# Efficient and Robust Estimation of Regression and Scale Parameters, with Outlier Detection

Harshit
Kiran
Saksham

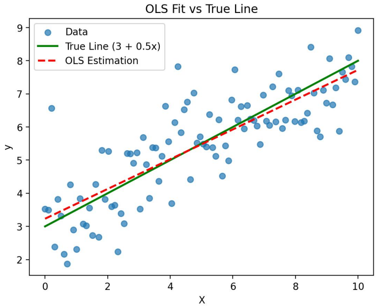
HSL613

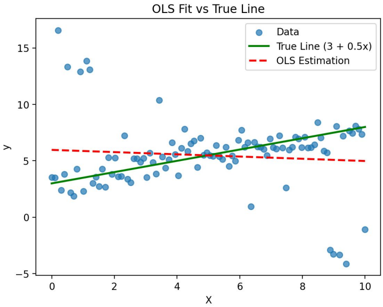
17-April-2025

# Contents

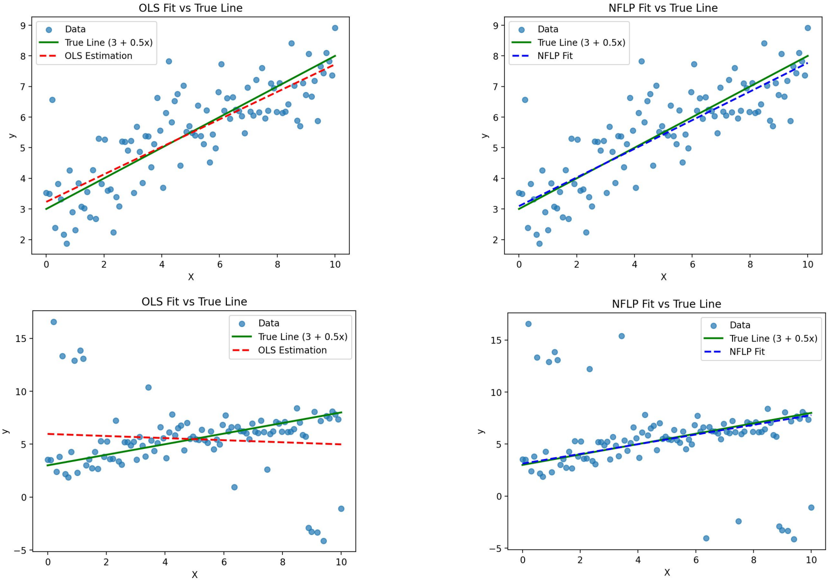
## OLS Estimators

- OLS have minimum variance among all Linear and Unbiased estimators.
- OLS are quite sensitive to outliers.



# Just for Motivation!!



# Framework

- We assume that the errors have a new distribution with heavier tails than the normal distribution, namely N-FLP Distribution.
- We estimate the parameters using Weighted Least Square(WLS) estimators which is achieved by adapting the EM algorithm in our methodology.

# N-FLP Distribution

The first step is to assume that the errors of the regression model have a *FLP*-contaminated normal distribution defined as

$$N - FLP(\omega, \mu, \sigma) = \omega N(\mu, \sigma^2) + (1 - \omega) FLP(\omega, \mu, \sigma)$$

where, the contaminating component is defined as Filtered Log Pareto(FLP) distribution.

$0 < \omega \leq 1$ is the proportion of normal observations,
$\mu \in \mathbb{R}$ is location parameter,
$\sigma > 0$ is the scale parameter.

# N-FLP Distribution(Cont.)

- The pdf of FLP distribution is given by

$$f_{FLP}(y \mid \omega, \mu, \sigma) = \begin{cases} 0 & \text{, if } |z| \leq \tau \text{ or } \omega = 1 \\ \dfrac{\omega}{\sigma(1-\omega)} \left[ \varphi(\tau) \dfrac{\tau}{|z|} \left( \dfrac{\log \tau}{\log |z|} \right)^{\lambda+1} - \varphi(z) \right] & \text{, if } |z| > \tau \text{ and } \omega < 1 \end{cases}$$

where $z = \left(\frac{y-\mu}{\sigma}\right)$.

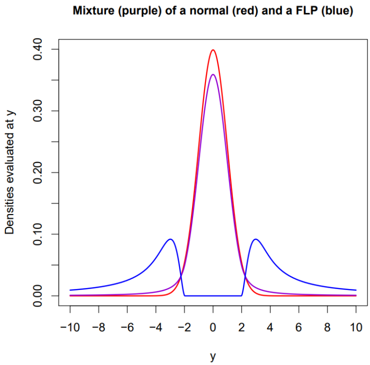
- Tail behaviour is controlled by $\lambda = \frac{2\omega}{(1-\rho\omega)} \varphi(\tau) \tau log(\tau) > 0$ with $\rho = 2\Phi(\tau) - 1$.
- The outlier region is controlled by $\tau > 1.69901$ defined as

$$\tau = g^{-1}(\omega) \text{ with } g(\tau) = \left( \rho + \frac{2\varphi(\tau)\tau log\tau}{(\tau^2 - 1)log\tau - 1} \right)^{-1}$$

# N-FLP Distribution(Cont.)

- The pdf of the N-FLP Distribution is thus given as

$$f_{N-FLP}(y \mid \omega, \mu, \sigma) = \begin{cases} \omega\sigma^{-1}\varphi(z) & \text{if } |z| \leq \tau \text{ or } \omega = 1, \\ \omega\sigma^{-1}\varphi(\tau)\dfrac{\tau}{|z|}\left(\dfrac{\log\tau}{\log|z|}\right)^{\lambda+1} & \text{if } |z| > \tau \text{ and } \omega < 1, \end{cases}$$



# Key Advantages of N-FLP Distribution

- It can adjust its shape depending on the number of outliers.
- For e.g., the tails of the distribution are heavier when $\omega = 0.5$ compared to when $\omega \geq 0.5$ and when $\omega = 1$, the distribution is simply Gaussian Distribution.

