LASSO

Data visualization

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-08-03

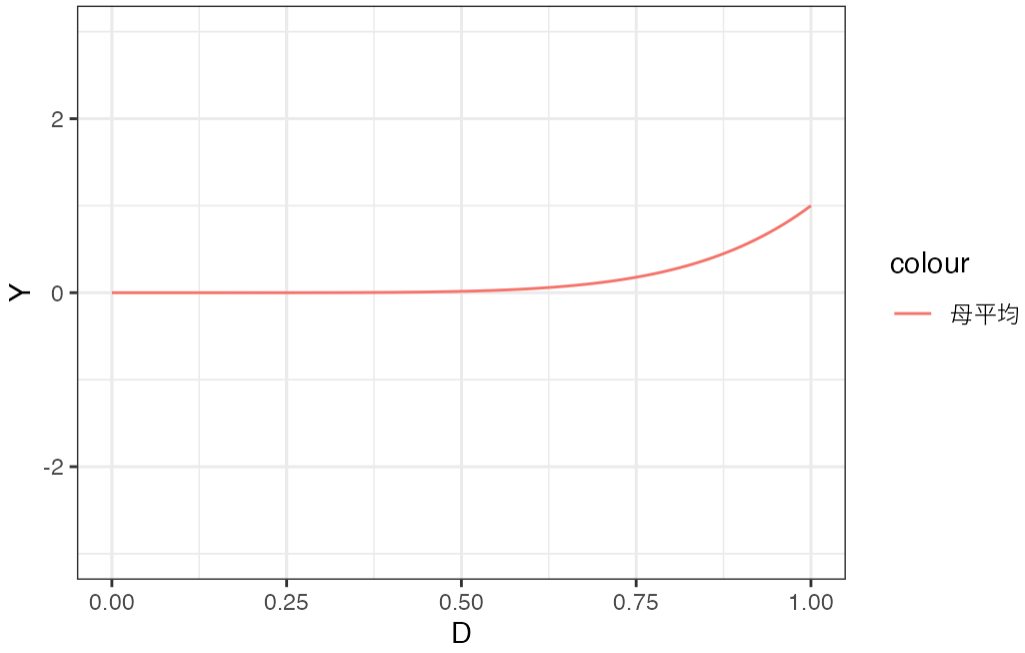
1 線型予測モデル

- 宿題の確認

1.1 復習: 理想の予測モデル

- 理想の予測モデルは、母平均
