527_project, M,P estimator

NetID: dichuan2

DICHUAN ZHENG

2024-12-01

Contents

Why M and P estimators?

M-estimator: Reduces the influence of large residuals by using robust loss functions, ensures that a few extreme outliers do not affect the model fit. P-estimator:

Designed to be even more robust than M-estimators, especially for datasets with a high proportion of outliers. Using robust statistical principles to minimize the effect of large residuals # M-estimator

Definition

An **M-estimator** minimizes a general loss function ρ instead of the sum of squared residuals used in ordinary least squares (OLS). It is defined as:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(y_i - X_i\beta)$$

where: - $\rho(x)$ is a robust loss function (e.g., Huber loss, Tukey's biweight). - y_i is the response variable, X_i is the predictor variable(s), and β represents the model parameters.

Estimation Procedure

The solution is typically found by solving the following first-order condition:

$$\sum_{i=1}^{n} \psi(y_i - X_i\hat{\beta})X_i = 0$$

where: - $\psi(x) = \frac{\partial \rho(x)}{\partial x}$ is the influence function that limits the impact of large residuals.

Common Loss Functions

1. **Huber Loss**:

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \le c \\ c|x| - \frac{1}{2}c^2, & \text{if } |x| > c \end{cases}$$

2. **Tukey's Biweight**:

$$\rho(x) = \begin{cases} c^2 \left(1 - \left(1 - \frac{x^2}{c^2}\right)^3\right), & \text{if } |x| \le c \\ c^2, & \text{if } |x| > c \end{cases}$$

P-estimator

Definition

A **P-estimator** is designed to provide even higher robustness than M-estimators. It minimizes a robust scale estimate of residuals while controlling for outlier contamination. It is often used in combination with S-estimators to achieve a balance between robustness and efficiency.

The P-estimator solves:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} w_i \cdot \rho(y_i - X_i\beta)$$

where: - w_i are robustness weights that adaptively reduce the impact of outliers.

Key Characteristics

1. P-estimators can handle a higher proportion of outliers than M-estimators.
2. They are less efficient in clean datasets but more robust in contaminated ones.