# Linear Regression: Vector Formulation

- The classical linear regression model is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

  where:
  - $\mathbf{y} \in \mathbb{R}^{n \times 1}$ is the response vector.
  - $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix with $n$ observations and $p$ predictors.
  - $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ is the vector of regression coefficients.
  - $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times 1}$ is the error vector.
- Each observation is modeled as:

$$y_i = \mathbf{x_i}^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n.$$

# Error Assumptions for OLS

- Under OLS, the errors $\epsilon_i$ are assumed to follow normal distribution i.e.,

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

- In the robust approach, the error $\epsilon_i$ are assumed to follow the mixture model, i.e.,

$$\epsilon_i \sim \omega \, \mathcal{N}(0, \sigma^2) + (1 - \omega) \, \mathrm{FLP}(\omega, 0, \sigma),$$

  where:
    - With probability $\omega$, $\epsilon_i$ is drawn from a normal distribution.
    - With probability $1 - \omega$, $\epsilon_i$ is drawn from a FLP distribution, which accounts for outliers.

# N-FLP Estimators

*The N-FLP estimators are given by*

$$\hat{\omega} = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i,$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{x}^T\mathbf{D}_{\hat{\pi}}\mathbf{x}\right)^{-1}\mathbf{x}^T\mathbf{D}_{\hat{\pi}}\mathbf{y} \quad \text{and}$$

$$\hat{\sigma}^2 = \frac{1}{\left(\sum_{i=1}^{n}\hat{\pi}_i - p\right)}\sum_{i=1}^{n}\hat{\pi}_i\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)^2,$$

*with*

$$\hat{\pi}_i \equiv \pi_{\hat{\omega}}(r_i) = \frac{\hat{\omega} f_{\mathcal{N}}(y_i \mid \mathbf{x}_i^T\hat{\boldsymbol{\beta}}, \hat{\sigma})}{f_{\mathcal{N}-\mathcal{FLP}}(y_i \mid \hat{\omega}, \mathbf{x}_i^T\hat{\boldsymbol{\beta}}, \hat{\sigma})} = \frac{\hat{\omega} f_{\mathcal{N}}(r_i \mid 0, 1)}{f_{\mathcal{N}-\mathcal{FLP}}(r_i \mid \hat{\omega}, 0, 1)},$$

where

$$r_i = \frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}}, \quad \mathbf{D}_{\hat{\pi}} := \operatorname{diag}(\hat{\pi}_1, \ldots, \hat{\pi}_n), \quad \text{and}$$

$$\mathbf{x} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T, \quad \mathbf{y} := (y_1, \ldots, y_n)^T.$$

# E-Step: Parameter Updates

- Given current parameters $\theta^{(t)} = (\omega^{(t)}, \beta^{(t)}, \sigma^{(t)})$, for each $i$ compute:

$$\pi_i^{(t)} = \frac{\omega^{(t)} f_N \left( y_i \mid x_i^T \beta^{(t)}, \sigma^{(t)} \right)}{f_{N-FLP} \left( y_i \mid x_i^T \beta^{(t)}, \sigma^{(t)} \right)},$$

where:

$$f_{N-FLP} \left( y_i \mid x_i^T \beta^{(t)}, \sigma^{(t)} \right) = \omega^{(t)} f_N \left( y_i \mid x_i^T \beta^{(t)}, \sigma^{(t)} \right) + \\ (1\text{-}\omega^{(t)}) f_{FLP} \left( y_i \mid \omega^{(t)}, x_i^T \beta^{(t)}, \sigma^{(t)} \right)$$

- Here, $f_N(\cdot)$ is the normal density and each $\pi_i^{(t)} \in \mathbb{R}^{n \times 1}$.

# M-Step: Parameter Updates

- Update the mixture weight:

$$\omega^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \pi_i^{(t)} \quad \in \mathbb{R}.$$

- Update the regression coefficients using weighted least squares:

$$\beta^{(t+1)} = \left(X^T D_\pi^{(t)} X\right)^{-1} X^T D_\pi^{(t)} y \quad \in \mathbb{R}^{p\times 1},$$

where $D_\pi^{(t)}$ is an $n \times n$ diagonal matrix with diagonal entries $\pi_i^{(t)}$.

- Update the scale parameter:

$$\sigma^{(t+1)2} = \frac{1}{\left(\sum_{i=1}^{n} \pi_i^{(t)} - p\right)} \sum_{i=1}^{n} \pi_i^{(t)} \left(y_i - x_i^T \beta^{(t+1)}\right)^2,$$

so that $\sigma^{(t+1)} \in \mathbb{R}_{>0}$.

# Iteration and Convergence

- **Iterate:** Alternate between the E-step and M-step until:

$$\|\theta^{(t+1)} - \theta^{(t)}\| < \epsilon,$$

where $\theta^{(t)} = \left(\omega^{(t)}, \beta^{(t)}, \sigma^{(t)}\right)$.