  - 無限大の事例数があれば、データ上の平均値が理想の予測モデル

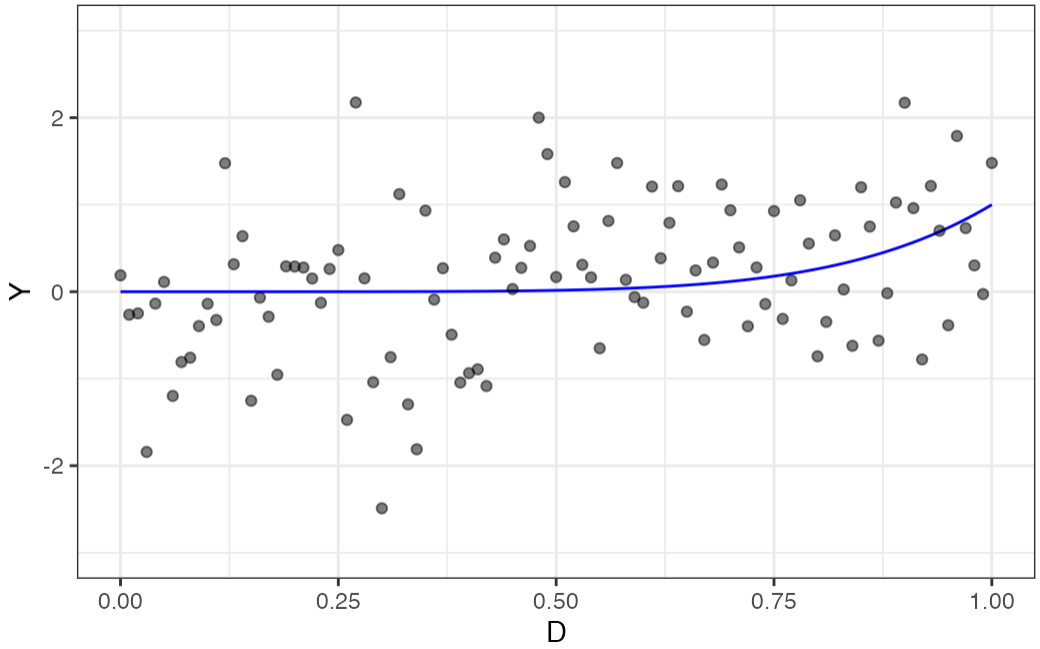
1.2 例: 母平均



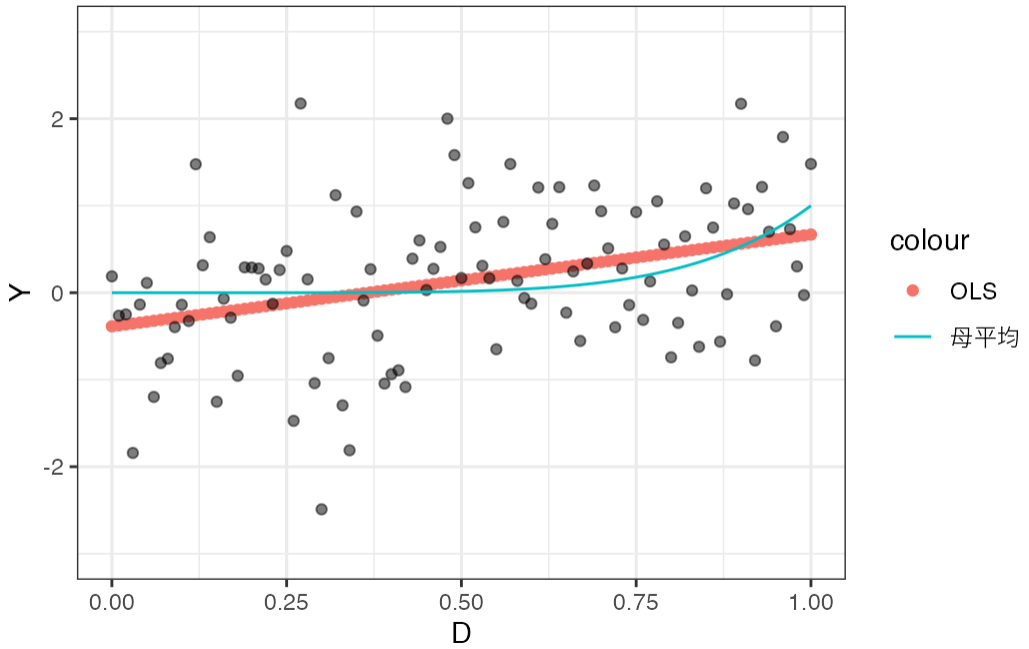
1.3 復習: データ

- 個人差がある限り、平均値から乖離する

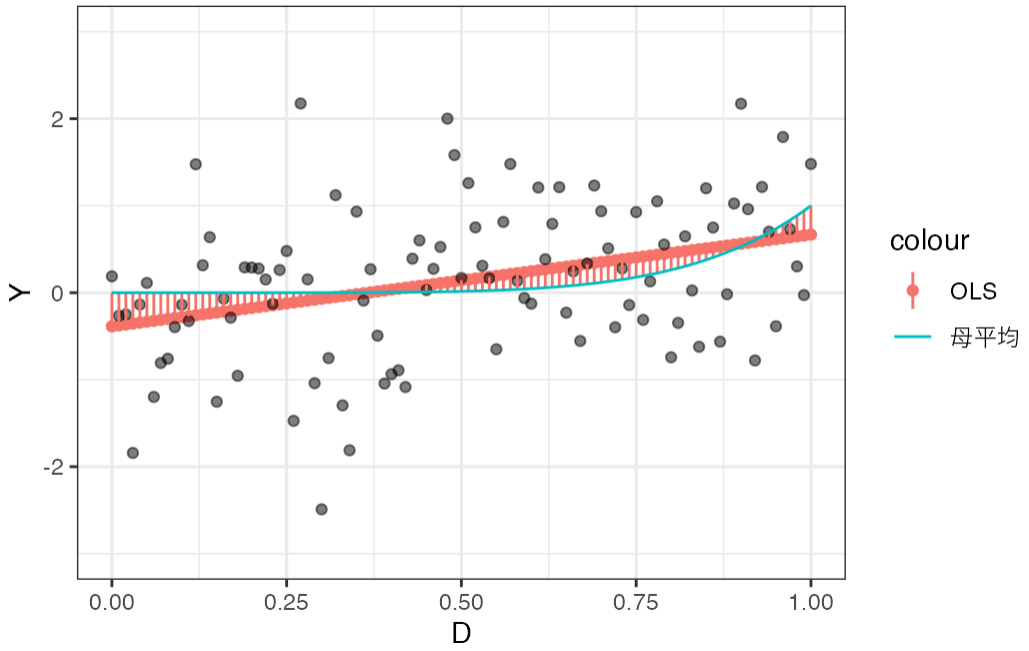
