Robust Statistics in Stata

Vincenzo Verardi (vverardi@fundp.ac.be)

FUNDP (Namur) and ULB (Brussels), Belgium
FNRS Associate Researcher

Based on joint work with C. Croux (KULeuven) and Catherine Dehon (ULB)

Overview

1. Introduction
2. Outliers in regression analysis
3. Overview of robust estimators
4. Stata codes
5. Conclusion

Introduction

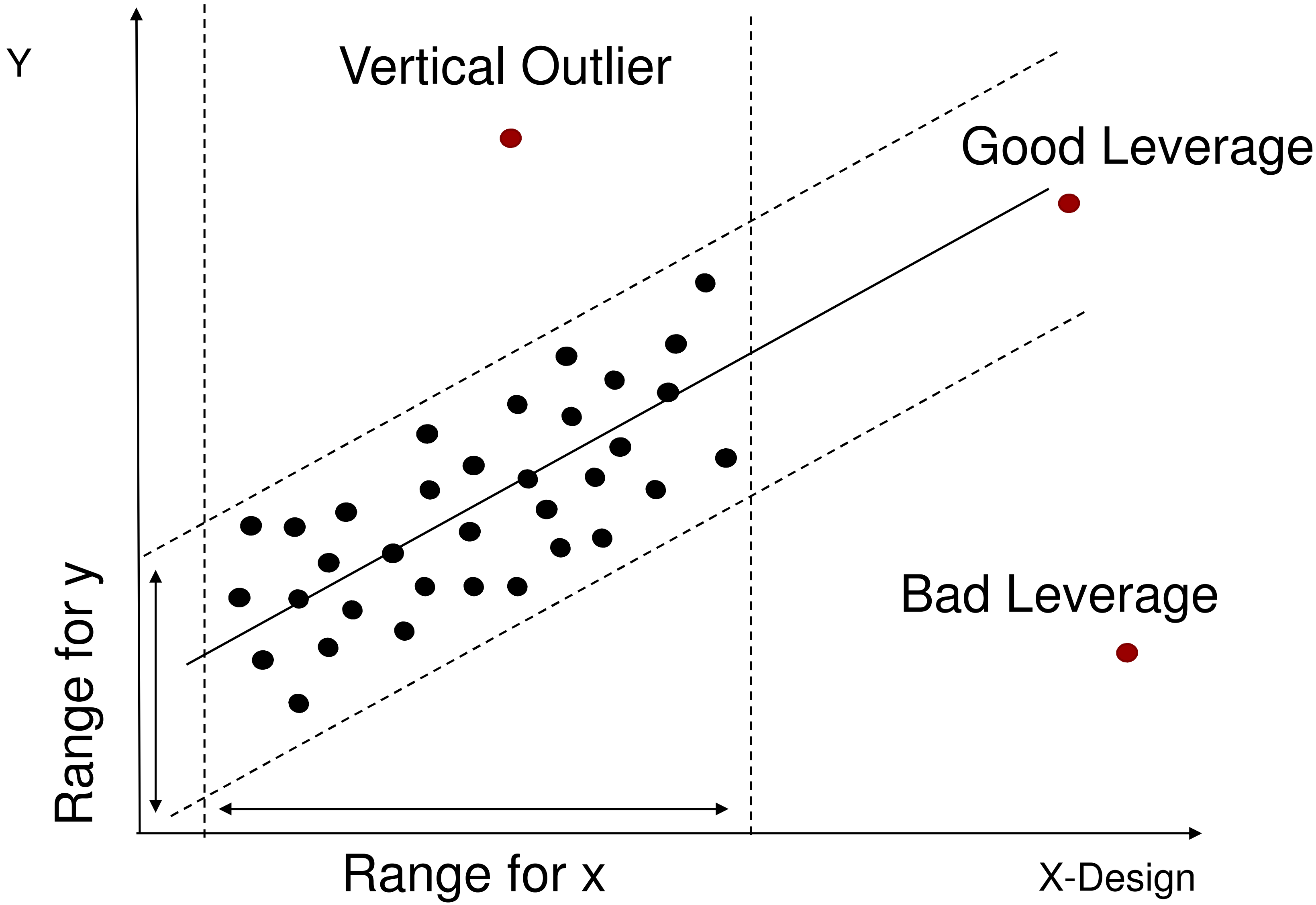
Outliers in regression analysis

Overview of robust estimators

Stata codes

Conclusion

Type of outliers in regression

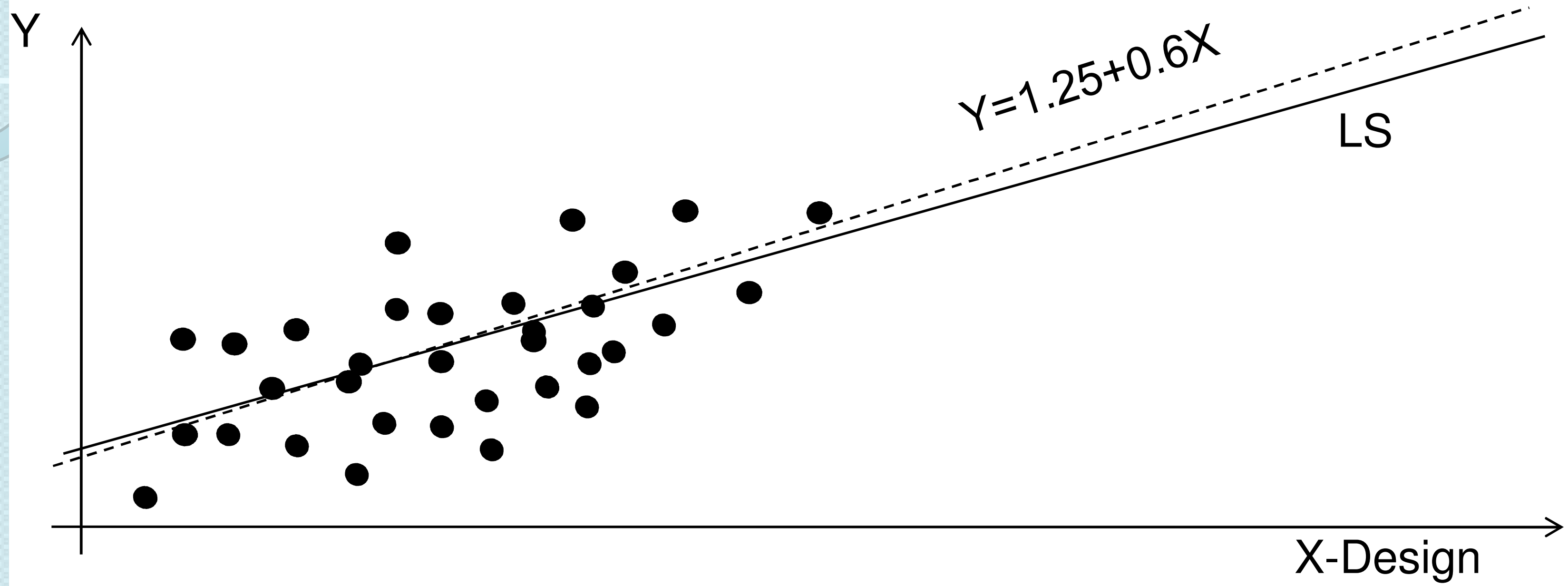


Outliers' influence

To illustrate the influence of outliers, we generate a dataset according to **Y=1.25+0.6X+ε**, where X and ε~N(0,1). We then contaminate the data with single outliers.

```
set obs 100
drawnorm X e
gen y=1.25+0.6*X+e
replace x= ...
```

Outliers in regression analysis



	Clean
Intercept t-stat	1.24 (10.76)
Slope t-stat	0.59 (4.96)

