

予測問題

機械学習

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

Table of contents

1	Getting Started	2
1.1	予測問題	2
1.2	アイディア	3
1.3	データ分割による評価	3
1.4	例	3
1.5	例	3
1.6	例	3
1.7	例	4
1.8	イメージ	5
1.9	典型的誤解	5
2	母集団を用いた整理	5
2.1	母集団	5
2.2	例	6
2.3	イメージ	6
2.4	数値例: データ	7
2.5	数値例: 予測モデル	7
2.6	数値例: 予測モデル	8
2.7	数値例: 予測モデルと母平均	8
2.8	数値例: Sampling Uncertainly	9
2.9	数値例: 伝統的アプローチ	9
3	モデルの評価	10
3.1	予測モデルの評価: 理想編	10
3.2	予測モデルの評価: 現実編	10
3.3	イメージ	11
3.4	補論: モデルを使い続けられるか?	11

4	極端なモデルの推定	11
4.1	理想の予測モデル	11
4.2	完璧な予測は可能か?	12
4.3	予測モデルの推定: 現実	12
4.4	現実的な方法: 単純平均法	12
4.5	現実的な方法: 丸暗記法	13
4.6	複雑なモデルは望ましいのか?	13
4.7	補論: 丸暗記が有効な場合	13
5	中庸なモデルの推定	14
5.1	数値例: Depth 1	14
5.2	数値例: Depth 3	15
5.3	数値例: Depth 6	15
5.4	数値例: Depth 30	16
5.5	モデル選択/集計	16
5.6	概要	16
5.7	イメージ: Bagging	17
5.8	イメージ: RandomForest	17
5.9	イメージ: RandomForest	18
5.10	RandomForest の利点	19
5.11	数値例: Random Forest	19
5.12	Big data は優れた予測モデルを保証するか?	19
5.13	数値例: 伝統的アプローチ (5000 事例)	20
5.14	数値例: Random Forest with large sample (5000 事例)	21
5.15	まとめ	21
5.16	補論: 実務への応用	21
5.17	補論: (例) 推薦アルゴリズム構築	22
5.18	補論: 予測性能が低い = 役に立たない?	22
5.19	Reference	22

1 Getting Started

1.1 予測問題

- 定義: 欠損情報 Y を、観察できる変数 X から予測するモデル $g(X)$ を構築する
 - 視聴履歴やいいね数 (X) → 好む未視聴動画 (Y)
 - メールの文名や件名 (X) → 迷惑メール (Y)
 - 事業の内容や財務状況 (X) → デフォルトリスク (Y)

- Y, X が共に観察できるデータを活用できれば、推定方法がある程度確立されている（教師付き学習）

1.2 アイディア

- 過去の $Y - X$ の間の定量的なパターンを、データから抽出し、予測に利用する
 - 「1 平米広い部屋は 100 万円高い価格で取引されがち」など
 - * そのデータ上にのみ偶然現れたパターンに注意する
- 予測の対象となる新しい事例と過去の事例（データ）との間に、一貫性が必要

1.3 データ分割による評価

- 予測モデルの性能をどのように評価するか
 - 実際に実装し、運用すれば、そのうち性能がわかるが...
 - * 遅すぎる場合も多い
- データをランダム 2 分割 (Training/Test) に分割し、Training データのみでモデルを作り、Test で評価する
- 典型的には、Test データについて、平均二乗誤差

$$(Y - \text{予測値})^2 \text{ の平均値}$$

1.4 例

1.5 例

- 単純平均価格モデル

```
ModelMean = lm(Price ~ 1, Train)
PredMean = predict(ModelMean, Test)
```

- 地域別平均価格モデル

```
ModelDistrict = lm(Price ~ District, Train)
PredDistrict = predict(ModelDistrict, Test)
```

