Chapter 3

Location and scale M-estimation

3.1 Parametric models

3.1.1 General theory

We assume that we have a sample $\{x_1, \ldots, x_n\}$ of n i.i.d. univariate random variables X_i. A parametric model assumes that $X_i \sim F_\theta$ with $\{F_\theta : \quad \theta \in \Theta\}$ a set of distribution functions. We will assume that the distribution has a density $f_\theta(x) = F'_\theta(x)$. An important goal of statistics is to find a good estimator $\hat{\theta}_n$ of θ, that is, a real function of the sample $\hat{\theta}_n(x_1, \ldots, x_n)$ such that $\hat{\theta}_n$ is close to the true parameter θ.

Recall the following concepts:

- The Mean Squared Error (MSE) of an estimator:

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{E}[(\hat{\theta}_n - \theta)^2].$$

  The MSE is often used as a criterion to measure the accuracy of an estimator.

It can be decomposed as

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{Var}[\hat{\theta}_n] + \mathrm{bias}(\hat{\theta}_n)^2$$

with

$$\mathrm{bias}(\hat{\theta}_n) = \mathrm{E}[\hat{\theta}_n] - \theta$$

and

$$\mathrm{Var}[\hat{\theta}_n] = \mathrm{E}[(\hat{\theta}_n - E[\hat{\theta}_n])^2].$$

An estimator is called unbiased if its bias is zero. For unbiased estimators the MSE equals the variance.

- Under some regularity conditions, the Fisher information of θ in the model is defined as

$$I_\theta = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_\theta(x)\right)^2\right].$$

The Cramér-Rao information inequality states that for any unbiased estimator $\hat{\theta}_n$ we have that

$$\mathrm{Var}(\hat{\theta}_n) \geqslant \frac{1}{nI_\theta}.$$

Thus

$$0 \leqslant \mathrm{eff}(\hat{\theta}_n) := \frac{1}{n\mathrm{Var}(\hat{\theta}_n)I_\theta} \leqslant 1.$$

The percentage $\mathrm{eff}(\hat{\theta}_n)$ is called the (finite sample) efficiency of $\hat{\theta}_n$. If the efficiency equals 1, then $\hat{\theta}_n$ has the smallest possible variance (and hence the smallest possible MSE) of all unbiased estimators.

- An estimator is called consistent if for all $\theta \in \Theta$ it holds that $\hat{\theta}_n \to_p \theta$. Here $\to_p$ denotes convergence in probability, that is, for any $\varepsilon > 0$ it holds that $P(|\hat{\theta}_n - \theta| \geq \varepsilon) \to 0$.

- An estimator is called asymptotically normal with asymptotic variance V_θ if $\sqrt{n}(\hat{\theta}_n - \theta) \to_d Y$ with $Y \sim N(0, V_\theta)$ where $\to_d$ denotes convergence in distribution. (That is, the cdf of $\sqrt{n}(\hat{\theta}_n - \theta)$ converges to the cdf of $N(0, V_\theta)$ in every point.)

  This implies that V_θ is the limit of $nVar[\hat{\theta}_n]$ and not the limit of $Var[\hat{\theta}_n]$.

  If the Fisher information I_θ exists, we define the asymptotic efficiency of an unbiased and asymptotically normal estimator as

$$0 \leqslant \mathrm{eff}(\hat{\theta}_n) := \frac{1}{V_\theta I_\theta} \leqslant 1.$$

  Furthermore, the Asymptotic Relative Efficiency (ARE) of two unbiased and asymptotically normal estimators $\hat{\theta}_A$ and $\hat{\theta}_B$ is defined as

$$\mathrm{ARE}(\hat{\theta}_A, \hat{\theta}_B) = \frac{V_{\hat{\theta}_B}}{V_{\hat{\theta}_A}}$$

  which can also be written as $\mathrm{eff}(\hat{\theta}_A)/\mathrm{eff}(\hat{\theta}_B)$ when I_θ exists.

- The likelihood function is the joint density

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

  considered as a function of θ.

- The Maximum Likelihood Estimator (MLE) is the estimator maximizing the likelihood function:

$$\hat{\theta}_n = \underset{\theta}{\mathrm{argmax}}\ L(x_1, \ldots, x_n; \theta).$$

3.1.2 Univariate location model

A simple example of a parametric model is the univariate location model. It assumes that $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d.) as

$$F_\mu(x) = F(x - \mu) \tag{3.1}$$

where $-\infty < \mu < +\infty$ is the unknown location parameter and F is a continuous distribution with density f, hence $f_\mu(x) = F'_\mu(x) = f(x - \mu)$. (Note that F_μ is only a model for the uncontaminated data. We do not model outliers.)

