予測問題
機械学習

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

2025-09-16

1 予測問題

1.1 目標

- 「データをモデルに要約する」をイメージとして掴む
  - 研究者主導アプローチの代表格である、OLS を紹介
- 正しい予測モデル評価法を学ぶ
  - 複雑なモデルが必ずしも高性能ではないことを確認

1.2 ロールプレイ

- 大手不動産会社にて、中古マンションの買取査定支援”AI”を開発しろ、と命じられた
  - 重要な情報は、当該物件の市場価格
    - 価格は、マンションの持つさまざまな属性(広さ、立地、築年数等)によって異なるため、予測困難
  - これらの属性から、取引価格を予測するモデルを構築したい

1.3 予測

- 日常的な予測: 自身が経験した/見聞きした事例の傾向から、なんとなく予想する
  - 港区の広い物件は、高い価格である傾向等
- データに基づく予測: 大規模なデータ上での傾向を要約し、予測モデル(“AI”)を推定
  - 国土交通省が提供する 不動産情報ライブラリ からデータを入手

1.4 データ

Price	Size	District	Tenure	Distance
94	40	千代田区	3	3
100	65	千代田区	12	4
130	65	千代田区	21	4
98	65	千代田区	16	4
58	40	千代田区	7	3

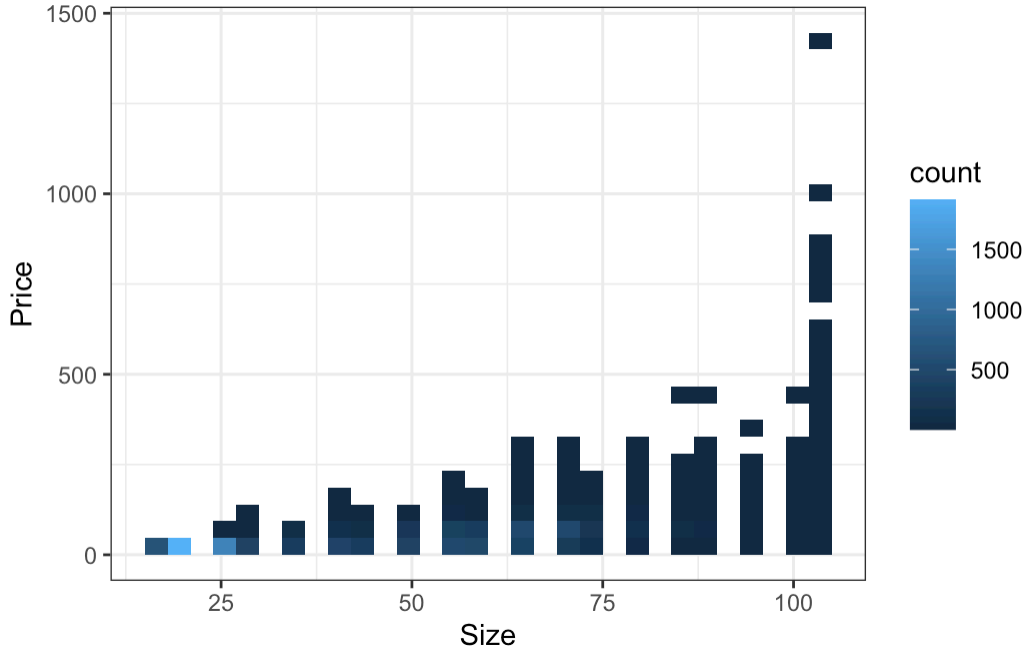
- Size = 部屋の広さ、Tenure = 築年数、Distance = 駅からの距離(分)

2 予測モデル

2.1 単純な予測モデル

- とりあえず、部屋の広さ(Size)のみから取引価格を予測するモデルを推定する
  - 予測モデル $= f(Size)$
    - Size を入力すれば、予測取引価格を自動計算してくれる