Outliers in regression analysis

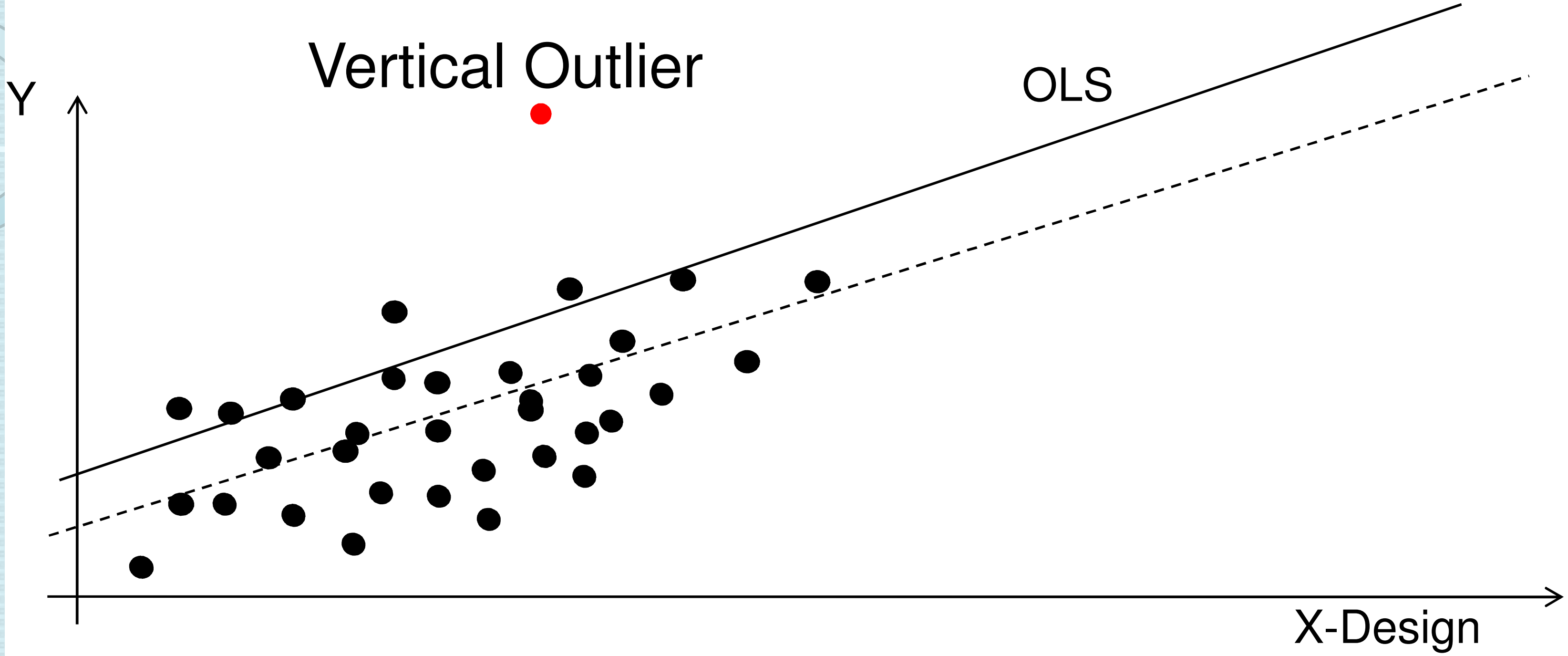
Introduction

Outliers in regression analysis

Overview of robust estimators

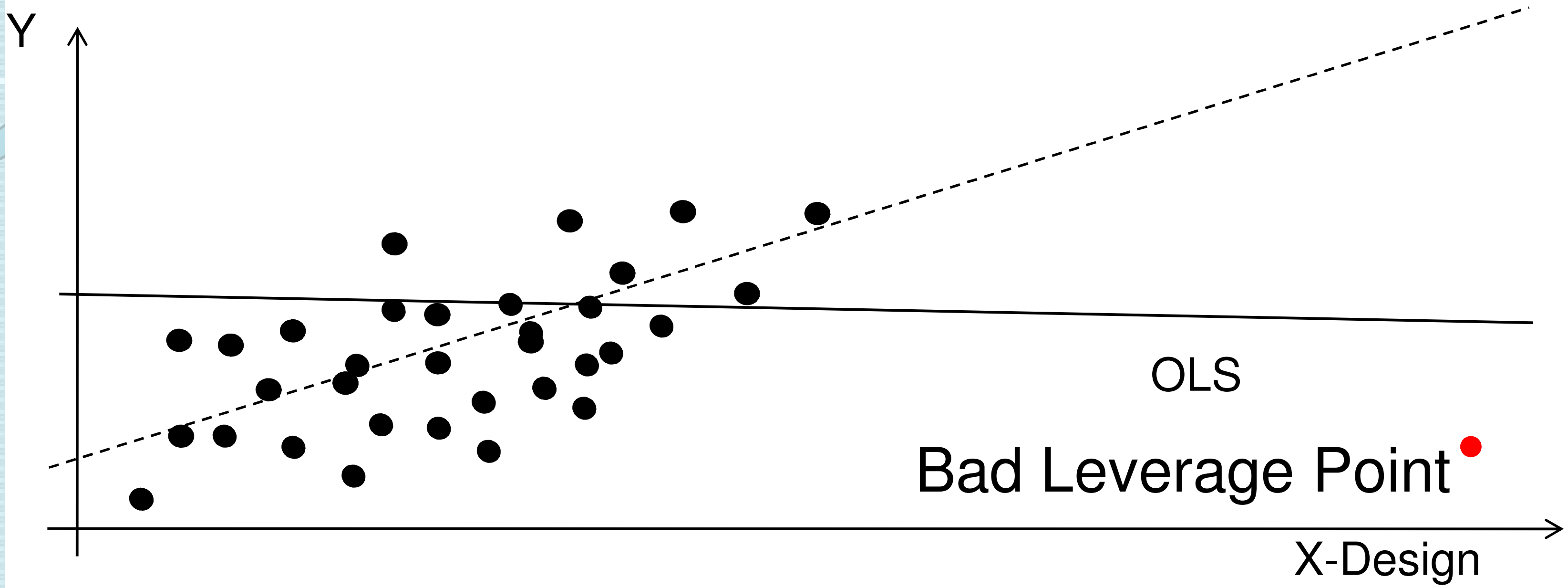
Stata codes

Conclusion



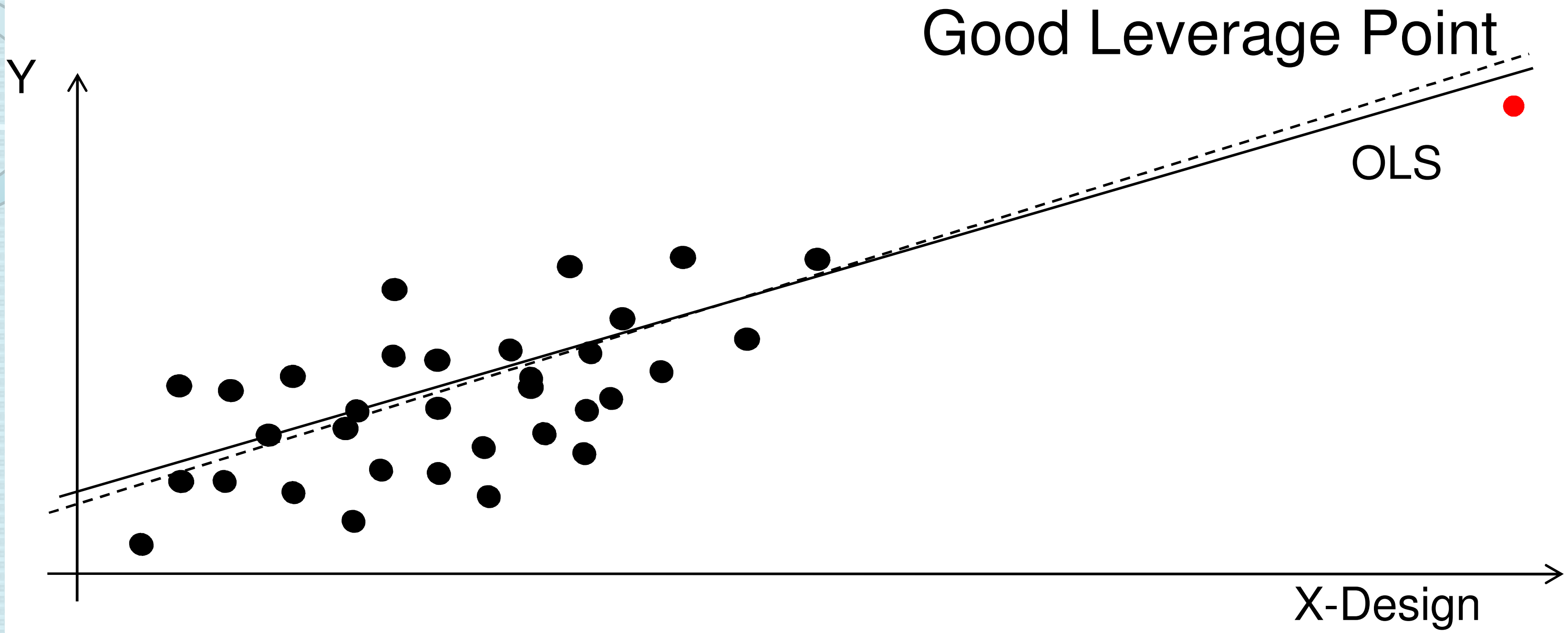
	Clean	Vertical
Intercept t-stat	1.24 (10.76)	2.24 (7.15)
Slope t-stat	0.59 (4.96)	0.67 (2.26)

Outliers in regression analysis



	Clean	Vertical	Bad leverage
Intercept t-stat	1.24 (10.76)	2.24 (7.15)	4.07 (6.99)
Slope t-stat	0.59 (4.96)	0.67 (2.26)	-0.42 (-9.02)

Outliers in regression analysis



	Clean	Vertical	Bad leverage	Good leverage
Intercept t-stat	1.24 (10.76)	2.24 (7.15)	4.07 (6.99)	1.25 (10.94)
Slope t-stat	0.59 (4.96)	0.67 (2.26)	-0.42 (-9.02)	0.57 (14.04)

Outliers in regression analysis

The objective of regression analysis is to figure out how a dependent variable is linearly related to a set of explanatory ones.

