.65	1.58	1.21
28	Pig	192.00	180.00	0.23	1.40	0.39	0.58

Data Set and # exp.	Type	Replication	Execution time	Ellementary set found	Excl. Set found	Reweigted Coef same	Excl. Set found	Reweigted Coef dev	Conf. int. identified
	LMS								
Wood.dat	Exact	38760	185						
20		500	4	5	11	11	9	2x%27	18
20		1000	8	14	14	14	6	6x%20	20
20		4000	34	17	17	17	3	1x%20	20
20		8000	68	20	20	20			20
20		20000	170	20	20	20			20
20		40000	341	20	20	20			20

Table 2: Table of experimenting on Wood.dat for LMS

	Type	Replication	Execution time	Ellementary set found	Excl. Set found	Reweigted Coef same	Excl. Set found	Reweigted Coef dev	Conf. int. identified
Brainlog.d	LMS								
20	Exact	378	185						
20	Random	15	1	9	20	20	20		20
20		30	1	6	20	20	20		20
20		60	1	9	20	20	20		20
20		150	1	16	20	20	20		20
20		300	2	17	20	20	20		20

	Type	*Replication*	*Execution time*	*Ellementary set found*	*Leverage pt found*	*Worst possible*
Brainlog.d	MVE					
20	Exact	77520	31			
20	Random	200	4	10	14	
20		400	9	18	19	
20		800	17	18	20	
20		2000	17	20	20	
20		4000	35	20	20	

	Type	*Replication*	*Execution time*	*Ellementary set found*	*Leverage pt found*	*Worst possible*
Wood.dat	MVE					
20	Exact	77520				
20	Random	250	5	0	13	5
20		500	11	0	17	2
20		2500	115	1	20	
20		5000	229	1	20	
20		10000				

Table 3.: Experiments on data set

	Income		*same houses*	*collar occupa.*	*public transp.*	*room/unit*	*unit struct.*	*(1958-1960)*	*air cond.*
N.Hanover	4572	21.8	1.1	42.2	0.4	4.7	78.8	82.2	24.1
Florence- GR.	4904	25.7	40.2	17.1	16.8	3.9	86.6	41.2	1.3
Kannapolis	5182	29.0	54.5	20.4	6.1	4.6	96.7	27.1	6.5
Brownsville	5306	22.6	35.3	39.1	3.2	4.6	93.5	46.4	26.0
East Los-A	5439	25.1	44.8	26.7	19.7	4.2	79.8	37.1	6.0
Bell Gardens	5567	24.4	26.1	67.6	1.5	3.8	89.8	59.6	30.3
Hempfield	5909	29.3	58.4	37.0	5.3	5.2	95.0	24.9	3.0
S. San Gabriel	6076	29.3	40.9	36.9	6.3	4.3	90.1	41.5	10.4
Essex	6160	24.8	46.7	34.5	6.7	5.4	78.5	37.5	12.0
Methuen	6278	34.6	63.1	38.2	6.7	5.3	72.6	19.9	3.6
Needham	9282	32.5	56.0	69.3	14.0	6.2	92.3	22.5	11.2
Teaneck	9518	33.0	63.0	62.9	29.4	6.1	80.6	17.9	26.9
Silver Springs	9540	31.7	48.6	76.2	11.0	5.7	68.9	33.7	40.6
Greenwich	9588	35.6	55.8	54.5	9.7	5.8	69.9	20.7	11 .6
West Hardford	9712	37.4	50.4	72.1	13.5	6.2	79.8	22.1	16.0
Cheltenham	9985	36.6	57.9	75.0	24.0	6.5	73.5	23.8	41.8
Mount Leb.	11108	36.9	49.1	62.8	25.5	6.2	81.4	25.4	14.0
Wellesley	11478	31.3	52.4	70.8	15.4	6.7	94.7	23.2	9.8
Lower Merion	12204	32.6	57.4	69.0	18.8	7.1	78.2	21.2	37.9
Bethesda	12357	31.4	36.3	82.6	8.5	6.5	82.3	37.4	41.7
Braintree, Mass.	7474	31 .0	60.7	52.2	7.7	5.8	88.8	19.4	6.7

