予測問題
機械学習

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

2025-09-09

1 予測問題

1.1 問題設定

- 観察できる情報 $X = [X_1, .., X_L]$ から、欠損している情報 Y を予想するタスク
  - 中古マンションの属性から、市場価格を予想する
  - 中期経営計画、有価証券報告、口コミサイトの情報から、企業の職場環境を予想する

1.2 アイディア

- $X =$ { 大学 } から出身都道府県Y を予想する
- 川田が予想する場合、$X = \{$**武蔵大学**$\}$ であれば、東京 と答える
  - 武蔵大学に通う学生は、東京圏出身者の割合が多いという背景知識を持つため

1.3 シンプルアイディア

- 年齢も予測に活用できる ($X =$ { 大学,年齢 }) から予想する
- もし 22 歳と 50 歳の武蔵大学出身者で、出身地が大きく異なり、それを知っているのであれば、予測を変える

1.4 問題

- 取り組む課題: 信頼できる背景知識がない場合に、どのように予測するか?
  - データから予測”モデル”を推定する

1.5 データ

- 日本人の一定数(例えば 1000 名)について、年齢、出身大学、出身都道府県を調査しデータ化する
  - 機械学習では、“教師”データとも呼ばれる

1.6 予測モデル

- 予測したい事例の年齢、出身大学 $(= X)$ を”代入”すれば、予想都道府県 $(= Y)$ を自動計算してくれるモデル (“計算式)
  - データから、予測モデルを推定する

1.7 まとめ

- 予測問題: 「データから予測モデルをどのように推定するか」が問題
- ある特徴Xを持つ集団のYの特徴を推定することが重要
  - 例: 「最近の武蔵大学出身者は、首都圏出身者が多い」

2 データの要約

2.1 基本アイディア

- 十分な事例数をもつデータであれば、以下が期待できる
  - データの特徴 $\underbrace{\simeq}_{\text{よく似ている}}$ 事例をランダムに抽出した集団(母集団)の特徴
    - ランダムサンプリングデータと呼ばれる

2.2 丸暗記予測モデル

- データ上での平均値を予測値とする方法
- (X 内での) Y の平均値: $X = x$ を満たす事例(例えば 30) $Y_1, .., Y_{30}$ について、以下で計算できる

$$f(x) = \frac{Y_1 + .. + Y_{30}}{\underbrace{30}_{\text{事例数}}}$$

2.3 丸暗記予測モデルの性質

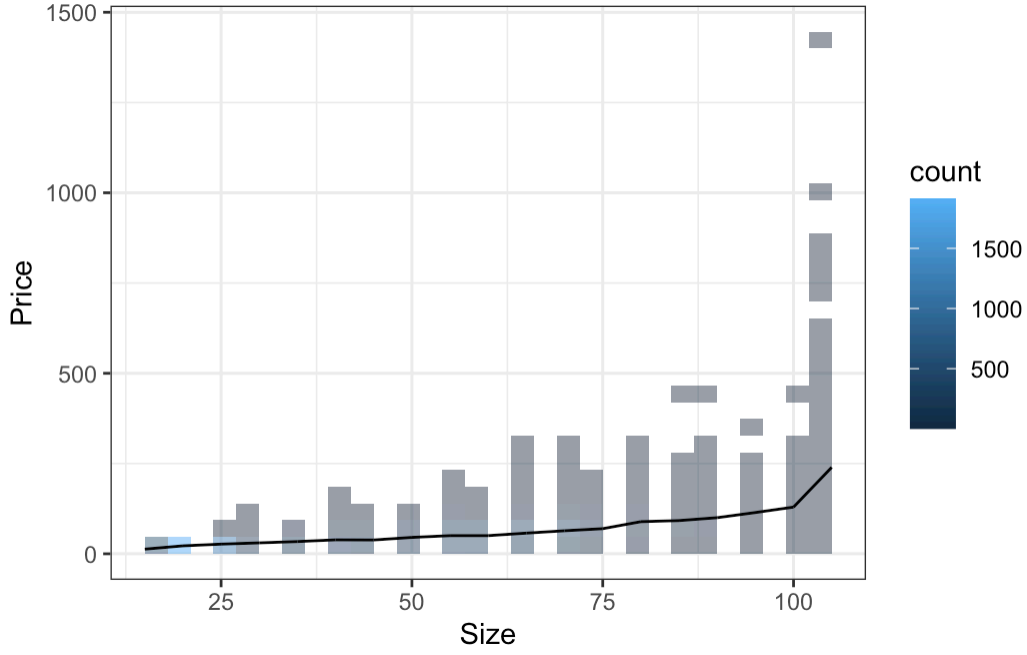
- 後述するように、母集団が予測対象で、事例数が極めて大きければ、
  - データ上の平均値 $=$ 母集団(予測対象) 上での平均値
    - 優れた予測モデル