Technically speaking, it consists in estimating the θ parameters in:

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_{p-1} x_{ip-1} + \varepsilon_i$$

to find the model that better fits the data.

Ordinary Least Squares (LS)

On the basis of the estimated parameters, it is then possible to fit the model and predict, $\hat{y}$ the dependent variable. The discrepancy between y and $\hat{y}$ is called the residual ($r_i = y_i - \hat{y}_i$).

The objective of LS is to minimize the sum of the squared residuals:

$$\hat{\theta}_{LS} = \underset{\theta}{\arg\min} \sum_{i=1}^{n} r_i^2(\theta) \text{ where } \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_{p-1} \end{bmatrix}$$

L_1-estimator

However, the squaring of the residuals makes LS very sensitive to outliers.

To increase robustness, the square function could be replaced by the absolute value (Edgeworth, 1887).

$$\hat{\theta}_{L_1} = \underset{\theta}{\arg\min} \sum_{i=1}^{n} \left| r_i(\theta) \right|$$

[`qreg` function in Stata]

M-estimators

Huber (1964) generalized this idea to a set of symmetric ρ functions that could be used instead of the absolute value to increase efficiency and robustness.

To guarantee scale equivariance, residuals are standardized by a measure of dispersion σ.

The problem becomes:

$$\hat{\theta}_M = \underset{\theta}{\arg\min} \sum_{i=1}^{n} \rho\left(\frac{r_i(\theta)}{\sigma}\right)$$

Introduction

Outliers in regression analysis

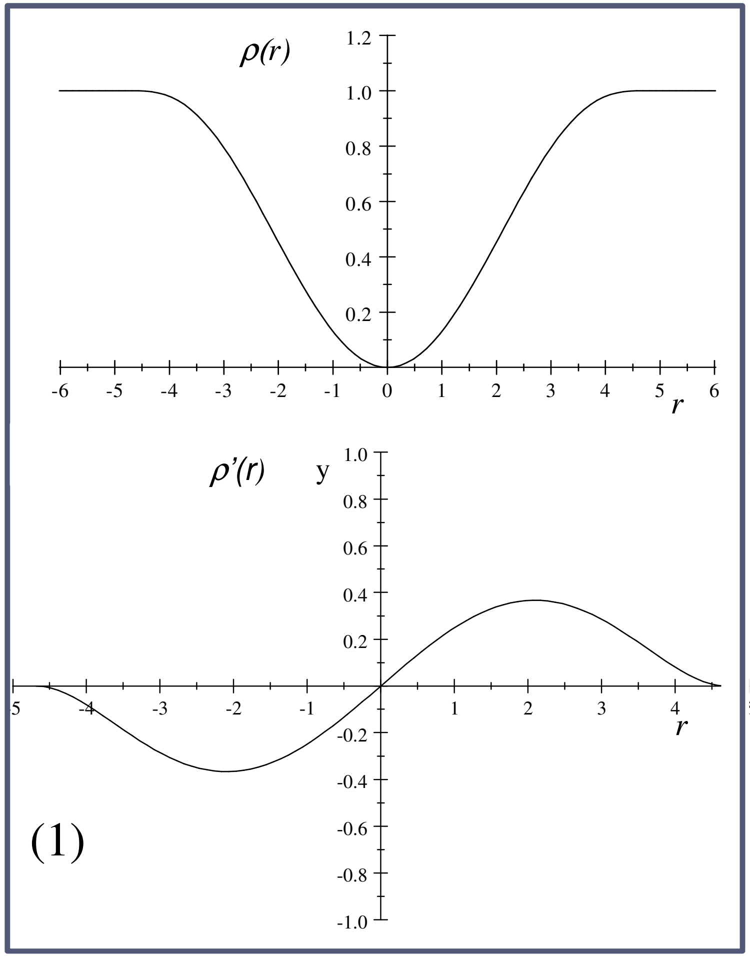
Overview of robust estimators

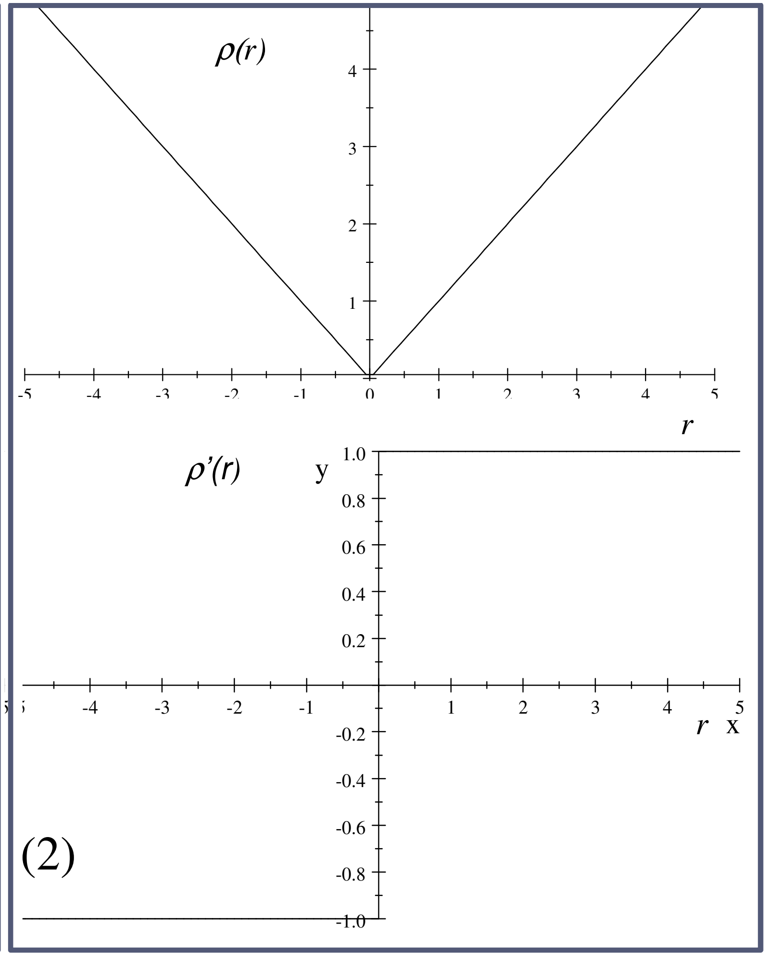
Stata codes

Conclusion

M-estimators

M-estimators can be redescending (1) or monotonic (2).





M-estimators

If σ is known, the practical implementation of M-estimators is straightforward. Indeed, by defining a weight:

$$w_i = \frac{\rho\left(r_i(\theta) / \sigma \right)}{r_i^2(\theta)}$$

the problem boils down to:

$$\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} w_i r_i^2(\theta)$$

M-estimators as WLS

$$\hat{\theta}_M = \underset{\theta}{\arg\min} \sum_{i=1}^{n} w_i r_i^2(\theta)$$

However:

1. Weights w_i are a function of θ that should thus be estimated iteratively

2. This iterative algorithm is guaranteed to converge (and yield a solution which is unique) only for monotonic M-estimators … which are not robust

3. σ is generally not known in advance

Stata's rreg command

The `rreg` command was created to tackle these problems. It works as follows:

1. It awards a weight zero to individuals with Cook distances larger than 1.

2. A "redescending" M-estimator is computed using the iterative algorithm starting from a monotonic M-solution.

3. σ is re-estimated at each iteration using the median residual of the previous iteration.

Stata's rreg command

Unfortunately, this command has not the expected robust properties:

1. Cook distances do not help identifying leverage points when (clustered) outliers mask one the other.

2. The preliminary monotonic M-estimator provides a poor initial candidate because of point 1.

3. σ is poorly estimated because of 1 and 2.