2.2 部屋の広さと取引価格の傾向

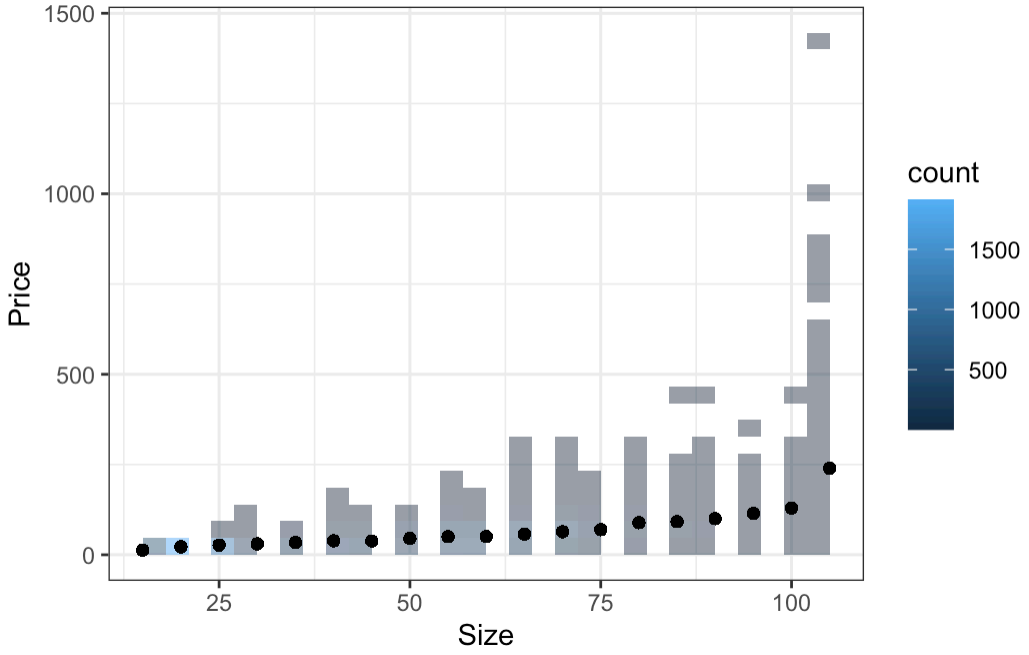


2.3 情報の要約

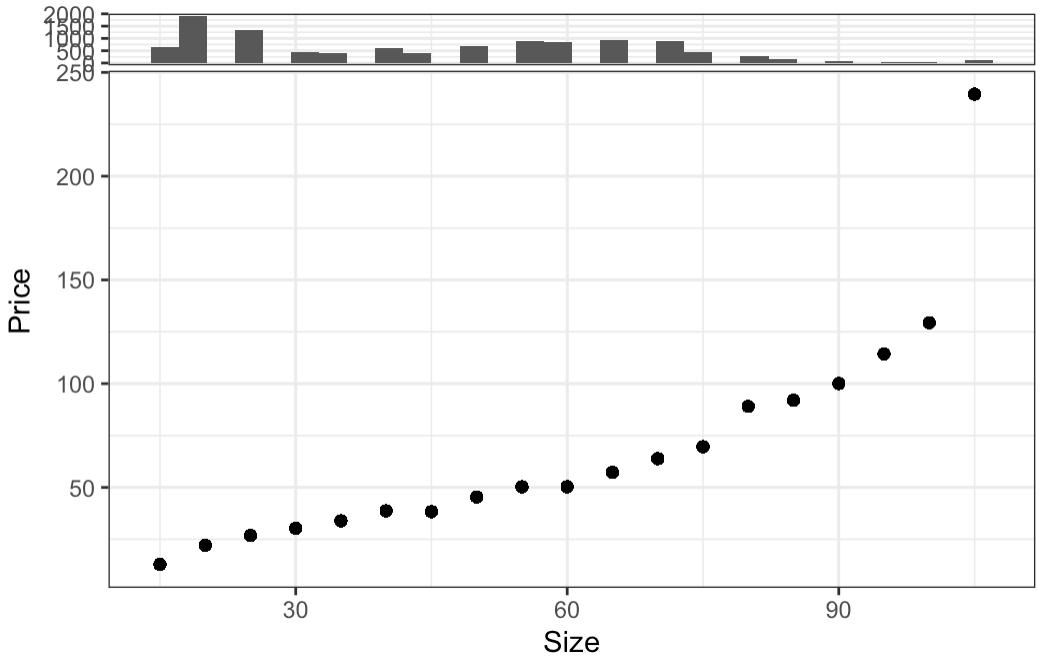
- 全く同じ Size でも、取引価格は大きく異なる

- Size 以外の要因が取引価格を左右
  - 最大値を予測価格とすると、他の要因が上振れた事例から予測しており、一般に不適切
    - 例: 男性の予測値
      - = 大谷翔平選手….?
- 代表的な方法は平均値

2.4 部屋の広さと平均取引価格の傾向



2.5 部屋の広さと平均取引価格の傾向



2.6 小規模事例の平均

- 平均値の計算に用いる事例が少なければ、観察できない要因の上振れ/下振れの影響を受けやすい
