Theil-Sen Estimator

An alternative to least squares regression

Srikanth K S

talegari.wikidot.com
sri.teach@gmail.com

We discuss,

We discuss,

- Basic idea of the classical *Least Squares Regression* in single predictor case.

We discuss,

- Basic idea of the classical *Least Squares Regression* in single predictor case.
- Disadvantages of *Least Squares Regression* with practical problems.

We discuss,

- Basic idea of the classical *Least Squares Regression* in single predictor case.
- Disadvantages of *Least Squares Regression* with practical problems.
- *Theil-Sen Estimator*

We discuss,

- Basic idea of the classical *Least Squares Regression* in single predictor case.
- Disadvantages of *Least Squares Regression* with practical problems.
- *Theil-Sen Estimator*
- Relative advantages of *Theil-Sen Estimator* over *Least Squares Regression*.

We discuss,

- Basic idea of the classical *Least Squares Regression* in single predictor case.
- Disadvantages of *Least Squares Regression* with practical problems.
- *Theil-Sen Estimator*
- Relative advantages of *Theil-Sen Estimator* over *Least Squares Regression*.
- An application: Using *Theil-Sen Estimator* as a outlier detection tool

We discuss,

- Basic idea of the classical *Least Squares Regression* in single predictor case.
- Disadvantages of *Least Squares Regression* with practical problems.
- *Theil-Sen Estimator*
- Relative advantages of *Theil-Sen Estimator* over *Least Squares Regression*.
- An application: Using *Theil-Sen Estimator* as a outlier detection tool
- Improvements over *Theil-Sen Estimator*

Least Squares Regression

We intend to *fit* a line to get a approximate idea of the *trend*.

Data $\longrightarrow$ set of pairs $(x_i, y_i), 1 \leq i \leq n$.

x_i's $\longrightarrow$ the predictor

y_i's $\longrightarrow$ the response.

Least Squares Regression

We intend to *fit* a line to get a approximate idea of the *trend*.

Data $\longrightarrow$ set of pairs $(x_i, y_i), 1 \leq i \leq n$.

x_i's $\longrightarrow$ the predictor

y_i's $\longrightarrow$ the response.

We intend to find b_0 and b_1 such that $\hat{y}_i := E(y_i|x_i) = b_0 + x_i b_1$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (the residual) is minimum.

Least Squares Regression

We intend to *fit* a line to get a approximate idea of the *trend*.

Data $\longrightarrow$ set of pairs $(x_i, y_i), 1 \leq i \leq n$.

x_i's $\longrightarrow$ the predictor

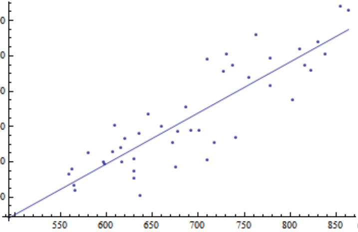
y_i's $\longrightarrow$ the response.

We intend to find b_0 and b_1 such that $\hat{y}_i := E(y_i|x_i) = b_0 + x_i b_1$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (the residual) is minimum.

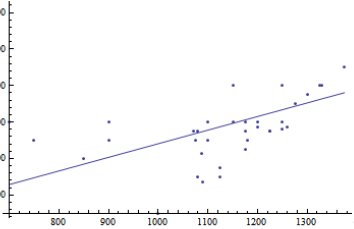
Standard optimization techniques yield,

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \qquad b_0 = \bar{y} - b_1\bar{x}$$

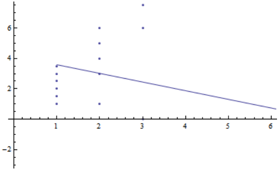
Some examples



(lsr07)

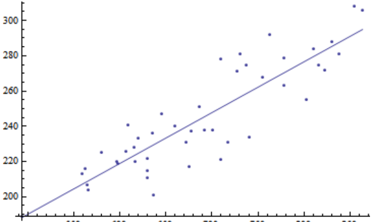


(lsr11)