Illustration

`qreg` and `rreg` are not robust methods:

Stata example:

```
set obs 100
drawnorm x1-x5 e
gen y=x1+x2+x3+x4+x5+e
replace x1=invnorm(uniform())+10 in 1/10
qreg y x*
rreg y x*
display e(rmse)
```

Command: qreg

```
Iteration  1:  WLS sum of weighted deviations =    117.31824

Iteration  1: sum of abs. weighted deviations =    119.64818
Iteration  2: sum of abs. weighted deviations =    117.18714
Iteration  3: sum of abs. weighted deviations =    117.04369
Iteration  4: sum of abs. weighted deviations =    116.65145
Iteration  5: sum of abs. weighted deviations =    116.01905
Iteration  6: sum of abs. weighted deviations =    116.01677

Median regression                                    Number of obs =        100
  Raw sum of deviations  202.8451 (about -.23892587)
  Min sum of deviations  116.0168                    Pseudo R2     =     0.4281
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.179877	.0536822	3.35	0.001	.0732897	.2864643
x2	.7547212	.1589944	4.75	0.000	.4390341	1.070408
x3	.949198	.16758	5.66	0.000	.616464	1.281932
x4	.8773521	.1624611	5.40	0.000	.5547817	1.199922
x5	.9931675	.1791938	5.54	0.000	.637374	1.348961
_cons	-.0009245	.1887648	-0.00	0.996	-.3757213	.3738724

Command: rreg

```
. rreg y x*

   Huber iteration 1:  maximum difference in weights =   .48417173
   Huber iteration 2:  maximum difference in weights =   .06025306
   Huber iteration 3:  maximum difference in weights =   .01572401
Biweight iteration 4:  maximum difference in weights =   .14759052
Biweight iteration 5:  maximum difference in weights =   .00770808

Robust regression                                      Number of obs =       100
                                                          F(  5,    94) =     33.28
                                                          Prob > F      =    0.0000
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.175267	.0514961	3.40	0.001	.0730203	.2775136
x2	.9241295	.1459845	6.33	0.000	.6342739	1.213985
x3	.9221296	.1569926	5.87	0.000	.6104172	1.233842
x4	.7781905	.1554807	5.01	0.000	.4694801	1.086901
x5	1.115836	.1639707	6.81	0.000	.790268	1.441403
_cons	-.0584287	.175098	-0.33	0.739	-.4060898	.2892325

```
. display e(rmse)
1.6151557
```

S-estimators

Robustness can be however achieved by tackling the problem from a different perspective.

Instead of minimizing the variance of the residuals (LS) a more robust measure of spread of the residuals could be minimized (Rousseeuw and Yohai, 1987).

The measure of spread considered here is an M-estimator of scale.

S-estimators

Intuition:

The variance is defined by:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2(\theta)$$ which can be rewritten:

$$1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i(\theta)}{\hat{\sigma}}\right)^2$$ hence LS looks for the

minimal $\hat{\sigma}$ that satisfies the equality.

<u>But the square function ...</u>

S-estimators

Replace the square by another ρ :

$$1 = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\theta)}{\hat{\sigma}^S}\right)$$

but for Gaussian data we want $\hat{\sigma}^S$ to be the standard deviation ($\Rightarrow$ correction)

$E_\Phi[\rho(u)]$ →

$$\delta = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\theta)}{\hat{\sigma}^S}\right)$$ ← M-estimator of scale …

The problem boils down to finding the $\hat{\theta}_S$ associated to the minimal $\hat{\sigma}^S$ that satisfies the equality

S-estimators

ρ is generally (Tukey Biweight):

$$\rho\left(\frac{r_i}{\sigma}\right)=\begin{cases}1-\left[1-\left(\dfrac{r_i/\sigma}{k}\right)^2\right]^3 \text{ if } \left|\dfrac{r_i}{\sigma}\right|\leq k\\[2ex] 1 \text{ if } \left|\dfrac{r_i}{\sigma}\right|>k\end{cases}$$

where for k=1.548 the BDP is 50% and the efficiency is 28%. For k=5.182 the efficiency is 96% but the BDP is 10%.

MM-estimators

To ensure robustness AND efficiency, Yohai (1987) proposes to estimate an M-estimator:

$$\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\theta)}{\sigma}\right)$$

where ρ is a 95% efficiency Tukey Biweight function and where σ is set equal to $\hat{\sigma}^S$, estimated using a high BDP S-estimator. The starting point for the iterations is $\hat{\theta}_S$.

Introduction

Outliers in regression analysis

Overview of robust estimators

Stata codes

Conclusion

Sregress and MMregress

```
. Sregress y x*
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.9755606	.1331711	7.33	0.000	.7096758	1.241445
x2	1.181668	.1296818	9.11	0.000	.9227498	1.440586
x3	.920803	.1450545	6.35	0.000	.6311923	1.210414
x4	.6578808	.1425573	4.61	0.000	.373256	.9425057
x5	.7086012	.1443784	4.91	0.000	.4203404	.9968621
_cons	.0339972	.1464742	0.23	0.817	-.2584479	.3264424

Scale parameter= 1.180746

```
. MMregress y x*
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	1.035236	.116956	8.85	0.000	.8026558	1.267815
x2	.8967535	.1108331	8.09	0.000	.6763498	1.117157
x3	1.005016	.1179203	8.52	0.000	.7705186	1.239513
x4	.9289665	.1197309	7.76	0.000	.6908684	1.167065
x5	.9892967	.1268872	7.80	0.000	.7369677	1.241626
_cons	-.1214685	.1284036	-0.95	0.347	-.3768131	.133876

Scale parameter= 1.180745

Stata codes

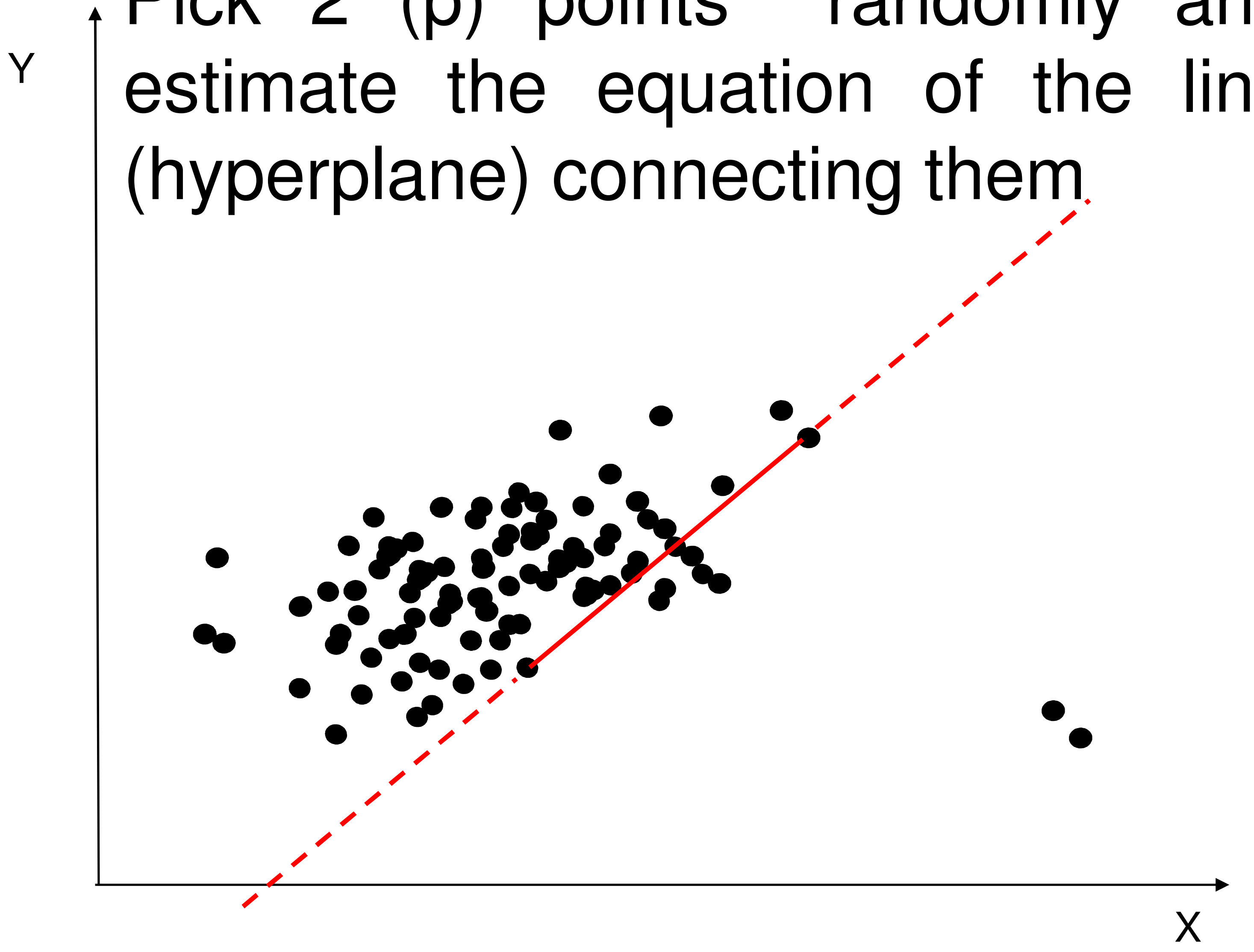
The implemented algorithm:

Salibian-Barrera and Yohai (2006)

1. P-subset
2. Improve the 10 best candidates (i.e. those with the 10 smallest $\hat{\sigma}^S$) using iteratively reweighted least squares.
3. Keep the improved candidate with the smallest $\hat{\sigma}^S$.