1.4 例: データ



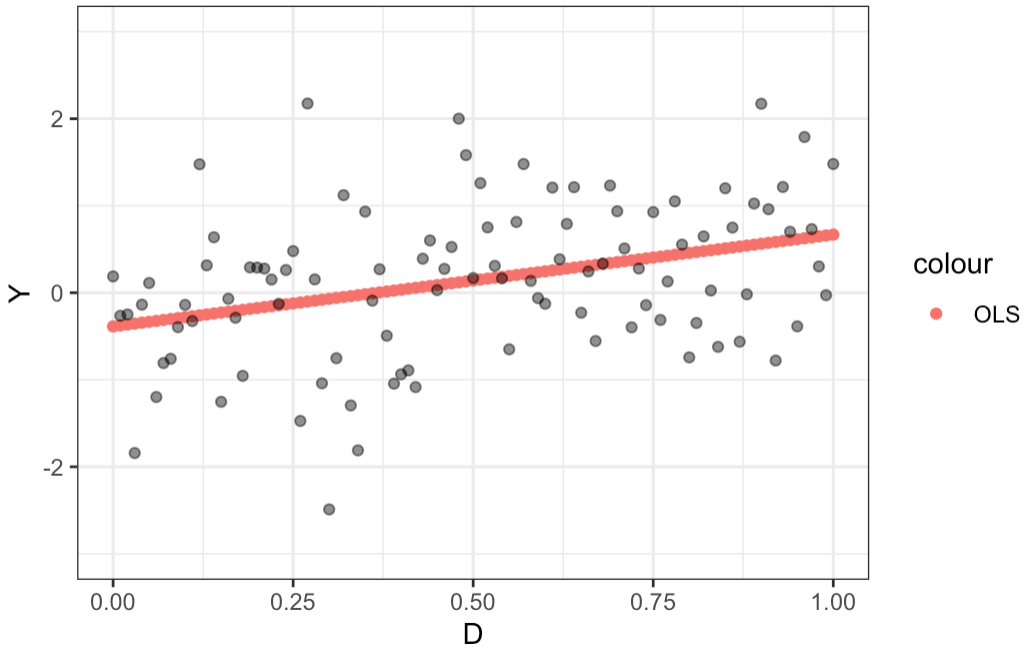
1.5 例: OLS



1.6 例: 母平均からの乖離



1.7 例: 実際の状況

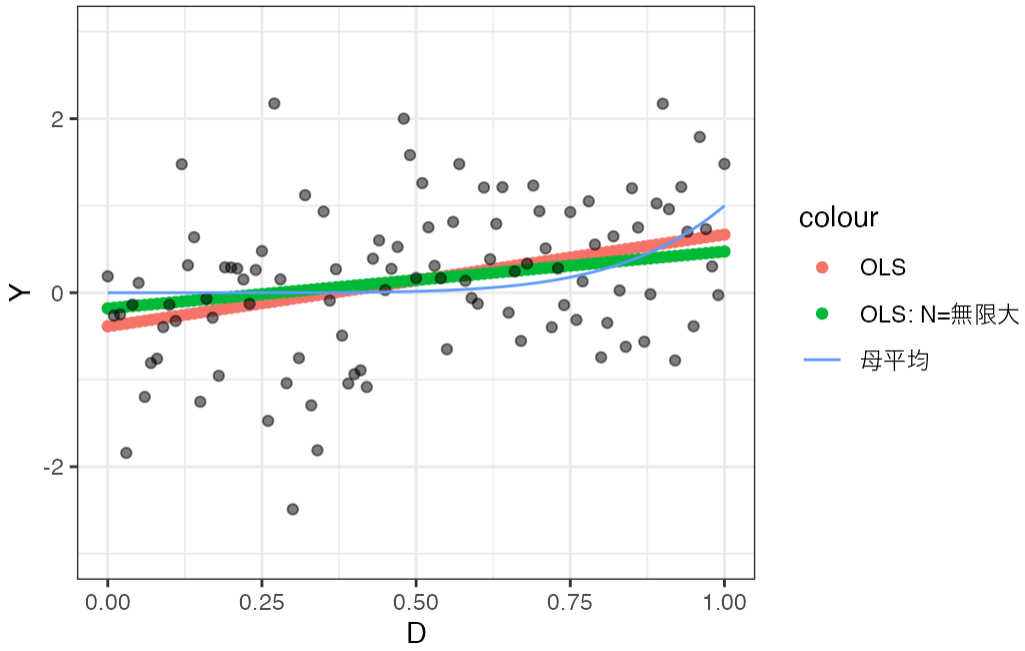


2 理論的性質

2.1 サンプルサイズ

- 単純なモデルの場合、