1.6 例

```
# A tibble: 20 x 4
  Price District PredDistrict PredMean
```

	<dbl>	<chr>	<dbl>	<dbl>
1	46	足立	29.9	45.7
2	37	葛飾	30.0	45.7
3	51	世田谷	48.5	45.7
4	88	港	93.8	45.7
5	40	江東	50.7	45.7
6	27	世田谷	48.5	45.7
7	29	練馬	37.6	45.7
8	24	品川	41.0	45.7
9	41	台東	36.4	45.7
10	100	港	93.8	45.7
11	21	練馬	37.6	45.7
12	6.7	台東	36.4	45.7
13	38	新宿	49.7	45.7
14	26	板橋	31.6	45.7
15	22	中央	53.0	45.7
16	70	荒川	40.8	45.7
17	20	足立	29.9	45.7
18	38	江戸川	39.8	45.7
19	29	世田谷	48.5	45.7
20	38	渋谷	56.2	45.7

1.7 例

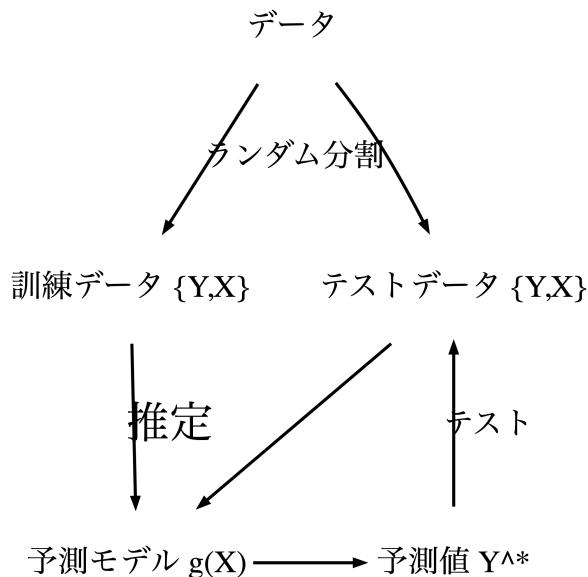
```
mean((Test$Price - PredMean)^2)
```

```
[1] 1317.481
```

```
mean((Test$Price - PredDistrict)^2)
```

```
[1] 1123.986
```

1.8 イメージ



1.9 典型的誤解

- 機械学習を用いて推定すれば、予測性能が高いモデルが獲得できる
- 複雑な現象を予測するためには、複雑なモデルの方が望ましい
- Big data を用いれば、理想的な予測モデルを推定できる
- 予測性能が悪いモデルは、役に立たない

2 母集団を用いた整理

- 母集団: 論点整理を目的とした概念

2.1 母集団

- 手元のデータに含まれる事例は、母分布(集団)から抽出されたと想定する
 - 本講義では、ランダムに選ばれるケースに集中する

* Random Sampling

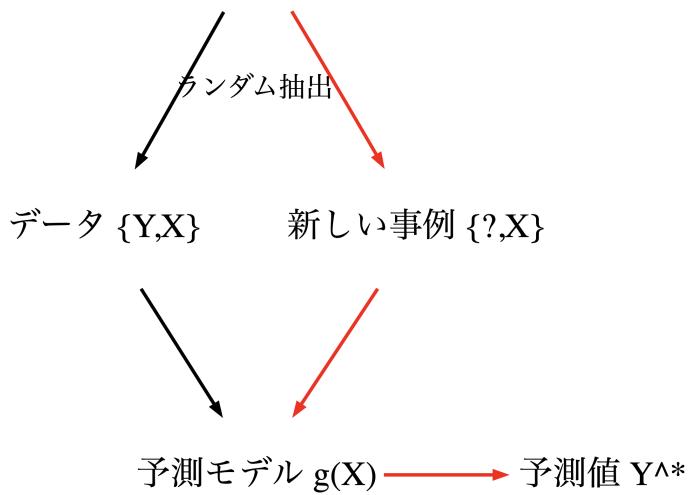
- 予測対象も、母分布からランダムに選ばれるとする
- 一般にデータと母分布は、完全一致しない