  - 岩手県水沢市出身 31 歳男性の事例は、大谷翔平選手のみ
  - 岩手県水沢市出身 31 歳男性の所得の平均値
    - ≃ 大谷翔平選手の所得….?

2.7 モデル化

- データの特徴を”大雑把に”捉える
- 代表例は線型モデル

$$Y \simeq \beta_0 + \beta_1 X_1 + ..$$

  - 極力データに適合するように、β の値を選ぶ

2.8 例

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

```

Call:
```

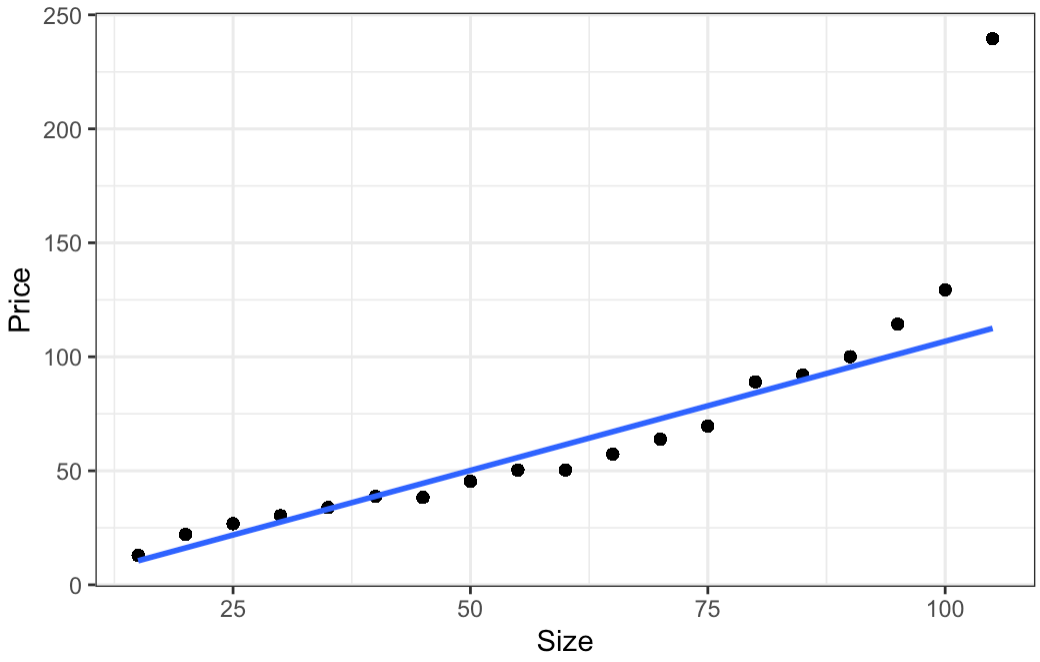
```
lm(formula = Price ~ Size, data = Data)

