予測問題

機械学習

川田恵介 (keisukekawata@iss.u-tokyo.ac.jp)

Table of contents

1 予測問題

- 観察できる情報 $X = [X_1, .., X_L]$ から、欠損情報 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

1.1 例. 出身都道府県予測

- $X =$ { 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、東京圏出身者が多いという背景知識を持つため

1.2 例. 出身都道府県予測

- 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、川田は予測を変える

1.3 例. 出身都道府県予測

- 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
- 日本人の一定数 (例えば 1000 名) から、年齢、出身大学、出身都道府県を調査しデータ化する
  - “教師” データ
- 教師データから、予測モデル f(年齢､出身大学) を推定する
  - 予測したい事例の年齢、出身大学を” 代入” すれば、予想都道府県を自動計算してくれるモデル

1.4 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」問題
- ある特徴 X を持つ集団の Y の平均的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: 平均

- データ分析の基本アイディア: ある集団について、十分な事例数をもつデータであれば、以下が成り立つと期待できる
  - 集団の特徴 $\underbrace{\simeq}_{\text{よく似ている}}$ データの特徴
  - データを要約し、その特徴を抽出する

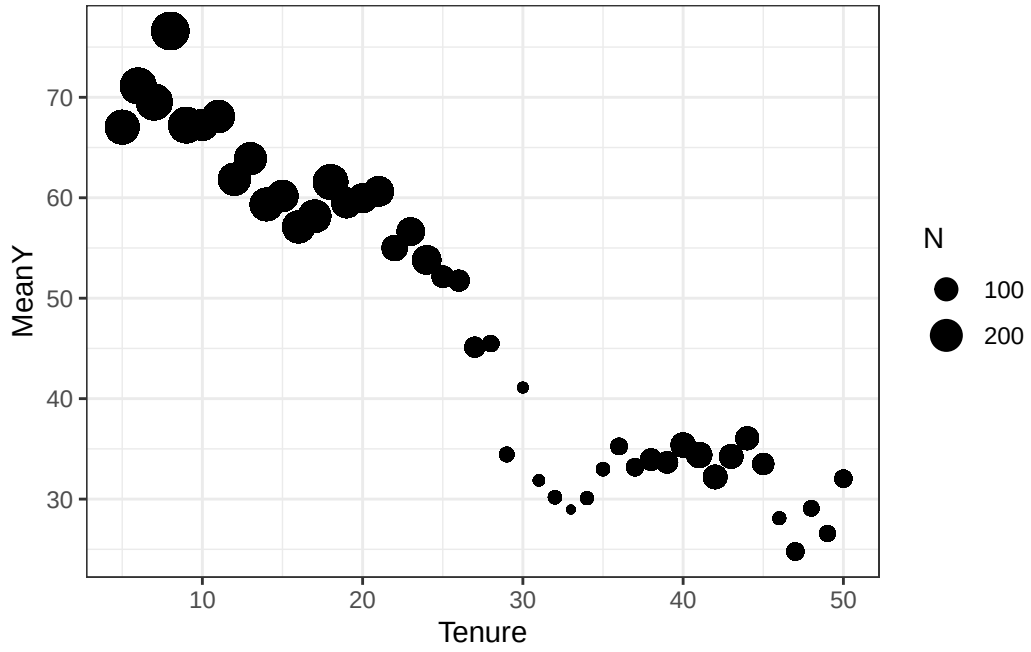
2.1 丸暗記予測モデル

- 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

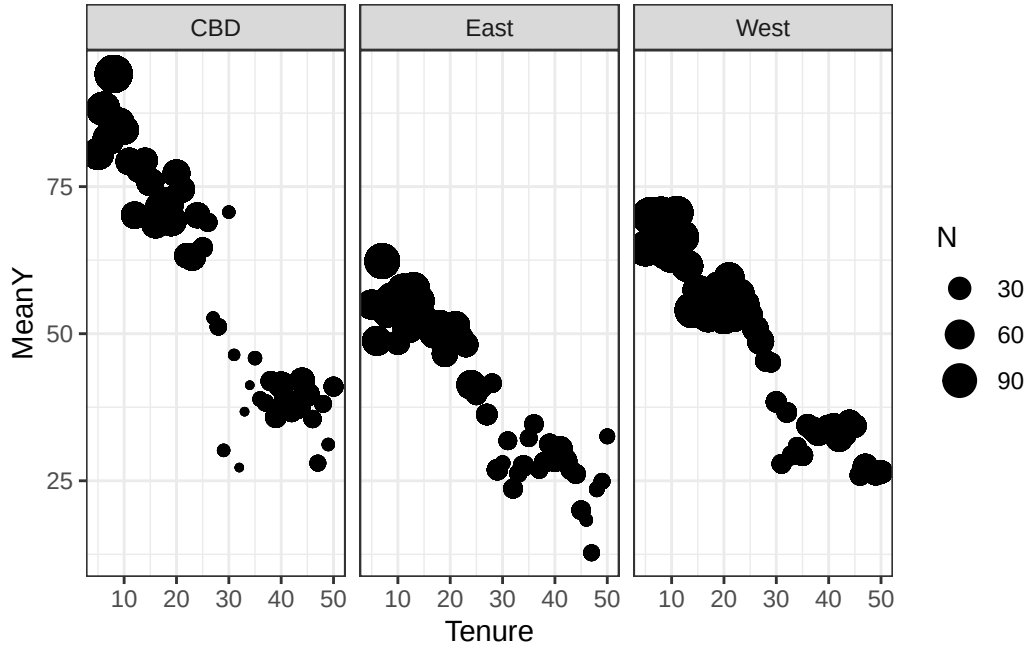
$$f(x) = \frac{Y_1 + .. + Y_{30}}{\underbrace{30}_{\text{事例数}}}$$

- 後述するように、全ての X について、事例数が十分に大きければ (例えば 10000)、極めて優れた予測モデル

2.2 実例: $X =$ Tenure



2.3 実例: $X =$ Tenure/Area



2.4 問題点

- 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: 30 歳の武蔵大学出身者について、1 名しか調査できない
  - 偶然” 香川県出身者” を調査すると、$f($ 武蔵大学,30$) =$ 香川という予測モデルを推定しまう
    - 調査事例が増えると、極端な事例の影響は低減できる

3 データの要約: 補助線

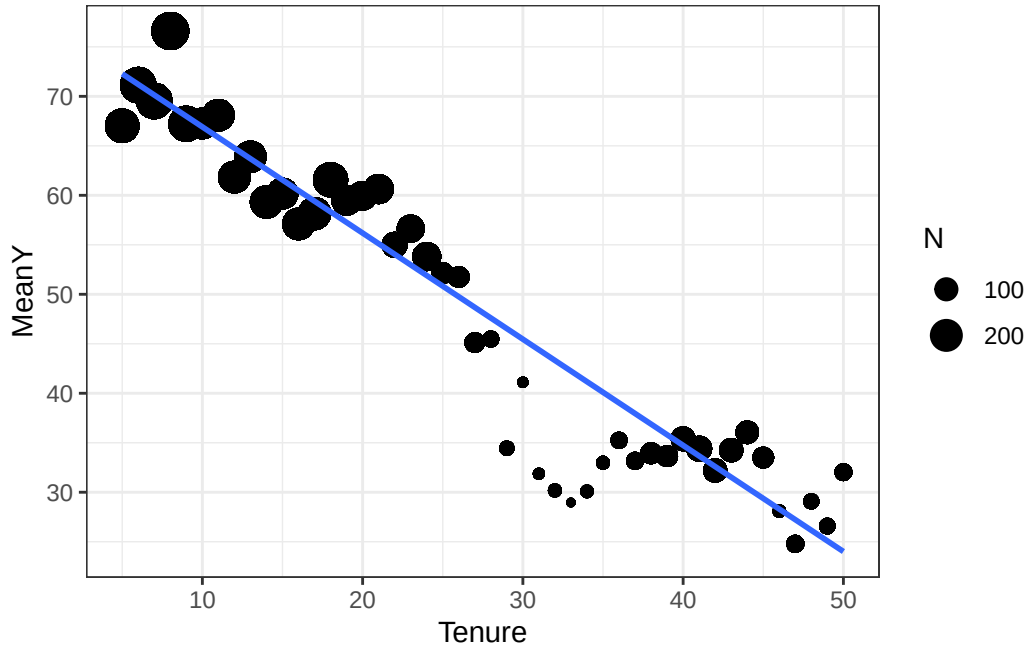
- 事例が少ないグループへの対処として、平均値そのものではなく、補助線 (線型モデル) を推定するアプローチが有力
- 最小二乗法 (OLS) で推定できる