- **Convergence:** If change in each parameter is less than the threshold, convergence has been reached and final estimates are :

$$\hat{\omega} \in \mathbb{R}, \quad \hat{\beta} \in \mathbb{R}^{p \times 1}, \quad \hat{\sigma} \in \mathbb{R}_{>0}.$$

- If multiple solutions are produced, select the one with the smallest $\hat{\sigma}$.

# First Estimation Method: MLE for Mixture Model

**Model:**

$$y_i \sim \omega\mathcal{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2) + (1-\omega)\mathcal{FLP}(\omega, \mathbf{x}_i^T\boldsymbol{\beta}, \sigma)$$

**Issue:** Both the normal and FLP (outlier) components share the same parameters $\boldsymbol{\beta}$ and $\sigma$.

**Why this is problematic:**

- **Outliers** (modeled by FLP) are *not trustworthy* — they should not influence core parameter estimation.

# "In-Between" Solution: Smarter Simplicity

**Steps:**

1. **Estimate Parameters via EM (Expectation Maximization):**
   - Use only the clean/normal component
   - Obtain estimates:

$$\hat{\omega}, \quad \hat{\beta}, \quad \hat{\sigma}$$

2. **Apply These Estimates to the FLP Component:**
   - Instead of estimating $\omega_0, \beta_0, \sigma_0$ separately,
   - Simply set:

$$\hat{\omega}_0 = \hat{\omega}, \quad \hat{\beta}_0 = \hat{\beta}, \quad \hat{\sigma}_0 = \hat{\sigma}$$

**Resulting Model:**

$$y_i \sim \hat{\omega} \cdot \mathcal{N}(x_i^\top \hat{\beta}, \hat{\sigma}^2) + (1 - \hat{\omega}) \cdot \mathrm{FLP}(\hat{\omega}, x_i^\top \hat{\beta}, \hat{\sigma})$$

**Why This Works**

No extra parameters are introduced — keeping the model smooth, symmetric, and less prone to overfitting.

# Robust Inference Using N-FLP Estimators

Once we obtain the N-FLP estimates of $\hat{\omega}, \hat{\boldsymbol{\beta}}, \hat{\sigma}$; we follow the steps given below for robust inference:

- We define a random vector $\mathbf{v} = (v_1, \ldots, v_n)^T$ of n independent latent binomial variables defined as:

$$v_i = \begin{cases} 1, & \text{if observation originates from the normal component,} \\ 0, & \text{if observation originates from the FLP component.} \end{cases}$$

- We first assume $\mathbf{v}$ to be known and proceed with inference using the normal observations only.
- Secondly, the latent variables $v_i$ are estimated by $\hat{\pi}_i$.

# Robust Inference Using N-FLP Estimators(Cont.)

Notice that, calculation of $\hat{\omega}, \hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ with OLS estimators on the normal observations only is done by:

- $\hat{\boldsymbol{\beta}}|\mathbf{v} = (\mathbf{x}^T\mathbf{D_v}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{D_v}\mathbf{y}$
- $\hat{\sigma}^2|\mathbf{v} = \frac{1}{(\sum_{i=1}^{n} v_i - p)} \sum_{i=1}^{n} v_i(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2$

  where, $\mathbf{D_v} = diag(v_1, \ldots, v_n)$
  and the MLE of $\omega$ is $\omega|\mathbf{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$

$$\mathrm{Var}(\hat{\beta} \mid \mathbf{v}) = (\mathbf{x}^T\mathbf{D_v}\mathbf{x})^{-1}\mathbf{x}^T\mathrm{Var}(\mathbf{D_v}\mathbf{y} \mid \mathbf{v})\mathbf{x}(\mathbf{x}^T\mathbf{D_v}\mathbf{x})^{-1} = \sigma^2(\mathbf{x}^T\mathbf{D_v}\mathbf{x})^{-1},$$

Since

$$\mathbf{D_v}\mathbf{y} \mid \mathbf{v} \overset{\mathcal{L}}{\sim} \mathcal{N}_n(\mathbf{D_v}\mathbf{x}\beta, \sigma^2\mathbf{D_v}).$$

The robust estimation of $\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathbf{x}^T\mathbf{D}_{\hat{\pi}}\mathbf{x})^{-1}$

# $1 - \alpha$ confidence intervals

we obtain $\hat{\boldsymbol{\beta}} \overset{\mathcal{L}}{\approx} \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{x}^T \mathbf{D}_{\hat{\pi}} \mathbf{x})^{-1})$ and

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(\mathbf{x}^T \mathbf{D}_{\hat{\pi}} \mathbf{x})^{-1}]_{j,j}}} \overset{\mathcal{L}}{\approx} t_{\hat{\omega} n - p} \quad \text{for } j = 1, \ldots, p, \quad \text{and}$$

$$\frac{(\hat{\omega} n - p)\hat{\sigma}^2}{\sigma^2} \overset{\mathcal{L}}{\approx} \chi^2_{\hat{\omega} n - p},$$

The robust $1 - \alpha$ confidence intervals for the parameters are given by

$$\beta_j \in \hat{\beta}_j \pm t_{\alpha/2; \hat{\omega} n - p}\, \hat{\sigma}\sqrt{[(\mathbf{x}^T \mathbf{D}_{\hat{\pi}} \mathbf{x})^{-1}]_{j,j}} \quad \text{and}$$

$$\frac{(\hat{\omega} n - p)\hat{\sigma}^2}{\chi^2_{\alpha/2; \hat{\omega} n - p}} \leq \sigma^2 \leq \frac{(\hat{\omega} n - p)\hat{\sigma}^2}{\chi^2_{1-\alpha/2; \hat{\omega} n - p}}$$