2.4 実例: $X =$ Size

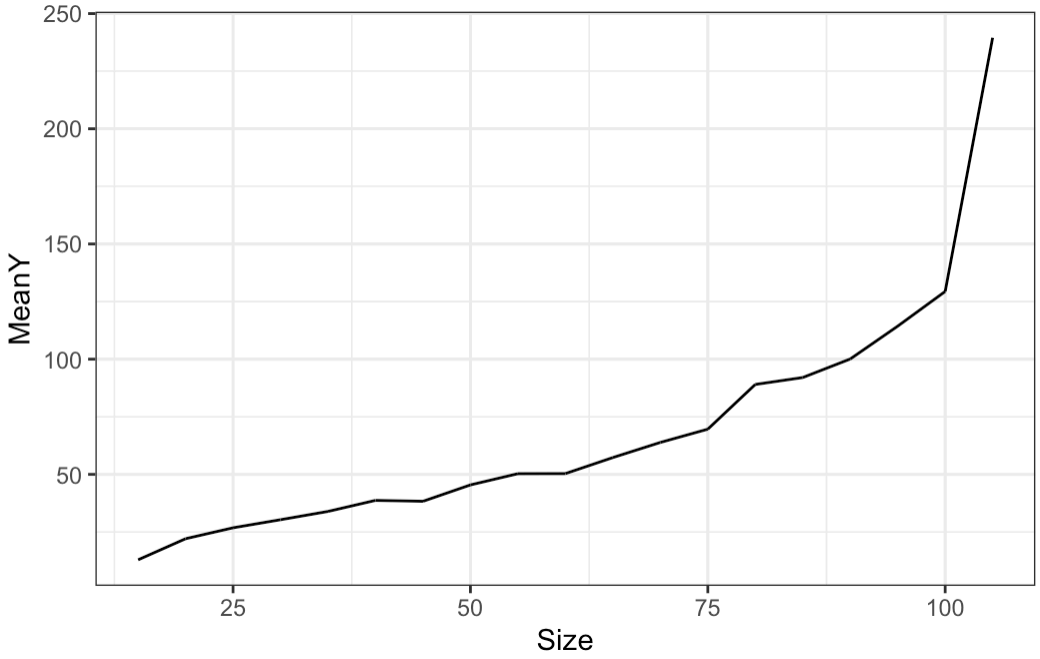
Size	MeanY	N
15	12.9	638
20	22.1	1913
25	26.8	1339

Size	MeanY	N
30	30.3	451
35	33.9	417
40	38.7	599
45	38.3	423
50	45.4	682
55	50.3	910
60	50.3	846

