Robust regression

Robust estimation of regression coefficients in linear regression model when orthogonality condition is breaking

Jiří Franc

Seminář strojového učení a modelování

Czech Technical University
Faculty of Nuclear Sciences and Physical Engineering
Department of Mathematics

Outline

The basic framework

Regression methods is one of the most widely used methods to cope with data analysis.

Definition

The multiple linear regression model is the model

$$Y_i = X_{i,1}\beta_1^0 + X_{i,2}\beta_2^0 + \cdots + X_{i,p}\beta_p^0 - e_i = X_i^T\beta^0 - e_i \qquad i = 1\ldots n.$$

or in the matrix notation

$$\mathbf{Y} = \mathbf{X}\beta^0 - \mathbf{e},$$

Where

- Y_i is a random sequence of response variables,
- $X_i{=}(X_{i,1}, X_{i,2}, \ldots, X_{i,p})^T$ is a random sequence of explanatory variables,
- $\beta^0{=}(\beta_1^0, \beta_2^0, \ldots, \beta_p^0)^T$ is a vector of unknown regression coefficients,
- e_i is a random sequence of unknown errors (disturbances).

Our main goal is to estimate the regression parameters β^0.

Note: Random explanatory variables X_i's is generally correlated with error terms e_i's.

The basic framework

Ordinary least squares method (OLS)

$$\hat{\beta}^{(OLS)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 = \arg\min_{\beta \in R^p} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$$

$$\hat{\beta}^{(OLS)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- In the certain conditions OLS is the best linear unbiased estimator (BLUE) of β^0.
- In the certain conditions OLS is the best estimator among all unbiased estimators (ordinary least squares method is best for multiple regression when the *iid* errors are normal distributed).
- OLS is not robust and consequently often gives false result for real data (even a single outlier can totally offset the OLS estimator).

The basic framework

Classical regression methods work well only under strict conditions and assumptions.

What if

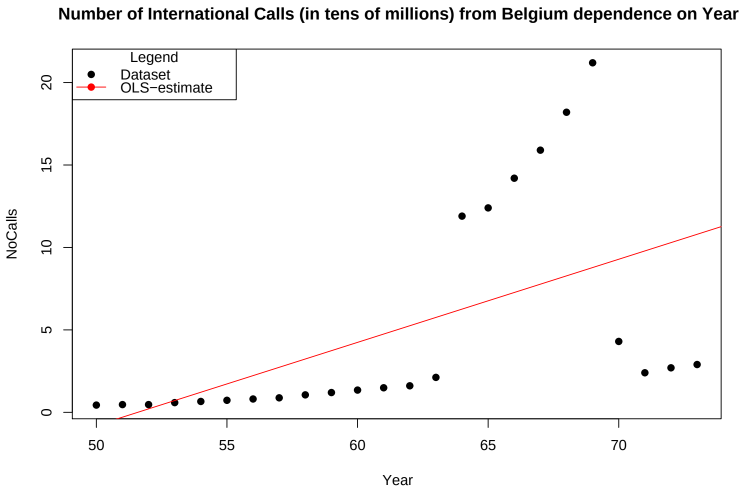
- wrong observations in the data set occur?
- assumptions are incorrect (e.g. orthogonality condition fails, $E[X_i\varepsilon_i] \neq 0$)?

The classical regression methods can be very misleading and the estimation can be totally damaged.

Methods, which can deal with outliers and some violation of basic assumptions, are called robust and they appeared for the first time in 1960s due to the works of J. W. Tukey , P. J. Huber or F. R. Hampel. Some of the most used robust regression estimators are M-Estimators, Least Trimmed Squares (LTS) or Weighted Least Squares (WLS).

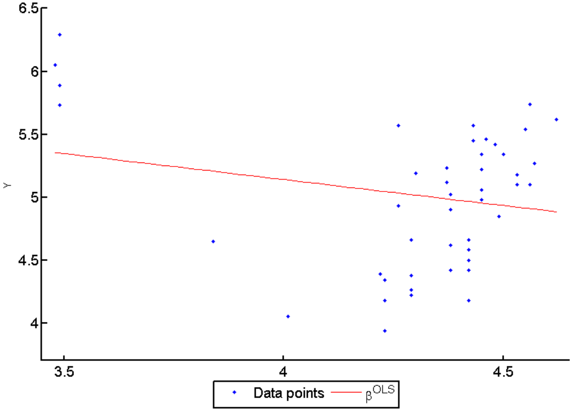
Motivation

An example of outliers in y-direction and OLS estimate.



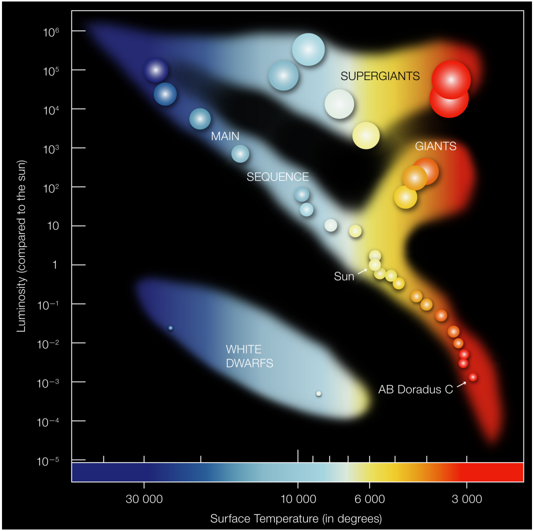
Motivation

Data points and OLS estimation for the Hertzsprung-Russell Diagram of the Star Cluster CYG OB1. The first variable is the logarithm of the effective temperature at the surface of the star and the second one is the logarithm of its light intensity.