# $R^2$ and Adjusted $\bar{R}^2$

$$R^2 = 1 - \frac{(\hat{\omega} n - p)\hat{\sigma}^2}{SSY} \quad \text{and} \quad \bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{SSY/(\hat{\omega} n - 1)},$$

where

$$SSY = \sum_{i=1}^{n} \hat{\pi}_i \left( y_i - \frac{\sum_{i=1}^{n} \hat{\pi}_i y_i}{\sum_{i=1}^{n} \hat{\pi}_i} \right)^2 .$$

# Outlier Detection

- The probability that an observation is *non-outlying* is given by:

$$\hat{\pi}_i = \frac{\hat{\omega} f_N(y_i | x_i^T \hat{\beta}, \hat{\sigma})}{f_{N-FLP}(y_i | x_i^T \hat{\beta}, \hat{\sigma})}$$

- An observation is flagged as an outlier if $\hat{\pi}_i < 0.5$.
- An adaptive threshold $\pi_{\hat{\omega}}^{-1}(0.5)$ is used instead of a fixed cutoff.

# Implementation

- Due to the non-convex nature of the robust estimation problem, we use Iterative Algorithm by using multiple initial values to select the best solution
- **Initialization:** Use ordinary least squares (OLS) estimates as starting values.
- **Perturbation:** For additional runs, perturb the initial $\beta$ values:

  $$\beta_{\text{start}} = \beta_{\text{OLS}} + \text{noise},$$

  where the noise $\sim \mathcal{N}(\cdot)$.
- **Selection:** From all runs, consider only solutions with $\omega > 0.5$.
- **Best Solution:** Select the solution with the smallest $\sigma$.

# Monte Carlo Simulations

- To compare the model statistically, the author proposes the following Monte Carlo simulation design:
  - Evaluate the efficiency of the estimator under normality (no contamination)
  - Evaluate robustness under contamination

# Efficiency Under Normality (Uncontaminated Model)

- The linear regression model studied in this section is:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon,$$

  where $\varepsilon \sim N(0, \sigma^2)$, with $p = 2$ (simple linear regression) and $p = 5$.
- Simulation Setup:
  - Dimensions: $p = 2$ (simple regression) and $p = 5$.
  - True parameters: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ ,$\sigma = 1$.
  - Explanatory variables: $x_{ij} \sim \mathcal{N}(0, 1)$ independently for $j = 2, \ldots, p$.
  - Sample sizes: $n = 50, 100, 200, 500$.
  - Runs: 100,000 Monte Carlo samples per setting.

# Metrics

- **Efficiency** measures the variance loss relative to OLS under a normal error model $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Relative efficiency (RE)** is defined as:

$$\mathrm{RE}(\hat{\theta}) = \left( \frac{\mathbb{E}[D_{\hat{\theta}_{\mathrm{OLS}}}(\theta)]}{\mathbb{E}[D_{\hat{\theta}}(\theta)]} \right)^2,$$

where $D_{\hat{\theta}}(\theta)$ is a distance metric from the true parameter $\theta$.

  - For regression coefficients:

$$D_{\hat{\beta}}(\beta) = \frac{1}{\sigma} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i^T \hat{\beta} - x_i^T \beta)^2}$$

  - For the scale parameter:

$$D_{\hat{\sigma}}(\sigma) = \left| \log \left( \frac{\hat{\sigma}}{\sigma} \right) \right|.$$

# Relative Efficiency for uncontaminated models

Table: Averaged over $n$'s

| **Model** | $p = 2,\ \hat{\beta}$ | $p = 2,\ \hat{\sigma}$ | $p = 5,\ \hat{\beta}$ | $p = 5,\ \hat{\sigma}$ |
|---|---|---|---|---|
| OLS | 1.000 | 1.000 | 1.000 | 1.000 |
| N-FLP | **0.992** | **0.928** | **0.990** | **0.920** |
| MM | 0.946 | 0.535 | 0.943 | 0.505 |
| M | 0.944 | 0.366 | 0.939 | 0.334 |
| REWLS | 0.899 | 0.502 | 0.851 | 0.450 |
| S | 0.293 | 0.535 | 0.281 | 0.506 |
| LMS | 0.137 | 0.207 | 0.141 | 0.190 |
| LTS | 0.126 | 0.305 | 0.136 | 0.256 |

# Robustness in the Presence of Outliers

- We compare the models with contaminated data.
- The Base Model remains unchanged:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

with $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ and $\sigma = 1$.

- Explanatory variables: $x_i \sim N(0, 1)$.
- Errors generated from $N(0, 1)$.

# Contamination

Contamination is introduced by replacing 10% or 20% of randomly chosen observations with outliers $(x_{io}, y_{io})$:

- **Low-leverage:** $x_{2o} \sim N(1,1)$, $y_o \sim N(\mu_0, 1)$.
- **High-leverage:** $x_{2o} \sim N(10,1)$, $y_o \sim N(\mu_0, 1)$.
- **Mixed:**
  - 60% High-leverage: $x_{2o} \sim \mathcal{N}(10,1)$, $y_o \sim \mathcal{N}(\mu_0, 1)$
  - 40% Low-leverage: $x_{2o} \sim \mathcal{N}(1,1)$, $y_o \sim \mathcal{N}(-\mu_0, 1)$
- $\mu_0$ varied in $[0, \mu_{\max}]$.

