Introduction to Robust Statistics

Klaus Nordhausen

Department of Mathematics and Statistics
University of Turku

Autumn 2015

Observational regression

So far we considered $\mathbf{x}$ as fixed. Let's change that and let it be random, this is then usally referred to as observational regression.

We make then the assumption that $\mathbf{x}$ and ϵ are independent.

For simplicity let us also assume for now, that $\mathbf{x}$ is numeric and that the distribution of $\mathbf{x}$ is not concentrated on any subspace ($P(\mathbf{a}'\mathbf{x} = 0) < 1$ for all $\mathbf{a}$).

Some results for LS for observational regression

Under the above conditions, LS is well-defined and it holds

$$E\left(\hat{\beta}|\mathbf{X}\right) = \boldsymbol{\beta} \quad \text{and} \quad \mathbf{COV}\left(\hat{\beta}|\mathbf{X}\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

where $\sigma^2 = \text{var}(\epsilon)$.

Furthermore, if ϵ is normally distributed, then the conditional distribution $\hat{\beta}$ given $\mathbf{X}$ is multivariate normal.

If ϵ is not normally distribution but $\boldsymbol{\mu}_{\mathbf{x}} = E(\mathbf{x})$ and $\mathbf{C}_{\mathbf{x}} = \mathbf{COV}(\mathbf{x})$ exist. Then

$$\hat{\boldsymbol{\beta}} \approx N\left(\boldsymbol{\beta}, \frac{\sigma^2}{n}\begin{pmatrix} 1 + \boldsymbol{\mu}_{\mathbf{x}}'\mathbf{C}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}} & \boldsymbol{\mu}_{\mathbf{x}}' \\ \boldsymbol{\mu}_{\mathbf{x}} & \mathbf{C}_{\mathbf{x}}^{-1} \end{pmatrix}\right).$$

LS and LAD for a simple data set

Let's fit LS and LAD to the following data set:

- $x_1 = 1, x_2 = 2, \ldots, x_{10} = 10, x_{11} = ??$
- $y_1 = 1, y_2 = 2, \ldots, y_{10} = 10, y_{11} = 15$

and x_{11} is then varying:

x_{11}	15	20	30	40	50	100	200	500
β_0^{LS}	0.00	1.19	2.75	3.54	4.00	4.82	5.18	5.38
β_1^{LS}	1.00	0.76	0.47	0.33	0.25	0.11	0.05	0.02
β_0^{LAD}	0.00	0.00	0.00	3.57	3.89	4.47	4.74	4.90
β_1^{LAD}	1.00	1.00	1.00	0.29	0.22	0.11	0.05	0.02

Hence high-leverage points in $\mathbf{x}$ dominate the estimate!

Generalized M-estimates

The most obvious way to limit the influence of high leverage $\mathbf{x}$ values in M-estimation is to downweight them, like for example

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right)\mathbf{x}_i w(d(\mathbf{x}_i)) = 0,$$

where w is a weight function and $d(\mathbf{x}_i)$ is some measure of largeness of $\mathbf{x}_i$.

In the model $y_i = \beta_0 + \beta_1 x_i + \epsilon$ one could choose

$$d(x_i) = \frac{|x_i - \hat{\mu}_x|}{\hat{\sigma}_x},$$

where $\hat{\mu_x}$ and $\hat{\sigma_x}$ are robust measures of location and scale. And to get a bounded influence the weight function must be such that $w(t)t$ is bounded.

Generalized M-estimates II

Alternatively if the weights are a function of the residuals and the predictors this is known as generalized M-estimation (GM-estimation) - $\boldsymbol{\beta}$ solves then

$$\sum_{i=1}^{n} \eta \left(d(\mathbf{x}_i), \frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} \right) \mathbf{x}_i = 0,$$

where $\hat{\sigma}$ is a simultaneously estimates M-estimate of scale.

The previous suggestion was $\eta(s, r) = w(s)\psi(r)$ and is called a Mallow's estimate.

The choice $\eta(s, r) = \psi(sr)/s$ is called the Hampel-Krasker-Welsch estimate where especially Huber's ψ function is popular.

(robust) Mahalanobis distance

We still need to generalize the “largeness” measure of x to the multivariate case $\mathbf{x}$.

The most popular measure is

$$d(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathbf{x}})' \hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathbf{x}}),$$

if $\hat{\boldsymbol{\mu}}_{\mathbf{x}}$ is the mean vector and $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}$ is the sample covariance matrix, this quantity is known as the squared Mahalanobis distance.

If the mean vector and the covariance matrix are replaced by robust multivariate location and scatter measures (discussed later) these are squared pseudo Mahalanobis distance.

IF of GM

Assume that ϵ is symmetric that $\hat{\boldsymbol{\mu}}_{\mathbf{x}} \to_p \boldsymbol{\mu}_{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{x}} \to_p \boldsymbol{\Sigma}_{\mathbf{x}}$.

The influence function of a GM estimate is then

$$IF((\mathbf{x}^*, y^*), GM, F) = \sigma\eta\left(d(\mathbf{x}^*), \frac{y^* - \mathbf{x}_0'\boldsymbol{\beta}}{\sigma}\right)\mathbf{B}^{-1}\mathbf{x}^*$$

where $\mathbf{B} = -E(\dot{\eta}(d(\mathbf{x}), \frac{\epsilon}{\sigma}))\mathbf{x}\mathbf{x}'$ with $\dot{\eta} = \frac{\partial\eta(s,r)}{\partial r}$.