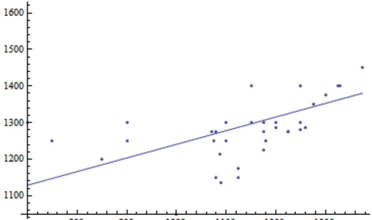


(lsrhetero)

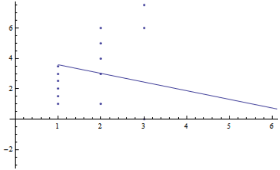
Some examples



(lsr07)



(lsr11)



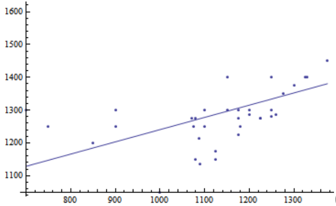
(lsrhetero)

- lsr07 - This is a fairly good fit.
- lsr11 - Outliers on the left top seem to have affected the fit.
- lsrhetero - The the difference in the variance among different x_i's has adversly affected the fit.

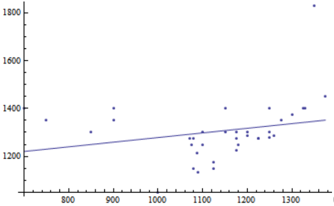
Least Squares Regression is sensitive to ...

Least Squares Regression is sensitive to ...

1. Outliers affect the fit adversely. Any method based on mean is bound to be affected by the outlier.



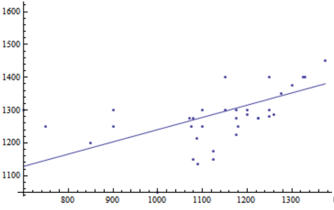
(lsr11)



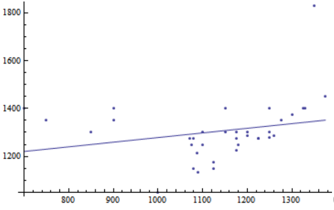
(lsr11mod)

Least Squares Regression is sensitive to ...

1. Outliers affect the fit adversely. Any method based on mean is bound to be affected by the outlier.



(lsr11)



(lsr11mod)

2. Heteroscedasticity - If the variance among y_i values corresponding to different x_i values differ, fit is affected.

Theil-Sen Estimator

offers a non-parametric (distribution-free) method to find a fit for *heteroscedastic data with outliers*[1]. It is named after Henri Theil(1950) and Pranab K. Sen(1968).

[1] allows 29% corruption of the data

[2] unless they have the same x-coordinate

Theil-Sen Estimator

offers a non-parametric (distribution-free) method to find a fit for *heteroscedastic data with outliers*[1]. It is named after Henri Theil(1950) and Pranab K. Sen(1968).

Computation: Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the data points.

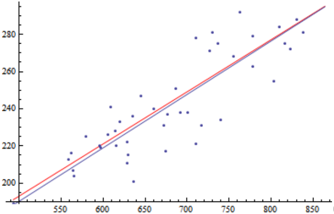
1. Find the slope of line connecting each pair of points[2].
2. The median m among the slopes is the slope of the 'fit' line.The median of the set $\{y_i - mx_i \quad | \quad 1 \leq i \leq n\}$ gives the intercept c the 'fit' line.

It competes well against non-robust least squares even for normally distributed data in terms of statistical power. It has been called "the most popular nonparametric technique for estimating a linear trend" (source:wikipedia).

