Introduction to Robust Statistics

Klaus Nordhausen

Department of Mathematics and Statistics
University of Turku

Autumn 2015

Robust regression

The plan is now to develop robust alternative for ordinary least squares regression.

The goals are the usual:

- Being efficient in a target model.
- Being also efficient in a neighborhood of the target model.

Model for robust regression

Follow the the general model formulation

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)$ are the regression coefficients including the intercept which means $\mathbf{x}$ is a vector where the first element is always 1.

Then for the error distribution we assume the symmetric density

$$\frac{1}{\sigma} f_0\left(\frac{\epsilon}{\sigma}\right)$$

which means we assume in a sample $y_1, \ldots, y_n$ that the observations are independent but not identically distributed and we have for y_i

$$y_i \sim \frac{1}{\sigma} f_0\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)$$

Equivariance for regression

So what are desirable properties a regression estimate should have?

Let $T(\mathbf{x}, y)$ be a p-dimensional regression functional, then desirable equivariance properties are:

- Regression equivariance:

$$T(\mathbf{x}, y + \boldsymbol{\gamma}'\mathbf{x}) = T(\mathbf{x}, y) + \boldsymbol{\gamma} \quad \text{for all p-vectors } \boldsymbol{\gamma}.$$

- Scale equivariance:

$$T(\mathbf{x}, \lambda y) = \lambda T(\mathbf{x}, y) \quad \text{for all } \lambda \neq 0.$$

- Affine equivariance:

$$T(\mathbf{A}'\mathbf{x}, y) = \mathbf{A}^{-1} T(\mathbf{x}, y) \quad \text{for all full rank matrices } \mathbf{A}.$$

Traditional approach

Assume for now, that the $\mathbf{x}$ are fixed, non-random and "full rank". Which means that outliers, can only occur in ϵ /y.

The traditional M approach for regression is then that $\hat{\boldsymbol{\beta}}$ must solve:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) \quad \text{and} \quad \sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) \mathbf{x}_i = 0,$$

where as usual the connection between ρ and ψ is $\rho' = \psi$ and $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ is the residual for observation i.

$\hat{\sigma}$ is an estimate of the scale of the error.

As previously, ρ and ψ are motivated by MLE, but need not to be MLE's for any distribution.

M-estimates of regression

It will be assumed that ρ and ψ fulfil the same criteria as the corresponding functions for location M-estimates.

Again we distinguish between monotone regression M-estimates and redescending regression M-estimates depending on the behavior of the corresponding ψ function.

Recall that for a monotone increasing ψ function the solution is unique but for redescending ψ functions might have "bad" roots.

Again however redescending estimates are considered to have a better trade-off between efficiency and robustness.

Monotone regression M-estimates are usually only used nowadays as starting points as will be elaborated later.

M-estimates of regression with known scale

Assume for a moment that the scale would be known.

Then it holds that

$$\hat{\boldsymbol{\beta}}_1 \to_p \boldsymbol{\beta}_1$$

And for large sample sizes

$$\hat{\boldsymbol{\beta}}_1 \approx N_p(\boldsymbol{\beta}_1, v(\mathbf{X}'\mathbf{X})^{-1}),$$

where $v = \sigma^2 \frac{E(\psi(\epsilon/\sigma)^2)}{(E(\psi'(\epsilon/\sigma)))^2}$.

This means that the efficiency at the normal model compared to least squares does not depend on the design $\mathbf{X}$.

M-estimates of regression with and the scale problem

Formulating least squares in the M-estimation framework means that

$$\rho(\hat{\boldsymbol{\beta}}) = \frac{r_i^2(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} \quad \text{and} \quad \psi(\hat{\boldsymbol{\beta}}) = \frac{r_i(\hat{\boldsymbol{\beta}})\mathbf{x}_i}{\hat{\sigma}},$$

and the MLE based on a double exponential distribution gives

$$\rho(\hat{\boldsymbol{\beta}}) = \frac{|r_i(\hat{\boldsymbol{\beta}})|}{\hat{\sigma}} \quad \text{and} \quad \psi(\hat{\boldsymbol{\beta}}) = \frac{\text{sign}(r_i(\hat{\boldsymbol{\beta}}))\mathbf{x}_i}{\hat{\sigma}}.$$

This means in those two cases, the scale is not needed to solve the equations for $\hat{\boldsymbol{\beta}}$!

However in the general case, again one needs either to plug in a scale estimate or estimate regression coefficient and scale simultaneously.

Least absolute deviation

The MLE estimate based on the double exponential distribution yields actually the counterpart of the median for regression and is known as least absolute deviation (LAD) estimate or L_1 estimate.

- The estimate seems older that the least squares estimate.
- LAD is regression, scale and affine equivariant.
- Solutions cannot be given in closed forms and are usually not unique. However many efficient algorithms exist (they are however quite sophisticated and advanced).
- The residuals in this model have a zero median.
- A property of the estimate is, that at least p residuals are zero.