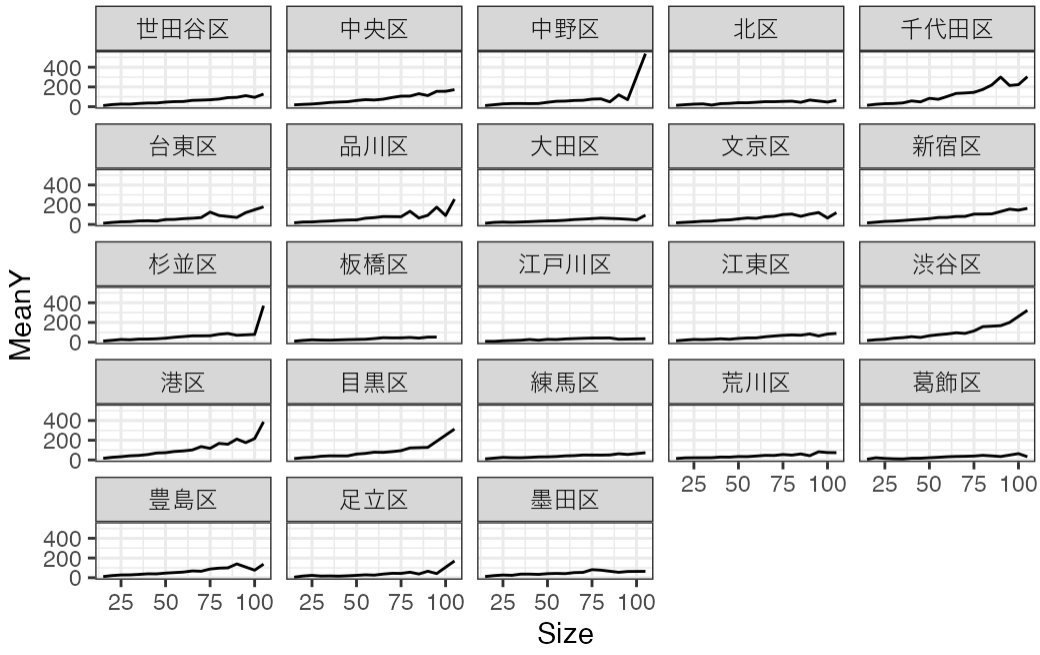
2.5 实例: $X =$ Size



2.6 実例: X = Size



2.7 実例: X = Size/District



2.8 課題

- X の組み合わせが増えると、よりきめの細かい予測ができる

- X の組み合わせが増えると、非常に少数の事例しか存在しないグループが発生する
- 集団とデータの特徴が大きく乖離するリスクが高い

- 例:「部屋の広さ、駅からの距離、築年数、立地する区が一致する事例がない」事例は、全体の 0.837 %

3 線型モデルによる要約

3.1 線型モデル

- 事例が少ないグループへの対処として、平均値そのものではなく、補助線(線型モデル)を推定するアプローチが有力
- 最小二乗法(OLS)で推定できる

3.2 補助線による予測モデル

- 平均値に”補助線”を引く
- 例: 平均値に最も適合する直線を引く: 以下を最小化するように直線の切片β_0と傾きβ_1を決める

$$\left(Y\text{の平均値} - \underbrace{\text{予測値}}_{\beta_0+\beta_1\times Size} \right)^2 \text{の総和}$$

