Approximate Median Regression via the Box-Cox Transformation

Garrett M. Fitzmaurice, Stuart R. Lipsitz, and Michael Parzen

INTRODUCTION

- Estimating median regression parameters using Gaussian estimating equations after applying a Box-Cox transformation to both the outcome and linear predictor.
- This proposed estimator is notably more efficient than the standard LAD estimator, despite a recognized loss of robustness.

ROBUSTNESS

- Provide reliable and accurate results even when the assumptions underlying the method are not perfectly met.

- Robust statistical methods are designed to be less sensitive to outliers, errors, or deviations from model assumptions compared to non-robust methods.

Median Regression

1 基礎概念

傳統的最小平方法（Least Squares Method）旨在最小化觀測值的預測值和實際值的平方差。而中位數迴歸（Median Regression）則針對中位數進行建模，這使它對於數據中的極端值不敏感，因此更具有穩健性（Robustness）。

OLS estimator

$$\text{minimize} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

LAD estimator

$$\text{minimize} \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

Median Regression

2 優勢與限制

Advantage

- 對異常值和極端值的穩健性（Robustness）
- 不受常態分佈假設的限制
- 對於因變數有偏態的數據更為適用

Disadvantage

- 若資料滿足常態的假設，則LAD的有效性（efficiency）與OLS相比較低

動機

The authors investigate the estimation after Box-Cox transformation for various distributions and compare the **bias** and **simulation variance** with the LAD estimator in this article.

Box-Cox transformation

為什麼要進行Box-Cox 轉換?

Taylor, J.M.G. (1985), "Power Transformations to Symmetry" showed that the Box-Cox transformation is generally the most suitable method for transforming to symmetry.

也就是Box-Cox 轉換可以將具有不同分配的數據轉換為更接近對稱分配的形式

$$U_i = u(Y_i, \lambda, c) = \begin{cases} \frac{(Y_i + c)^\lambda - 1}{\lambda} & \text{if} \quad \lambda \neq 0, \\ \log(Y_i + c) & \text{if} \quad \lambda = 0; \end{cases}$$

Y_i為第i個獨立觀察值的連續應變變數

c為一固定常數來確保$Y_i + c > 0$

λ為一需要被估計的未知參數

Median Regression via Box-Cox transformation

在一組數據經由Box-Cox轉換後，需要估計中位數迴歸線中的參數 $\hat{\beta}_0$、$\hat{\beta}_1$。

而估計方法又根據原數據的分配有不同方式:

- 當經Box-Cox轉換後的數據服從常態分配時，用MLE來估計參數(Log-Normal distribution)。
- 當經Box-Cox轉換後的數據不服從常態分配時，用Quasi-Likelihood Estimator來估計參數。

Monte Carlo Simulation

Performed a simulation study with four different specifications of the distribution of $Y_i|x_i$: Log-normal, Exponential, Gamma, and Pareto

For Gamma distribution : $Y_i = \beta_0 + \beta_1 x_i = 6.5 + 1.0x_i$
For Log-normal, Exponential, Pareto : $Y_i = \beta_0 + \beta_1 x_i = 6.0 + 1.0x_i$

Consider sample size $n = 80, 160, 320$ and let x_i take on values 1.0, 1.5, 2.0, 2.5, with each value represented by 25% of the sample.

Performed 2,500 simulation replications

Log-normal

$\log(Y_i) \sim N(\log(6.5 + x_i), 1)$

Theoretical estimates of the medians obtained via the optimal Box-Cox transformation are “unbiased” for $n = 80, 160, 320$

Simulation result:

- Relative Bias(%) ≈ 1%
- The simulation variances of the LAD estimates are 50-75% larger than those for the estimates from the Box-Cox transformation.

Exponential

$$Y_i|x_i \sim Exp(\theta = \frac{\log 2}{\beta_0 + \beta_1 x_i})$$

The average skewness of the residuals from the regression of the Box-Cox transformed Y_i over all simulation replications was small (skewness= -0.05)

Exponential

Simulation result:

Table 1. Results of simulation study for estimated medians at each value of x with Y_i exponential.

		$n = 80$				$n = 160$				$n = 320$			
		x				x				x			
	Approach	1	1.5	2	2.5	1	1.5	2	2.5	1	1.5	2	2.5
Relative Bias (%)	Box-Cox	−0.74	−0.59	−0.45	−0.34	−1.06	−1.33	−1.56	−1.77	−1.43	−1.48	−1.52	−1.56
	LAD	3.73	2.68	1.74	0.92	1.15	0.88	0.65	0.44	0.28	0.44	0.59	0.72
	MLE	0.17	−0.23	−0.58	−0.90	−0.12	0.01	0.12	0.22	−0.07	0.01	0.08	0.14
Simulation Variance	Box-Cox	2.68	1.29	1.56	3.48	1.32	0.67	0.81	1.74	0.68	0.33	0.39	0.86
	LAD	4.69	1.93	2.31	5.82	2.22	0.99	1.23	2.96	1.07	0.49	0.59	1.39
	MLE	2.13	0.95	1.17	2.77	1.05	0.47	0.58	1.40	0.52	0.24	0.29	0.68
Coverage Probability[a]	Box-Cox	96.0	96.4	95.9	95.8	95.3	94.5	95.2	95.3	95.4	94.4	94.7	95.2
	LAD	95.2	96.5	96.2	95.0	94.0	95.0	94.4	94.4	95.0	95.6	95.0	95.6
	MLE	93.2	93.4	93.9	93.3	94.3	95.1	94.6	94.1	94.4	94.5	94.6	94.6

