

# Enveloped Huber Regression

刘盛

2025 年 12 月 21 日

UESTC

## Enveloped Huber Regression

Zhou, Le, R. Dennis Cook, and Hui Zou. "Enveloped huber regression." Journal of the American Statistical Association 119.548 (2024): 2722-2732.

# Classical Linear Regression

We observe data pairs  $\{(x_i, y_i)\}_{i=1}^n$ , where

- $x_i \in \mathbb{R}^p$  is a vector of predictors,
- $y_i \in \mathbb{R}$  is a response variable.

The classical linear regression model assumes

$$y_i = \beta_0 + x_i^\top \beta + \varepsilon_i,$$

where  $\varepsilon_i$  is a random error term.

Ordinary Least Squares (OLS): OLS estimates  $\beta$  by minimizing the squared loss:

$$\min_{\beta_0, \beta} \sum_{i=1}^n (y_i - \beta_0 - x_i^\top \beta)^2.$$

## Limitations of OLS

- Highly sensitive to outliers and extreme observations.
- Efficiency relies on light-tailed (e.g., Gaussian) errors.
- Performance degrades under heavy-tailed or contaminated data.

# Huber Regression: A Robust Alternative

Real-world data often contain

- outliers,
- heavy-tailed noise,
- deviations from Gaussian assumptions.

Huber Loss Function

Huber regression replaces the squared loss with a robust loss:

$$\rho_k(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq k, \\ k|r| - \frac{1}{2}k^2, & |r| > k, \end{cases}$$

where  $r = y_i - \beta_0 - x_i^\top \beta$ .

Key Properties

- Quadratic for small errors  $\Rightarrow$  high efficiency.
- Linear for large errors  $\Rightarrow$  robustness to outliers.
- Interpolates between least squares and absolute deviation regression.

# Huber Influence Function

From optimization to estimation

Huber regression is an M-estimator characterized by the first-order condition

$$\sum_{i=1}^n \psi_k(r_i)(1, x_i^\top)^\top = 0,$$

where  $\psi_k(r) = \frac{d}{dr} \rho_k(r)$ .

Huber influence function

$$\psi_k(r) = \begin{cases} r, & |r| \leq k, \\ k \operatorname{sign}(r), & |r| > k. \end{cases}$$

Key insight

- Influence function determines how each observation affects estimation.
- Bounded influence  $\Rightarrow$  robustness to outliers.
- Robustness is achieved by controlling influence, not by removing data.

# Conditional Mean for Huber Regression

Assume the robust conditional center is linear in predictors:

$$E_\rho[y | x] = \mu^* + x^\top \beta^*.$$

- This is the linear Huber regression model.
- Note: we are modeling a robust center, not necessarily the usual mean.

Let  $\varepsilon = y - \mu^* - x^\top \beta^*$ . Then the model can be written as

$$y = \mu^* + x^\top \beta^* + \varepsilon \quad \text{with} \quad E[\psi_k(\varepsilon) | x] = 0,$$

where  $\psi_k(\cdot) = \rho'_k(\cdot)$  is the Huber influence function.

a weak (robust) condition

- We do not require  $E[\varepsilon | x] = 0$  or Gaussian errors.
- We only require the Huber score to be centered:  $E[\psi_k(\varepsilon) | x] = 0$ .
- Conditional heteroscedasticity (error depending on  $x$ ) is allowed.

# Conditional Mean for Huber Regression

For a given  $x$ , the usual conditional mean can be characterized by

$$E[y | x] = \arg \min_{u \in \mathbb{R}} E[(y - u)^2 | x].$$

- Interpretation: choose a single number  $u$  that best represents  $y$  given  $x$ , measured by squared error.
- Issue: squared loss puts huge weight on large deviations (outliers / heavy tails).

Replace squared loss by the Huber loss  $\rho_k(\cdot)$  and define the Huber  $\rho$ -mean:

$$E_\rho[y | x] = \arg \min_{u \in \mathbb{R}} E[\rho_k(y - u) | x].$$

- Same idea: pick a representative  $u$ .
- Different criterion: robust loss  $\rho_k$  reduces sensitivity to extremes.

# Motivation of Enveloped Huber Regression

- The predictor vector  $x \in \mathbb{R}^p$  may be high-dimensional.
- Many predictor directions may be irrelevant for explaining  $y$ .
- Irrelevant directions inflate variance and reduce efficiency.

Robust conditional center

$$E_\rho[y \mid x] = \arg \min_{u \in \mathbb{R}} E[\rho(y - u) \mid x]$$

- A robust alternative to the usual conditional mean.
- Less sensitive to extreme observations.

Linear Huber regression assumption

$$E_\rho[y \mid x] = \mu^* + x^\top \beta^*$$

- Same linear form as classical regression.
- Different notion of “center”.