2.2 例

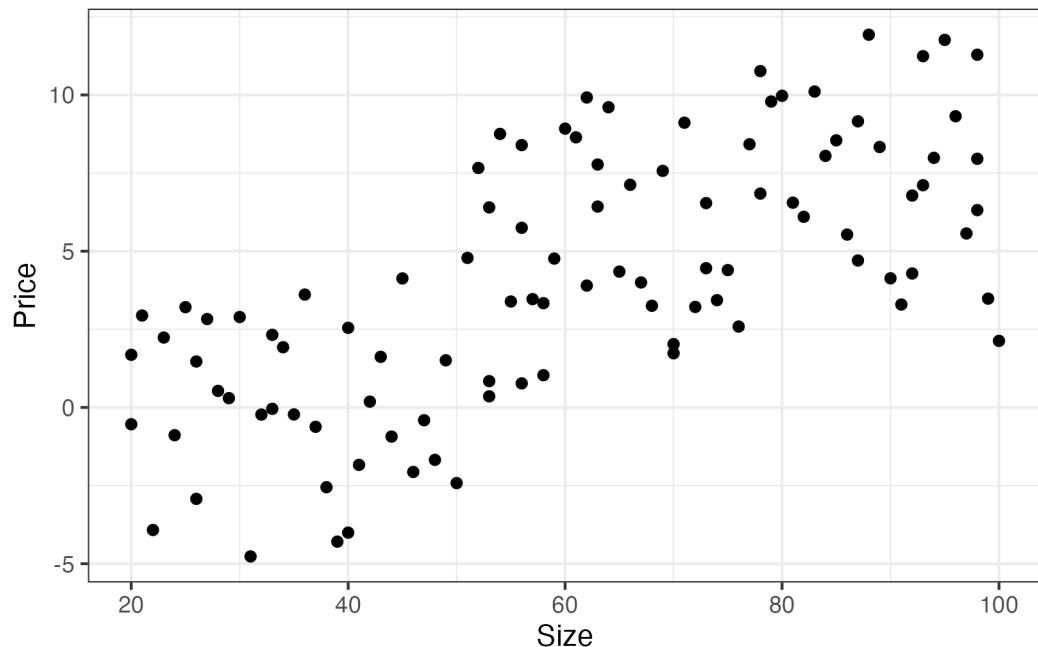
- ユーザーの需要調査 (どのようなサービスを望むか)
 - データ: サービス利用者の中からランダムに選ばれた 200 名 (例: 20 代が 2 割)
 - * 母集団: 全サービス利用者 (例: 20 代が 2.2 割)
 - * 予測対象: X (年齢など) のみ判明している利用者
- 不動産取引価格の予測モデル
 - データ: 2022 年第二四半期のすべての取引事例
 - * 母集団: 潜在的な全取引
 - * 予測対象: まだ取引されていない物件