Ross, Pa.	7475	31.1	52.6	4.0	12.9	5.7	86.1	28.6	6.5
Elmont, N.Y.	7494	31.1	68.2	44.2	33.	5.6	88.5	16.6	19.0
Framingham,Mss	7495	29.1	44.5	53.0	5.9	5.6	79.8	32.7	21.0
Arlington, Mass.	7538	34.8	61.8	60.4	27.7	5.8	56.8	22.3	7.8
Natick, Mass.	7550	28.7	57.8	55.7	9.1	5.9	81.6	23.7	5.4
Ewing, N.J.	7597	31.1	56.0	48.7	6.8	5.6	91.5	23.0	24.0
Middietown. Pa.	7656	22.6	19.2	56.3	7.9	6.1	99.2	35.8	24.3
Catonsville, Md.	7662	32.0	50.5	64.5	14.9	6.0	82.3	25.9	14.8
Hamden, Conn.	7741	35.5	59.4	55.3	13.7	5.7	84.5	21.4	10.6

Table 4: Data for medium income families.

City	Tem.	Lat.	Long.	Alt.
Mobile, Ala.	61	30	88	5
Montgomery, Ala.	59	32	86	160
Juneau, Alaska	30	58	134	50
Phoenix, Ariz.	64	33	112	1090
Little Rock, Ark.	51	34	92	286
Los Angeles, Calif.	65	34	118	340
San Francisco, Calif.	55	37	122	65
Denver, Col.	42	39	104	5280
New Haven, Conn.	37	41	72	40
Wilmington, Del.	41	39	75	135
Washington, D.C.	44	38	77	25
Jacksonville, Fla.	67	38	81	20
Key West, FIa.	74	24	81	5
Miami, Fla.	76	25	80	10
Atlanta, Ga.	52	33	84	1050
Honolulu, Hawaii	79	21	157	21
Boise, Idaho	36	43	116	2704
Chicago, III.	33	41	87	595
Indianapolis, Ind.	37	39	86	710
Des Moines, lowa	29	41	93	805
Dubuque, lowa	27	42	90	620
Wichita, Kansas	42	37	97	1290
Louisville, Ky.	44	38	85	450
New Orleans, La.	64	29	90	5
Portland, Maine	32	43	70	25
Baltimore, Md.	44	39	76	20
Boston, Mass.	37	42	71	21
Detroit, Mich.	33	42	83	585
Sault Ste. Marie, Mich.	23	46	84	650
Minn.-St. Paul, Minn.	22	44	93	815
St. Louis, Missouri	40	38	90	455
Helena, Montana	29	46	112	4155

Omaha, Nebraska	32	41	95	1040
Concord, N.H.	32	43	71	290
Atlantic City, N.J.	43	39	74	10
Albuquerque, N.M.	46	35	106	4945
Albany, N.Y.	31	42	73	20
New York, N.Y.	40	40	73	55
Charlotte, N.C.	51	35	80	720
Raleigh, N.C.	52	35	78	365
Bismarck, N.D.	20	46	100	1674
Cincinnati, Ohio	41	39	84	550
Cleveland, Ohio	35	41	81	660
Oklahoma City, Okla.	46	35	97	1195
Portland, Ore.	44	45	122	77
Harrisburg, Pa.	39	40	76	365
Philadelphia, Pa.	40	39	75	100
Charlestown, S.C.	61	32	79	9
Rapid City, S.D.	34	44	103	3230
Nashville, Tenn.	49	36	86	450
Amarillo, Tx.	50	35	101	3685
Galveston, Tx.	61	29	94	5
Houston, Tx.	64	29	95	40
Salt Lake City, Utah	37	40	111	4390
Burlington, Vt.	25	44	73	110
Norfolk, Va.	50	36	76	10
Seattle-Tacoma, Wash.	44	47	122	10
Spokane, Wash.	31	47	117	1890
Madison, Wisc.	26	43	89	860
Milwaukee, Wisc.	28	43	87	635
Cheyenne, Wyoming	37	41	104	6100
San Juan	81	18	66	35