# Not All Predictor Information Matters

Central idea

The robust conditional center  $E_\rho[y | x]$  may depend on  $x$  only through certain linear combinations.

Important clarification

- This is not variable selection.
- It is selection of directions (linear subspaces).

We seek to identify and retain only the material predictor directions.

Subspace decomposition

Let  $S \subset \mathbb{R}^p$  be a subspace. Decompose:

$$x = P_S x + Q_S x$$

- $P_S x$ : projection onto  $S$  (material part)
- $Q_S x$ : projection onto  $S^\perp$  (immaterial part)

Interpretation

- $P_S x$  contains information relevant for predicting  $y$ .
- $Q_S x$  contains noise that does not affect the robust center.

# Separation of Predictor Variation

Assumption

$$\text{cov}(P_S x, Q_S x) = 0$$

- Ensures material and immaterial parts are statistically separable.
- Prevents loss of useful information when discarding  $Q_S x$ .
- Standard structural assumption in envelope methodology.

Noise should not contaminate signal.

Condition (b): Dependence Only on Material Part Key modeling assumption

$$E_\rho[y \mid x] = E_\rho[y \mid P_S x]$$

Interpretation

- Once  $P_S x$  is known,  $Q_S x$  adds no information.
- The immaterial part does not affect the robust conditional center.

This defines what we mean by “irrelevant” predictor directions.

# Equivalent Characterization

$$E_\rho[y \mid x] = E_\rho[y \mid P_S x] \iff \beta^* \in S$$

Geometric intuition

- $\beta^*$  defines the signal direction.
- If  $\beta^* \in S$ , orthogonal directions cannot affect  $x^\top \beta^*$ .
- Hence only  $P_S x$  matters.

Multiple valid subspaces

$$\mathcal{E}_{\Sigma_x^*}(\beta^*) = \bigcap \{S : S \text{ satisfies (a) and } \beta^* \in S\}$$

- Smallest subspace containing all relevant information.
- Dimension  $u \leq p$ .

Let  $\Gamma \in \mathbb{R}^{p \times u}$  span the envelope, and  $\Gamma_0$  span its orthogonal complement.

Model structure

$$\beta^* = \Gamma \eta, \Sigma_x^* = \Gamma \Omega \Gamma^\top + \Gamma_0 \Omega_0 \Gamma_0^\top$$

- Regression signal lies in a  $u$ -dimensional subspace.
- Predictor variation is decomposed accordingly.

# Enveloped Huber Regression Model

Complete model

$$\begin{aligned}y_i &= \mu^* + x_i^\top \beta^* + \varepsilon_i, \\E[\psi(\varepsilon_i) | x_i] &= 0, \\\beta^* &= \Gamma\eta, \\\Sigma_x^* &= \Gamma\Omega\Gamma^\top + \Gamma_0\Omega_0\Gamma_0^\top.\end{aligned}$$

Key takeaway

EHR = Huber regression + envelope structure

If  $\Gamma$  were known, estimation reduces to:

$$\min_{\mu, \eta} \sum_{i=1}^n \rho(y_i - \mu - \eta^\top \Gamma^\top x_i)$$

Consequences

- Estimation in  $u$  dimensions instead of  $p$ .
- Less noise from irrelevant directions.
- Smaller asymptotic variance.

# Turn the Model into Moment Conditions

Given data  $\{(x_i, y_i)\}_{i=1}^n$ , estimate the unknown parameters in the EHR model:

$$y_i = \mu^* + x_i^\top \beta^* + \varepsilon_i, \quad E[\psi(\varepsilon_i) | x_i] = 0, \quad \beta^* = \Gamma\eta.$$

Parameter:

$$\theta^* = (\mu^*, \eta^*, \Gamma^*, \Omega^*, \Omega_0^*, \mu_x^*), \quad \Gamma^{*\top} \Gamma^* = I_u.$$

Huber regression implies a robust moment condition

Let  $r_i(\mu, \eta, \Gamma) = y_i - \mu - \eta^\top \Gamma^\top x_i$ . The model condition  $E[\psi(\varepsilon) | x] = 0$  yields unconditional moments:

$$E\left[\psi(r_i) \begin{pmatrix} 1 \\ x_i \end{pmatrix}\right] = 0.$$

Envelope structure also constrains predictor moments

With  $\Sigma_x = \Gamma\Omega\Gamma^\top + \Gamma_0\Omega_0\Gamma_0^\top$ , the mean and covariance of  $x$  satisfy:

$$E[x] = \mu_x, \quad E[(x - \mu_x)(x - \mu_x)^\top] = \Sigma_x.$$

Collect all moments into one vector:

$$E[g(Z_i, \theta)] = 0, \quad Z_i = (y_i, x_i).$$

# Moment Conditions in Enveloped Huber Regression

The estimating equations used in EHR can be written as

$$G_n(\theta) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \psi(y_i - \mu - x_i^\top \beta) \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\ \text{vech}(\Sigma_x) - \text{vech}\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(x_i - \mu_x)^\top\right) \\ \mu_x - \bar{x} \end{pmatrix} = 0. \quad (3.1)$$