3.3 実例

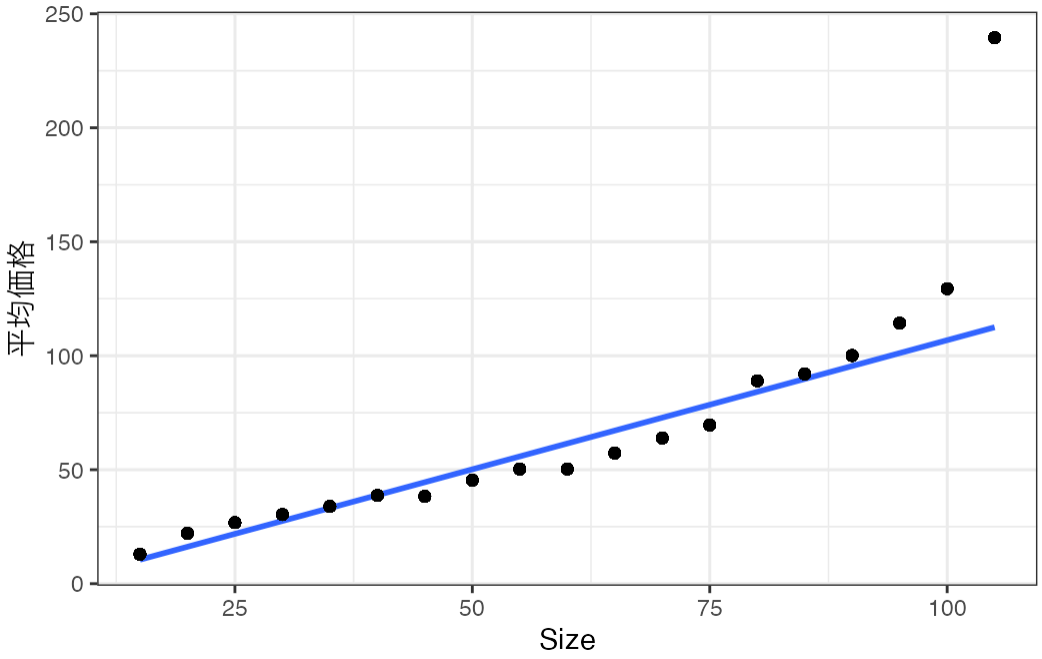
```
lm(Price ~ Size,
   Data)
```

```

Call:
lm(formula = Price ~ Size, data = Data)

Coefficients:
(Intercept)         Size
     -6.463        1.133
```

3.4 実例



3.5 実例

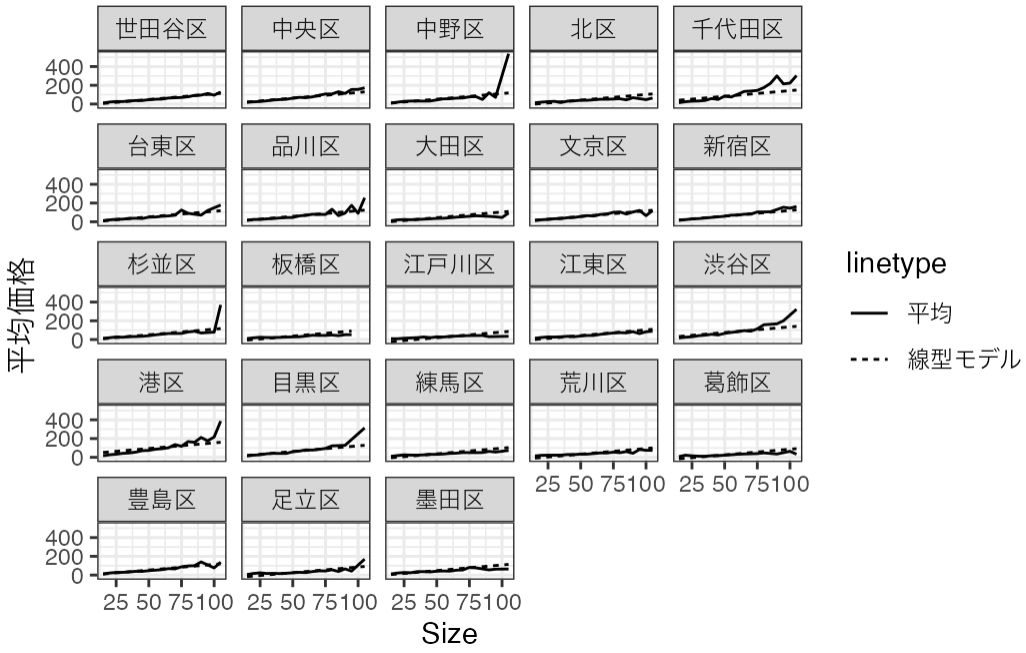
```
lm(Price ~ Tenure + District,
   Data)
```

```

Call:
lm(formula = Price ~ Tenure + District, data = Data)

Coefficients:
     (Intercept)            Tenure    District中央区    District中野区
         66.3344           -0.6465            3.1178          -11.4065
    District北区  District千代田区    District台東区    District品川区
        -18.8800           18.4564          -15.9634           -5.4329
  District大田区    District文京区    District新宿区    District杉並区
        -20.9095           -4.9296           -6.5910          -10.0386
  District板橋区  District江戸川区    District江東区    District渋谷区
        -22.7377          -17.3206           -6.7586           13.6454
    District港区    District目黒区    District練馬区    District荒川区
         39.0324            4.5394          -19.1794          -15.7669
  District葛飾区    District豊島区    District足立区    District墨田区
        -24.1394          -14.4339          -20.7257          -21.9710
```