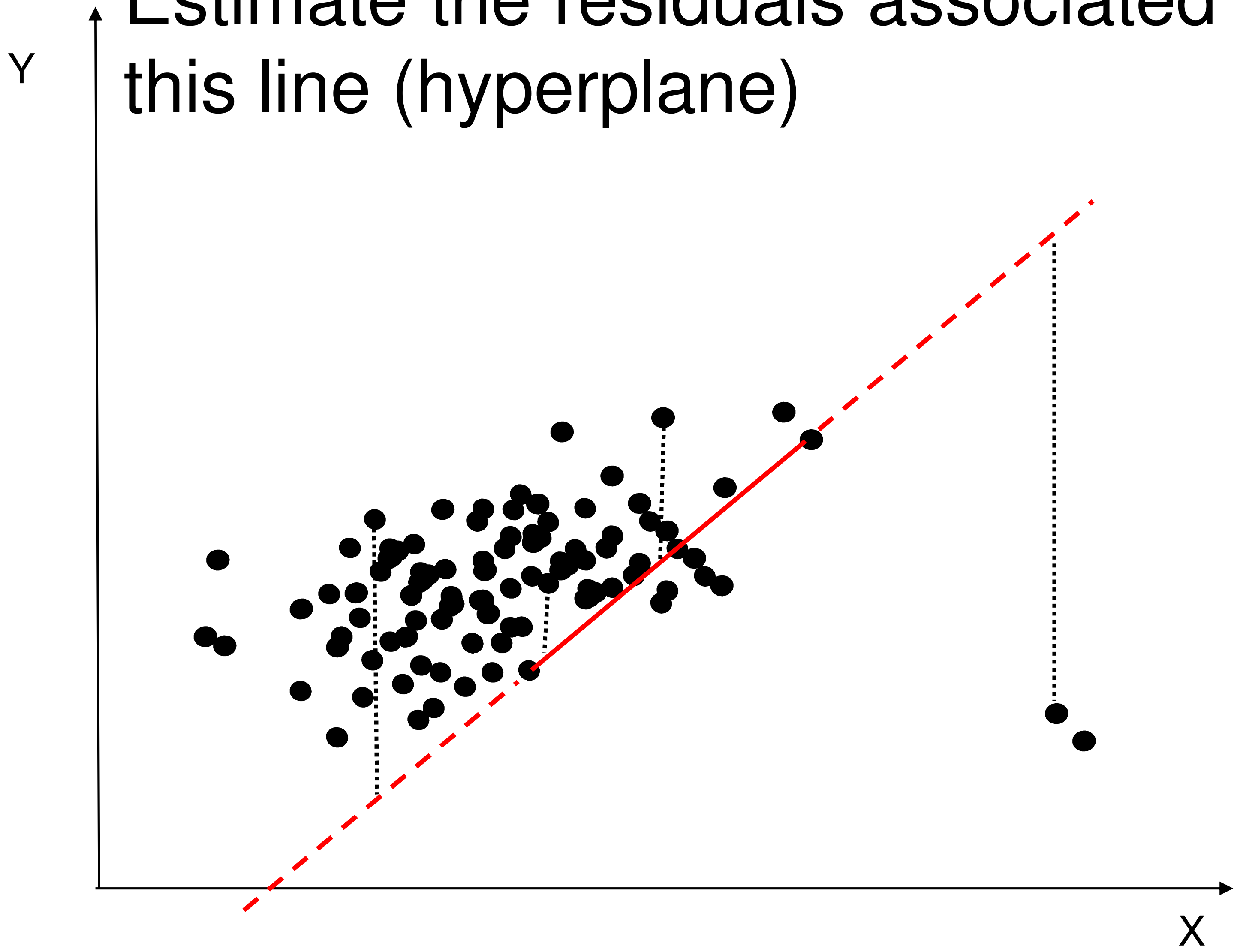
P-subset (p=2)

Pick 2 (p) points randomly and estimate the equation of the line (hyperplane) connecting them



P-subset (p=2)

Estimate the residuals associated to this line (hyperplane)



Introduction

Outliers in regression analysis

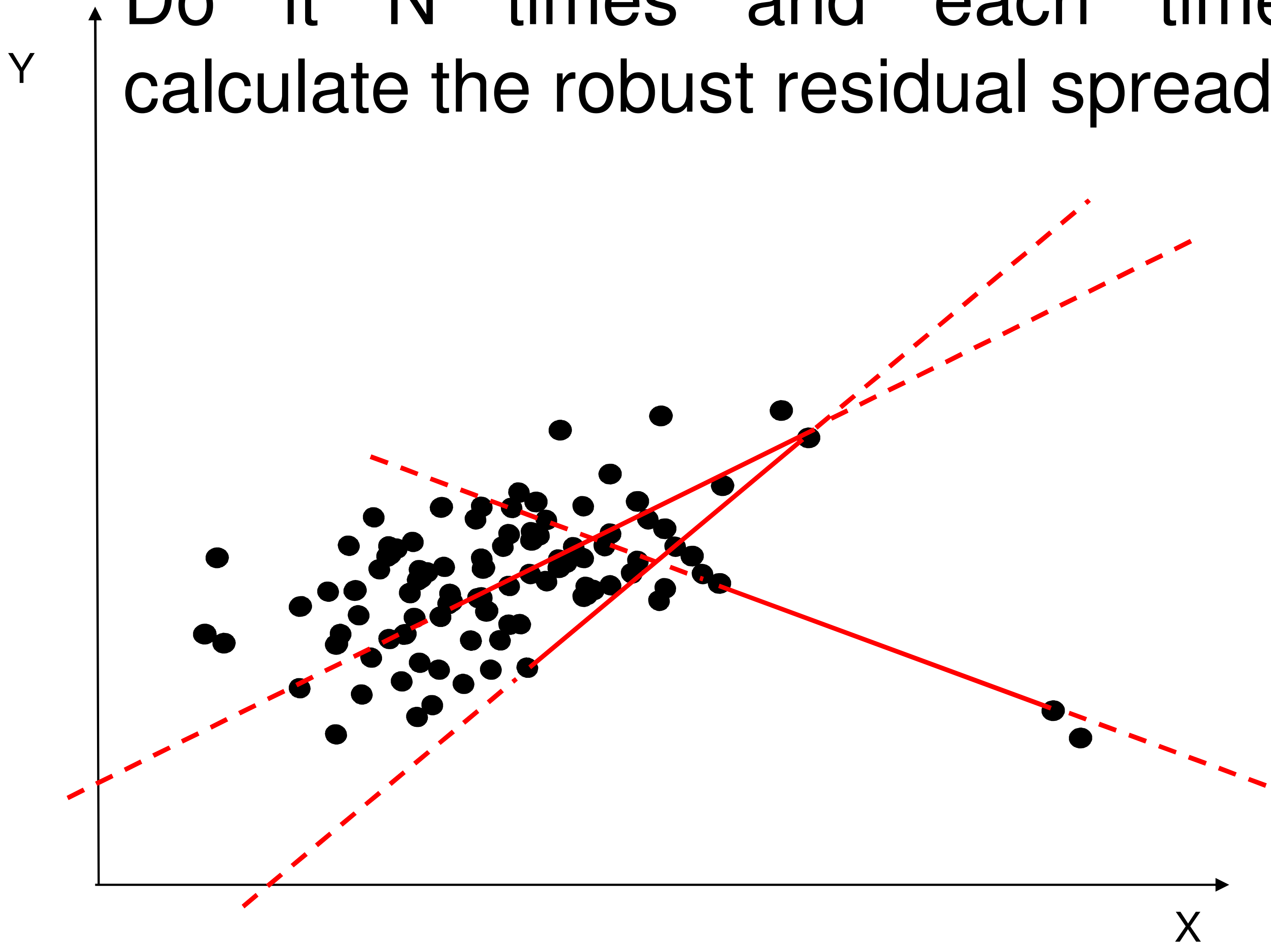
Overview of robust estimators

Stata codes

Conclusion

P-subset (p=2)

Do it N times and each time calculate the robust residual spread



P-subset (p=2)

Take the 10 regression lines (hyperplanes) associated with the smallest robust spreads and run the iterative algorithm described previously to improve the initial candidate.