2.3 イメージ

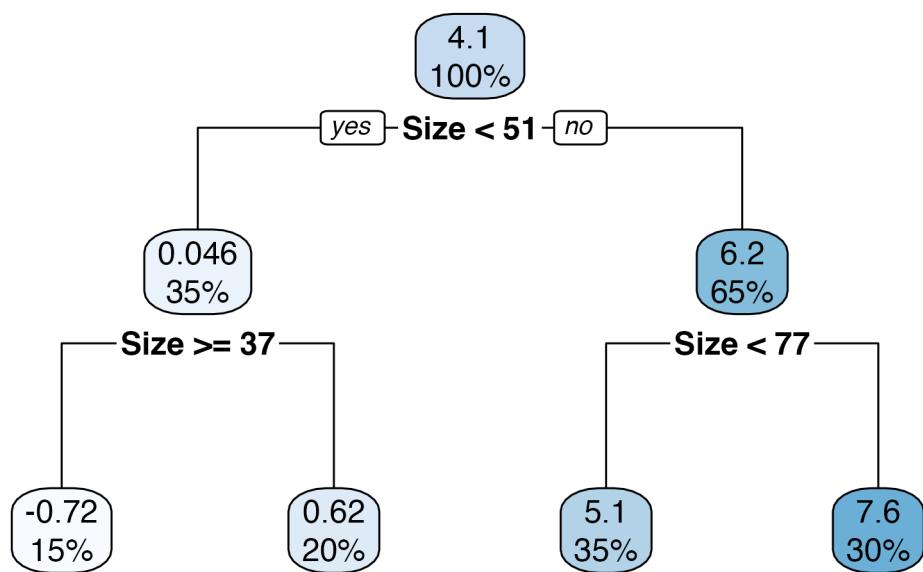
母集団: 無数の事例 $\{Y, X\}$



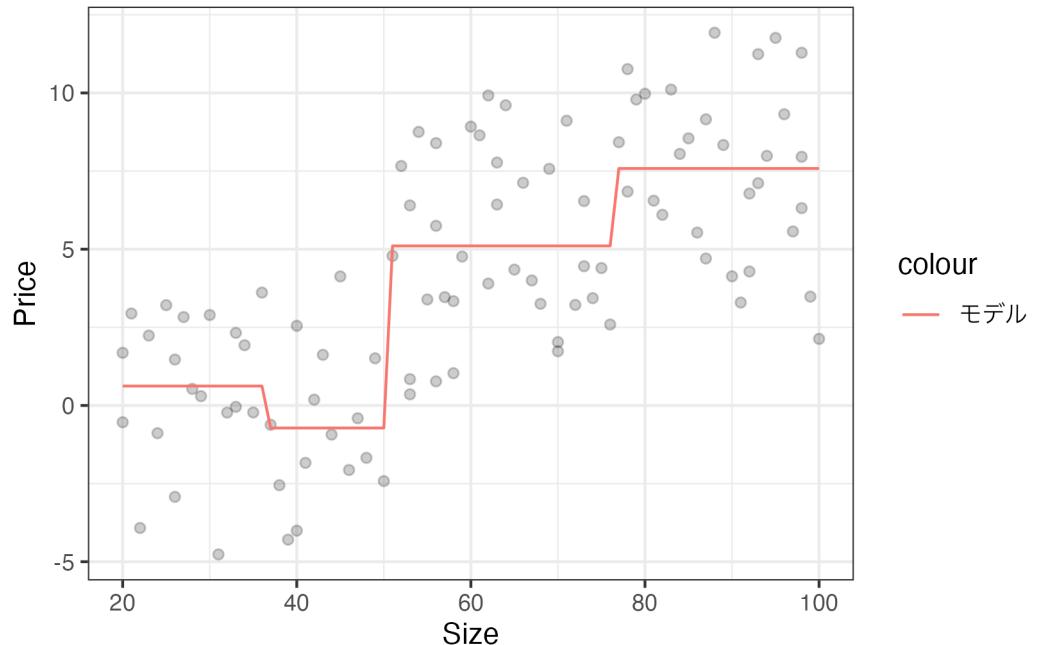
2.4 数値例: データ



2.5 数値例: 予測モデル

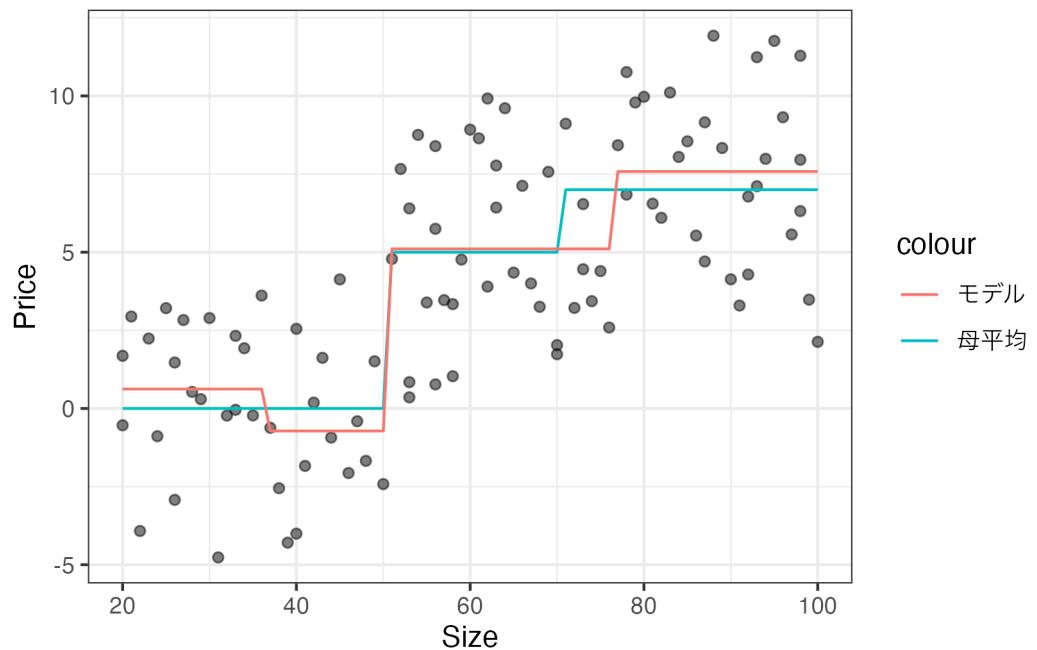


2.6 数値例: 予測モデル



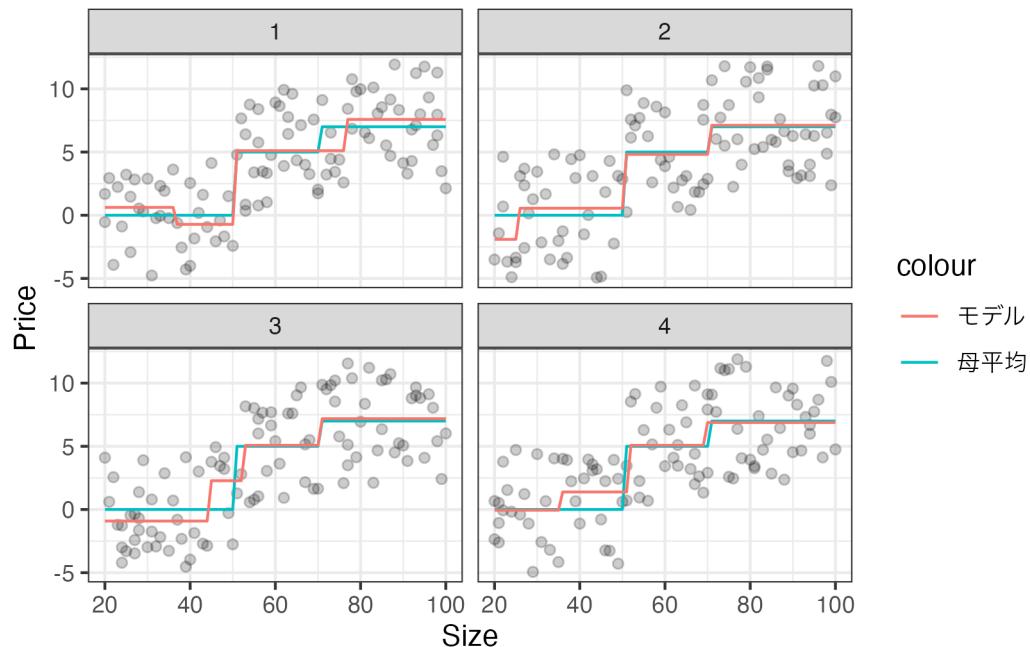
2.7 数値例: 予測モデルと母平均

- ここから想像の世界