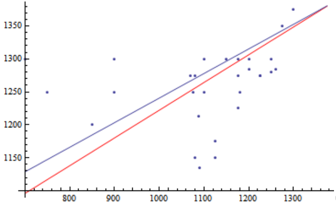
[1] allows 29% corruption of the data

[2] unless they have the same x-coordinate

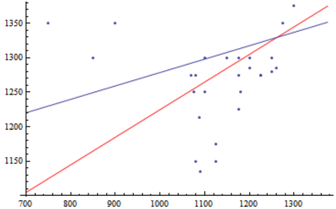
Comparing fits of *Least Squares Regression* and *Theil-Sen Estimator*



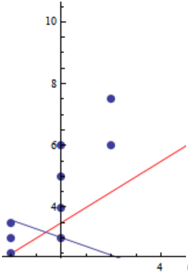
(07lsr)



(11lsr)



(11lsrmod)



(lsrhetero)

A Summary of comparative advantages

Condition	Least Squares regression	Theil-Sen Estimator
Presence of Outliers	Not great	Handles well.
Heteroscedasticity	Not great	Handles Well.
Robustness	Even the presence of a single outlier affects the fit.	allows 29% data corruption. That is, with a sufficiently large sample size, about 29% of the points must be altered to make the estimate arbitrarily large or small.
Confidence interval (95%)	Based on Normal or student's T. This might not give an accurate CI.	The middle 95% of the slopes form the CI
Complexity	$O(n)$	$O(n^2)$, although randomized algorithms reduce it to $O(n \log n)$
Asymptotic efficiency (sample size for a reasonable fit)	Not great.	High.

Theil-Sen Estimator to detect outliers and fit

[3] farther than median + interquantilerange

Theil-Sen Estimator to detect outliers and fit

Detecting outliers in a 2D (and higher) non-trivial task. We use *Theil-Sen Estimator* 's insensitivity to outliers to obtain a two-stepped process to remove outliers and get a *better* fit.

[3] farther than median + interquantilerange

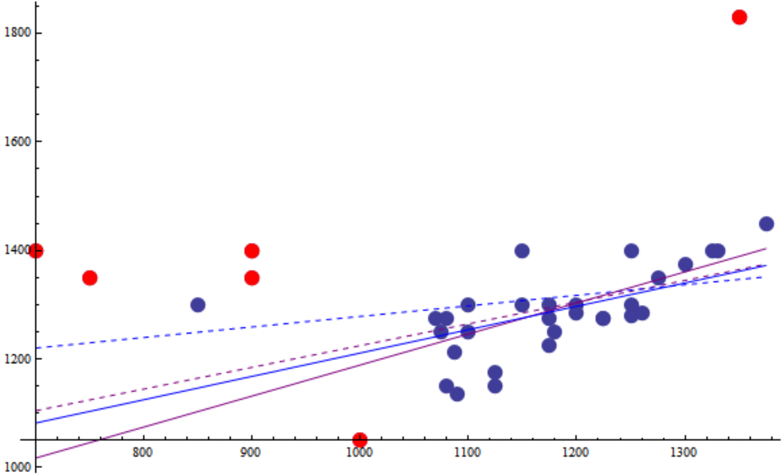
Theil-Sen Estimator to detect outliers and fit

Detecting outliers in a 2D (and higher) non-trivial task. We use *Theil-Sen Estimator* 's insensitivity to outliers to obtain a two-stepped process to remove outliers and get a *better* fit.

1. Identify points that lie at a *large*[3] distance from the theil-sen line as *outliers*.
2. Compute *Theil-Sen Estimator* (or *Least Squares Regression*) after removing this data.

[3] farther than median + interquantilerange

lsr11mod



(lsr11mod)