Analysis the Dataset

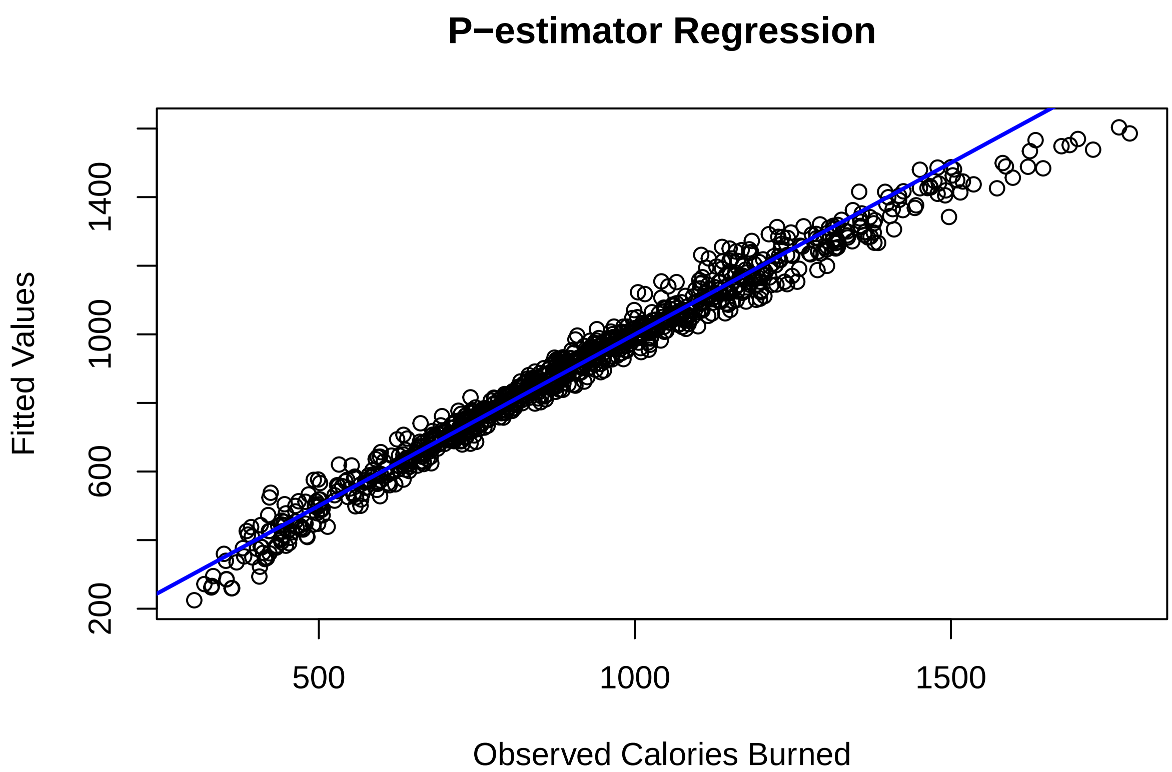
```
##
## Call: rlm(formula = Calories_Burned ~ ., data = gym_data, method = "M")
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.758  -23.392   -1.135   22.859  182.247
##
## Coefficients:
##                               Value     Std. Error t value
## (Intercept)                   -966.7005   80.7598  -11.9701
## Age                             -3.3626    0.0970  -34.6612
## GenderMale                      82.2678    4.2344   19.4284
## Weight..kg.                     -0.8572    0.4720   -1.8162
## Height..m.                      89.4294   43.4451    2.0584
## Max_BPM                          0.0897    0.1023    0.8770
## Avg_BPM                          6.1062    0.0820   74.4962
## Resting_BPM                      0.3039    0.1607    1.8912
## Session_Duration..hours.       710.8228    5.4578  130.2402
## Workout_TypeHIIT                -1.9409    3.3707   -0.5758
## Workout_TypeStrength            -2.3849    3.2459   -0.7347
## Workout_TypeYoga                -5.1514    3.3083   -1.5571
```

```
## Fat_Percentage                    -0.2618    0.3106    -0.8428
## Water_Intake..liters.             -1.9311    3.0008    -0.6435
## Workout_Frequency..days.week.      0.6175    2.3565     0.2621
## Experience_Level                   0.0661    3.6790     0.0180
## BMI                                2.9298    1.4362     2.0400
##
## Residual standard error: 34.39 on 956 degrees of freedom
```