3.6 実例



3.7 OLS: 曲線

- 平均値に最も適合する”曲線”を引くこともできる:
  - 例: 以下を最小化するように補助線を決める

$$\left(Yの平均値 - \underbrace{予測値}_{\beta_0+\beta_1 Size+\beta_2 Size^2} \right)^2 の総和$$

  - 例: 以下を最小化するように補助線を決める

$$\left(Yの平均値 - \underbrace{予測値}_{\beta_0+\beta_1 Size+..+\beta_4 Size^4} \right)^2 の総和$$

3.8 実例

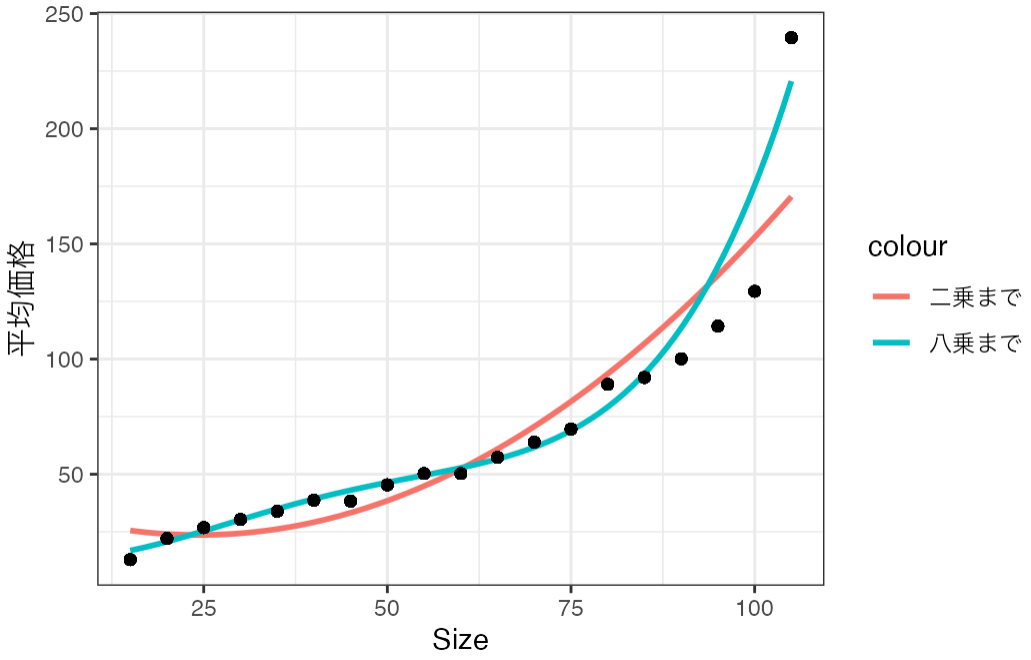
```
lm(Price ~ poly(Size,4), # 4乗まで加える
   Data)
```

```

Call:
lm(formula = Price ~ poly(Size, 4), data = Data)
```

```
Coefficients:
   (Intercept)  poly(Size, 4)1  poly(Size, 4)2  poly(Size, 4)3  poly(Size,
4)4
         45.24         2703.44         1216.89          896.92
315.60
```