[a] Coverage Probability of 95% Confidence Intervals

Gamma

$$Y_i | x_i \sim Gamma(k = 0.5, \theta = \frac{\{\log(2)\}}{\beta_0 + \beta_1 x_i})$$

This is an example of a skewed, heavy-tailed distribution where
even the Box-Cox transformed Y_i is likely to show skewness. (skewness$= -0.1$)

Gamma

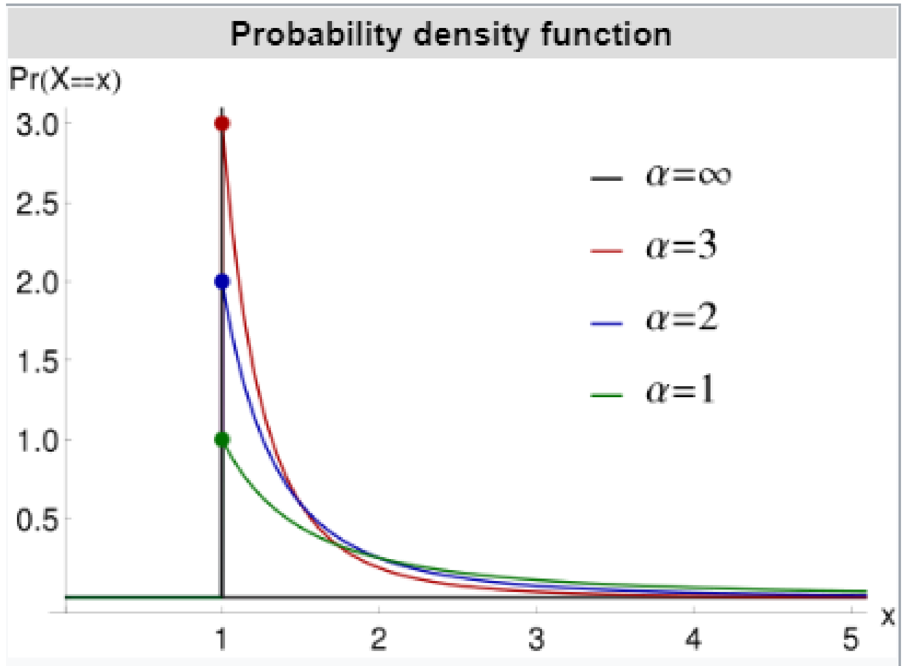
Simulation result:

Table 2. Results of simulation study for estimated medians at each value of x with Y_i Gamma.

		$n = 80$				$n = 160$				$n = 320$			
		x				x				x			
	Approach	1	1.5	2	2.5	1	1.5	2	2.5	1	1.5	2	2.5
Relative Bias (%)	Box-Cox	−2.24	−3.38	−4.38	−5.27	−5.02	−5.24	−5.43	−5.60	−5.47	−5.48	−5.49	−5.50
	LAD	7.54	5.57	3.86	2.34	3.28	2.83	2.45	2.11	0.43	0.87	1.25	1.58
Simulation Variance	Box-Cox	5.72	2.94	3.48	7.34	2.79	1.36	1.63	3.59	1.36	0.70	0.84	1.79
	LAD	10.37	4.57	5.60	13.47	5.07	2.29	2.83	6.69	2.34	1.05	1.35	3.25
Coverage Probability[a]	Box-Cox	96.0	95.5	95.4	96.1	95.6	94.5	94.1	95.1	94.3	92.9	92.6	94.1
	LAD	95.1	96.8	96.8	95.7	94.6	95.4	96.1	94.6	95.6	96.2	95.4	95.1

[a] Coverage Probability of 95% Confidence Intervals

Average bias
4.9%
2.8%



Pareto Type I probability density functions for various α

Pareto

$Y_i | x_i \sim Pareto(\alpha, x_m = 1)$

The Pareto is an example of an extremely skewed, heavy-tailed distribution where even the Box-Cox transformed Y_i is likely to show very discernible skewness. (skewness= 0.29)

Simulation result:

- The proposed estimator yields badly biased estimates of median.
- the LAD estimator of (β_0, β_1) is far more robust and almost unbiased when the sample sizes are large.

Monte Carlo Simulation Summary

- The simulation study suggest that the proposed estimator is relatively robust to “modest degrees of asymmetry” in the distribution of Y_i after a Box-Cox transformation.

- The relative bias is less than 5% and of comparable to LAD estimator

- The proposed method provides discernibly more efficient estimates than the standard LAD estimator

- However, when there is strong asymmetry in the distribution of Y_i , the proposed estimator can yield badly biased estimates.

Conclusion

- Compared to the LAD estimation method, the Box-Cox transformation demonstrates higher efficiency but comes at the cost of reduced robustness.

- The Box-Cox transformation focuses on the symmetry of the transformed data distribution, and the lower the symmetry of the transformed data distribution, the more likely biased estimates may arise.

- The LAD estimation method exhibits robustness in handling outliers, while the Box-Cox transformation may be sensitive to extreme values and anomalies in the transformed data.

Thank you!