## Impacts

Low-leverage outliers affect the intercept, High-leverage ones distort both slope and intercept, and mixed outliers create conflicting influences. Adjusting outlier severity ensures methods are resilient for hard as well as easy to detect distortions.

# Test Metrics

**Performance Metrics:**

- For $\beta$, $\mathbb{E}[D_{\hat{\beta}}(\beta)]$ where $D_\beta(\hat{\beta})$ is calculated as before:

$$D_\beta(\hat{\beta}) = \frac{1}{\sigma}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i^T\hat{\beta} - x_i^T\beta\right)^2}.$$

- Similarly for $\sigma$, $\mathbb{E}[D_{\hat{\sigma}}(\sigma)]$ where $D_\sigma(\hat{\sigma})$:

$$D_\sigma(\hat{\sigma}) = \left|\log\left(\frac{\hat{\sigma}}{\sigma}\right)\right|.$$

# Average Deviation from Best Estimate

Table: Average deviation from the best estimate

| Model | Avg Deviation (Beta) | Avg Deviation (Sigma) |
|---|---|---|
| N-FLP | **0.020** | **0.009** |
| REWLS | 0.044 | 0.066 |
| MM | 0.071 | 0.077 |
| S | 0.107 | 0.078 |
| LTS | 0.226 | 0.027 |
| LMS | 0.250 | 0.067 |
| M | 0.601 | 0.200 |
| OLS | 0.936 | 0.515 |

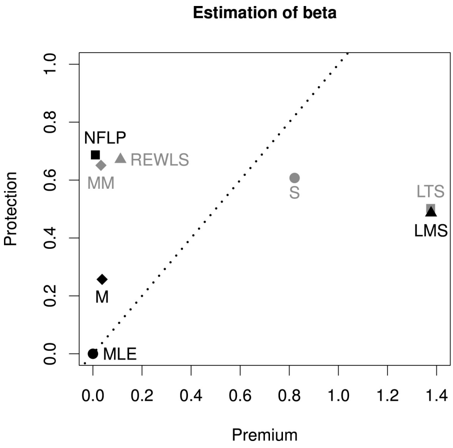
# Efficiency vs. Robustness Trade-off

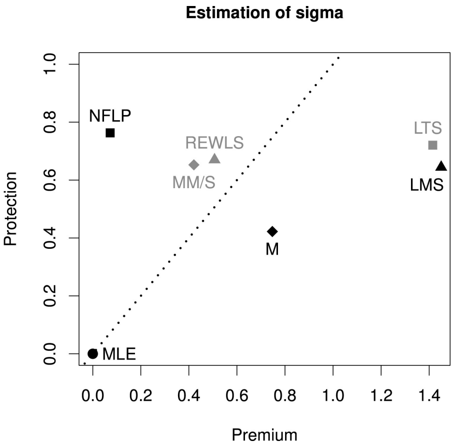
We now quantify the trade-off between efficiency (under normality) and robustness (under contamination).

- **Premium**: Represents the cost of using a robust estimator instead of the OLS estimator in the absence of outliers.
- **Protection**: Represents the gain in the presence of outliers when using a robust estimator.
- Ideal estimators have low Premium and high Protection (top-left corner of Protection-Premium plot).

# RESULTS

For the sample size of 50 we can plot:



# Economic Interpretations

- To gain economic interpretations from our model we fit the Mincer Earning's Function on a real world dataset.
- The canonical semi-logarithmic form is:

$$\ln(Y_i) = \beta_0 + \beta_1 S_i + \beta_2 X_i + \beta_3 X_i^2 + \epsilon_i$$

where:

- $\ln(Y_i)$ is the natural logarithm of wage income of individual $i$ (`INCWAGE`)
- $S_i$ is the years of education for individual $i$ (`EDUC`).
- $X_i$ is the potential experience for individual $i$ (`Exp`).
- $X_i^2$ is experience squared (`Exp2`).
- $\epsilon_i$ is the error term
- $\beta_o = ln(Y_o)$ where $Y_o$ is the earnings of someone with no education and no experience.

# Dataset

- **DataSet:** Integrated Public Use Microdata Series (IPUMS) USA dataset 2019 with the following variables
  - `AGE`: Age of the individual.
  - `EDUC`: Years of education completed.
  - `INCWAGE`: Income from wages and salaries.
- **Sampling:** The dataset is quite large, to maintain computational feasibility, we draw a random sample comprising 20% of the original data. (`seed = 613` for reproducibility) . This results in a dataset with approximately 320,000 observations.
- For labor market experience ($\texttt{EXP}$), we use the proxy $\texttt{EXP} = \texttt{AGE}$ - $\texttt{EDUC}$
- **Data Cleaning:**
  - Individuals with zero reported wage income ($\texttt{INCWAGE} = 0$) are excluded, as the Mincer model uses the logarithm of wages.
  - Observations with missing values in any of the key variables (`AGE`, `EDUC`, `INCWAGE`) are removed via listwise deletion.

# Estimated Coefficient Vectors $\hat{\beta}$

The estimated parameter vector $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]^T$ corresponds to `[INTERCEPT, EDUC, EXP, EXP2]`.