3.1 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

$$(Y\text{の平均値} - \underbrace{\quad\text{直線}\quad}_{\beta_0+\beta_1 Size})^2\text{の総和}$$

3.2 実例



3.3 OLS: 重回帰

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

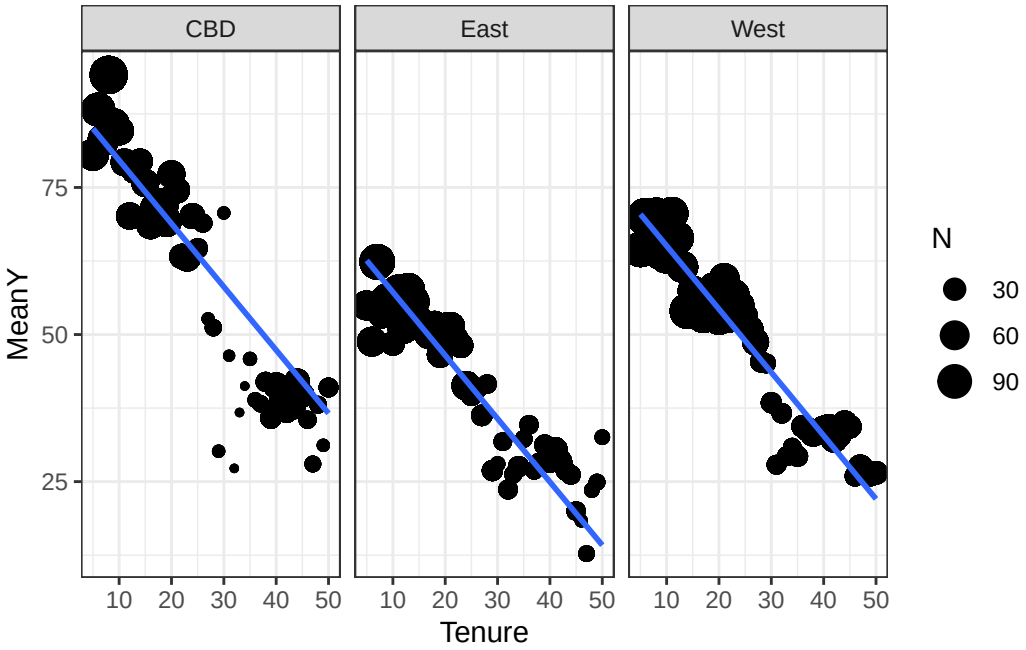
$$(Y\text{の平均値} - \underbrace{\text{補助線}}_{\beta_0+\beta_1 Size+\beta_2 Size^2+\beta Area})^2\text{の総和}$$