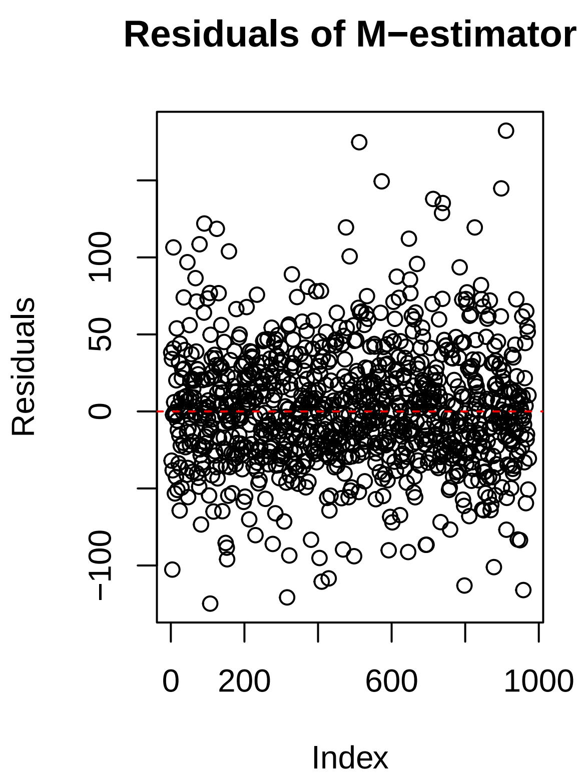

```
## [1] 1543.91
```

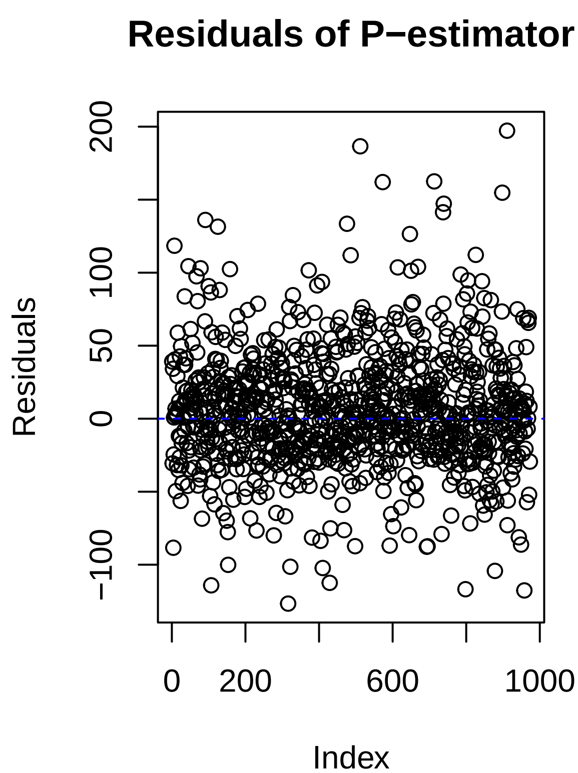
```
##
## Call:
## lmrob(formula = Calories_Burned ~ ., data = gym_data, method = "S")
##  \--> method = "S"
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.632  -18.857    3.301   27.238  197.239
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -870.91957  121.42784  -7.172 1.48e-12 ***
## Age                            -3.08847    0.13849 -22.302  < 2e-16 ***
## GenderMale                     81.53791    6.41026  12.720  < 2e-16 ***
## Weight..kg.                    -0.37247    0.70096  -0.531    0.595
## Height..m.                     46.29317   64.70402   0.715    0.474
## Max_BPM                         0.03907    0.15657   0.250    0.803
## Avg_BPM                         5.85889    0.13075  44.809  < 2e-16 ***
## Resting_BPM                     0.39064    0.24507   1.594    0.111
## Session_Duration..hours.      705.66285    8.66079  81.478  < 2e-16 ***
## Workout_TypeHIIT               -1.13985    5.13523  -0.222    0.824
## Workout_TypeStrength            0.81928    4.95517   0.165    0.869
```

```
## Workout_TypeYoga               -1.14183    5.17927  -0.220    0.826
## Fat_Percentage                 -0.13587    0.47654  -0.285    0.776
## Water_Intake..liters.          -5.32298    4.40471  -1.208    0.227
## Workout_Frequency..days.week.  -1.50224    3.60250  -0.417    0.677
## Experience_Level                6.09505    5.41836   1.125    0.261
## BMI                             1.86139    2.13657   0.871    0.384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 34.2
## Multiple R-squared:  0.9927, Adjusted R-squared:  0.9926
##
## Robustness weights:
##  157 observations c(4,7,16,24,35,45,51,67,70,78,82,90,91,100,104,106,107,108,116,119,125,130,137,140
##   are outliers with |weight| = 0 ( < 0.0001);
##  30 weights are ~= 1. The remaining 786 ones are summarized as
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0003622 0.4516000 0.7626000 0.6658000 0.9356000 0.9989000
## Algorithmic parameters:
##       tuning.chi                bb        tuning.psi        refine.tol
##        1.548e+00         5.000e-01         4.685e+00         1.000e-07
##          rel.tol         scale.tol         solve.tol          zero.tol
##        1.000e-07         1.000e-10         1.000e-07         1.000e-10
##      eps.outlier             eps.x warn.limit.reject warn.limit.meanrw
##        1.028e-04         3.620e-10         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s          k.max
##            500             50              2              1            200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0           2000
##                   psi           subsampling                   cov
##            "bisquare"         "nonsingular"             ".vcov.w"
## compute.outlier.stats
##                   "S"
## seed : int(0)
```




```
## [1] 1643.805
## M-estimator Residual Standard Error: 39.29262
## P-estimator Residual Standard Error: 40.54387
```