**OLS Results:**

$$\hat{\boldsymbol{\beta}}_{OLS} = \begin{pmatrix} 6.61114612 \\ 0.13616902 \\ 0.12645805 \\ -0.00173565 \end{pmatrix}$$

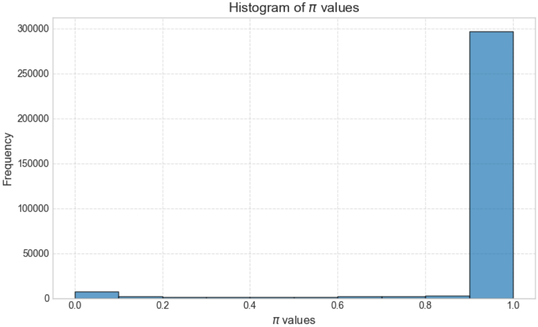
**NFLP Results:**

$$\hat{\boldsymbol{\beta}}_{NFLP} = \begin{pmatrix} 6.67985607 \\ 0.14408746 \\ 0.12249732 \\ -0.00168146 \end{pmatrix}$$

Intercept term: `(OLS, NFLP): (6.61114612, 6.67985607)` $\rightarrow$ estimated $Y_o$: `(743.33, 796.20)`

# NFLP Results

- For NFLP we get $\omega = 0.9564$, so the algorithm detects almost 5% observations as outliers
- The NFLP method also provides an associated value $\pi_i$ for each observation $i$, this is the estimated probability that $i^{th}$ observation has a normal error term.
- **Distribution of $\pi_i$ Values:**



# Outlier Filtering

- **Outlier Filtering Strategy:** We want to explore the economic characteristics of potential outliers identified by NFLP.
  - Define a threshold $\tau \in [0, 1]$.
  - Observations with $\pi_i < \tau$ are classified as "Outliers".
  - Observations with $\pi_i = 1$ are classified as "Normal".
  - By varying $\tau$, we can examine groups of observations with different extremities according to the NFLP metric.
  - The thresholds considered are $\tau \in \{1.0, 0.75, 0.50, 0.25, 0.01\}$.

# Comparison: Outlier Group Sizes ($N_{\text{Outliers}}$)

| Threshold ($\tau$) | $N_{\text{Outliers}}$ | Percentage (%) |
|---|---:|---:|
| 1.00 | 58,131 | 18.27 |
| 0.75 | 17,616 | 5.54 |
| 0.50 | 13,518 | 4.25 |
| 0.25 | 9,883 | 3.11 |
| 0.01 | 3,813 | 1.20 |

**Population:** 318,084

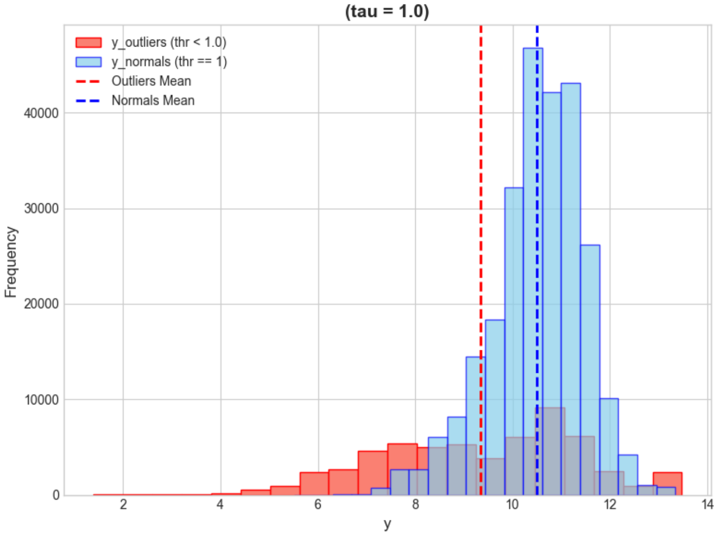
# Comparison: Mean Characteristics ($\bar{X}_{\text{OUTLIER}}$)

| Threshold ($\tau$) | $\bar{1}$ | $\overline{\texttt{EDUC}}$ | $\overline{\texttt{EXP}}$ | $\overline{\texttt{EXP}^2}$ |
|---|---|---|---|---|
| 1.00 | 1.00 | 13.52 | 29.58 | 1168.98 |
| 0.75 | 1.00 | 13.14 | 30.80 | 1309.07 |
| 0.50 | 1.00 | 13.17 | 31.11 | 1331.50 |
| 0.25 | 1.00 | 13.26 | 31.51 | 1350.93 |
| 0.01 | 1.00 | 13.70 | 32.94 | 1408.62 |
| **Population** | **1.00** | **13.83** | **29.24** | **1094.55** |

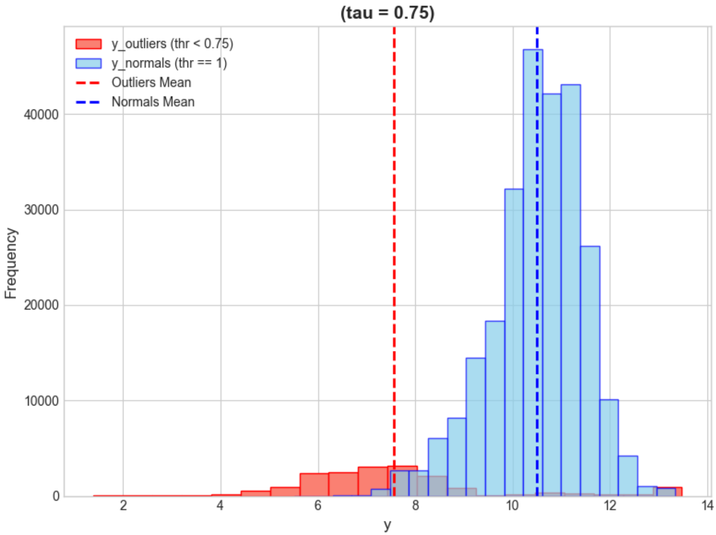
# Comparison: Log-Wage ($y_{\text{OUTLIER}}$) Distribution