- - - - Theil-Sen line before outlier removal

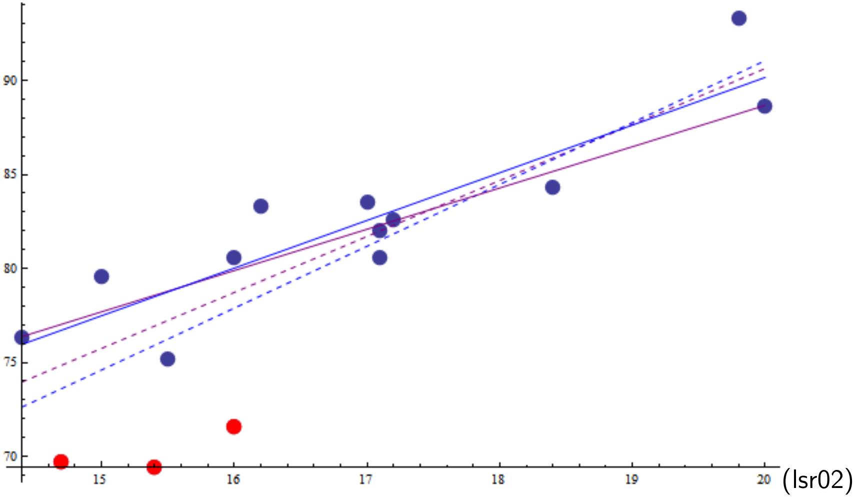
—— Theil-Sen line after outlier removal

Outlier Points

- - - - Least squares line before outlier removal

—— Least squares line after outlier removal

lsr02



- - - - Theil-Sen line before outlier removal | —— Theil-Sen line after outlier removal | Outlier Points

- - - - Least squares line before outlier removal | —— Least squares line after outlier removal

Code

Code

```
theilsenline[data_] := Module[{l, pairsofpoints = {}, slopesofpairs = {}, dummy, theilslope, theilintercept},
   l = Length[data];
   pairsofpoints = orderedpairs[data];
   slopesofpairs = DeleteCases[If[#[[2, 1]] - #[[1, 1]] ≠ 0, (#[[2, 2]] - #[[1, 2]])/(#[[2, 1]] - #[[1, 1]]), dummy] & /@ pairsofpoints,
     dummy];
   theilslope = Median[slopesofpairs];
   theilintercept = Median @@ {((#[[2]] - theilslope #[[1]]) & /@ data)};
   Return[{theilslope, theilintercept}];
  ];
```

Code

```
theilsenline[data_] := Module[{l, pairsofpoints = {}, slopesofpairs = {}, dummy, theilslope, theilintercept},
   l = Length[data];
   pairsofpoints = orderedpairs[data];
   slopesofpairs = DeleteCases[If[#[[2, 1]] - #[[1, 1]] ≠ 0, (#[[2, 2]] - #[[1, 2]])/(#[[2, 1]] - #[[1, 1]]), dummy] & /@ pairsofpoints,
     dummy];
   theilslope = Median[slopesofpairs];
   theilintercept = Median @@ {((#[[2]] - theilslope #[[1]]) & /@ data)};
   Return[{theilslope, theilintercept}];
  ];
```

```
perpdist[slope_, intercept_, point_] := Abs[(slope (point[[1]]) - point[[2]] + intercept)/Sqrt[1 + slope^2]]
```

Code

```
theilsenline[data_] := Module[{l, pairsofpoints = {}, slopesofpairs = {}, dummy, theilslope, theilintercept},
   l = Length[data];
   pairsofpoints = orderedpairs[data];
   slopesofpairs = DeleteCases[If[#[[2, 1]] - #[[1, 1]] ≠ 0, (#[[2, 2]] - #[[1, 2]])/(#[[2, 1]] - #[[1, 1]]), dummy] & /@ pairsofpoints,
     dummy];
   theilslope = Median[slopesofpairs];
   theilintercept = Median @@ {((#[[2]] - theilslope #[[1]]) & /@ data)};
   Return[{theilslope, theilintercept}];
  ];
```

```
perpdist[slope_, intercept_, point_] := Abs[(slope (point[[1]]) - point[[2]] + intercept)/Sqrt[1 + slope^2]]
```

```
outliers[data_] := Module[{distances = {}, q = {}, outliers = {}, dummy},
   distances = perpdist[theilsenline[data][[1]], theilsenline[data][[2]], #] & /@ data;
   q = Quartiles[distances];
   outliers = If[distances[[#]] >= q[[2]] + (q[[3]] - q[[1]]), #, dummy] & /@ Range[Length[distances]];
   Return[data[[#]] & /@ DeleteCases[outliers, dummy]];
  ];
```