3.4 実例

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

Coefficients:
(Intercept)         Size     AreaEast     AreaWest
      3.965        1.145      -31.493      -21.194
```

3.5 実例



3.6 OLS: 曲線

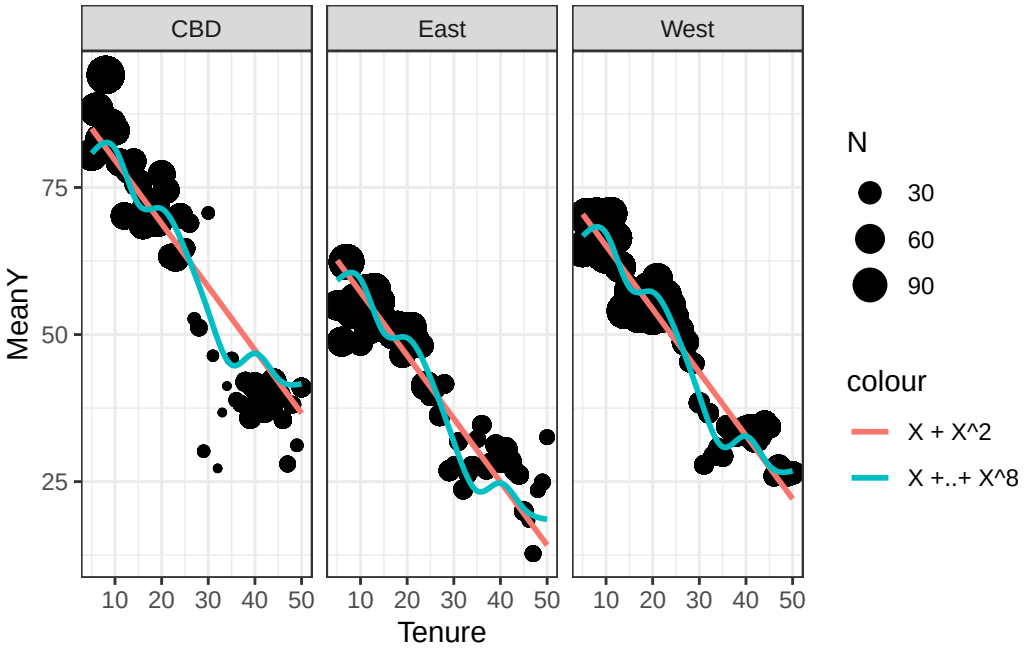
- 平均値に最も適合する” 曲線” を引く:
  - 例: 以下を最小化するように補助線を決める

$$(Yの平均値 - \underbrace{\text{補助線}}_{\beta_0+\beta_1 Size+\beta_2 Size^2})^2の総和$$

  - 例: 以下を最小化するように補助線を決める

$$(Yの平均値 - \underbrace{\text{補助線}}_{\beta_0+\beta_1 Size+..+\beta_8 Size^8})^2の総和$$

3.7 実例



4 予測モデルの性能評価

- 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、まぐれあたりはしうるので、平均的に上手くいくかどうかを測定したい

4.1 理想のテスト

- 評価用の事例を新しく (大量に) 入手できれば、理想的なテストができる
  - 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

$$(Y - f(X))^2 \text{の新しいデータについての平均}$$

  - 評価用の新しい事例を収集するのは難しい

4.2 望ましくないテスト

- 新しい事例なしでテストできないか?
- 「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
  - 「非常に高い予測性能を持つ」と誤って評価してしまう
- 有名な警句:「Double dipping (２度漬け) には注意」

4.3 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(武蔵大学) = 香川県, f(東京大学) = 大阪府 と予測するモデルを作る
  - 直感的に予測性能は低い

4.4 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

4.5 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

4.6 重大な注意点: 仮説創設と検証の分離

- “仮説を作る際に用いた事例は、仮説の検証に使用できない”
- 伝記、インタビュー、自己啓発本を読む際にも注意

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 例

Price	Size	District	OLS	Mean	Error: OLS	Error: Mean
27	55	江東区	32	44	25	289
55	70	江東区	69	81	196	676
64	75	板橋区	53	108	121	1936
49	55	足立区	57	28	64	441
29	65	足立区	34	38	25	81
30	60	足立区	28	36	4	36
55	85	葛飾区	36	35	361	400

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 ~ factor(Tenure)*factor(District),
```

```
  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] 559.1068
```

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

```
[1] 628.7327
```

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、集団の傾向とおおきおおき

4.11 Reference