| Threshold ($\tau$) | Mean $y_{\text{OUT}}$ | Median $y_{\text{OUT}}$ |
|---|---|---|
| 1.00 | 9.347 (11,464.37) | 9.547 (14,002.62) |
| 0.75 | 7.563 (1,925.61) | 7.244 (1,399.68) |
| 0.50 | 7.213 (1,356.95) | 6.908 (1,000.24) |
| 0.25 | 6.841 (935.42) | 6.685 (800.31) |
| 0.01 | 5.945 (381.83) | 5.991 (399.81) |
| **Population** | **10.29 (29,436.77)** | **10.55 (38,177.43)** |

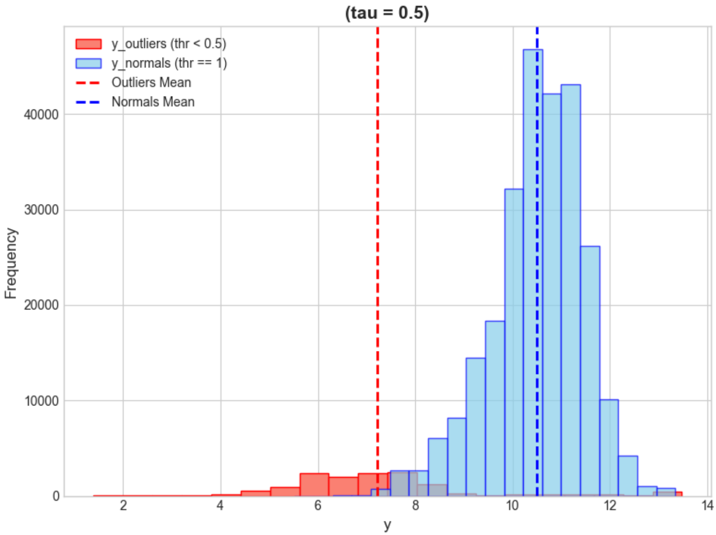
# Tau = 1.0


(tau = 1.0)
y_outliers (thr < 1.0)
y_normals (thr == 1)
Outliers Mean
Normals Mean
Frequency
40000
30000
20000
10000
0
2
4
6
8
10
12
14
y

# Tau = 0.75

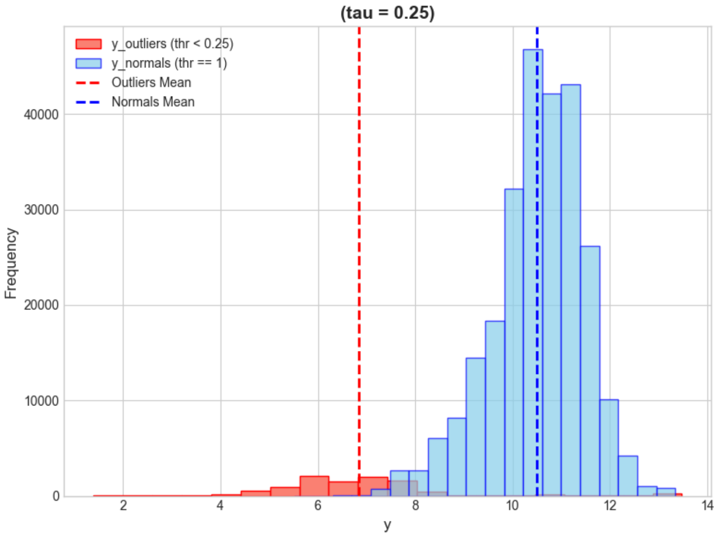

(tau = 0.75)
y_outliers (thr < 0.75)
y_normals (thr == 1)
Outliers Mean
Normals Mean
Frequency
40000
30000
20000
10000
0
2
4
6
8
10
12
14
y

# Tau $= 0.50$



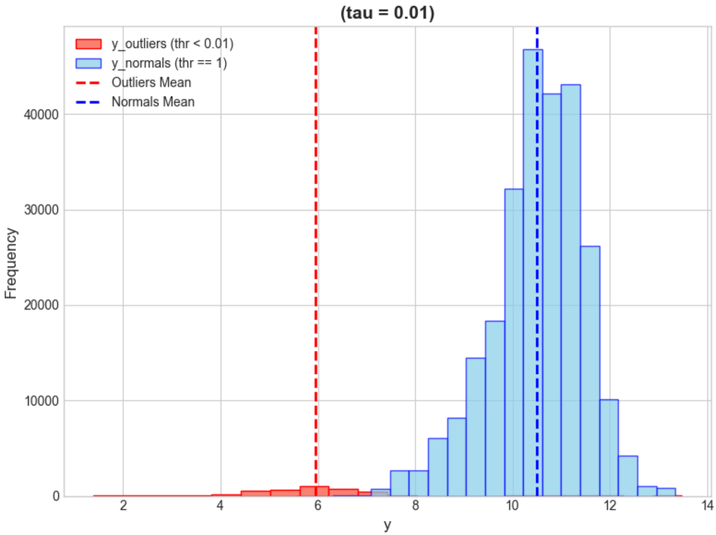
(tau = 0.5)
y_outliers (thr < 0.5)
y_normals (thr == 1)
Outliers Mean
Normals Mean
Frequency
40000
30000
20000
10000
0
2
4
6
8
10
12
14
y