The regression line (hyperplane) associated with the smallest refined robust spread will be the estimated S.

Number of subsets

The minimal number of subsets we need to have a probability (*Pr*) of having at least one clean if α% of outliers corrupt the dataset can be easily derived:

Contamination: α %

Number of subsets

The minimal number of subsets we need to have a probability (*Pr*) of having at least one clean if α% of outliers corrupt the dataset can be easily derived:

$$(1-\alpha)$$

Will be the probability that one random point in the dataset is not an outlier

Number of subsets

The minimal number of subsets we need to have a probability (*Pr*) of having at least one clean if α% of outliers corrupt the dataset can be easily derived:

$$(1-\alpha)^p$$

Will be the probability that none of the p random points in a p-subset is an outlier

Number of subsets

The minimal number of subsets we need to have a probability (*Pr*) of having at least one clean if α% of outliers corrupt the dataset can be easily derived:

$$1-(1-\alpha)^p$$

Will be the probability that at least one of the p random points in a p-subset is an outlier

Number of subsets

The minimal number of subsets we need to have a probability (*Pr*) of having at least one clean if α% of outliers corrupt the dataset can be easily derived:

$$\left[1-(1-\alpha)^{p}\right]^{N}$$

Will be the probability that there is at least one outlier in each of the N p-subsets considered (i.e. that all p-subsets are corrupt)

Number of subsets

The minimal number of subsets we need to have a probability (*Pr*) of having at least one clean if α% of outliers corrupt the dataset can be easily derived:

$$1-\left[1-(1-\alpha)^{p}\right]^{N}$$

Will be the probability that there is at least one clean p-subset among the N considered

Number of subsets

The minimal number of subsets we need to have a probability (*Pr*) of having at least one clean if α% of outliers corrupt the dataset can be easily derived:

$$\mathrm{Pr} = 1 - \left[1 - (1-\alpha)^p\right]^N$$

Rearranging we have:

$$N^* = \left\lceil \frac{\log(1-\mathrm{Pr})}{\log(1-(1-\alpha)^p)} \right\rceil$$

Drawback

If several dummies are present, the algorithm might lead collinear samples.

To solve this we programmed the MS-estimator (out of the scope here). Idea:

Drawback

If several dummies are present, the algorithm might lead collinear samples.

To solve this we programmed the MS-estimator (out of the scope here). Idea:

$$y = \underbrace{X_1}_{discrete}\theta_1 + \underbrace{X_2}_{continuous}\theta_2 + \varepsilon$$

$$\begin{cases} \theta_1^{MS} = \arg\min_{\theta_1} \sum_{i=1}^{n} \rho([y_i - X_2\hat{\theta}_2^{MS}] - X_1\theta_1) \\ \theta_2^{MS} = \arg\min_{\theta_2} \hat{\sigma}^S([y_i - X_1\hat{\theta}_1^{MS}] - X_2\theta_2) \end{cases}$$

Identify outliers

To properly identify outliers, in addition to robust (standardized) residuals, we need an assessment of the outlyingness in the design space (x variables).

This is generally done by calling on Mahalanobis distances:

$$MD = \sqrt{(x_i - \mu)\Sigma^{-1}(x_i - \mu)'}$$

That are known to be distributed as a $\sqrt{\chi_p^2}$ for Gaussian data.

Leverage points

However MD are not robust since they are based on classical estimations of μ (location) and Σ (scatter).

This drawback can be easily solved by using robust estimations μ and Σ.

Minimum Covariance Determinant

A well suited method for this is MCD that considers several subsets containing (generally) 50% of the observations and estimates μ and Σ on the data of the subset associated with the smallest covariance matrix determinant.

Intuition …

Generalized Variance

The generalized variance proposed by Wilks (1932), is a one-dimensional measure of multidimensional scatter. It is defined as $GV = \det(\Sigma)$.

In the 2x2 case it is easy to see the underlying idea:

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \text{ and } \det(\Sigma) = \sigma_x^2\sigma_y^2 - \sigma_{xy}^2$$

Spread due to covariations (σ_{xy}^2)

Raw bivariate spread ($\sigma_x^2\sigma_y^2$)

Fast-MCD Stata code

The implemented algorithm:

Rousseeuw and Van Driessen (1999)

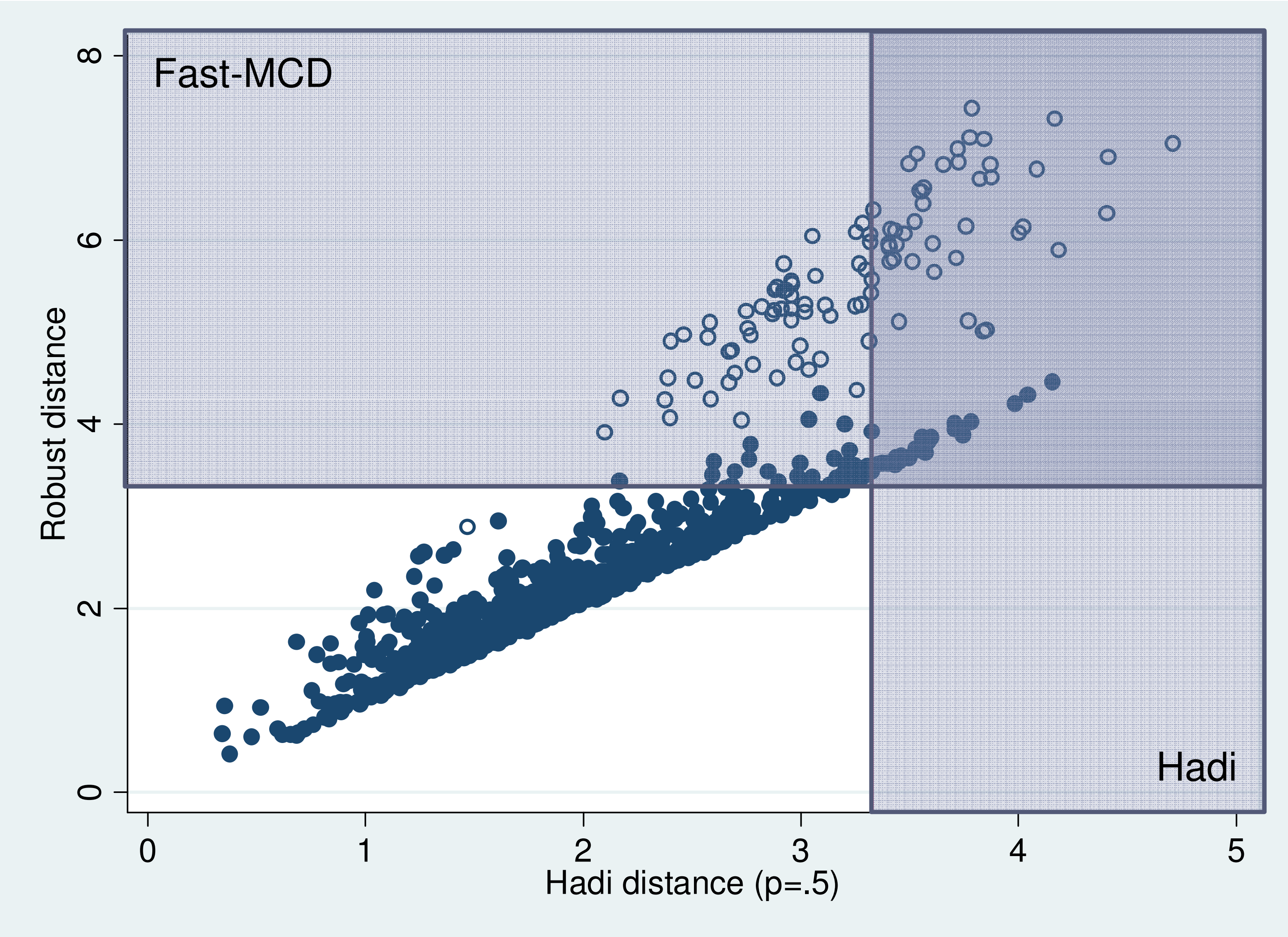
1. P-subset
2. Concentration (sorting distances)
3. Estimation of robust μ_{MCD} and Σ_{MCD}
4. Estimation of robust distances:

$$RD = \sqrt{(x_i - \hat{\mu}_{MCD})\hat{\Sigma}_{MCD}^{-1}(x_i - \hat{\mu}_{MCD})'}$$