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

- Size $= 50$ の物件の予測取引価格は、

$$-6.463 + 1.133 \times \underbrace{Size}_{=50} \simeq 50.2$$

2.9 例



2.10 曲線モデル

- 直線以外を当てはめることも容易
- $$Y \simeq \beta_0 + \beta_1 X_1 + \beta_2 \underbrace{X_2}_{X_1^2}$$

2.11 例

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

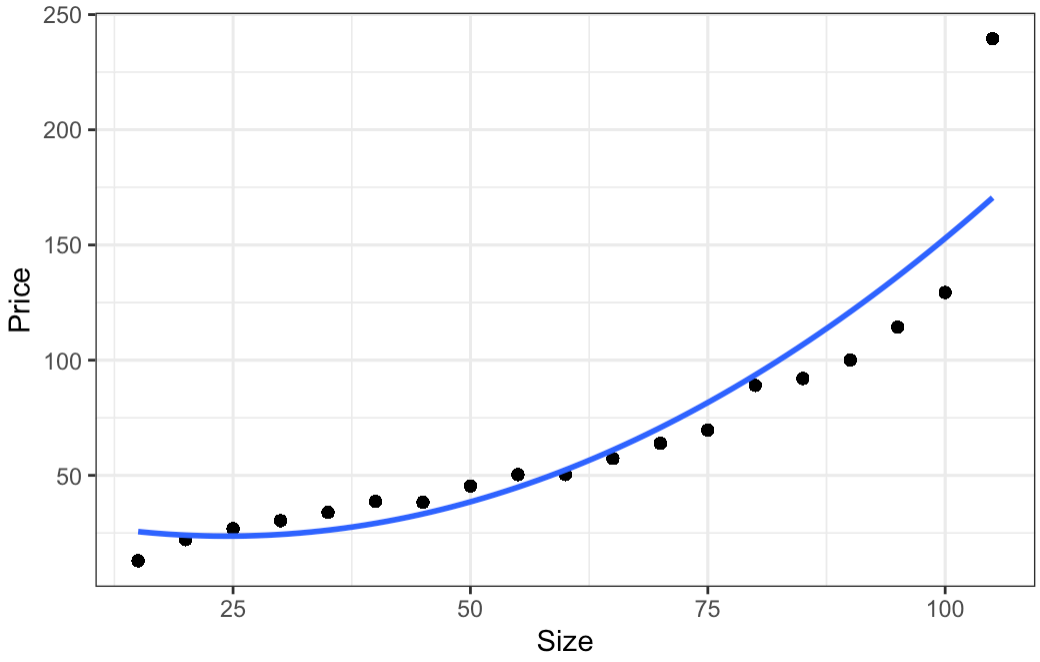
```