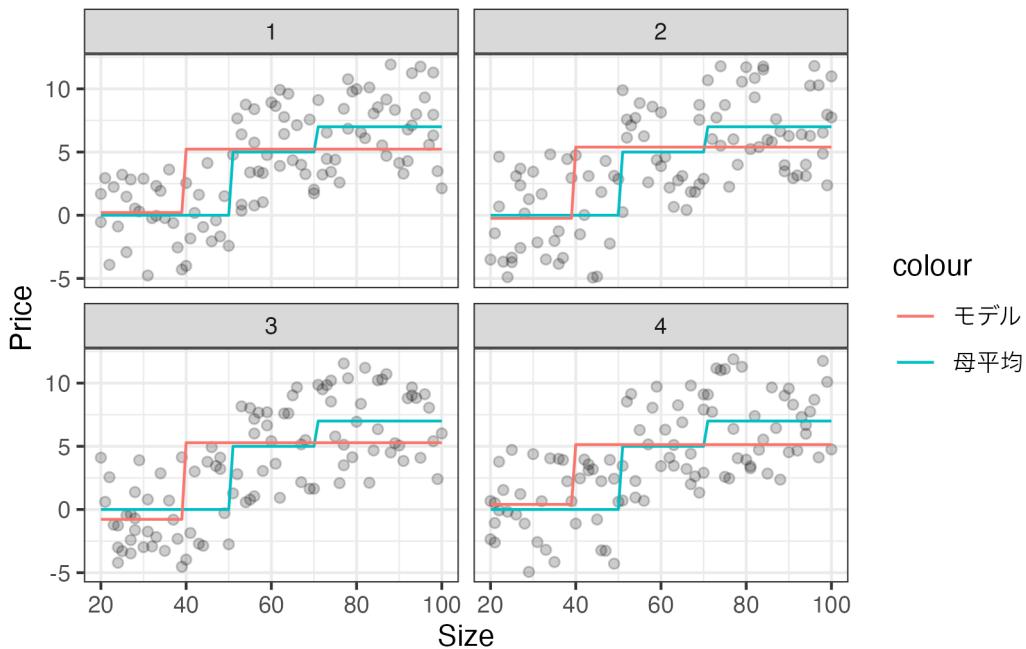
2.8 数値例: Sampling Uncertainty

- 4名の分析者 (ID 1-4) が独立してデータを収集し、モデルを推定したとすると、



2.9 数値例: 伝統的アプローチ

- Size = 40 で分割



- データとは無関係に、size = 40 で予測値が変化する

3 モデルの評価

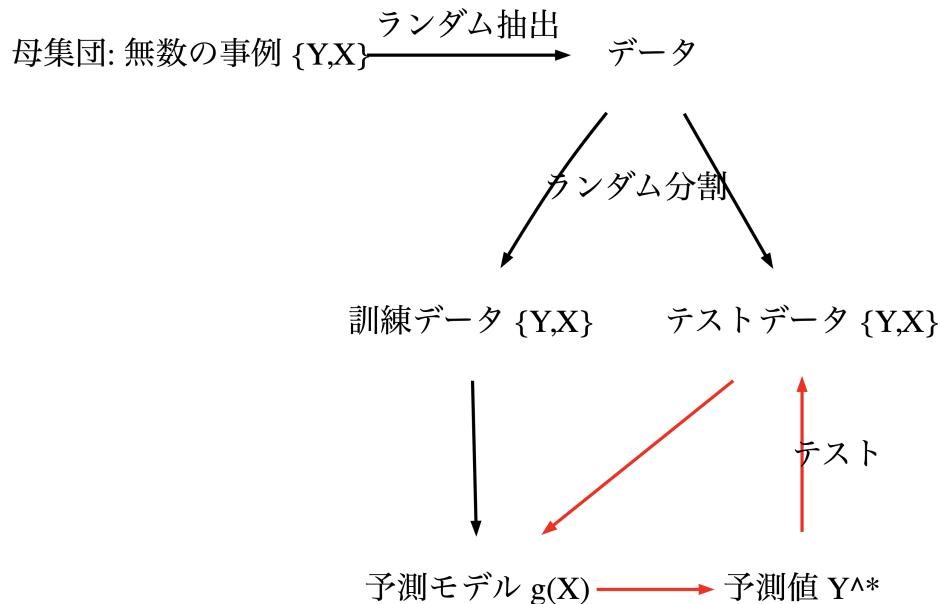
3.1 予測モデルの評価: 理想編

- どんな予測もまぐれあたりしうる
 - 平均的にうまく行くモデルを採用したい
- 予測モデルを母分布に適用して評価する: 典型的には二乗誤差の母平均値を用いる $E[(Y - g(X))^2]$

3.2 予測モデルの評価: 現実編

- 母分布は観察できないので、母分布に適用した際の予測性能は”推定する必要”がある
 - 母分布からランダムに選ばれた事例でテストしたい
- 最も典型的な評価法は、データ分割: データをランダムに2分割(訓練/テストデータ)し、訓練データのみでモデルを推定し、テストデータで評価

3.3 イメージ



3.4 補論: モデルを使い続けられるか?