# Tau $= 0.25$



(tau = 0.25)
y_outliers (thr < 0.25)
y_normals (thr == 1)
Outliers Mean
Normals Mean
Frequency
40000
30000
20000
10000
0
2
4
6
8
10
12
14
y

# Tau = 0.01



(tau = 0.01)
y_outliers (thr < 0.01)
y_normals (thr == 1)
Outliers Mean
Normals Mean
Frequency
0
10000
20000
30000
40000
2
4
6
8
10
12
14
y

# Effect of Outlier Extremity

- For a sample of 1000 observations, we contaminate 10% as outliers of varying extremity.
- For this we source the outliers and normal observations from the following indices:

$$\mathcal{I}_{\text{out}}(\tau) := \{\, i \mid \pi_i < \tau \,\},$$
$$\mathcal{I}_{\text{norm}} := \{\, i \mid \pi_i = 1 \,\}.$$

$$\tau \in \{1,\ 0.75,\ 0.5,\ 0.25,\ 0.01\}.$$

- Error metric (SSE):

$$\mathrm{SSE}(\hat{\beta}, \beta) = \sum_{j=1}^{p} (\hat{\beta}_j - \beta_j)^2$$

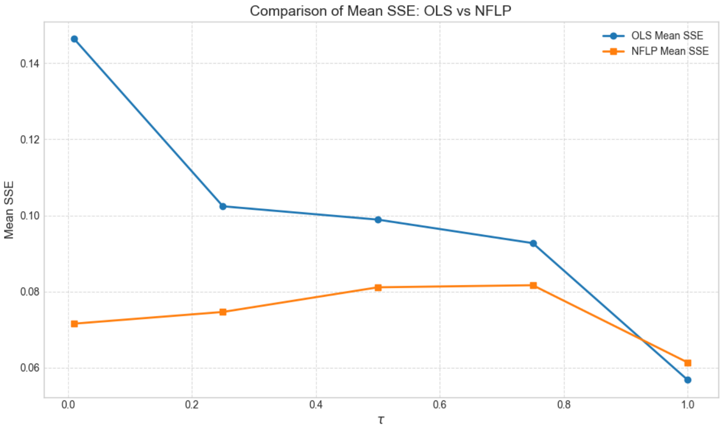
# Metrics

- For each threshold $\tau$, compute mean SSE over 500 iterations:

$$\overline{\mathrm{SSE}}_{\mathsf{OLS}}(\tau) = \frac{1}{500} \sum_{i=1}^{500} \mathrm{SSE}\left(\hat{\beta}^{(i)}_{\mathsf{OLS}}, \beta\right)$$

$$\overline{\mathrm{SSE}}_{\mathsf{NFLP}}(\tau) = \frac{1}{500} \sum_{i=1}^{500} \mathrm{SSE}\left(\hat{\beta}^{(i)}_{\mathsf{NFLP}}, \beta\right)$$

# Results



Comparison of Mean SSE: OLS vs NFLP
OLS Mean SSE
NFLP Mean SSE
Mean SSE
0.14
0.12
0.10
0.08
0.06
0.0
0.2
0.4
0.6
0.8
1.0
τ

# Effect of Outlier Population

- We take a sample of 1000 observations with varying outlier percentage $p$ over set $\{10, 15, \ldots, 40, 45\}$
- These outliers are sourced from

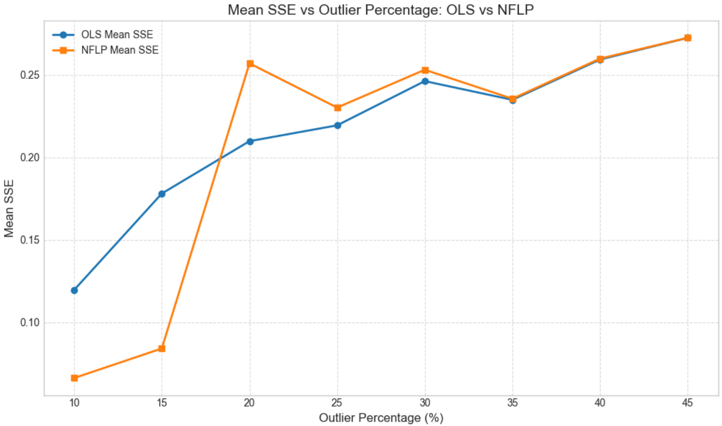
$$\mathcal{I}_{\text{out}} := \{\, i \mid \pi_i < 0.1 \,\}$$

- For each $p$, mean SSE over 500 iterations:

$$\overline{\mathrm{SSE}}_{\text{OLS}}(p) = \frac{1}{500} \sum_{i=1}^{500} \mathrm{SSE}\left(\hat{\beta}^{(i)}_{\text{OLS}}, \beta\right)$$

$$\overline{\mathrm{SSE}}_{\text{NFLP}}(p) = \frac{1}{500} \sum_{i=1}^{500} \mathrm{SSE}\left(\hat{\beta}^{(i)}_{\text{NFLP}}, \beta\right)$$

# Results



Mean SSE vs Outlier Percentage: OLS vs NFLP
OLS Mean SSE
NFLP Mean SSE
Mean SSE
0.25
0.20
0.15
0.10
10
15
20
25
30
35
40
45
Outlier Percentage (%)

# References I

Desgagné, A. (2021). *Efficient and robust estimation of regression and scale parameters, with outlier detection*. Computational Statistics and Data Analysis, 155, 107114.