Motivation

Hertzsprung-Russell Diagram



Robust regression

The main aims of robust statistics:

- description of the structure best fitting the bulk of the data.
- identification of deviating data points (outliers) or deviating substructures for further treatment.
- identification of highly influential data points (leverage points) or at least warning about them.
- deal with unsuspected serial correlations.

Two ways how to deal with regression outliers:

- **Regression diagnostics**: where certain quantities are computed from the data with the purpose of pinpointing influential points, after which these outliers can be removed or corrected.
- **Robust regression:** which tries to devise estimators that are not so strongly affected by outliers.

Robust regression

Influence function (IF)

Hampel (1968) introduced the approach to robustness based on the IF. The IF measures the infinitesimal influence of an observation situated at the point x on the value of the estimator (functional) T and allows to study local robustness properties (another terms derived from the *IF* are Gross Error Sensitivity, Local Shift Sensitivity or Rejection point).

Breakdown point

The breakdown point is a global measure of reliability (tell us when an estimator "still gives some relevant information").
Let $D = \{(X_{1,1}, \ldots, X_{1,p}, Y_1), \ldots, (X_{n,1}, \ldots, X_{n,p}, Y_n)\}$ be a sample of n data points, and let T be a regression estimator so that $\hat{\beta} = T(D)$. Consider all possible corrupted samples D' that are obtained by replacing any m of the original data points by arbitrary values.
Let the maximum *bias* that can be caused by such a contamination be
$bias(m, T, D) = \sup_{Z'} \| T(D') - T(D) \|$
The breakdown point of the estimator T at the sample D is defined as

$$\varepsilon_n^*(T, D) := \min\{\frac{m}{n};\ bias(m, T, D) \text{ is infinite}\}$$

Robust estimators

The set of requirements which we demand:

- consistency
- reasonably high *efficiency*
- *scale* and *regression equivariance*
- quite low gross-error sensitivity
- low local shift sensitivity
- finite rejection point
- controllable breakdown point
- existence of an algorithm with acceptable complexity and reliability of evaluation

M-estimators

M-estimators are based on the idea of replacing the squared residuals used in OLS estimation by another function of the residuals.

M-estimators

$$\hat{\beta}^{(M)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(r_i(\beta)),$$

where ρ is a symmetric function with a unique minimum at zero.

Differentiating this expression with respect to the regression coefficients yields:

M-estimators

$$\sum_{i=1}^{n} \psi(r_i) X_i = 0,$$

where ψ is the derivative of ρ. The M-estimate is obtained by solving this system of p equations.

M-estimators

- *OLS*, L_1 are also M-estimators with $\psi(t) = t$ for *OLS* and $\psi(t) = sgn(t)$ for L_1 estimate.
- M-estimators are unfortunately not scale equivariant even if they are regression equivariant. Hence one has to studentizate the M-estimators by an estimate of scale of disturbances $\hat{\sigma}$ necessarily.

$$\hat{\beta}^{(M)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right),$$

One possibility is to use the median absolute deviation (*MAD*):

$$\hat{\sigma} = C \operatorname{median}_i \left(\left| r_i - \operatorname{median}_j(r_j) \right| \right),$$

where C is a correction factor which depends on the distribution. For normally distributed data C = 1.4826.

M-estimators

The influence function with respect of Y_0 can by bounded by choice of ψ , but the influence function of M-estimators is unbounded in respect of $\mathbf{X_0}$. The breakdown point of M-estimators is 0% due to the vulnerability to leverage points.

Maronna and Yoahai (1981) showed, under certain conditions, that M-estimators are consistent and asymptoticly normal.

M-estimators

- **Huber minimax M-estimator**

$$\psi(t) = \begin{cases} t & \text{if } \ t < b \\ b\,\mathrm{sgn}(t) & \text{if } \ t \geq b \end{cases}$$

where b is a constant.

- **Andrew M-estimator**

$$\psi(t) = \begin{cases} \sin(t) & \text{if } \ -\pi \leq |t| < \pi \\ 0 & \text{otherwise} \end{cases}$$

- **Tukey M-estimator**

$$\psi(t) = \begin{cases} t\left(1 - \left(\frac{t}{c}\right)^2\right)^2 & \text{if } \ |t| < c \\ 0 & \text{otherwise} \end{cases}$$

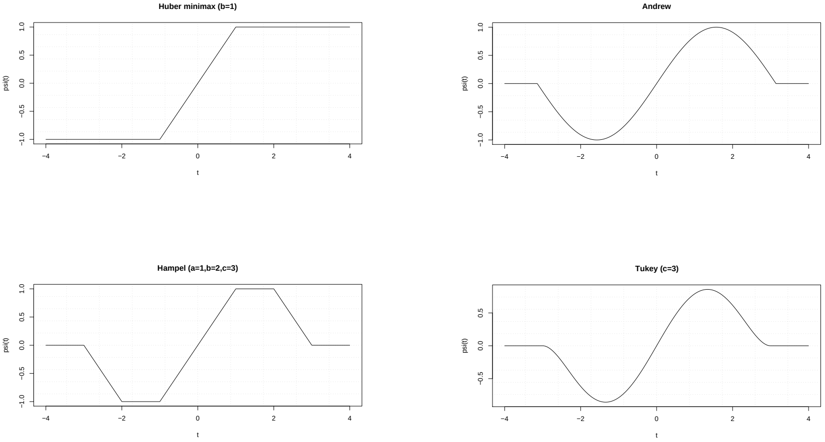
where c is a constant.