Code

```
theilsenline[data_] := Module[{l, pairsofpoints = {}, slopesofpairs = {}, dummy, theilslope, theilintercept},
  l = Length[data];
  pairsofpoints = orderedpairs[data];
  slopesofpairs = DeleteCases[If[#[[2, 1]] - #[[1, 1]] ≠ 0, (#[[2, 2]] - #[[1, 2]])/(#[[2, 1]] - #[[1, 1]]), dummy] & /@ pairsofpoints,
    dummy];
  theilslope = Median[slopesofpairs];
  theilintercept = Median @@ {((#[[2]] - theilslope #[[1]]) & /@ data)};
  Return[{theilslope, theilintercept}];
 ];
```

```
perpdist[slope_, intercept_, point_] := Abs[(slope (point[[1]]) - point[[2]] + intercept)/Sqrt[1 + slope^2]]
```

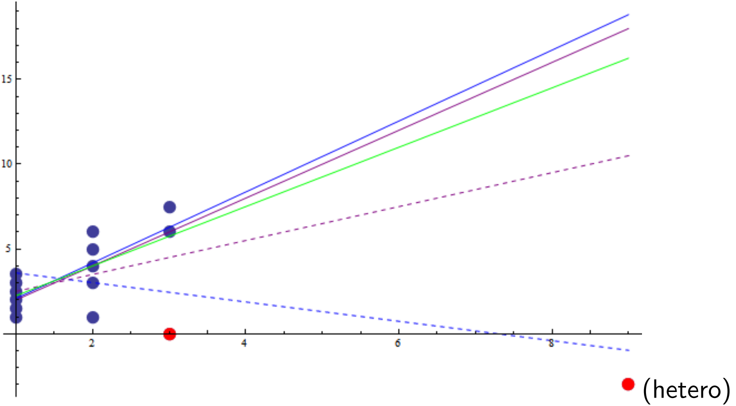
```
outliers[data_] := Module[{distances = {}, q = {}, outliers = {}, dummy},
  distances = perpdist[theilsenline[data][[1]], theilsenline[data][[2]], #] & /@ data;
  q = Quartiles[distances];
  outliers = If[distances[[#]] >= q[[2]] + (q[[3]] - q[[1]]), #, dummy] & /@ Range[Length[distances]];
  Return[data[[#]] & /@ DeleteCases[outliers, dummy]];
 ];
```

R users can use the package *mblm* by Lukasz Komsta.

Improvements over *Theil-Sen Estimator*

Improvements over *Theil-Sen Estimator*

- A variation of the TheilSen estimator due to **Siegel (1982)** determines, for each sample point, the median m_i of the slopes of lines through that point, and then determines the overall estimator as the median of these medians. A higher breakdown point, 50%, holds for the repeated median estimator of Siegel.



- A different variant pairs up sample points by the rank of their x-coordinates (the point with the smallest coordinate being paired with the first point above the median coordinate, etc.) and computes the median of the slopes of the lines determined by these pairs of points.

- A different variant pairs up sample points by the rank of their x-coordinates (the point with the smallest coordinate being paired with the first point above the median coordinate, etc.) and computes the median of the slopes of the lines determined by these pairs of points.
- Variations of the Theil-Sen estimator based on weighted medians have also been studied, based on the principle that pairs of samples whose x-coordinates differ more greatly are more likely to have an accurate slope and therefore should receive a higher weight.

References

1. Wikipedia Page (http://en.wikipedia.org/wiki/Theil%E2%80%93Sen_estimator)
2. Some results on extensions and modications of the TheilSen regression estimator - Rand R. Wilcox (British Journal of Mathematical and Statistical Psychology (2004), 57, 265280)
3. The Theil-Sen Estimators In Linear Regression - Hanxiang Peng
4. Basic Statistics: Understanding Conventional Methods and Modern Insights - Rand Wilcox

Thank you.

The presentation is available at

https://sites.google.com/a/cmrit.ac.in/srikanth/talks