Equivalent representation

Let  $z_i = (y_i, x_i^\top)^\top$ . Then

$$G_n(\theta) = \frac{1}{n} \sum_{i=1}^n g(z_i; \theta),$$

where

$$g(z_i; \theta) = \begin{pmatrix} g_1(z_i; \theta) \\ g_2(z_i; \theta) \\ g_3(z_i; \theta) \end{pmatrix},$$

with

$$g_1(z_i; \theta) = \psi(y_i - \mu - x_i^\top \beta) \begin{pmatrix} 1 \\ x_i \end{pmatrix}, \text{ robust regression moments}$$

$$g_2(z_i; \theta) = \text{vech}(\Sigma_x) - \text{vech}\left((x_i - \mu_x)(x_i - \mu_x)^\top\right), \text{ covariance moments for predictors}$$

$$g_3(z_i; \theta) = \mu_x - x_i \cdot \text{mean moments for predictors}$$

# Sample Moments and the GMM Objective

Sample moment vector

$$G_n(\theta) = \frac{1}{n} \sum_{i=1}^n g(Z_i, \theta).$$

if there exists  $\theta$  with  $G_n(\theta) = 0$ , we could solve the equations exactly.

typically the system is over-identified or has no exact solution, so we choose  $\theta$  that makes  $G_n(\theta)$  as small as possible.

GMM optimization objective

$$\hat{\theta} = \arg \min_{\theta \in \Theta_{\text{env}}} Q_n(\theta), \quad Q_n(\theta) = G_n(\theta)^\top W_n G_n(\theta)$$

where  $\Theta_{\text{env}}$  encodes the envelope constraints (e.g.,  $\beta = \Gamma\eta$ ).

$G_n(\theta)$  is a vector. If we cannot make it exactly zero, we minimize its weighted squared length:

$$\|G_n(\theta)\|_{W_n}^2 = G_n(\theta)^\top W_n G_n(\theta).$$

- Different moment conditions have different scales and noise levels.
- $W_n$  downweights noisy moments and upweights reliable ones.

Special case:  $W_n = I$  gives an unweighted least-squares fit of moments.

# What Is “The Solution” in Practice? (Outputs)

The optimization returns estimated parameters

$$\hat{\theta} = (\hat{\mu}, \hat{\eta}, \hat{\Gamma}, \hat{\Omega}, \hat{\Omega}_0, \hat{\mu}_x).$$

From these we report the primary quantities of interest:

- Regression coefficient:

$$\hat{\beta} = \hat{\Gamma} \hat{\eta}.$$

- Estimated envelope subspace:

$$\hat{\mathcal{E}} = \text{span}(\hat{\Gamma}).$$

(Only the subspace is identifiable, not the specific basis.)

- Predictor covariance under the envelope:

$$\hat{\Sigma}_x = \hat{\Gamma} \hat{\Omega} \hat{\Gamma}^\top + \hat{\Gamma}_0 \hat{\Omega}_0 \hat{\Gamma}_0^\top.$$

# Reparameterization: From $\Gamma$ to an Unconstrained Matrix $A$

Why optimizing over  $\Gamma$  is tricky

- Constraint:  $\Gamma^\top \Gamma = I_u$  (semi-orthogonal).
- Non-identifiability:  $\Gamma Q$  spans the same subspace for any orthogonal  $Q$ .
- Therefore the real target is the point on the Grassmann manifold (the set of  $u$ -dimensional subspaces in  $\mathbb{R}^p$ ).

Partition  $\Gamma$  as

$$\Gamma = \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \end{pmatrix}, \quad \Gamma_1 \in \mathbb{R}^{u \times u}.$$

Assume  $\Gamma_1$  is invertible and define

$$A = \Gamma_2 \Gamma_1^{-1} \in \mathbb{R}^{(p-u) \times u}.$$

Then

$$\Gamma = \begin{pmatrix} I \\ A \end{pmatrix} \Gamma_1, \quad \Rightarrow \quad \text{span}(\Gamma) = \text{span} \begin{pmatrix} I \\ A \end{pmatrix}.$$

Key benefit

- $A$  is unconstrained (ordinary Euclidean parameter).
- We can optimize over  $A$  instead of constrained  $\Gamma$ .