- **Hampel M-estimator**

$$\psi(t) = \begin{cases} t & \text{if } \ |t| < a \\ a\,\mathrm{sgn}(t) & \text{if } \ a \leq |t| < b \\ \frac{c-|t|}{c-b}\,\mathrm{sgn}(t) & \text{if } \ b \leq |t| < c \\ 0 & \text{otherwise} \end{cases}$$

where a, b and c are constants.

M-estimators



GM-estimators

Generalized M-estimators are introduced in order to bounding the influence function of outlying X_i's by means of some weight function w.

GM-estimators

$$\hat{\beta}^{(GM)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} w(X_i) \frac{\rho(r_i(\beta))}{\hat{\sigma}}$$

The definition can be rewrite to

$$\sum_{i=1}^{n} w(X_i)\psi\left(\frac{r_i}{\hat{\sigma}}\right) X_i = \mathbf{0}.$$

Unfortunately Maronna, Buston and Yohai (1979) showed that the breakdown point of GM-estimators can be no better than a certain value that decrease as a function of p^{-1}, where p is the number of regression coefficients.

GM-estimators

The algorithm of Iteratively reweighted least squares with GM-estimates based on some ψ function is following.

1. The first elementary estimate $\hat{\beta}^{(OLS)}$ of β^0.
2. Count the residuals $r_i(\hat{\beta}) = Y_i - \hat{Y}_i = Y_i - X_i^T\hat{\beta} \quad i = 1 \ldots n$.
3. Count the estimate $\hat{\sigma}$ of σ.
   (e.g. MAD: $\hat{\sigma} = 1.4826 \operatorname*{median}_i \left(\left| r_i - \operatorname*{median}_j (r_j) \right| \right)$)
4. Count the weights w_i.
   (e.g. Andrew's ψ function: $w_i = \frac{\psi(\frac{r_i}{\hat{\sigma}})}{\frac{r_i}{\hat{\sigma}}}$)
5. Update the estimate $\hat{\beta}$ by performing a weighted least squares with the weights w_i
   Calculate $\hat{\beta}^{(WLS)} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$
6. Go back to item 2 and iterate until convergence

Another robust estimators

- **R-estimation**: procedure based on the ranks of the residuals.
- **S-estimators**: procedure derived from a scale statistic in an implicit way.
- **MM-estimators**: high-breakdown and high-efficiency estimators, where the initial estimate is obtained with an S-estimator, and it is then improved with an M-estimator.
- **Least median of squares (LMS)**: probably the first really applicable 50% breakdown point estimator introduced by Rousseeuw (1984).

$$\hat{\beta}^{(LMS)} = \arg\min_{\beta \in \mathbb{R}^p} \left(\underset{i}{med} (r_i^2(\beta)) \right).$$

LTS-estimator

Least trimmed squares estimator - Rousseeuw (1984)

$$\hat{\beta}^{(LTS)} = \arg\min_{\beta \in \mathbb{R}^p} = \sum_{i=1}^{h} r^2_{(i)}(\beta),$$

where $r^2_{(1)} \leq \ldots \leq r^2_{(n)}$ are the ordered squared residuals.

- There always exists a solution for the LTS-estimator.
- The LTS estimator is regression equivariant , scale equivariant and affine equivariant.
- If $p > 1$, $h = [n/2] + [(p+1)/2]$ then the breakdown point of the LTS-estimator is
$$\varepsilon^* := ([n-p]/2 + 1)/n.$$
- The LTS can be very sensitive to a very small change of data or to a deletion of even one point from data set (i.e. small change of data can really cause a large change of the estimate).

Implicit weighting and Least weighted squares (LWV)

For any $\beta \in \mathcal{R}^p$ define the ith rank residual as $r_i(\beta) = Y_i - X_i^T \beta$ and $r^2_{(h)}(\beta)$ denotes the h-th order statistic among the squared residuals:

Method of the Least Weighted Squares (LWS) - Víšek (2000)

$$\hat{\beta}^{(LWS,w,n)} = \arg\min_{\beta \in \mathcal{R}^p} \sum_{i=1}^{n} w_i r^2_{(i)}(\beta) = \arg\min_{\beta \in \mathcal{R}^p} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r^2_{(i)}(\beta),$$

where weights w_i are defined by the weight function $w : \langle 0,1\rangle \to \langle 0,1\rangle$, which is absolutely continuous, $w(0) = 1$ and non-increasing with the derivative $w'(t)$ bounded from below by the constant $(-L)$.