Fast-MCD vs hadimvo

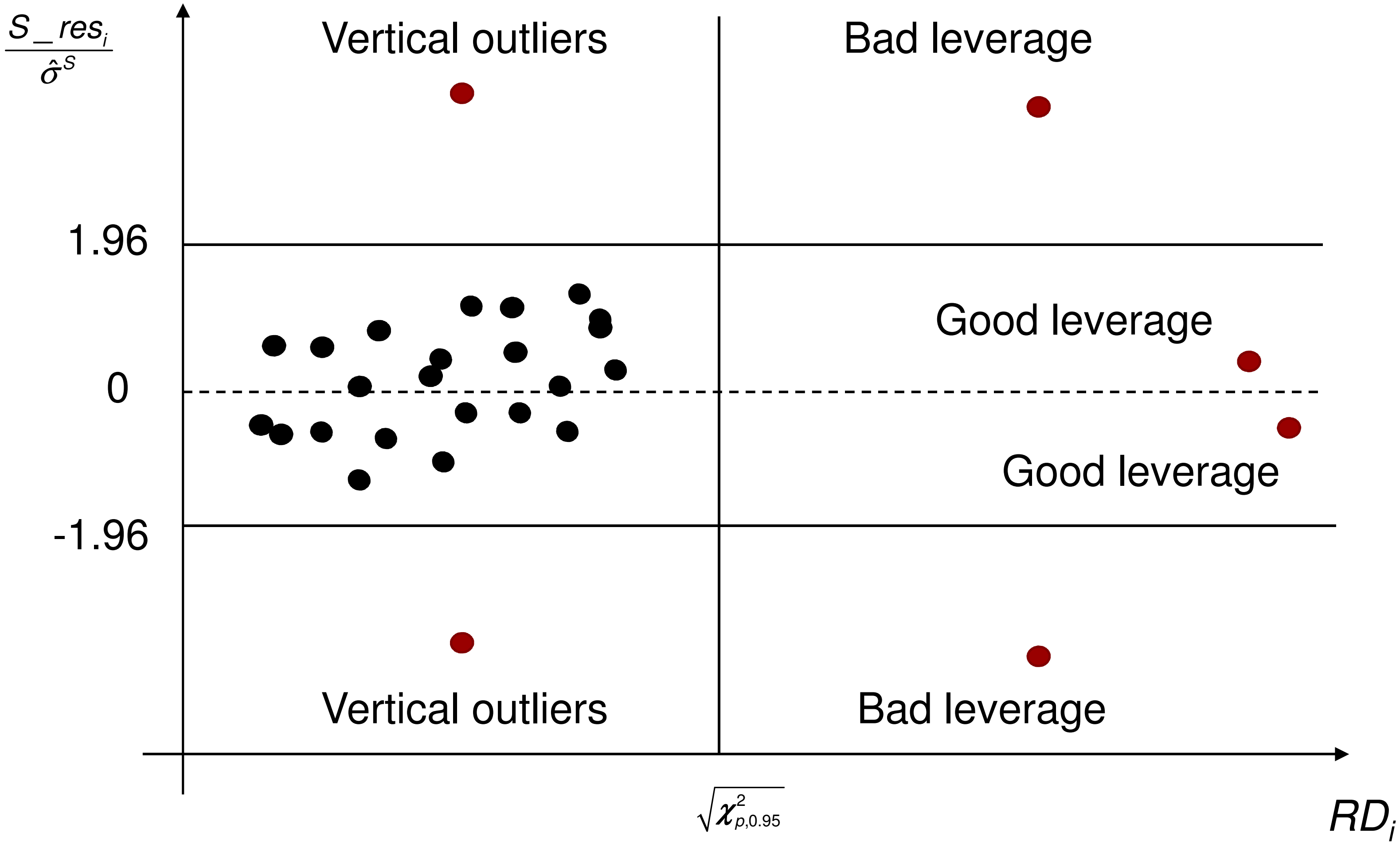
```
clear
set obs 1000
local b=sqrt(invchi2(5,0.95))
drawnorm x1-x5 e
replace x1=invnorm(uniform())+5 in 1/100
gen outlier=0
replace outlier=1 in 1/100
mcd x*, outlier
gen RD=Robust_distance
hadimvo x*, gen(a b) p(0.5)
Scatter RD b
```

Illustration



Identify outliers in regression

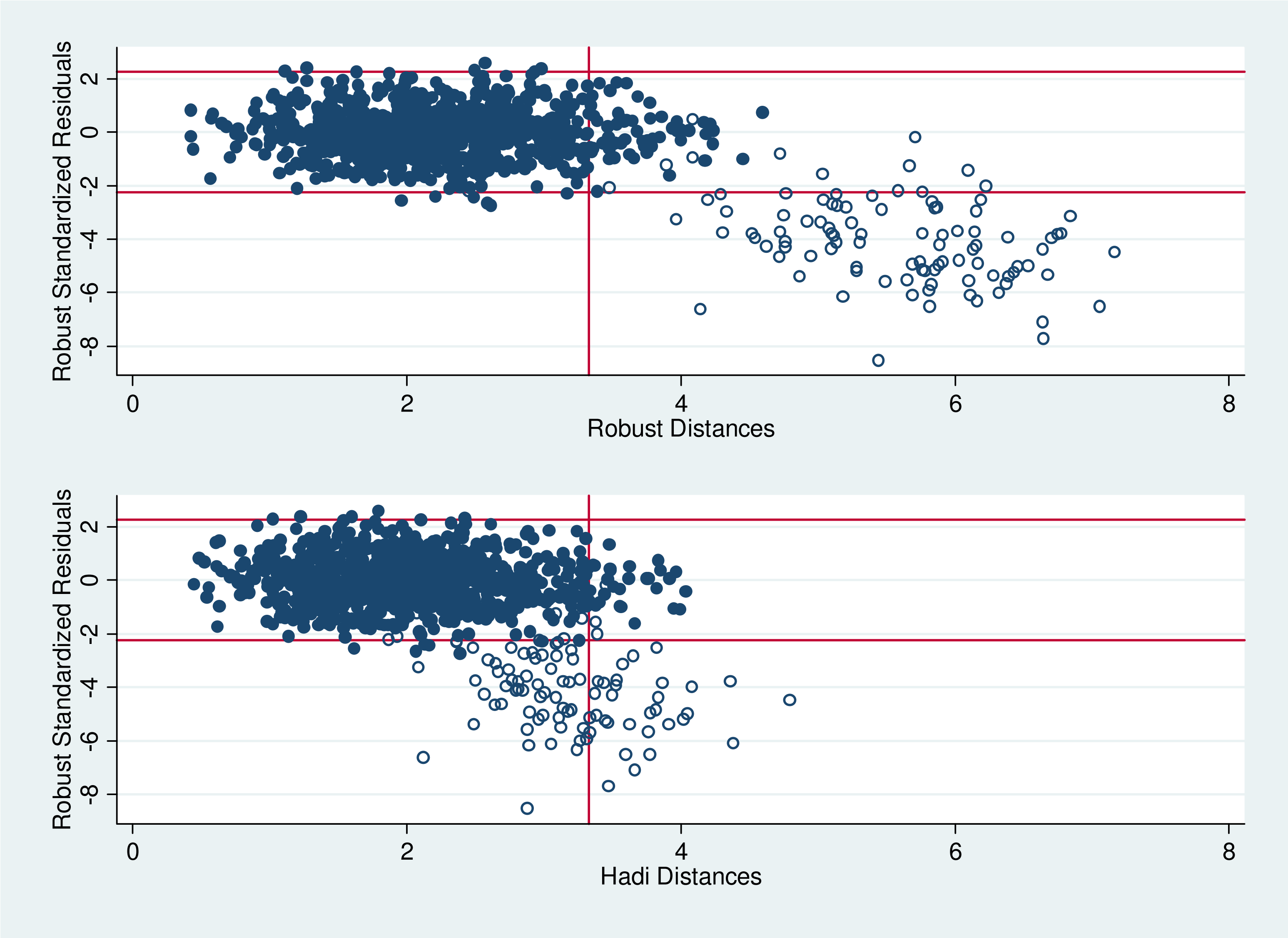
(Rousseeuw and Van Zomeren, 1990)



Illustration

```
clear
set obs 1000
local b=sqrt(invchi2(5,0.95))
drawnorm x1-x5 e
gen y=x1+x2+x3+x4+x5+e
replace x1=invnorm(uniform())+5 in 1/100
gen noise=1 in 1/100
Sregress y x*, outlier
mcd x*, outlier
hadimvo x*, gen(a b)
```

