Relative Efficiency of Higher Order Norm-Based Estimators over the Least Squares Estimator

With special emphasis on L_4 norm based estimator

Authors:

Gopal K. Basak, Samarjit Das, Arijit De, Atanu Biswas

Presented by: Gaurav Kumar Rampuria

(BS, Statistics & Data Science, IIT Kanpur)

Motivation & Overview

The Fundamental Question

"Why are we minimizing the residual sum of squares? Why not higher order loss functions?"

Least squares (LS) dominates statistical practice - primarily due to its simplicity, optimal properties and robustness to any distributional assumption. However, this doesn't imply it is always the best estimator to use in all situations.

- This paper studies the performance of the estimator that minimizes the loss function based on L_{2k} (for $k \geq 2$) against the estimator that minimizes the loss function of L_2 order (or the least squares estimator).
- A detailed simulation study verifies the effectiveness of this decision rule. Also, the superiority of the L4 based estimator is demonstrated in a real life data set as well as a constructed dataset.

Model and Methodology

Linear Regression Framework

$$Y = X\theta + \varepsilon \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} F,\ E[\varepsilon] = 0,\ Var(\varepsilon) = \sigma^2$$

where Y is $n \times 1$ response vector, X is $n \times k$ Design matrix.
Competing Estimators:

$$\hat{\theta}_{L_2} = \arg\min \sum_{i=1}^{n} \varepsilon_i^2 \quad \text{(OLS)} \tag{1}$$

$$\hat{\theta}_{L_4} = \arg\min \sum_{i=1}^{n} \varepsilon_i^4 \tag{2}$$

Both are M-estimators with consistency and asymptotic normality:

$$\sqrt{n}(\hat{\theta}_{OLS} - \theta_0) \xrightarrow{d} N(0, \sigma^2 S^{-1}) \tag{3}$$

$$\sqrt{n}(\hat{\theta}_{L_4} - \theta_0) \xrightarrow{d} N\left(0, \frac{\mu_6 - \mu_3^2}{9\mu_2^3} S^{-1}\right) \tag{4}$$

Main Theoretical Result

Theorem (Efficiency Condition)

The L_4 estimator performs better than the OLS estimator iff

$$\frac{\mu_6 - {\mu_3}^2}{{\mu_2}^3} < 9$$

Special Case: *Symmetric Distributions, i.e, $\mu_3 = 0$:*

$$\frac{\mu_6}{\mu_2^3} < 9$$

General Result

For L_{2k} vs L_2 comparison:

$$\frac{\mathrm{Var}(\varepsilon^{2k-1})}{(2k-1)^2(\mathrm{Var}(\varepsilon))^{2k-1}} < 1$$

OLS vs L_4 for Selected Distributions

Distribution	$\mu_6/9\mu_2^3$ Value	L_2 vs L_4
Uniform $(-a, a)$	$\frac{3}{7} \approx 0.4286$	L_4
Normal (μ, σ^2)	$\frac{15}{9} = 1.6667$	L_2
Laplace (λ)	$\frac{6!}{72} > 1$	L_2
Truncated Normal	< 1 for $c \in (0, 2.33)$	L_4
Raised Cosine	0.8926	L_4
Gaussian Mixture	< 1 (Means $> 1.058\sigma$ apart)	L_4
Pearsonian Family	< 1 for $a > -3.2$	L_4

Table: Comparison of L_4 and L_2 Estimators for Various Distributions

- Use L_4 when $\mu_6/9\mu_2^3 < 1$.
- L_4 performs better for thin-tailed or bounded error distributions.
- L_2 (OLS) is better for normal and heavy-tailed distributions.

Empirical Decision Rule

Test Statistic

Define $v = \frac{\mu_6 - \mu_3^2}{\mu_2^3}$. Then:

$$T = \frac{\sqrt{n}(\hat{v} - 9)}{s} \xrightarrow{d} N(0,1)$$

where $s^2 = \widehat{\mathrm{Var}(v)}$

Decision Rule:

- $H_0 : v \geq 9$ (prefer L_2) vs $H_1 : v < 9$ (prefer L_4)
- Use 0-1 loss function to minimize misclassification risk
- Asymptotically valid test with known distribution

Practical Implementation

Simple to compute from sample moments with established asymptotic properties

Simulation Study

Setup: We find the risk (expected loss), using the 0-1 loss function. For this, we ran 10,000 Monte Carlo replications for each of the the 3 scenarios. The following **error distributions** considered:

- Mixtures of Student-t with varying component separation.
- Mixtures of Beta (asymmetric and symmetric).
- Mixtures of Gaussians, including near-boundary cases.

Key Findings

- *L_4-favorable distributions*: Risk of misclassification $\rightarrow 0$ as $n \uparrow$ (e.g., well-separated t mixtures).
- *OLS-favorable distributions*: Rule correctly selects OLS nearly 100% for large n (e.g. single normal or Laplace mixtures).
- *Borderline cases* (e.g. symmetric Beta at boundary): risk 50% even at $n = 5000$, reflecting inherent ambiguity.
- Across all panels, risk consistently decreases with sample size, demonstrating asymptotic consistency of the decision rule.
- Decision rule performs excellently when means are well-separated

Real Data: Land Possession in India

Dataset

NSSO 68th round (2011-2012): 98,483 rural households with non-zero land possession

Key Findings:

- Bimodal distribution suggests "vanishing middle-class" phenomenon
- Test statistic: 5.30 (95% CI: 8.91, 9.09) $\Rightarrow$ Prefer L_4
- Standard errors uniformly smaller for L_4 estimates
- Pseudo R^2: 0.148 (L_4) vs 0.098 (L_2)

Conclusion: In this large real-data application, L_4 provided substantially tighter inference and better fit, consistent with the theoretical criterion.

Constructed Example: Comparison of L_4 vs L_2

Scenario:

A linear regression setup where the dependent variable (Y) is rounded to the nearest integer. Independent variables (X_1, X_2) are measured without rounding errors.
True Model: $Y^* = 8 + X_1 + 2X_2$, where X_1, X_2 are randomly generated, and $Y = 5 \cdot \text{floor}(Y^*/5)$.

Simulation

Sample size: $n = 40$, repeated for 5000 simulations.

Estimator	**Intercept**	**Slope 1 (X_1)**	**Slope 2 (X_2)**
L_2	7.013	1.115	1.924
L_4	6.548	1.021	1.962

- L_4 was preferred 90% of the time.
- Pseudo-R^2 was consistently higher for L_4, indicating better model fit.

Conclusion

- We have shown that, under non-normal error distributions, minimizing the fourth-power loss (L_4) can yield substantially lower estimator variance than ordinary least squares.
- A simple, empirically testable criterion—using sample residual moments—is provided to determine when to prefer the L_4 estimator over OLS.
- Extensive simulations confirm the decision rule's consistency: it reliably selects the more efficient estimator as sample size grows.
- Real-data applications (rounding-error regression and NSSO landholding data) demonstrate practical efficiency gains and tighter inference under L_4.
- This work invites further exploration of convex combinations of loss functions, robustness under outliers, and extension to other M-estimation contexts.