The goal is to find a good estimator $\hat{\mu}_n$ of μ. Most location estimators satisfy the following desirable properties: the estimator is called *location equivariant* if for all $b \in \mathbb{R}$ it holds that

$$\hat{\mu}_n(x_1 + b, \ldots, x_n + b) = \hat{\mu}_n(x_1, \ldots, x_n) + b$$

and it is called *scale equivariant* if for all $a \neq 0$

$$\hat{\mu}_n(ax_1, \ldots, ax_n) = a\hat{\mu}_n(x_1, \ldots, x_n).$$

This guarantees that estimates are independent of the measurement units.

Often f is assumed to be symmetric. The simplest location model assumes that F is normal (gaussian) with mean 0 and known standard deviation σ, that is

$$X_i \sim N(\mu, \sigma^2).$$

If this assumption holds *exactly*, then the sample *mean*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is an optimal estimator: it is the MLE, it is unbiased, asymptotically normal, and has the highest possible efficiency (100%).

The sample *median* is another well known location estimator:

$$\text{Med} = \begin{cases} X_{(\frac{n+1}{2})} & \text{if} \quad n \text{ is odd} \\ \frac{1}{2}\left(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right) & \text{if} \quad n \text{ is even} \end{cases}$$

with $X_{(i)}$ the ith order statistic. Under the assumption of normality the sample median has a lower efficiency than the sample mean, namely $2/\pi \approx 64\%$. But the situation may be reversed if the model distribution F has longer tails.

3.1.3 Scale model

The scale model states that the i.i.d. random variables X_i satisfy

$$X_i \sim F_\sigma \quad \text{with} \quad F_\sigma(x) = F(\frac{x}{\sigma}) \tag{3.2}$$

where $\sigma > 0$ is the unknown scale parameter. As before F is a continuous distribution with density f, but now

$$f_\sigma(x) = F'_\sigma(x) = \frac{1}{\sigma} f(\frac{x}{\sigma}) \ .$$

Most scale estimators $\hat{\sigma}_n$ are *location invariant*, i.e. for all $b \in \mathbb{R}$ it holds that

$$\hat{\sigma}_n(x_1 + b, \dots, x_n + b) = \hat{\sigma}_n(x_1, \dots, x_n) \tag{3.3}$$

and they are typically scale equivariant, i.e. for all $a \neq 0$

$$\hat{\sigma}_n(ax_1, \dots, ax_n) = |a|\hat{\sigma}_n(x_1, \dots, x_n). \tag{3.4}$$

The most common scale estimator is the sample *standard deviation*

$$\text{Stdev}(x_1, \dots, x_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \ .$$

Other well-known scale estimators include the *Inter Quartile Range* (IQR) defined as

$$\mathrm{IQR}(x_1,\ldots,x_n) = x_{(\lceil 0.75n \rceil)} - x_{(\lfloor 0.25n \rfloor)}$$

and the *Median Absolute Deviation* (MAD):

$$\mathrm{MAD}(x_1,\ldots,x_n) = \underset{i}{\mathrm{median}}(|x_i - \mathrm{Med}(x_1,\ldots,x_n)|) \ .$$

If F is a gaussian distribution, Stdev is a consistent estimator of the parameter σ, and it is also the most efficient estimator. Note that neither the MAD nor the IQR are consistent the way they are defined above. Under normality one can prove that the MAD converges to $c\sigma$ and the IQR to $2c\sigma$ with $c = \Phi^{-1}(0.75) = 0.6745$. We typically use the normalized MAD given by

$$\mathrm{MADN}(x_1,\ldots,x_n) = \frac{\mathrm{MAD}(x_1,\ldots,x_n)}{\Phi^{-1}(0.75)}.$$

The factor $(\Phi^{-1}(0.75))^{-1} = 1.4826$ is called a consistency factor. Some people prefer the notation MAD to include this factor, others do not. When using software one should watch out which convention is used. In R, *mad* refers to the version with the factor included.

3.2 The sensitivity curve

Let $\{F_\theta\}$ be a parametric model. For a sample $(x_1, \ldots, x_n)$ and an estimator $\hat{\theta}_n$ the *sensitivity curve* in the point z is defined by

$$\mathrm{SC}(z) = (n+1)(\hat{\theta}_{n+1}(x_1, \ldots, x_n, z) - \hat{\theta}_n(x_1, \ldots, x_n)). \tag{3.5}$$