3.9 実例



3.10 実例

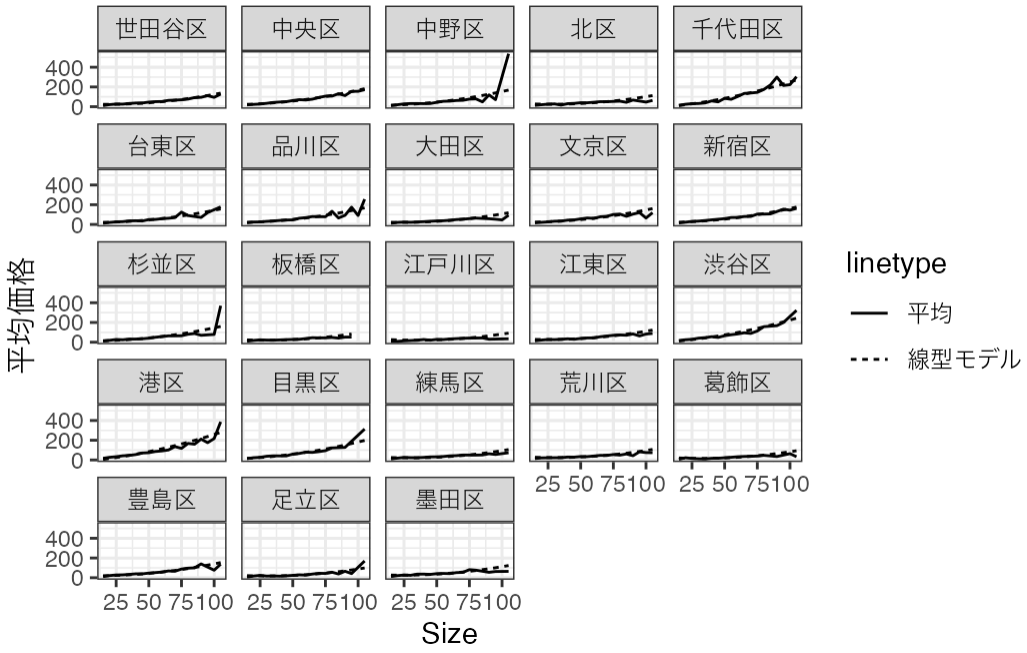
```
lm(Price ~ (Size + District)^2 + I(Size^2),
   Data)
```

```
Call:
lm(formula = Price ~ (Size + District)^2 + I(Size^2), data = Data)

Coefficients:
          (Intercept)                   Size         District中央区
              27.49616              -0.47996              -7.48756
       District中野区          District北区       District千代田区
             -14.89608               8.33559             -35.78432
       District台東区         District品川区         District大田区
              -5.98287              -6.93843               1.26406
```

```
      District文京区         District新宿区         District杉並区
            -3.26512               -7.32800              -13.87600
      District板橋区       District江戸川区         District江東区
             5.29723               10.27034                7.43258
      District渋谷区           District港区         District目黒区
           -30.84930              -44.07787              -19.04409
      District練馬区         District荒川区         District葛飾区
             6.20795                9.40464                3.29372
      District豊島区         District足立区         District墨田区
            -6.43399                3.45692                5.99473
           I(Size^2)    Size:District中央区    Size:District中野区
             0.01455                0.49868                0.43727
   Size:District北区  Size:District千代田区    Size:District台東区
            -0.31719                1.61221                0.24465
  Size:District品川区    Size:District大田区    Size:District文京区
             0.36019               -0.19477                0.25263
  Size:District新宿区    Size:District杉並区    Size:District板橋区
             0.43748                0.34405               -0.37274
Size:District江戸川区    Size:District江東区    Size:District渋谷区
            -0.53284               -0.20967                1.30389
    Size:District港区    Size:District目黒区    Size:District練馬区
             1.77377                0.75656               -0.36752
  Size:District荒川区    Size:District葛飾区    Size:District豊島区
            -0.38278               -0.45032                0.21566
  Size:District足立区    Size:District墨田区
            -0.40075               -0.19018
```