  - 事例数が増えたとしても、母平均とは乖離する可能性が高い
  - 事例数が少なくても、推定結果が大きく変化しない可能性が高い

2.2 例: サンプルサイズ



2.3 モデルの複雑化

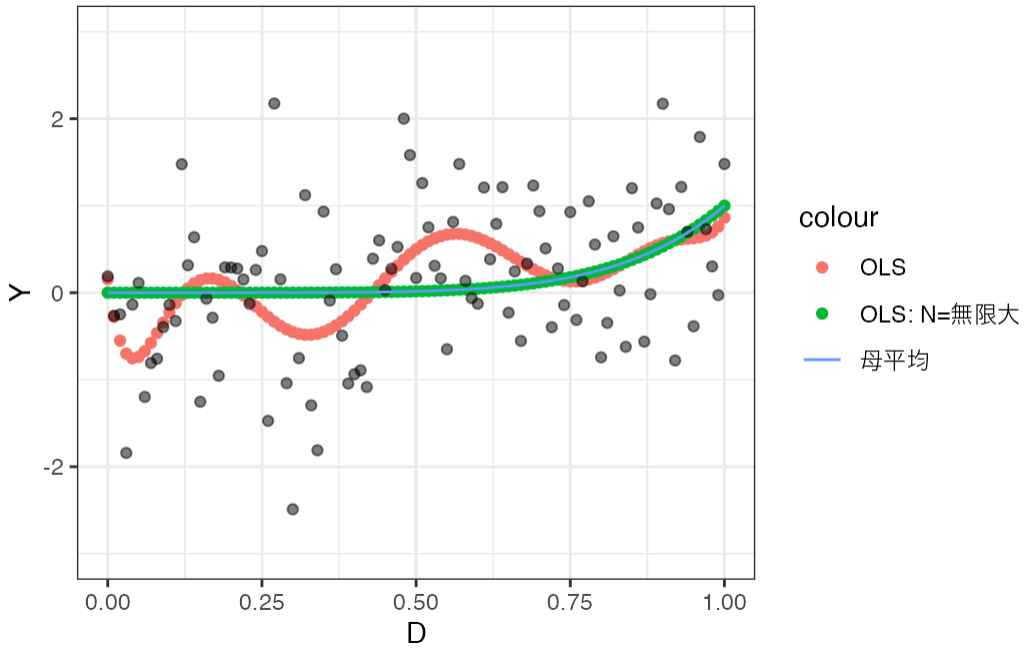
- 例: 同じ X について、モデルは容易に複雑化できる
- $X = [Size, Tenure]$ について、