Limiting distribution of GM

Under the same conditions as for the IF, it can be shown that an GM estimate is asymptotically normal distributed with covariance matrix

$$\mathbf{COV}(\hat{\beta}) = \sigma^2 \mathbf{B}^{-1'} \mathbf{C} \mathbf{B}^{-1},$$

where

$$\mathbf{C} = E\left(\eta\left(d(\mathbf{x}), \frac{y - \mathbf{x}'\beta}{\sigma}\right)^2 \mathbf{x}\mathbf{x}'\right).$$

Pros and Cons of GM

A GM estimate has many attractive properties:

- if $\eta(s, r)s$ is bounded, then the IF is bounded.
- it has a positive BP if $\eta(s, r)s$ is bounded.
- it is easy to compute.

However

- The efficiency depends on the distribution of $\mathbf{x}$ - if $\mathbf{x}$ is heavy tailed it cannot be both very efficient and very robust.
- BP is less than 0.5, especially for large p.
- The robust estimates for location and scatter of $\mathbf{x}$ must be affine equivariant.

Hence GM estimates are usually only recommended for small p data sets.

M-estimates with bounded ρ function

Let's return to "normal" M-estimates assuming $\mathbf{x}$ and y might contain outliers.

Our estimate to consider now is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho \left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}} \right)$$

where ρ is bounded and hence ψ is redescending. $\hat{\sigma}$ is a robust plugin scale estimate.

The estimating equations

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

have therefore multiple solutions and only one corresponds to the "good" solution.

BP of M-estimates with bounded ρ function

Write $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ as the estimate $\hat{\beta}$ is a function of $\mathbf{Z}$.

The BP is defined as $\epsilon^* = m^*/n$ with

$$m^* = \max \left\{ m \leq 0 : \hat{\beta}(\mathbf{Z}_m \text{ bounded } \forall \; \mathbf{Z}_m \in \mathcal{Z}_m) \right\},$$

where $\mathcal{Z}_m$ is the set of all data sets having $n - m$ elements in common with $\mathbf{Z}$.

It is easy to show that the BP of monotone M-estimates are then zero as large x_i dominate.

The maximum BP is however still the same as then one with fixed $\mathbf{x}$ and we'll show later how to attain it.

IF of M-estimates with bounded ρ function

Assume shortly that σ is known, then

$$IF(\mathbf{z}^*, T, F) = \frac{\sigma}{b}\psi\left(\frac{y^* - \mathbf{x}^{*\prime}\boldsymbol{\beta}}{\sigma}\right) E(\mathbf{x}\mathbf{x}')^{-1}\mathbf{x}^*$$

where $b = E(\psi'(\epsilon/\sigma))$.

So the IF is always unbounded, however the behaviour for monotone ψ's is quite different to that from redescending ones.
If σ needs to be estimated and ϵ is symmetric, then in the IF we can replace σ by the asymptotic value of $\hat{\sigma}$.

Hence this is different from GM - one cannot get a bounded IF but a better BP!

IF differences for monotone vs redescending

The difference in the IF is the following:

- For monotone ψ the IF tends for a fixed $\mathbf{x}^*$ to infinity whenever y^* tends to infinity.
- If ψ is redescending in such a way that $\psi(x) = 0$ for $|x| > k$ then the IF will tend to infinity only when $\mathbf{x}^*$ tends to infinity and at the same time $|y^* - \mathbf{x}^{*\prime}\boldsymbol{\beta}|/\sigma \leq k$.

Limiting distribution of M-estimates with bounded ρ function

Assume $\hat{\sigma} \rightarrow_p \sigma$ and that $\mathbf{x}$ has finite variance.
Then the M-estimate is consistent asymptotically normal with

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d N_p\left(v\mathbf{V}_{\mathbf{x}}^{-1}\right)$$

where $\mathbf{V}_{\mathbf{x}} = E(\mathbf{x}\mathbf{x}')$ and

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

Hence, as long as $\mathbf{x}$ has finite variance, the distribution does not depend on $\mathbf{x}$!

Summary of monotone M-estimates in the regression case

A monotone M-estimate is attracted by a high leverage point to go through that point. One the one side, if this point fits the model the fit improves, but if not the total fit fails!

Hence $\mathbf{x}$ with heavy tails are very problematic!

Strategy to obtain high BP and high efficiency

So finding **the** absolute minimum of

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right)$$

is quite challenging if p is large - but we will see that finding **some** good local minimum might already give a high BP and good efficiency at the normal model!

The strategy is then to start from a reliable starting point, apply the IRWLS algorithm. Similarly, the starting point is used to obtain a robust scale estimate $\hat{\sigma}$.

Just that for now we have no good starting point as the LAD is not a reliable option.

Schematic view

Assume for now we have a method for a good starting point, then the steps would be:

1. Compute an initial consistent estimate for $\boldsymbol{\beta}$ with high BP but possible low efficiency at the normal model.
2. Compute a robust scale measure based on the estimated residuals
3. Solve iteratively $\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) \mathbf{x}_i$ based on the initial value where ψ comes from a bounded ρ.

This way, we can achieve a high BP with a high efficiency at the normal model!

Requirements

What conditions in these steps should be met?

- The initial estimate must have all the desired equivariance properties.
- The ρ function must be bounded.
- The scale estimate must be a M-functional of the form

$$\frac{1}{n}\sum_{i=1}^{n}\rho_s\left(\frac{r_i}{\hat{\sigma}}\right) = 0.5$$

  and hence have a BP of 0.5.
- The scale must be normalized for the normal model. Usually a bisquare scale estimate is used with $c_0 = 1.56$.