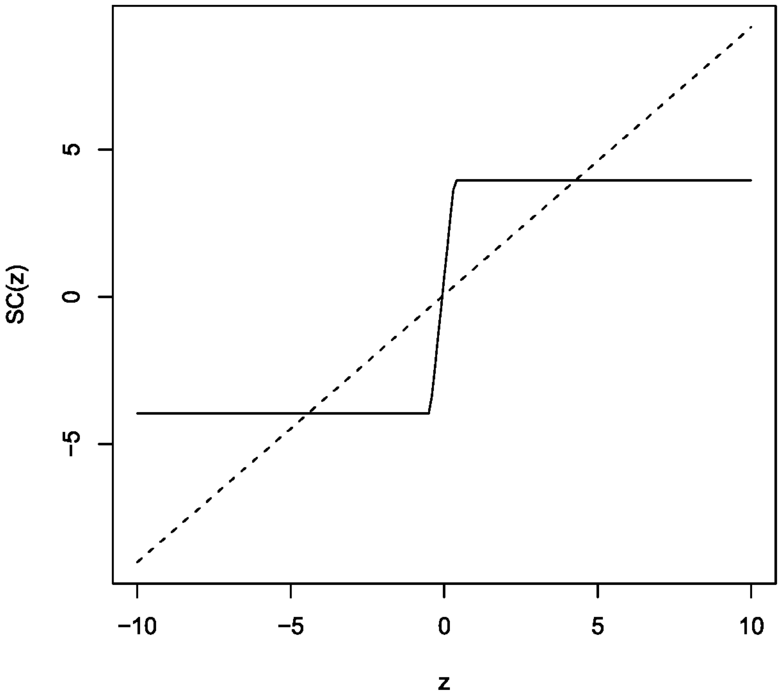
Consider for example the location model. The following function computes the sensitivity curve of the sample mean for a given sample on a grid between -10 and 10.

```
scmean <- function(sample) {
  n <- length(sample)
  sc <- rep(0,201) # "repeat": array with 201 elements, all zero
  for (z in -100:100){
        sc[z+101] <- n*(mean(c(sample,z/10))-mean(sample))
  }
  sc   # the output
}
```

A similar function can be written for the sample median. Both functions can be drawn as follows.

```
> sample <- rnorm(10)   # randomly generated from N(0,1)
> res <- scmean(sample)
> plot((-100:100)/10,res,xlab="z",ylab="SC(z)")
> resmed <- scmedian(sample)
> points((-100:100)/10,resmed)
```

The sample mean is clearly more sensitive to outliers than the sample median. Its sensitivity curve is unbounded as $|z|$ increases. The sensitivity curve of the sample median is bounded, reflecting the fact that the addition of a single point z to a sample cannot have an arbitrarily large impact.



3.3 The finite-sample breakdown value

If the sensitivity curve is bounded, then one observation cannot have an arbitrarily large impact on the estimator. But what happens when more than one point is altered? To measure this, the finite-sample breakdown value [11] plays an important role.

In a given sample $\mathbf{x} = (x_1, \ldots, x_n)$, replace any m data points $x_{i_1}, \ldots, x_{i_m}$ by arbitrary values $y_1, \ldots, y_m$. Call the new data set $(z_1, \ldots, z_n)$. Let Θ be the parameter space of the parametric model. The *finite-sample breakdown value* of an estimator $\hat{\theta}_n$ is

$$\varepsilon_n^*(\hat{\theta}_n; \mathbf{x}) = \min\left\{\frac{m}{n} : \hat{\theta}_n(z_1, \ldots, z_n) \text{ approaches the boundary of } \Theta\right\}. \tag{3.6}$$

In most cases ε_n^* does not depend on the sample $\mathbf{x}$ (for samples satisfying some weak conditions).

Consider first the univariate location model. Then $\Theta = \mathbb{R}$ and

$$\varepsilon_n^*(\hat{\mu}_n; \mathbf{x}) = \min\left\{\frac{m}{n} : \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} |\hat{\mu}_n(z_1, \ldots, z_n)| = \infty\right\}. \tag{3.7}$$

In this situation:

- The breakdown value of the sample mean equals $1/n$.
- For any location equivariant location estimator $\hat{\mu}_n$ it holds that
$$\varepsilon_n^*(\hat{\mu}_n, \mathbf{x}) \leqslant \frac{1}{n}\left\lfloor\frac{n+1}{2}\right\rfloor \approx \frac{1}{2}. \tag{3.8}$$
- The sample median attains this upper bound.

In the univariate scale model we have that $\Theta =]0, \infty[$. The boundary of Θ is the pair $\{0, +\infty\}$. The breakdown value with respect to 0 is called the *implosion breakdown* *value*. The one with respect to $+\infty$ is called the *explosion breakdown value*. In this setting:

- The breakdown value of the sample standard deviation equals $1/n$ because it explodes easily.
- For any equivariant scale estimator $\hat{\sigma}_n$ it holds that

$$\varepsilon_n^*(\hat{\sigma}_n, \mathbf{x}) \leqslant \frac{\lfloor n/2 \rfloor}{n} \approx \frac{1}{2}.$$

- The MAD attains this upper bound at samples without ties.