Illustration

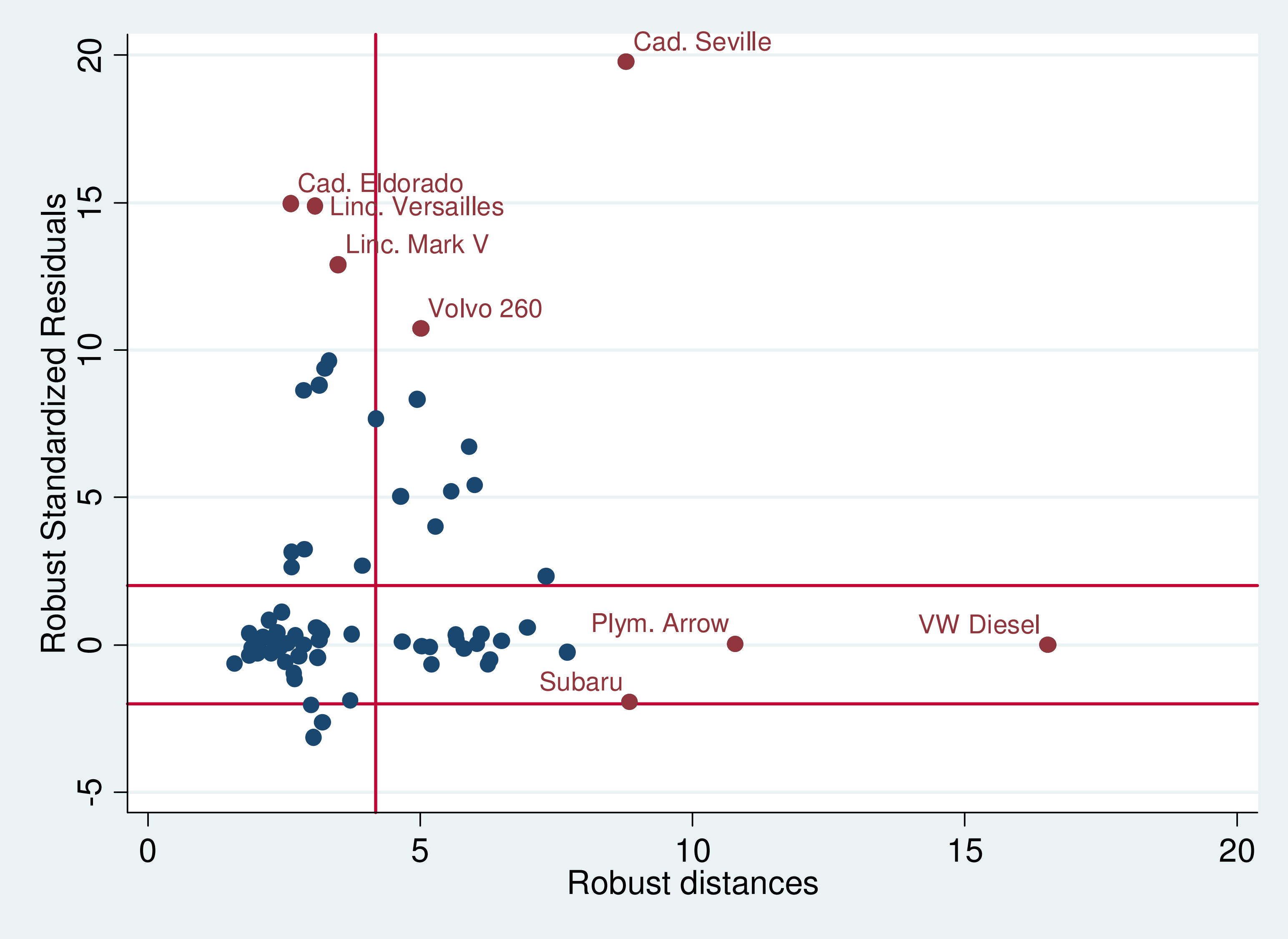


Example

```
webuse auto

xi: Sregress  price mpg headroom trunk
weight   length   turn   displacement
gear_ratio foreign i.rep78, outlier

mcd  mpg  headroom  trunk  weight  length
turn displacement gear_ratio, outlier

Scatter S_stdres Robust_distance
```

```
gen w1= invnormal(0.975)/abs(S_stdres)
replace w1=1 if w1>1
gen w2= sqrt(invchi2(r(N),0.95))/RD
replace w2=1 if w2>1
gen w=w1*w2
```

Example



Introduction

Outliers in regression analysis

Overview of robust estimators

Stata codes

Conclusion

Example

S

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-48.35603	21.18847	-2.28	0.029	-91.51553	-5.196522
headroom	-291.452	94.40745	-3.09	0.004	-483.7537	-99.15036
trunk	182.7921	26.66287	6.86	0.000	128.4816	237.1025
weight	1.188093	.3610366	3.29	0.002	.4526852	1.9235
length	-38.58704	11.50622	-3.35	0.002	-62.02444	-15.14965
turn	-6.398393	29.59498	-0.22	0.830	-66.68139	53.8846
displacement	3.427948	2.286095	1.50	0.144	-1.228675	8.084571
gear_ratio	568.3984	315.6108	1.80	0.081	-74.4799	1211.277
foreign	-132.9538	272.893	-0.49	0.629	-688.8187	422.9111
_Irep78_2	90.42532	358.4681	0.25	0.802	-639.7504	820.601
_Irep78_3	-784.8107	339.6177	-2.31	0.027	-1476.589	-93.03208
_Irep78_4	-309.2105	353.9961	-0.87	0.389	-1030.277	411.856
_Irep78_5	610.7227	376.5768	1.62	0.115	-156.3391	1377.785
_cons	6102.548	1666.071	3.66	0.001	2708.872	9496.224

LS

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-43.948	85.07476	-0.52	0.608	-214.4416	126.5456
headroom	-689.3982	400.1119	-1.72	0.091	-1491.24	112.444
trunk	74.29435	100.4034	0.74	0.462	-126.9186	275.5073
weight	4.667033	1.464867	3.19	0.002	1.731373	7.602693
length	-80.65842	43.41116	-1.86	0.069	-167.6563	6.339501
turn	-143.7061	129.3259	-1.11	0.271	-402.881	115.4688
displacement	12.70613	8.774824	1.45	0.153	-4.87901	30.29127
gear_ratio	115.0845	1269.769	0.09	0.928	-2429.59	2659.759
foreign	3064.515	1061.906	2.89	0.006	936.4084	5192.622
_Irep78_2	1353.801	1721.302	0.79	0.435	-2095.765	4803.366
_Irep78_3	955.4354	1618.354	0.59	0.557	-2287.818	4198.689
_Irep78_4	976.6333	1664.928	0.59	0.560	-2359.957	4313.224
_Irep78_5	1757.997	1804.181	0.97	0.334	-1857.663	5373.657
_cons	9969.75	7135.813	1.40	0.168	-4330.739	24270.24

Example

S

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-48.35603	21.18847	-2.28	0.029	-91.51553	-5.196522
headroom	-291.452	94.40745	-3.09	0.004	-483.7537	-99.15036
trunk	182.7921	26.66287	6.86	0.000	128.4816	237.1025
weight	1.188093	.3610366	3.29	0.002	.4526852	1.9235
length	-38.58704	11.50622	-3.35	0.002	-62.02444	-15.14965
turn	-6.398393	29.59498	-0.22	0.830	-66.68139	53.8846
displacement	3.427948	2.286095	1.50	0.144	-1.228675	8.084571
gear_ratio	568.3984	315.6108	1.80	0.081	-74.4799	1211.277
foreign	-132.9538	272.893	-0.49	0.629	-688.8187	422.9111
_Irep78_2	90.42532	358.4681	0.25	0.802	-639.7504	820.601
_Irep78_3	-784.8107	339.6177	-2.31	0.027	-1476.589	-93.03208
_Irep78_4	-309.2105	353.9961	-0.87	0.389	-1030.277	411.856
_Irep78_5	610.7227	376.5768	1.62	0.115	-156.3391	1377.785
_cons	6102.548	1666.071	3.66	0.001	2708.872	9496.224

Furthermore:

LS_R^2=0.61 LS_RMSE=2031

S_R^2=0.82 S_RMSE=402

Commands

```
Sregress varlist [if exp] [in range] [,
e(#) proba(#) noconstant outlier test
replic(#) setseed(#)]

MMregress varlist [if exp] [in range]
[, e(#) proba(#) noconstant outlier eff
replic(#)]

mcd varlist [if exp] [in range] [, e(#)
p(#) trim(#) outlier finsample]

MSregress varlist [if exp] [in range] ,
dummies(dummies) [ e(#) proba(#)
noconstant outlier test]
```

Conclusion

The available methods to identify (and treat) outliers in Stata are not fully efficient

The proposed commands should be helpful to deal with outliers in:

1. Regression analysis
2. Multivariate analysis (PCA, etc)
3. Available from vverardi@fundp.ac.be