3.11 実例



3.12 複雑なモデルの問題点

- 少数事例が持つデータ上の特徴を反映した補助線が引かれる
  - 極端な特徴を持つ事例であれば、集団の特徴からは乖離する
- 非常に複雑なモデル = 平均値と同じ予測をもたらす
  - 補助線を用いる意味がなくなる

4 予測モデルの性能評価

4.1 性能評価の重要性

- 予測モデルを実務に実装する前に、その予測精度を測定する必要がある
  - どんな予測であったとしても、まぐれあたりはする
  - 安定的な予測性能を測定したい

4.2 理想の性能テスト

- 評価用の新規事例を大量に入手できれば、理想的なテストが可能
  - X からYを予測するモデルをデータから推定し、**新しい追加事例を収集し**どの程度当たるか確かめる
- もし可能であれば、代表的な評価指標を計算すれば良い。例えば二乗誤差

$(Y - \text{予測値})^2$の新しいデータについての平均

- 評価用の新しい事例を収集するのは難しい

4.3 望ましくないテスト

- 新しい事例を用いずに、テストできないか?
- 「モデルを推定した事例を、テストにも再利用」したくなるが、間違えた方法
  - 予測ではなく、“確認”であり、過度に高い評価になってしまう
- 有名な警句: 「Double dipping (２度漬け) には注意」

4.4 例

- ２事例のみからなる(しょぼい)データから予測モデルを推定する

Y	X
香川県	武蔵大学
大阪府	東京大学

- $f(\text{武蔵大学}) = \text{香川県}$ と予測するモデルを作る
  - 直感的に予測性能は低い

4.5 例: 新しい事例によるテスト

- 武蔵大学の学生から新しく 10 事例を収集し、モデルをテストすると

X	Y	予測値
武蔵大学	東京都	香川県
武蔵大学	東京都	香川県
武蔵大学	東京都	香川県
武蔵大学	東京都	香川県
武蔵大学	千葉県	香川県

- まったく当てはまらないことがわかる

4.6 例: 同じ事例によるテスト

- 同じ事例に当てはめると

X	Y	予測値
武蔵大学	香川県	香川県

- 一見完璧に当てはまるが、予測ではなく、“確認”しているだけ

4.7 データ分割によるテスト

- データを２分割 (訓練/テスト) にランダムに分割する
  - 訓練: 予測モデルを推定する
  - テスト: 予測性能を評価する

4.8 実例

Price	Size	District	OLS	Error: OLS
28.0	20	新宿区	55	729.00
150.0	75	文京区	51	9801.00
43.0	55	品川区	45	4.00
33.0	40	品川区	45	144.00
70.0	55	目黒区	45	625.00
30.0	25	目黒区	43	169.00
29.0	30	目黒区	43	196.00
48.0	60	豊島区	41	49.00
6.5	15	板橋区	21	210.25
30.0	60	足立区	31	1.00
24.0	80	葛飾区	29	25.00

4.9 実例

```
set.seed(11)

Group = sample(1:2, nrow(Data), replace = TRUE) # データの分割

FitOLS = lm(
  Price ~ Tenure + District,
  Data,
  subset = Group == 1) # OLSモデルの推定

FitMean = lm(
```

```
  Price ~ 1,
  Data,
  subset = Group == 1) # 平均値の推定

mean((Data$Price - predict(FitOLS,Data))[Group == 2]^2) # OLSのテスト
```

```
[1] 1805.888
```

```
mean((Data$Price - predict(FitMean,Data))[Group == 2]^2) # 平均値のテスト
```

```
[1] 2109.792
```

4.10 実例

- 平均値の方が、OLS よりも予測力が低い
  - 事例数が少なく、集団の傾向との乖離が大きい

4.11 Takeaway

- データの持つ煩雑な情報をモデルに集約し、予測に活用
  - 理論的にも望ましい性質を持つ(次回)
- モデルの予測性能を評価するためには、新しい事例が必要
  - 典型的なアプローチは、事前にデータを一部残しておく