Implicit weighting and Least weighted squares (LWV)

For any $i \in \{1, \ldots, n\}$ let's denote by $\pi(\beta, i)$ the random rank of the i-th residual as

$$\pi(\beta, i) = j \in \{1, \ldots, n\} \qquad \Leftrightarrow \qquad r_i^2(\beta) = r_{(j)}^2(\beta)$$

Method of the Least Weighted Squares (LWS)

$$\hat{\beta}^{(LWS,w,n)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} w\left(\frac{\pi(\beta, i) - 1}{n}\right) r_i^2(\beta).$$

Normal equations for the Least Weighted Squares

$$\sum_{i=1}^{n} w\left(\frac{\pi(\beta, i) - 1}{n}\right) X_i(Y_i - X_i^T\beta) = 0.$$

The problem, how to find the LWS estimator $\hat{\beta}^{(LWS,w,n)}$, is equal to the problem, how to find "the best" classical weighted least squares $\hat{\beta}^{(WLS,w(best),n)}$ among $n!$ possibilities.

The basic framework

In econometrics, the explanatory variables are frequently assumed to be correlated with the random errors $p\lim\left(\frac{1}{n}\mathbf{X}^T\mathbf{e}\right) \neq \mathbf{0}$
If we now apply LS, LTS or LWS estimators, we get an inconsistent estimate.

Model in which the explanatory variables are measured with a random error

We suppose that $Y_i = Y_{0i} - \varepsilon_i, \; X_i = X_{0i} - \theta_i$
and that there exists $\beta^0 \in \mathbb{R}^{p\times 1}$ such that $Y_i + \varepsilon_i = (X_i + \theta_i)\beta^0, \; i = 1\dots n$.

Assuming usually that $\mathbb{E}\left[\varepsilon_i\right] = 0$, $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2 \in (0,\infty)$ and $\mathbb{E}\left[\theta_i\right] = 0$, $\mathbb{E}\left[\theta_i\theta_i^T\right] = \Sigma_\theta$ nonsingular and $\mathbb{E}\left[\theta_i\varepsilon_i\right] = 0$. If we consider now classical regression model

$$Y_i = X_{0i}\beta^0 - \varepsilon_i = (X_i + \theta_i)\beta^0 - \varepsilon_i = X_i\beta^0 + \theta_i\beta^0 - \varepsilon_i = X_i\beta^0 + e_i,$$

we can easily find out that orthogonality condition is broken.

$$\mathbb{E}\left[X_ie_i\right] = \mathbb{E}\left[(X_{0i} - \theta_i)\cdot(\theta_i\beta^0 - \varepsilon_i)\right] = -\Sigma_\theta\beta^0.$$

There are two possibilities how to cope with such a cases when the orthogonality condition is broken and in addition the data set contains outliers.

Instrumental variables (IV)

In econometrics, the explanatory variables $\mathbf{X}$ are usually assumed to be correlated with the random error $\mathbf{e}$, (i.e. $p\lim\left(\frac{1}{n}\mathbf{X}^T\mathbf{e}\right) \neq \mathbf{0}$).

Suppose there are some variables $\mathbf{Z}$, called instruments, that are uncorrelated with $\mathbf{e}$ ($E\left[Z_i e_i\right] = 0$) and the matrix of correlations between the variables in $\mathbf{X}$ and the variables in $\mathbf{Z}$ is of maximum possible rank ($E\left[Z_i X_i^T\right] = \Sigma_{XZ},\ \mathrm{rank}(\Sigma_{XZ}) = p$).

$$\hat{\beta}^{(IV)} = (\mathbf{Z}^T\mathbf{X})^{-1}\mathbf{Z}^T\mathbf{Y} = \beta^0 + \left(\frac{1}{n}\sum_{i=1}^{n} Z_i X_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \epsilon_i\right) \xrightarrow[n\to\infty]{P} \beta^0.$$

Normal equations for the Instrumental variables

$$\sum_{i=1}^{n} Z_i(Y_i - X_i^T\beta) = 0.$$

How to find proper instruments?

Instrumental weighted variables (IWV)

Method of the Instrumental weighted variables (IWV) - Víšek (2007)

Let $\mathbf{Z}$ be any array of proper instrumental variables, then the instrumental weighted variables estimator $\hat{\beta}^{(IWV,w,n)}$ is defined by the solution of normal equations

$$\mathbb{NE}_{Z,n}(\beta) = \sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) Z_i \left(Y_i - X_i^T \beta\right) = 0,$$

If we compute all permutations $\pi \in \mathcal{P}_n$

$$\hat{\beta}^{(WIV,n,W(\pi))} = (\mathbf{Z}^T\mathbf{W}(\pi)\mathbf{X})^{-1}\mathbf{Z}^T\mathbf{W}(\pi)\mathbf{Y},$$

and we find the permutation π_{best} defined as

$$\pi_{best} = \arg\min_{\pi\in\mathcal{P}_n} \sum_{i=1}^{n} w\left(\frac{\pi_i - 1}{n}\right)\left(Y_i - X_i^T\hat{\beta}^{(WIV,n,W(\pi))}\right)^2$$

