9. Robust regression

Least squares regression

$$\hat{\theta} = \text{argmin}_\theta \sum_{i=1}^{n}(y_i - x_i^T\theta)^2 = \text{argmin}_\theta \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \text{argmin}_\theta \sum_{i=1}^{n}\hat{\epsilon}_i^2$$

Why least squares regression?
- Historic (used since 1800)
- The least squares estimator $\hat{\theta} = (X^TX)^{-1}X^Ty$ has a closed form solution, and is simple to compute
- If $y = X\theta + \epsilon$ and $\epsilon \sim N_n(0, \sigma^2 I)$:
  - ◆ Least squares estimator $=$ MLE
  - ◆ Least squares estimator has smallest variance among all unbiased estimators (Gauss-Markov)

Sensitivity to outliers

- Least squares estimators are sensitive to
  - ◆ Outliers in the y-direction (see picture on board)
  - ◆ Outliers in the x-direction (see picture on board)
- Possible solutions:
  - ◆ Good diagnostics (Cook's distance) - use least squares regression with and without the influential points
  - ◆ Robust regression - move away from least squares regression

---

Robust regression - methods

- L_1 regression (=Least Absolute Deviations (LAD) regr.)
- Huber regression
- Mallows regression
- Schweppe regression
- Least Median of Squares (LMS) regression

L_1 **regression**

$$\hat{\theta} = \mathsf{argmin}_\theta \sum_{i=1}^{n} |y_i - x_i^T \theta|$$

- Older than LS: Boscovich (1760), Laplace (1789)
- Did not become popular, since the solution cannot be written in closed form (no problem anymore with modern computers)
- In location model: L_1 regression gives median of the data
- Is more robust against outliers in the y-direction, but still very sensitive to outliers in the x-direction
- Is very inefficient when the errors are normally distributed

Huber regression

$$\hat{\theta} = \mathsf{argmin}_\theta \sum_{i=1}^{n} \rho_c(y_i - x_i^T \theta),$$

where

$$\rho_c(u) = \begin{cases} u^2/2 & \text{if } |u| \le c \\ c(|u| - c/2) & \text{if } |u| \ge c \end{cases}$$

- Compromise between L_1 and L_2 regression:
  - ◆ $c = \infty \Rightarrow L_2$ regression (=least squares)
  - ◆ $c = 0 \Rightarrow L_1$ regression (use $\rho_c(u) = |u|$)
- Idea: penalize small residuals quadratically, and large residuals linearly
- Computation of the estimator: see board.
- c should be chosen suitably with respect to residuals.
- Estimator is computed in R using iterated weighted least squares.

Mallows/Schweppe regression

$$\sum_{i=1}^{n} \eta\left(x_i, \frac{y_i - x_i^T \hat{\theta}}{\hat{\sigma}}\right) x_i = 0$$

■ Mallows:

$$\eta(x, r) = \min\left(1, \frac{a}{\|Ax\|}\right) \psi_c(r)$$

■ Schweppe:

$$\eta(x, r) = \frac{1}{\|Ax\|} \psi_c(\|Ax\| r)$$

■ $\|Ax\|$ is a measure of leverage of x, for example $\|Ax\|^2 = a x^T (X^T X)^{-1} x$, but then robust version
■ $\psi_c = \rho'(c)$, from Huber regression

Breakpoint

The breakpoint of an estimator = the maximum proportion of observations that can be changed without changing the estimator
■ Breakpoint of average: 0
■ Breakpoint of median: 1/2
■ Breakpoint of Least Squares regression: 0
■ Breakpoint of L_1 and Huber: 0 (in x-direction)
■ Breakpoint Mallows/Schweppe: $\leq 1/p$

LMS regression

$$\hat{\theta} = \text{argmin}_\theta \text{median}((y_i - x_i^T \theta)^2)$$

■ See picture on slide
■ Hampel (1975), Rousseeuw (1984)
■ Breakpoint is approximately 0.5
■ Difficult to compute because of many local minima
■ Inefficient when statistical errors are normally distributed (convergence rate $n^{-1/3}$! This can be improved by replacing the median by an α-truncated mean

Some closing thoughts (see Faraway Ch 13)

- Robust estimators protects against long-tailed errors, but not against problems with model choice and variance structure. These latter problems are often more serious than non-normal errors
- Robust estimators give $\hat{\theta}$ and possibly standard errors without inferential methods. One can use bootstrap for inference.
- Robust methods can be used in addition to least squares. There is cause to worry if the two estimators differ a lot.

10 / 10