Robust Regression

University of California, San Diego
Instructor: Armin Schwartzman

Alcohol and tobacco expenses in Great Britain

- The `alcoholtobacco` dataset contains the average weekly household spending, in British pounds, on tobacco products and alcoholic beverages for each of the 11 regions of Great Britain (1981).
- We fit a simple linear model using least squares

$$\text{Tobacco} = \beta_0 + \beta_1 \text{ Alcohol} + \text{Error}$$

  The t-test for $\beta_1 = 0$ is not significant.
- We now fit the model without `11:Northern.Ireland`, and the t-test for $\beta_1 = 0$ is highly significant.
- Observation `11:Northern.Ireland` is therefore influential.

  (The other observations do not have nearly the same impact.)

Least squares regression

- We have data $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, to which we want to fit a linear model:

$$\mathbb{E}(y|\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$$

- Fitting this model by least squares amounts to solving:

$$\min_{\mathbf{b}} \sum_{i=1}^{n} e_i(\mathbf{b})^2$$

  where

$$e_i(\mathbf{b}) = y_i - \mathbf{b}^\top \mathbf{x}_i$$

  is the ith residual for the coefficient vector $\mathbf{b}$.
- If the errors are i.i.d. normal with mean zero, i.e., with density

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\varepsilon^2/(2\sigma^2)}$$

  the least squares method coincides with maximum likelihood estimation.
- *What if the errors have a different distribution?*

Least absolute regression

- Suppose the errors are i.i.d. double-exponential with mean zero:

$$f(\varepsilon) = \frac{\lambda}{2} e^{-\lambda|\varepsilon|}$$

- The MLE corresponds to solving

$$\min_{\mathbf{b}} \sum_{i=1}^{n} |e_i(\mathbf{b})|$$

  This method is called least absolute regression (LAR).

  (In general, there is no closed form expression for the estimates.)

- LAR (also called L_1 regression) is more robust to outliers (in response) than OLS (also called L_2 regression), simply because the exponential distribution has heavier tails than the normal distribution.

M-estimation

- Suppose the errors have density

$$f(\varepsilon) \propto e^{-\rho(\varepsilon/s)}$$

  $\rho(t) \geq 0$ for all $t \in \mathbb{R}$, and $s > 0$ is a scale parameter.

- The MLE corresponds to solving

$$\min_{\mathbf{b}} \sum_{i=1}^{n} \rho\left(\frac{e_i(\mathbf{b})}{s}\right)$$

- Assuming ρ is differentiable with derivative $\rho' = \psi$, looking for critical values amounts to solving

$$\sum_{i=1}^{n} \psi\left(\frac{e_i(\mathbf{b})}{s}\right) \mathbf{x}_i = \frac{1}{s} \sum_{i=1}^{n} w\left(\frac{e_i(\mathbf{b})}{s}\right) e_i(\mathbf{b}) \mathbf{x}_i = 0$$

  where $w(r) = \psi(r)/r$ is the weight function.

- Examples include Huber's ψ, Hampel's ψ, and Tulkey's ψ (biweight).

Examples

- OLS ρ is:

$$\rho(r) = r^2$$

  with weights $w(r) = 2$ (constant).

- Huber's ρ is:

$$\rho(r) = \begin{cases} r^2 & \text{if } |r| \leq c \\ 2cr - c^2 & \text{otherwise} \end{cases}$$

  with $c = 1.345$ optimizing the asymptotic relative efficiency.

- Tukey's ρ (aka bisquare or biweight) is:

$$\rho(r) = \begin{cases} \left(1 - \left(1 - \frac{|r|}{c}\right)^2\right)^3 & \text{if } |r| \leq c \\ 1 & \text{otherwise} \end{cases}$$

  with $c = 4.685$ optimizing the asymptotic relative efficiency.

