Statistics 540, Robust Estimation

Robust Estimation

Problem ...

The usual introductory setting for robust estimation is the idealized location problem

$$y_i = \mu + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} F, \quad i = 1, \ldots, n,$$

where F denotes a distribution of errors which is typically assumed to be symmetric about zero ($P(X \leq x) = P(X > -x)$, or $F(x) = 1 - F(-x)$). This symmetry forces various estimators like the mean and median to have a common target. Both the mean $E\ y_i$ (assuming it exists) and median are equal to μ. Let $\sigma^2 = \text{Var}(y_i) = \text{Var}(\epsilon_i)$ denote the common variance which may be infinite. The goal is to estimate μ from the n observations y_i, with an emphasis on non-gaussian errors.

Problems with the sample average ...

Although the sample average $\overline{Y} = \sum Y_i/n$ has a variety of optimality properties, it fares poorly (in the sense of high sampling variance) when the data are prone to outliers. For example, suppose that F is the Cauchy distribution with density

$$f(x) = \frac{1}{\pi(1 - x^2)},$$

which is also Student's t with one degree of freedom. The expectation $E\ Y$ does not exist for the Cauchy, but one can still compute $\overline{Y}$ from data. It has infinite variance. In contrast, the population median $F^{-1}(1/2)$ is well-defined, and the sample median Y_{med} has asymptotic variance (let n be even)

$$\begin{aligned} \text{Var}(Y_{\text{med}}) &= \text{Var}(\hat{F}^{-1}(\frac{1}{2})) \\ &= \text{Var}(F^{-1}(U_{(n/2)})) \qquad 0 < U_{(1)} \leq \cdots \leq U_{(n)} < 1 \\ &\to \frac{1}{f^2(\mu)} \frac{1}{4n} \end{aligned} \tag{1}$$

Contaminated distributions ...

OK, so the Cauchy is a bit extreme. A more reasonable model, perhaps, for outliers in data is the contamination model in which one observes

$$Y_i \sim (1 - p)N(\mu, \sigma^2) + pN(\mu, c\sigma^2), \quad 0 \leq p \leq 1,$$

and the contamination constants p and c are on the order of $p = 0.05$ and $c = 10$ (*i.e.*, 5% contamination with 10 times larger variance). In this case, its easy to see that

$$\text{Var}(\overline{Y}) = \sigma^2 \frac{(1 - p) + p\ c}{n}\ .$$

The expression (1) for the variance of the median still holds, but with $f(\mu)$ determined from the mixture of normals. As $c \to \infty$ for some $p > 0$, $\text{Var}(\overline{Y}) \to \infty$ whereas $\text{Var}(Y_{\text{med}})$ remains finite.

Task ...

Construct a plot showing the values of p and c where $\mathrm{Var}(\overline{Y}) = \mathrm{Var}(Y_{\mathrm{med}})$.

Trade-offs and the trimmed mean ...

So why not use the median all of the time. Simple: From (1), the sample median pretty inefficient compared to the sample average when the data *are* Gaussian (100 $2/\pi \approx 64\%$ efficient). Robust statistic seek a balance which behaves like $\overline{Y}$ for 'good' data, but downweights outliers. For example, a trimmed mean is the average of an inner fraction of the observed data,

$$\overline{Y}_\alpha = \sum_{i=k}^{n-k+1} Y_{(i)}/(n-2k)$$

where $k = \alpha\ n$ (rounded to an integer). That is, $\overline{Y}_\alpha$ is the average after 'trimming' $100\alpha\%$ from each tail of the data. The trimmed mean is like a mean in the center of the data, but like a median at the extremes (the extremes are only used to determine which observations are at the center of the data).

Influence functions ...

A nice way to think about an estimate of location $\hat{\mu}$ is as the solution of a scoring equation (estimating equation)

$$\sum \psi(Y_i - \hat{\mu}) = 0\ , \tag{2}$$

where ψ denotes a scoring function known as the *influence function*. If indeed $\rho = \partial \log f/\partial\mu$ is *the* score function, we obtain the usual MLE. To obtain $\overline{Y}$, we have $\rho(x) = x$; to get Y_{med}, use $\rho(x) = \mathrm{signum}(x)$. The influence function of the trimmed mean $\overline{Y}_\alpha$ is a combination of these two.

Having seen these influence functions, one immediate thinks of making them "smoother". Tukey went even further, and introduced a class of so-called redescending ψ functions that don't even count outliers at all, but rather eliminate them entirely from consideration. The most famous of these is known as the *biweight* influence function

$$\psi_b = \begin{cases} (1 - (\frac{x}{s})^2)^2, & |x| < s, \\ 0 & \text{otherwise}\ . \end{cases}$$

The scaling term is quite important. As a practical matter, in applications, $s = 7 \times (\mathit{medianabsolutedeviation})$ gives an estimator which is 95% efficient when the data is normal.

Computing the robust estimator ...

When you've got a closed form expression, there's no problem. The biweight, however, is defined implicitly from (2). This leads to an application of Newton's method. But beware, since ψ_b is redescending, the biweight estimator is not uniquely defined.

Alternatively, and very cleverly, you can make this problem a lot easier with a simple insight. Namely, write $\psi(x) = xW(x)$ so that (2) becomes

$$\sum \psi(Y_i - \hat{\mu}) = \sum (Y_i - \hat{\mu}) W(Y_i - \hat{\mu}) = \sum w_i (Y_i - \hat{\mu}) \; . \tag{3}$$

Given the weights w_i, this expression defines $\hat{\mu}$ as simply a weighted average of the responses. However, we need $\hat{\mu}$ to obtain the weights. What do we do? Iterate. Use some initial estimator $\hat{\mu}_0$ to obtain weights w_i^0 and compute the weighted mean $\hat{\mu}_1$. Start over, and continue the iterations. This popular algorithm is known as *iteratively reweighted least squares* (IRLS).

Variance estimates ...

What should be used for estimating $\mathrm{Var}(\hat{\mu})$? It is a weighted mean, and we have expressions for the variance of these, but are they appropriate here?

References ...

Tierney offers some additional description of a robust regression in Section §5.6.2 (page 173). A nice overview of the area motivated with some examples is Hoaglin, Mosteller, and Tukey (*Understanding Robust and Exploratory Data Analysis*, 1983, Wiley) and a more technical treatment (with a great introduction) is Hampel, Ronchetti, Rousseeuw, and Stahel (*Robust Statistics*, 1986, Wiley).