Call:
```

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

Coefficients:
(Intercept)         Size    I(Size^2)
   36.94330     -1.09684      0.02256
```

2.12 例



2.13 重回帰

- 複数の属性も容易に導入できる
  - 平均値の場合、事例数がより少なくなり、非現実的
- 例

$$Price \simeq$$

$$\beta_0 + \beta_1 Size + \beta_2 Tenure + ... + \beta_3 \text{千代田} + ..$$

  - 千代田: 千代田区であれば 1、それ以外であれば 0

2.14 例

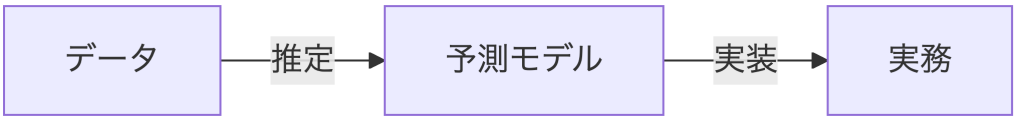
```
lm(Price ~ Size + Tenure + District, Data)
```

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

Coefficients:
     (Intercept)              Size            Tenure    District中央区
          2.6467            1.2127           -0.6799           11.1456
  District中野区      District北区  District千代田区    District台東区
          3.7414           -9.9308           33.2732           -0.1924
  District品川区    District大田区    District文京区    District新宿区
          8.1725           -4.4987            8.0578           11.4954
  District杉並区    District板橋区  District江戸川区    District江東区
          2.8550          -12.4476          -27.5717           -7.2599
  District渋谷区      District港区    District目黒区    District練馬区
         26.5507           43.4248           15.3891          -12.1884
  District荒川区    District葛飾区    District豊島区    District足立区
        -15.5613          -22.1682            3.6266          -22.6756
  District墨田区
         -5.1848
```

2.15 Takeaway

- 研究者が指定した線型モデルに、大量の属性情報を集約する
  - OLS は、データに当てはまるようにモデルを推定する
  - 性能の良いモデルであれば、実務に実装できる



3 予測モデルの性能評価

3.1 推定と評価

- 予測モデルは、どの程度機能するのか?
  - 実務上極めて重要
- 事後評価: 実際に実装し業務に活用しながら、確かめる
  - 予測に失敗した場合の被害が軽微な場合は活用可能

3.2 事後評価

1. 予測モデルを推定
2. 実際の事例に応用し、実際の価格 (例: 70)と予測価格 (例: 60) を新規に取集

3. 実際と予測価格の乖離を測定: 典型的には

$$(\textbf{実際の価格} - \textbf{予測価格})^2 = 100$$

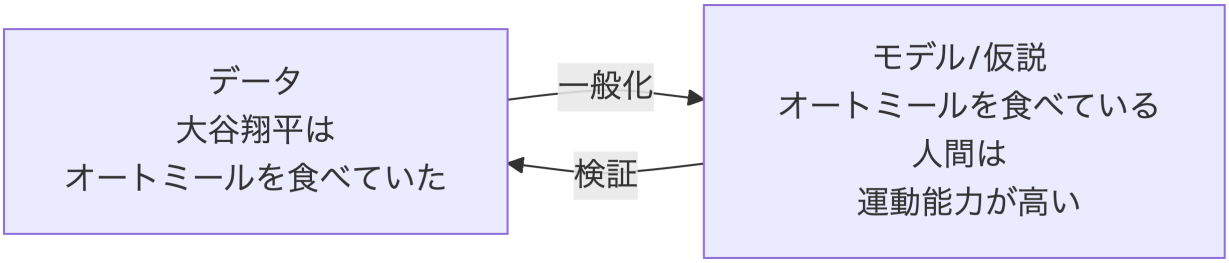
3.3 事前評価

- 実務に実装する前に、その予測精度を測定したい
- 課題は、
  - 予測の評価に用いるための、新規の事例が存在しない

3.4 不適切な事前評価

- 「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」

3.5 例: 自己啓発本的レトリック



3.6 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

出身地 Y	所属大学 X
香川県	**武蔵大学**
大阪府	**東京大学**

- $f(\textbf{武蔵大学}) = \textbf{香川県}$ と予測モデルを推定
  - 直感的に予測性能は低い

3.7 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

所属大学 X	出身地 Y	予測値
武蔵大学	**東京都**	**香川県**

所属大学 X	出身地 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	千葉県	香川県

- まったく当てはまらないことがわかる

3.8 例: 同じ事例によるテスト

- 同じ事例に当てはめると

所属大学 X	出身地 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

3.9 推奨される事前評価

- データを２分割 (訓練/テスト) にランダム分割する
  - 訓練: 予測モデルを推定する
  - テスト: 予測性能を評価する
- まぐれあたりによる過大/過小評価を避けるために、テストにも十分な事例数を割く必要がある
  - 典型的には２割程度をテストに配分する

3.10 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

Price	Size	District	OLS	Error: OLS
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

3.11 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
```

3.12 実例

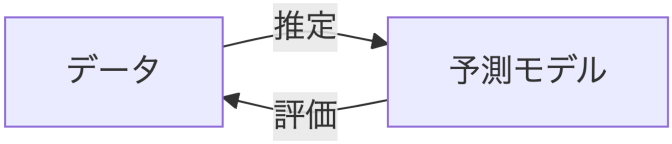
- 平均値の方が、OLS よりも予測力が低い
  - 事例数が少なく、集団の傾向との乖離が大きい

3.13 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

3.14 不適切な事前評価



3.15 適切な事前評価