The Newton-Raphson method

- Suppose that $F : \mathbb{R}^p \to \mathbb{R}^p$ is differentiable and we want to find a root of F in some domain $\Omega \subset \mathbb{R}^p$, meaning,

$$\text{find } \mathbf{b} \in \Omega \text{ such that } F(\mathbf{b}) = 0$$

  For us,

$$F(\mathbf{b}) = \sum_{i=1}^{n} \psi\left(\frac{e_i(\mathbf{b})}{s}\right) \mathbf{x}_i$$

- The Newton (or Newton-Raphson) method is an iterative method that consists in linearizing F at each step:

  1. Initialize at some $\mathbf{b}_0 \in \Omega$.
  2. Iterate until convergence: solve for b_{t+1} in the following the linear system

$$J_F(\mathbf{b}_t)(\mathbf{b}_{t+1} - \mathbf{b}_t) = -F(\mathbf{b}_t)$$

     where J_F is the Jacobian matrix of F

- The rationale is based on a Taylor expansion

$$F(\mathbf{b}_{t+1}) \approx F(\mathbf{b}_t) + J_F(\mathbf{b}_t)(\mathbf{b}_{t+1} - \mathbf{b}_t)$$

  Pretending that F is linear, and given $\mathbf{b}_t$, we set the right-hand side to zero and solve for $\mathbf{b}_{t+1}$.

- Note that the method is also known as iteratively reweighted least-squares (IRLS) in statistics.

Estimating the scale parameter

- In general, we need to estimate the scale parameter s.

  (This is not the case of OLS or LAR.)

- A popular estimate for s is the median absolute deviation (MAD) of some residuals:

$$\mathrm{MAD}(e_1, \ldots, e_n) = \mathrm{Median}(|e_1 - m|, \ldots, |e_n - m|)$$

  where $m = \mathrm{Median}(e_1, \ldots, e_n)$. We then choose

$$\widehat{s} = c\ \mathrm{MAD}(e_1, \ldots, e_n), \quad c = 1.4826$$

- Another option is

$$\widehat{s} = c\ \mathrm{Median}(|e_1|, \ldots, |e_n|),$$

  where the constant c may be different.

- The residuals may come from LS, or a more robust method like LAR or LMS (see below).

Break-down point

- The break-down point is the smallest fraction of anomalous data that can cause the estimator to move towards infinity without bound.

  A high break down point may be desirable.

- The smallest break-down point is $1/n$, that of all the methods seen so far.

Least median of squares (LMS)

- This method has the highest possible break-down point (50%).

- It consists in solving

$$\min_{\mathbf{b}}\ \mathrm{Median}(e_1(\mathbf{b})^2, \ldots, e_n(\mathbf{b})^2)$$

  (This optimization problem is not convex.)

Least trimmed sum of squares (LTS)

- Also has a high break-down point (50% for $h = n/2$ below).

- It consists in solving

$$\min_{\mathbf{b}} \sum_{i=1}^{h} e_{(i)}(\mathbf{b})^2$$

  where $|e_{(1)}(\mathbf{b})| \leq \cdots \leq |e_{(n)}(\mathbf{b})|$ are the ordered residuals.

  (This optimization problem is not convex.)

Efficiency

- Note that the estimator that returns $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ regardless of the data is very robust, but is completely useless, so robustness needs to be combined with a measure of how well this estimator does.
- Robust estimators put less weight on outliers, so effectively they use a smaller sample size than OLS, which uses the entire sample.
- Asymptotic relative efficiency (ARE) of an estimator is the sample size ratio required to obtain the same variance as OLS.
- For example, the ARE of LMS is 0.64. This means that its variance is $1/0.64 = 1.57$ times larger than OLS. It needs 57% more samples to achieve the same variance.
- ARE is asymptotic, so these claims are approximate for large sample sizes.

Efficiency

- The procedures above gain robustness at the cost of efficiency. The goal is to design a robust method (high break-down point) that performs well in the classical setting (high efficiency).
  - Typically, M-estimators have high efficiency but low breakdown point.
  - LMS and LTS have low efficiency and high breakdown point.

Inference for robust regression

- M-estimators are asymptotically normal under some conditions (satisfied here) and so is the LTS estimator. Various estimates for the asymptotic covariance matrix are available.
- In practice, it might be safer to use a computer-intensive method, e.g., the bootstrap, to perform inference, e.g., build confidence intervals for the coefficients.

  (Note however that the bootstrap can fail in high dimensions. This is an ongoing research area.)