$$\beta_0 + \beta_1 Size + \beta_2 Tenure$$

$$\underbrace{+\beta_3 Size^2 + \beta_4 Tenure^2}_{\text{交差項}}$$

$$\underbrace{\beta_5 (Size \times Tenure)}_{\text{交差項}}$$

2.4 例: $Y \sim X + X^2$



2.5 実装

```
Call:
lm(formula = Price ~ poly(Size, 2) + poly(Tenure, 2) + Size:Tenure,
    data = data)

Coefficients:
     (Intercept)    poly(Size, 2)1    poly(Size, 2)2  poly(Tenure, 2)1
        64.50855         3710.84191         1168.83982          371.78073
poly(Tenure, 2)2       Size:Tenure
       423.80494           -0.02057
```

2.6 Takeaway

- 事例数に応じて、適したモデルは異なる
- より複雑なモデル (β の数が多いモデル) を推定すると、データへの適合度が改善するが
  - ハズレ値の影響を受けやすく、良い予測モデルとは限らない

2.7 Takeaway

モデルの性質	予測値 - Population OLS	Population OLS - 母平均	予測値-母平均
単純	小さい傾向		?
非常に複雑		小さい傾向	?

3 データ主導のアプローチ

3.1 OLS の限界

- 良い予測モデルを推定するためには、適切な複雑性を持つモデルを研究者が設定する必要がある
  - 多くの応用で難しい

3.2 LASSO

- (二乗項や交差項を含む)複雑な線型モデル $g(X) = \beta_0 + .. + \beta_L X_L$ を以下を最小化するように推定する

$$\text{平均二乗誤差} + \underbrace{\lambda}_{TuningParameter:>0} \times \underbrace{(|\beta_1| + .. + |\beta_L|)}_{\text{複雑性の測定値}}$$

- 複雑性に”課税 (税率 λ)“することで、単純なモデル「 $g(X) = \beta_0 = Y$ の平均値」 に近づける

3.3 実装例

```
lm(Price ~ poly(Size,2) + poly(Tenure,2) + Size:Tenure, data)
```

```

Call:
lm(formula = Price ~ poly(Size, 2) + poly(Tenure, 2) + Size:Tenure,
    data = data)

Coefficients:
     (Intercept)    poly(Size, 2)1    poly(Size, 2)2  poly(Tenure, 2)1
        64.50855        3710.84191        1168.83982         371.78073
poly(Tenure, 2)2       Size:Tenure
       423.80494          -0.02057
```

```
hdm::rlasso(Price ~ poly(Size,2) + poly(Tenure,2) + Size:Tenure, data)
```