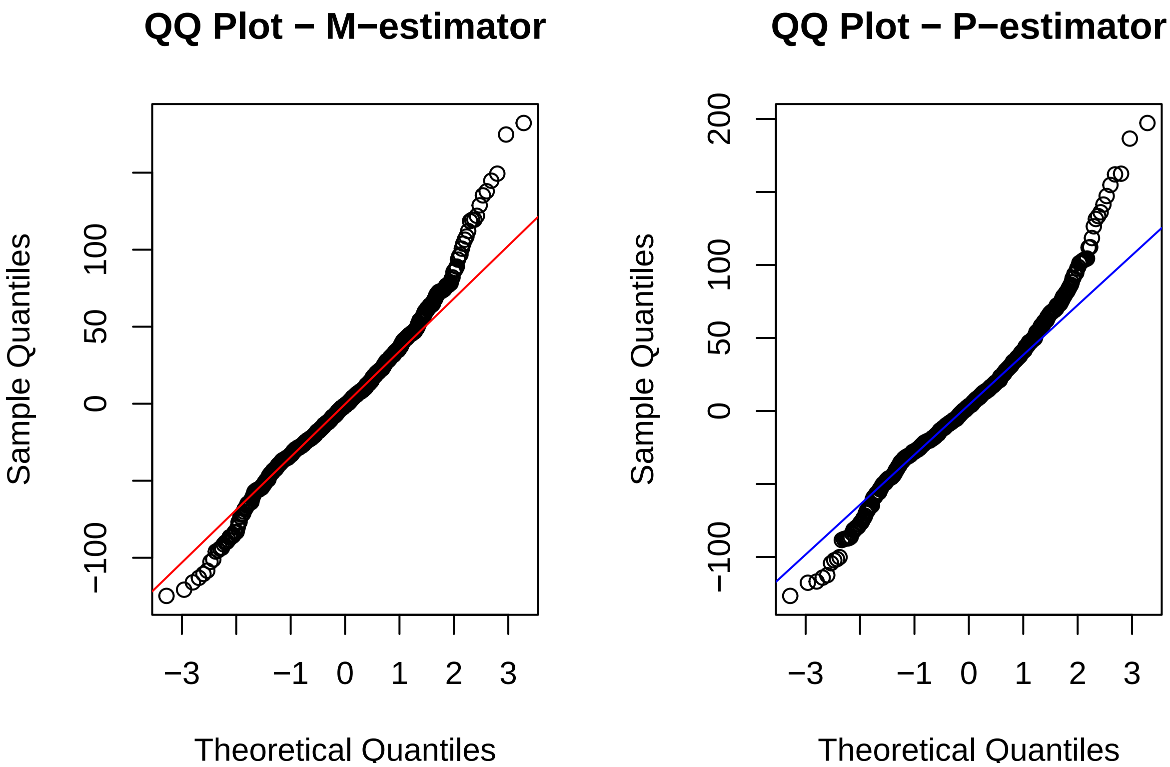



```
##        AIC        BIC
```

```
##   9939.092 10022.058
```

```
##       AIC       BIC
## 10000.09 10083.06
```

Both AIC and BIC suggest that M model is better.




```
##
##   Shapiro-Wilk normality test
##
## data:  residuals_m
## W = 0.9798, p-value = 2.282e-10
```

```
##
##   Shapiro-Wilk normality test
##
## data:  residuals_p
## W = 0.97283, p-value = 1.626e-12
```

The result show the residuals of both M and P estimators are normally distributed.