M-estimation with plug in scale

The natural way it to obtain somehow a "good" estimate of the residuals and estimate then robustly the residual scale.

As LAD does not require a scale estimate it is the natural "first" estimate to obtain the residuals.

As the residuals of the LAD estimate have median zero and many zero residuals it is custom to estimate the scale as

$$\hat{\sigma} = \frac{1}{0.675} \operatorname{med}(|r_i| \; | r_i \neq 0).$$

The zero residuals are usually excluded to avoid to underestimate σ in case p is large.

M-estimation with plug in scale II

Assume that your initial scale estimate is equivariant and that $\hat{\sigma} \to_p \sigma$.

Also then for large sample sizes

$$\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, v(\mathbf{X}'\mathbf{X})^{-1}),$$

where $v = \sigma^2 \frac{E(\psi(\epsilon/\sigma)^2)}{(E(\psi'(\epsilon/\sigma)))^2}$.

For practical purposes v can be estimates as

$$\hat{v} = \hat{\sigma}^2 \frac{n}{n-p} \frac{\text{ave}\left\{\psi(r_i/\hat{\sigma})^2\right\}}{\left(\text{ave}\left\{\psi'(r_i/\hat{\sigma})\right\}\right)^2}.$$

M-estimation with plug in scale III

Hence having a way of estimation the regression parameters as well as of the parameters of the limiting distribution inference like p-values and confidence intervals is straight forward.

For example $\hat{\beta}_i \approx N(\beta_i, \hat{v}(\mathbf{X}'\mathbf{X})^{-1}_{ii})$.

M-estimation with plug in scale IV

The general recommendation for M-regression is then usually:

- Use the LAD as initial estimate.
- Get a robust estimate of scale.
- Plug the scale into a redescending ψ function.

Then your estimate will give basically zero weight to large residuals.

M-estimation with plug in scale V

Assume that the selected ψ function is smooth and let the weight function be

$$u(r) = \begin{cases} \psi(r)/r & r \neq 0 \\ \psi'(r) & r = 0 \end{cases}$$

Hence in this case a iteratively reweighted least squares (IRWLS) algorithm can be used, where at iteration k

- $w_{i,k} = u((y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_k)/\hat{\sigma})$
- solve $\sum_{i=1}^{n} w_{i,k}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})$ to obtain $\boldsymbol{\beta}_{k+1}$

Convergence is usually defined as

$$\max_i \left\{ |(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_k) - (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{k+1})| \right\} < \epsilon\hat{\sigma}$$

M-estimation with simultaneous estimation of scale

As in the location case, this approach is less used in this context as it is considered less robust.

The main idea is to solve simultaneously

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \rho_s\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) = \delta$$

For smooth ψ the IRWLS algorithm can also be used just that also $\hat{\sigma}$ is updated each round.

General position

Recall that we are still in case where $\mathbf{X}$ is non-random. Hence the only "place" where something can happen in the moment is $\boldsymbol{\epsilon}/\mathbf{y}$.

We still have to consider however $\mathbf{X}$.

Let $k^*(\mathbf{X})$ be the maximum number of design points $\mathbf{x}_i$ lying in a lower dimensional subspace

$$k^*(\mathbf{X}) = \max\{\#(\theta'\mathbf{x}_i = 0) : \theta \neq \mathbf{0}\}.$$

Consider the case when fitting a straight line, then $k^*(\mathbf{X})$ is the number of identical design points $\mathbf{x}_i$.

If $k^*(\mathbf{X}) = p - 1$ it is said that $\mathbf{X}$ is in general position.

This means that at most p-1 points lie on a hyperplane.

BP for regression equivariant estimates

It can be shown, that for all regression equivariant estimates

$$\epsilon^* \leq \frac{m^*_{\text{max}}}{n},$$

where

$$m^*_{\text{max}} = \left[\frac{n - k^* - 1}{2}\right] \leq \left[\frac{n - p}{2}\right]$$

For designs which consist not only of zeros and ones the BP cannot obtain the maximum.

L-estimates for regression

We have seen earlier that L-statistics are appealing simple estimates in the location case. Recall that also quantiles can be seen as special cases.

It is however rather challenging to extend the concept to the regression case.
Consider for $\alpha \in (0,1)$

$$\rho_\alpha(x) = \begin{cases} \alpha x & \text{if} \quad x \geq 0 \\ -(1-\alpha)x & \text{if} \quad x < 0 \end{cases}$$

Then it is easy to show that the solution of

$$\underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\alpha(x_i - \mu)$$

gives μ as the sample α-quantile.

L-estimates for regression II

This approach can be nicely extended to the regression case defining the α-quantile regression as

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\alpha(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}).$$

This is then known as quantile regression. Naturally LAD is a special case here with $\alpha = 0.5$.

Quantile regression is considered to give a more general picture of the whole distribution as "standard" regression and is considered especially informative in the case of heteroscedastic data.

In `R` quantile regression is available in the `quantreg` package using function `rq`.

Example for quantile regression