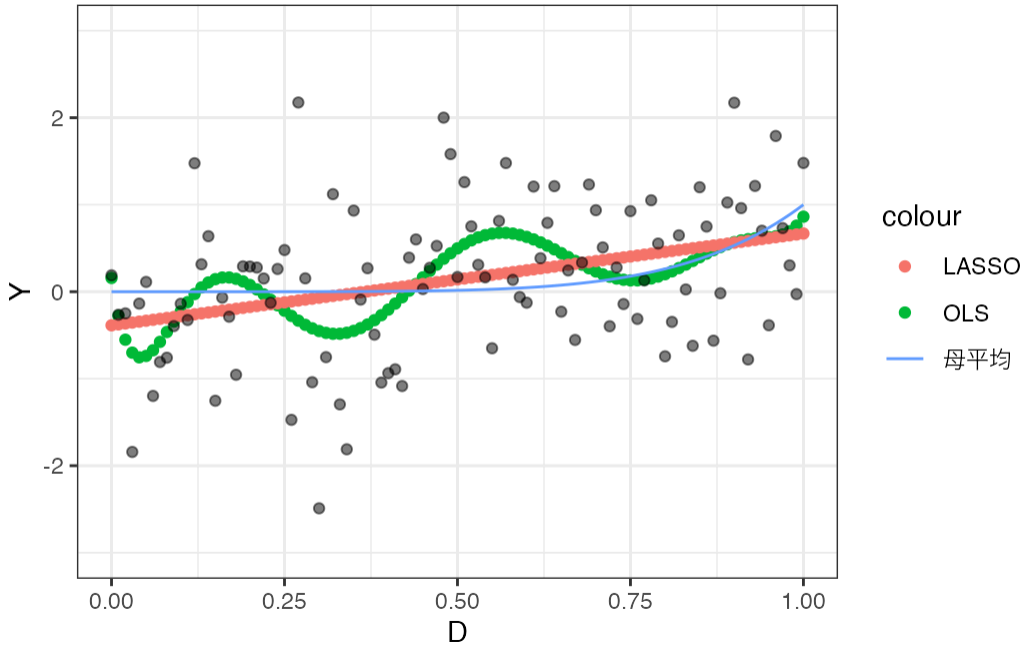
```
Call:
rlasso.formula(formula = Price ~ poly(Size, 2) + poly(Tenure,
    2) + Size:Tenure, data = data)

Coefficients:
     (Intercept)    poly(Size, 2)1    poly(Size, 2)2  poly(Tenure, 2)1
           45.24           2753.65           1196.06           -913.88
poly(Tenure, 2)2       Size:Tenure
          399.30              0.00
```

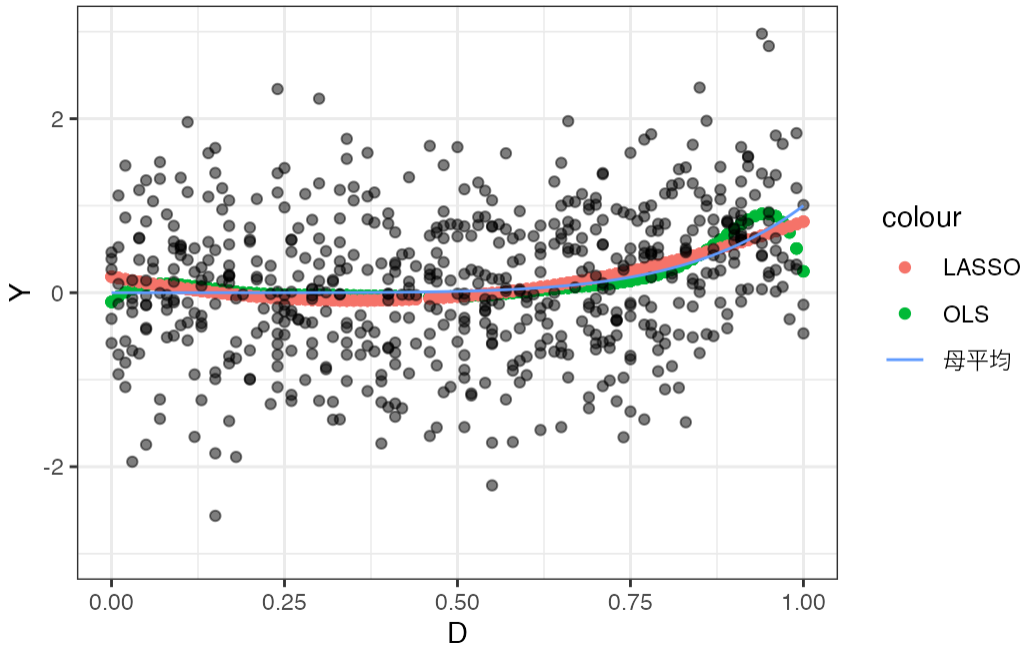
3.4 λ の選び方

- 推定された予測モデルと母平均の乖離を小さくするように設定したい
- いくつかの方法が提案
  - 赤池情報基準の利用 (gamlr)
  - 交差推定の利用 (glmnet)
  - 理論指標の利用 (hdm)
    - 詳細は CausalMLBook 第 3 章 参照