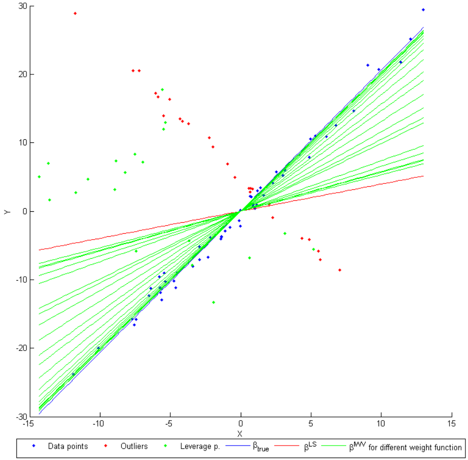
then it holds

$$\hat{\beta}^{(IWV,n,w)} = \hat{\beta}^{(WIV,n,W(\pi_{best}))} = (\mathbf{Z}^T\mathbf{W}(\pi_{best})\mathbf{X})^{-1}\mathbf{Z}^T\mathbf{W}(\pi_{best})\mathbf{Y}.$$

Let basic conditions be fulfilled. Then the sequence $\left\{\hat{\beta}^{(IWV,n,w)}\right\}_{n=1}^{+\infty}$ of the solutions of normal equations $\mathbb{NE}_{Z,n}(\beta) = 0$ is weakly consistent. Proof: Víšek (2007), another approach Franc (2009).

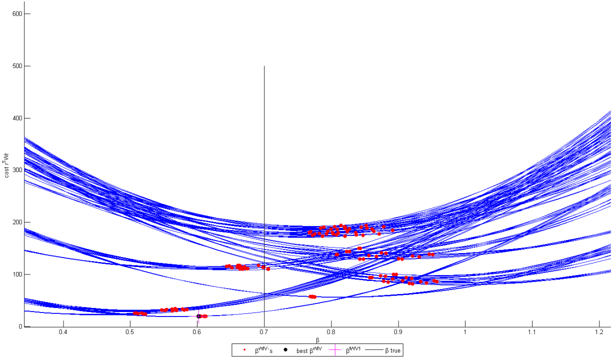
Selection of weighting function

The example of different regression lines $\hat{\beta}^{(IWV,n,W(h))}$, for varying parameter $h = \frac{n}{2}, \ldots, n$.



Global minimum of weighted instrumental variables

Convex curves (120 different parabolas) that show the dependence of the cost function (sum of weighted squared residuals) on parameter $\beta \in R$ for certain weights with minima in $\hat{\beta}^{(WIV,n,W(\pi))}$.



Algorithms

Classical algorithm is based on the idea of iterative re-weighting. The $j+1$th iteration of the *IWV* estimator is obtained as:

$$\hat{\beta}_{(j+1)}^{\left(IWV,n,\mathbf{W}\left(\hat{\beta}_{(j)}^{(IWV,n)}\right)\right)} = \left(\mathbf{Z}^T\mathbf{W}\left(\hat{\beta}_{(j)}^{(IWV,n)}\right)\mathbf{X}\right)^{-1}\mathbf{Z}^T\mathbf{W}\left(\hat{\beta}_{(j)}^{(IWV,n)}\right)\mathbf{Y},$$

where as the initial estimate $\hat{\beta}_{(0)}^{(IWV,n)}$ we can consider the simple OLS estimator of p randomly picked different observations and

$$W(\beta) = \mathrm{diag}\left\{w_1, w_2, \ldots, w_n\right\} \quad \mathrm{s} \quad w_i = w\left(\frac{\pi(\beta, i) - 1}{n}\right).$$

Another types of algorithms are based on on theory of simulated annealing and use Metropolis-Hastings algorithm for Markov Chain - Monte Carlo (MCMC) or on genetic algorithms.

The quality of the estimation consists not only in the choice of Instruments.

Total Least Squares

The Total Least Squares method is viewed as a tool for deriving approximate linear static models and is sometimes called Orthogonal Regression or Errors-in-variables model.

Given an overdetermined set of n linear equations $\mathbf{Y} \approx \mathbf{X}\beta$ in p unknowns β.

- the Ordinary Least Squares problem seeks to

$$\hat{\beta}^{(OLS,n)} = \min_{\beta \in \mathbb{R}^{\boldsymbol{p}}, \varepsilon \in \mathbb{R}^{\boldsymbol{n}}} \|\varepsilon\|_2 \quad \text{subject to} \quad \mathbf{Y} + \varepsilon = \mathbf{X}\beta.$$

- the Data Least Squares problem seeks to

$$\hat{\beta}^{(DLS,n)} = \min_{\beta \in \mathbb{R}^{\boldsymbol{p}}, \Theta \in \mathbb{R}^{\boldsymbol{n} \times (\boldsymbol{p})}} \|\Theta\|_F \quad \text{subject to} \quad \mathbf{Y} = (\mathbf{X} + \Theta)\beta.$$

- the Total Least Squares problem seeks to

$$\hat{\beta}^{(TLS,n)} = \min_{\beta \in \mathbb{R}^{\boldsymbol{p}}, [\varepsilon, \Theta] \in \mathbb{R}^{\boldsymbol{n} \times (\boldsymbol{p}+1)}} \|[\varepsilon, \Theta]\|_F \quad \text{subject to} \quad \mathbf{Y} + \varepsilon = (\mathbf{X} + \Theta)\beta.$$