# Profiling/Updating Blocks: How the Minimization Is Carried Out

Conceptual block structure

$$\zeta = (\mu, \eta, A, \Omega, \Omega_0), \quad \theta = \text{env}(\zeta), \quad Q_n(\zeta) = G_n(\text{env}(\zeta))^\top \hat{W} G_n(\text{env}(\zeta)).$$

A typical practical routine:

- ① Set  $\hat{\mu}_x = \bar{x}$  (profile out).
- ② Initialize  $A$  (subspace) using a sensible method (e.g., PLS / robust start).
- ③ Given  $A$  (thus  $\Gamma$ ), update  $(\mu, \eta)$  by minimizing  $Q_n$  w.r.t.  $(\mu, \eta)$ .
- ④ Update  $(\Omega, \Omega_0)$  to best match the covariance block.
- ⑤ Update  $A$  to further reduce  $Q_n$ .
- ⑥ Iterate until  $Q_n$  stabilizes (local minimum).

In the paper: derivative-free optimization (e.g., Nelder–Mead) is used for the nonconvex minimization.

# Error Distributions in the Simulation Study

Error distributions considered

- $N(0, 1)$  (Gaussian): Benchmark case where classical assumptions hold.
- $t_3$  (Student- $t$ ): Heavy-tailed distribution with finite mean but large variance.
- Mixnorm: Normal mixture  $0.9N(0, 1) + 0.1N(0, 25)$  introducing outliers.
- Laplace(0, 1): Sharp peak and heavier tails than Gaussian.
- Shifted Gamma (2, 2): Skewed and heavy-tailed errors.
- Cauchy(0, 1): Extremely heavy-tailed; variance does not exist.

## Simulations

**Table 2.** Comparison of estimation MSE ( $\times 10^{-2}$ ) for simulation models described in (6.1).

$\epsilon$	EHR	ENV	HR	PLS	LS
$N(0, 1)$	0.131 (0.003)	0.129 (0.003)	10.99 (0.24)	0.129 (0.003)	10.22 (0.22)
$t_3$	0.142 (0.004)	0.290 (0.121)	17.29 (0.43)	0.170 (0.006)	29.90 (1.14)
mixnorm	0.138 (0.003)	0.189 (0.007)	15.24 (0.42)	0.190 (0.007)	35.42 (0.93)
$Laplace(0, 1)$	0.140 (0.004)	0.226 (0.077)	15.28 (0.35)	0.149 (0.005)	20.02 (0.46)
$sGamma(2, 2)$	0.654 (0.036)	2.829 (2.204)	240.63 (5.43)	0.656 (0.037)	235.63 (5.54)
$Cauchy(0, 1)$	6.21 (1.58)	$5.52 \times 10^6$ $(5.48 \times 10^6)$	42.61 (1.37)	$5.16 \times 10^4$ $(4.70 \times 10^4)$	$7.42 \times 10^6$ $(7.37 \times 10^6)$

NOTE: The results are based on 100 replications. The standard errors are listed in the parentheses ( $\times 10^{-2}$ ).  $u$  is fixed at the true value.  $n$  is fixed at 500. "mixnorm" stands for the normal mixture  $0.9N(0, 1) + 0.1N(0, 25)$ .

## Simulations

**Table 3.** Comparison of estimation MSE ( $\times 10^{-2}$ ) for simulation model (6.2)  $y_i = \mu^* + \mathbf{x}_i^\top \beta^* + \sigma(\mathbf{x}_i)\tilde{\epsilon}_i$ .

$\tilde{\epsilon}$	EHR	ENV	HR	PLS	LS
$N(0, 1)$	0.114 (0.002)	0.120 (0.002)	3.41 (0.09)	0.121 (0.002)	6.25 (0.14)
$t_3$	0.120 (0.003)	0.169 (0.019)	5.04 (0.14)	0.151 (0.005)	18.34 (0.97)
mixnorm	0.118 (0.002)	0.236 (0.064)	4.72 (0.12)	0.173 (0.009)	22.13 (0.87)
$Laplace(0, 1)$	0.116 (0.002)	0.182 (0.045)	4.14 (0.10)	0.137 (0.003)	12.24 (0.33)
$sGamma(2, 2)$	0.398 (0.034)	2.290 (1.207)	88.32 (2.19)	0.538 (0.044)	150.86 (3.71)
$Cauchy(0, 1)$	3.791 (0.934)	$1.00 \times 10^7$ $(1.00 \times 10^7)$	11.54 (0.36)	$8.64 \times 10^4$ $(8.53 \times 10^4)$	$1.35 \times 10^7$ $(1.34 \times 10^7)$

NOTE: The scale function is  $\sigma(\mathbf{x}) = \frac{x_1+x_{24}}{4}$ . The results are based on 100 replications. The distributions of  $\tilde{\epsilon}$  are listed in the first column. The standard errors are