3.5 例: LASSO: 101 事例



3.6 例: LASSO (600 事例)



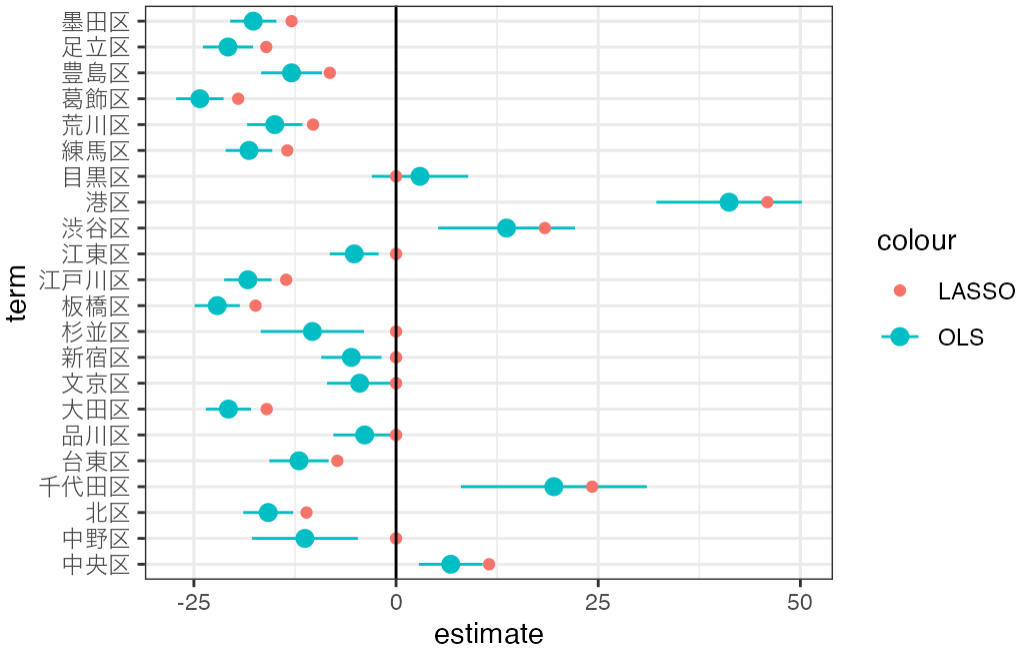
3.7 変数選択

- LASSO 法は、“データ主導の変数選択”を伴う手法であると強調される
  - 係数値 β が厳密に”0”となる変数が生じる
    - 該当する変数が、モデルから”除外された”と解釈できる
  - OLS では生じない

3.8 データ主導の問題点

- 推定されているモデルから除外されている $\neq$ Population OLS においても除外される
  - 限られた事例数に対処するために、“しょうがなく”単純化している
- LASSO や回帰木などの推定値については、信頼区間を計算するのが難しい
  - データ主導のモデル化のせいで、データの特徴と推定結果が複雑な関係性を持ってしまい、推定結果について安定的な性質が成り立ちにくい

3.9 例 $Price \sim 0 + Ditrict$



3.10 Takeaway

- データ主導の線型モデル推定
  - 二乗項や交差項により複雑化したモデルを、データ手動で単純化する
- 予測性能が改善する応用が存在する
  - 信頼区間が計算しにくく、母集団の特徴を推論する手法としては使いにくい
  - 例外は 100 % 近い予測性能を発揮するケースだが、市場/社会分析では望み薄

3.11 Reference

Bibliography