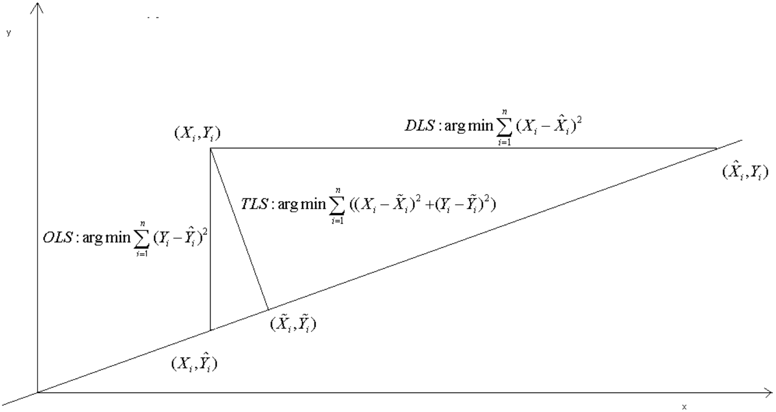
The norm $\| \ \|_F$ is called the Frobenius norm

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{n} \sum_{i=1}^{p} x_{ij}^2} = \sqrt{\operatorname{trace}(\mathbf{X}^T \mathbf{X})} = \sqrt{\sum_{i=1}^{\min\{n,p\}} \sigma_i^2} = \sqrt{\sum_{i=1}^{\operatorname{rank}(\mathbf{X})} \sigma_i^2},$$

where σ_i's are the singular values of the matrix $\mathbf{X}$.

LS x DLS x TLS fitting

The comparison of OLS, DLS and TLS estimate.



Total Least Squares

TLS minimizes the sum of the squared orthogonal distances from the data points to the fitting hyperplane.

$$\begin{aligned}\hat{\beta}^{(TLS,n)} &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\sum_{i=1}^{n}\frac{\left|\nu^T(A-p_i)\right|^2}{\|\nu\|^2} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\sum_{i=1}^{n}\frac{\left|\left[\beta^T,-1\right]\begin{bmatrix}X_i\\ Y_i\end{bmatrix}\right|^2}{\|[\beta^T,-1]\|^2}\\ &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\frac{1}{1+\|\beta\|^2}\sum_{i=1}^{n}|Y_i-X_i\beta|^2 = \underset{\beta\in\mathbb{R}^p}{\arg\min}\frac{\|\mathbf{Y}-\mathbf{X}\beta\|}{\sqrt{1+\|\beta\|^2}}.\end{aligned}$$

where A is arbitrary point from the fitting hyperplane ρ and $\nu = \left[\beta^T,-1\right]^T$ is the normal vector of ρ.

SVD

Singular Value Decomposition Theorem

Let us consider $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\operatorname{rank}(X) = r$ then there exist orthonormal matrices $\mathbf{U} = [u_1, \ldots, u_r] \in \mathbb{R}^{n \times r}$ and $\mathbf{V} = [v_1, \ldots, v_r] \in \mathbb{R}^{p \times r}$ such that

$$\mathbf{X} = \mathbf{U} \Sigma \mathbf{V}^T, \quad \Sigma = \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\} \in \mathbb{R}^{r \times r},$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.

The dyadic expansion (decomposition) of the matrix $\mathbf{X}$ is following

$$\mathbf{X} = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

Numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ are square roots of nonzero eigenvalues of matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ related to eigenvectors $\{u_1, \ldots, u_r\}$ and $\{v_1, \ldots, v_r\}$.

Golub-Van Loan Theorem

Suppose that the matrix $[\mathbf{Y}, \mathbf{X}]$ has full column rank.

Theorem

Let the singular value decomposition of $[\mathbf{X}, \mathbf{Y}] = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ *and* $\sigma_{min}(\mathbf{X})$ *be the smallest singular value of* $\mathbf{X}$. *If* $\sigma_{min}(\mathbf{X}) > \sigma_{p+1}$, *then the TLS solution*

$$\hat{\beta}^{(TLS,n)} = -\frac{1}{v_{p+1,p+1}} \left[v_{1,p+1}, \ldots, v_{p,p+1}\right]^T$$

exists and is the unique solution to $\mathbf{Y_0} = \mathbf{X_0}\beta$ *and the corresponding TLS correction matrix is given by*

$$[\varepsilon, \Theta] = \sigma_{p+1} u_{p+1} v_{p+1}^T.$$

Golub-Van Loan Theorem

Suppose that the matrix $[\mathbf{Y}, \mathbf{X}]$ has full column rank.

Theorem

Let the singular value decomposition of $[\mathbf{X}, \mathbf{Y}] = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ *and* $\sigma_{min}(\mathbf{X})$ *be the smallest singular value of* $\mathbf{X}$*. If* $\sigma_{min}(\mathbf{X}) > \sigma_{p+1}$*, then the TLS solution*

$$\hat{\beta}^{(TLS,n)} = -\frac{1}{v_{p+1,p+1}} \left[v_{1,p+1}, \dots, v_{p,p+1}\right]^T$$

exists and is the unique solution to $\mathbf{Y}_0 = \mathbf{X}_0 \beta$ *and the corresponding TLS correction matrix is given by*

$$[\varepsilon, \Theta] = \sigma_{p+1} u_{p+1} v_{p+1}^T.$$

Since singular vectors v_i's are eigenvectors of the matrix $[\mathbf{Y}, \mathbf{X}]^T [\mathbf{Y}, \mathbf{X}]$, then $\hat{\beta}^{(TLS,n)}$ satisfies

$$\hat{\beta}^{(TLS,n)} = (\mathbf{X}^T\mathbf{X} - \sigma_{p+1}^2 \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