Table 5: Data for US cities temperature

Country	Population	Mediun Income
Maricopa	663510	5896
Pima	265660	5690
Pinal	62673	4412
Cochise	55039	5107
Yuma	46235	5360
Coconino	41857	5398
Navajo	37994	4237
Apache	30438	2832
Yavapai	28912	5197
Gila	25245	5087
Graham	14045	4593
Greenlee	11059	5168
SantaCruz	10808	4620
Mohave	7736	5111
Mississippi	70174	2725
Crittenden	47564	2506
Craighead	47303	3408
Phillips	43997	2360
StFrancis	33303	1973
Poinsett	30834	2591
Greene	25198	2654
Clay	21258	2633
Lee	21001	1710
Cross	19551	2480
Pulaski	242980	4935
White	32795	2893
Lonoke	24551	2708
Faulkner	24303	2968
Arkansas	23355	3348
Jackson	22843	2995
Independence	20048	2502
Monroe	17327	2162
Lawrence	17267	2255
Conway	15430	2751
Woodruff	13954	1902
Randolph	12520	2497
Prairie	10515	2853
Cleburne	9059	2137
Izard	6766	2699
Fulton	6657	1886
Sharp	6319	1902
Stone	6294	1740
Perry	4927	2217
Sebastian	66685	3089
Washington	55797	3683
Benton	36272	3180
Crawford	21318	3122
Pope	21177	3046
Boone	16116	2837
Logan	15957	2376
Johnson	12421	2484
Yell	11940	2600
Carroll	11284	2555

Sonoma	147325	5725
Marin	146820	8110
Merced	90448	4806
S.LuisObiapo	81011	5659
Imperial	72105	5507
Yolo	65727	6240
Placer	56468	5989
Kings	49954	4957
Sutter	33380	5670
Siskiyou	32885	5558
ElDorado	29390	6603
Tehama	25305	5589
Glenn	17245	5290
SanBenito	15396	5538
Lake	13786	4438
Calaveras	10289	5824
Amador	9900	5636
Trinity	9706	6210
Tulare	166403	4815
Butte	82030	5408
Shasta	59468	5989
Madera	40466	4596
Yuba	33859	5031
Nevads	20911	5419
Tuolumne	14404	5602
Lassen	13597	5861
Colusa	12075	5604
Inyo	11689	5837
Plumas	11260	5834
Modoc	8308	5709
Mariposa	5064	4704
Sierra	2247	5863
Mono	2213	6321

Table 6 Data for population size and medium income

Education	*Income*
4	6281
4	10516
6	6898
6	8212
6	11144
8	8618
8	10011
8	12405
8	14664
10	7472
10	11598
10	15336
11	10186
12	9771
12	12444
12	14213
12	16908
12	18347
13	19546
14	12660
14	16326
15	12772
15	17218
16	12599
16	14852
16	19138
16	21779
17	16428
17	20018
18	16526
18	19414
20	18822

Table7 :Education years vs. Income

Franklin	10213	2611
Baxter	9943	2800
Madison	9068	1928
Searcy	8124	2066
Scott	7297	2168
VanBuren	7228	1968
Marion	6041	2210
Newton	5963	1666
Jefferson	81373	3671
Union	49518	4361
Garland	46697	3511
Miller	31686	3372
Ouachita	31641	3686
Saline	28956	4483
Columbia	26400	3438
Ashley	24220	3432
HotSprings	21893	3881
Clark	20950	3127
Desha	20770	2430
Hempstead	19661	2676
Chicot	18990	2013
Drew	15213	2614
Lincoln	14447	1911
Bradley	14029	3069
Polk	11981	2694
Lafayette	11030	2245
Howard	10878	3033
Nevada	10700	2538
Dallas	10522	2809
Sevier	10156	3089
LittleRiver	9211	2725
Grant	8294	2985
Pike	7864	2614
Cleveland	6944	2363
Calhoun	5991	2394
Montgomery	5370	2572
LosAngles	6038771	7046
SanDiego	1033011	6545
Alameda	908209	6786
SanFrancisco	740316	6717
SantaClara	642315	7417
Sacramento	502775	7100
SanMat,o	444387	8103
ContraCosta	409030	7327
SanJoaquin	249919	5889
Ventura	199138	6466
Monterey	198351	5770
SantaBarbara	168962	6833
Solano	134597	6190
Humboldt	104892	6282
SantaCruz	84219	5325
Napa	65890	6524
Mendocino	51058	5803
DelNorte	17771	6277
Orange	7039Z5	7219
SanBernardin	503591	5998
Fresno	365945	6603
Riverside	30619t	5693
Kern	291981	5933
Stanislaus	157294	5260