Golub-Kahan bidiagonalization

The computational stability and speed can by improved by using the Golub-Kahan bidiagonalization to the matrix $[\mathbf{X}, \mathbf{Y}]$. This concept is called core problem and has been developed by Paige and Strakoš (2006). The idea is to find by the help of GKB two orthonormal matrices $\mathbf{P}, \mathbf{Q}$ such that

$$\mathbf{P}^T [\mathbf{Y}, \mathbf{X}\mathbf{Q}] = \begin{bmatrix} b_1 & \mathbf{A}_{11} & 0 \\ 0 & 0 & \mathbf{A}_{22} \end{bmatrix}$$

where the matrix $\mathbf{A}_{11}$ is lower bidiagonal with nonzero bidiagonal elements, has full column rank, its singular values are simple and has minimal dimensions, $\mathbf{A}_{22}$ has maximal dimensions and the first elements of all left singular vectors of $\mathbf{A}_{11}$, are nonzero. These properties guarantee that the subproblem $b_1 \approx \mathbf{A}_{11}\beta_{11}$ has minimal dimensions and contains all necessary and sufficient information for solving the original problem $\mathbf{Y} \approx \mathbf{X}\beta$. All irrelevant and redundant information is contained in $\mathbf{A}_{22}$.

Total Least Trimmed Squares (TLTS)

TLTS minimizes the sum of the h smallest squared orthogonal distances of data points p_i's from the pth dimensional fitting hyperplane $\rho(\beta)$.
The j-th orthogonal distances is denoted by d_j and defined by

$$d_j = \frac{|Y_j - X_j\beta|^2}{1 + \|\beta\|^2}.$$

Total Least Trimmed Squares

$$\hat{\beta}^{(TLTS,n)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{h} d_{(i)}^2,$$

where h is an optional parameter satisfying $\frac{n}{2} \leq h \leq n$ and $d_{(i)}^2$ is the i-th least squared orthogonal distance, i.e. for any $\beta \in \mathbb{R}^p$

$$d_{(1)}^2(\beta) \leq d_{(2)}^2(\beta) \leq \ldots \leq d_{(n)}^2(\beta).$$

Total Least Weighted Squares (TLWS)

The infinite local sensitivity of TLTS can be improved by adding some continuous weighting function and multiply the distances by a weights from $\langle 0, 1\rangle$.

Total Least Weighted Squares

$$\begin{aligned}\hat{\beta}^{(TLWS,w,n)} &= \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} w\left(\tfrac{i-1}{n}\right) d_{(i)}^2(\beta) = \\ &= \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) d_i^2(\beta),\end{aligned}$$

where weights w_i are defined by the weight function $w : \langle 0, 1\rangle \rightarrow \langle 0, 1\rangle$, which is absolutely continuous, $w(0) = 1$ and non-increasing with the derivative $w'(t)$ bounded from below by a constant $(-L)$, where $L \geq 0$ and $\pi(\beta, i)$ is the random rank of the i-th residual.

Mixed LS - TLS

If the linear modeling problem $\mathbf{Y} \approx \mathbf{X}\beta$ contains the intercept or some columns of $\mathbf{X}$ are known exactly, the TLS solution does not give the accurate estimation. The generalization of the TLS approach is called mixed least squares - total least squares problem.

$$\mathbf{Y} \approx \mathbf{X}\beta, \quad \mathbf{Y} \in \mathbb{R}^{n}, \ \mathbf{X} \in \mathbb{R}^{n \times p}, \ n > p,$$

$$\text{partition} \quad \begin{array}{ll} \mathbf{X} = \left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right] & \mathbf{X}^{(1)} \in \mathbb{R}^{n \times p_1}, \ \mathbf{X}^{(2)} \in \mathbb{R}^{n \times p_2} \\ \beta^{T} = \left[\beta^{(1)\,T}, \beta^{(2)\,T}\right] & \beta^{(1)} \in \mathbb{R}^{p_1}, \ \beta^{(2)} \in \mathbb{R}^{p_2} \end{array}$$

and assume that the columns of $\mathbf{X}^{(1)}$ are error free and $p_1 + p_2 = p$.

LS-TLS problem

$$\hat{\beta}^{(LS-TLS,n)} = \min_{\beta \in \mathbb{R}^{p}, [\varepsilon, \Theta] \in \mathbb{R}^{n \times (p_2+1)}} \left\| [\varepsilon, \Theta] \right\|_F$$

$$\text{subject to} \quad \mathbf{Y} + \varepsilon = \mathbf{X}^{(1)}\beta^{(1)} + (\mathbf{X}^{(2)} + \Theta)\beta^{(2)}.$$

Mixed LS - TLS

Let a matrix $\left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right]$ be given, have full column rank and columns of $\mathbf{X}^{(1)}$ are error free. Suppose that $0 < p_1 < p$ and compute the QR factorization of

$$\left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}\right] = \mathbf{Q} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \mathbf{R}_{\mathbf{Y}_1} \\ 0 & \mathbf{R}_{22} & \mathbf{R}_{\mathbf{Y}_2} \end{bmatrix}.$$

Then compute the ordinary TLS solution $\hat{\beta}^{(TLS,n-p_1)}$ of $\mathbf{R}_{\mathbf{Y}_2} \approx \mathbf{R}_{22}\beta$ which gives the last p_2 components of $\hat{\beta}^{(LS-TLS,n)}$. The first p_1 components we obtain from the solution of following equation

$$\mathbf{R}_{11}\hat{\beta}^{(LS,p_1)} = \mathbf{R}_{\mathbf{Y}_1} - \mathbf{R}_{12}\hat{\beta}^{(TLS,n-p_1)}.$$

The mixed LS-TLS solution is $\hat{\beta}^{(LS-TLS,n)} = \left[\hat{\beta}^{(LS,p_1)}, \hat{\beta}^{(TLS,n-p_1)}\right]$.

Unfortunately this universal estimator is not robust and gives misleading results when outliers occur.

Robustified mixed LS - TLS

To compute the robustified mixed LS-TLS estimation we need to identify the influential points from both parts and downweight them.

Let us compute the squared vertical distances of each data point from the $p_1 + 1$ dimensional hyperplane given by LS solution and squared orthogonal distances of each data point from the $p_2 + 1$ dimensional hyperplane given by TLS solution. Discard $n - h$ outermost points. Compute ordinary mixed LS-TLS solution only for remaining data points. Repeat these two steps until convergence. This estimation can be called mixed Least Trimmed Squares - Total Least Trimmed Squares.

Present and future work

Verify the properties of *Least Trimmed Squares - Total Least Trimmed Squares* and *Least Weighted Squares - Total Least Weighted Squares* estimation trough more simulations.

Prove some theoretical properties of these estimators such as consistency.

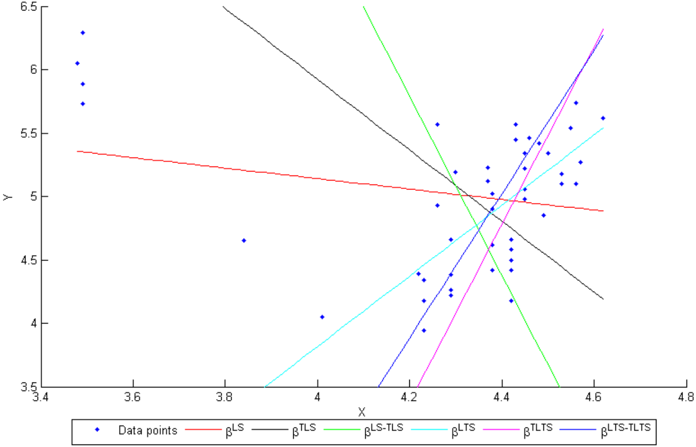
Testing on real data set

Data for the Hertzsprung-Russell Diagram of the Star Cluster CYG OB1, which contains 47 stars in the direction of Cygnus. The first variable is the logarithm of the effective temperature at the surface of the star and the second one is the logarithm of its light intensity.

Estimation of Hertzsprung-Russell diagram data set						
	$\hat{\beta}^{LS}$	$\hat{\beta}^{TLS}$	$\hat{\beta}^{LS-TLS}$	$\hat{\beta}^{LTS}$	$\hat{\beta}^{TLTS}$	$\hat{\beta}^{LTS-TLTS}$
β_1	6.7935	17.1124	35.4293	-7.3095	-26.0518	-19.9323
β_2	-0.4133	-2.7973	-7.0574	2.7816	7.0074	5.6710

Testing on real data set

Data points and various estimation lines for the Hertzsprung-Russell Diagram of the Star Cluster CYG OB1.



References

- Golub, G. and Van Loan, C. (1980), An analysis of the total least squares problem. SIAM J. Numerical Analysis 17, 883 -893.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986), Robust Statistics: The Approach Based on Influence Functions, J. Wiley, New York.
- Rousseeuw, P. J. and Leroy, A. M. (1987), Robust Regression and Outlier Detection. J. Wiley, New York.
- Van Huffel, S. and Vandewalle, J. (1991), The Total Least Squares Problem: Computational Aspects and Analysis, SIAM Philadelphia.
- Jurečková, J. (2001), Robustní statistické metody, Karolinum, Prague.
- Víšek, J. Á. (2001), Regression with high breakdown point. Robust 2000, 324 - 356.
- Paige, C. C. and Strakoš, Z. (2006),Core problems in linear algebraic systems, SIAM Journal on Matrix Analysis and Applications 27, 861 - 875
- Markovsky, I. and Van Huffel, S. (2007), Overview of total least squares methods, Signal Processing 87, number 10, 2283 - 2302
- Víšek, J. Á. (2009), Consistency of the instrumental weighted variables, Annals of the Institute of Statistical Mathematics 61, number 3, 543 - 578.
- Franc, J. (2009), Robustified instrumental variables, Master thesis, FNSPE, Czech Technical University in Prague, Prague.
- Franc, J. (2010), Robustified Total Least Squares, Doktorandské dny 2010, Proceedings of Workshop 2010, p. 47 - 58, Prague.

Thank you for attention.