- モデル推定に用いるデータの母分布と、予測対象の母分布が異なれば、予測性能は悪化する
 - 高性能モデルでも、時代の変化/大きな社会的ショックの発生等の結果、性能が悪化しうる
 - * 通常の設備と同様に、経年劣化/破壊されうる
- 定期的に予測性能を計り直し、再推定する必要がある
 - Concept-Drift と呼ばれる

4 極端なモデルの推定

- 予測モデルの推定の原理を理解するために、極端な推定方法で何が生じるのか理解する

4.1 理想の予測モデル

- もし母集団をすべて活用して、モデル推定できるのであれば、理想の予測モデルは母平均 $E[Y|X]$
 - 予測誤差を平均二乗誤差で測定するのが前提

- ・「各グループ内での平均的傾向」よりも優れた予測はあり得ない

4.2 完璧な予測は可能か?

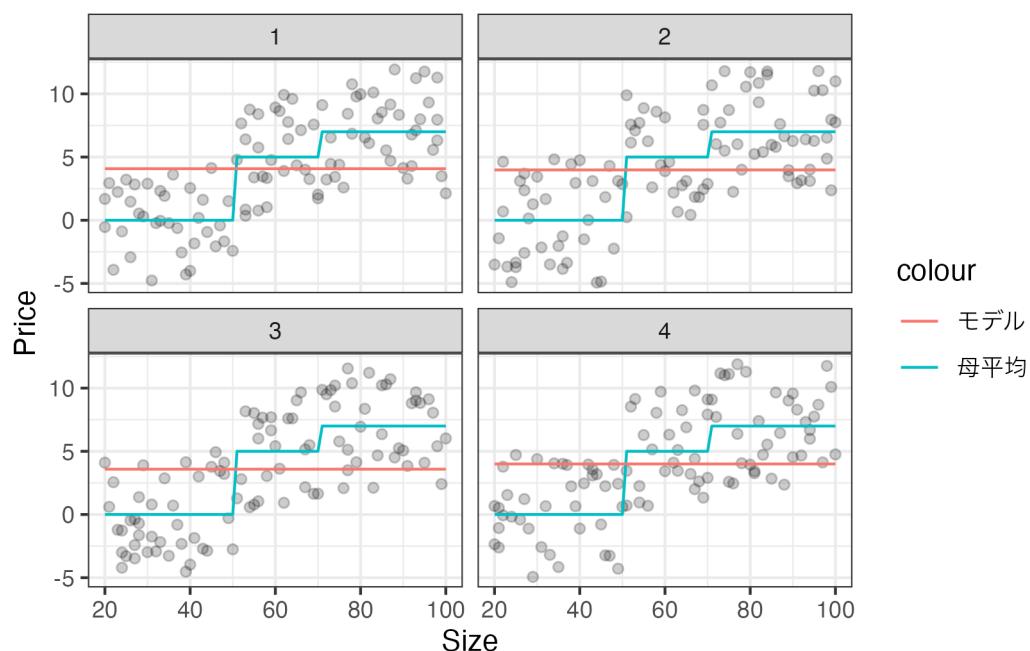
- ・予測不可能な部分 $Y - E[Y|X]$ が発生
 - 削減不可能な誤差
- ・予測対象となる母集団(社会)、 Y/X が決まった時点で自動的に定まる
 - X 内で個人差がある現象の予測は難しい
 - * 人間行動やその相互作用で決まる現象(例: 消費、貯蓄、出生、就業、賃金、等)は大きい傾向

4.3 予測モデルの推定: 現実

- ・母集団は観察できないので、データで置き換える必要がある
 - 単純なモデルを推定すると、母平均の複雑さを捉えられない
 - 複雑なモデルを推定すると、データが偶然持った特徴を母平均の特徴と混同してしまう

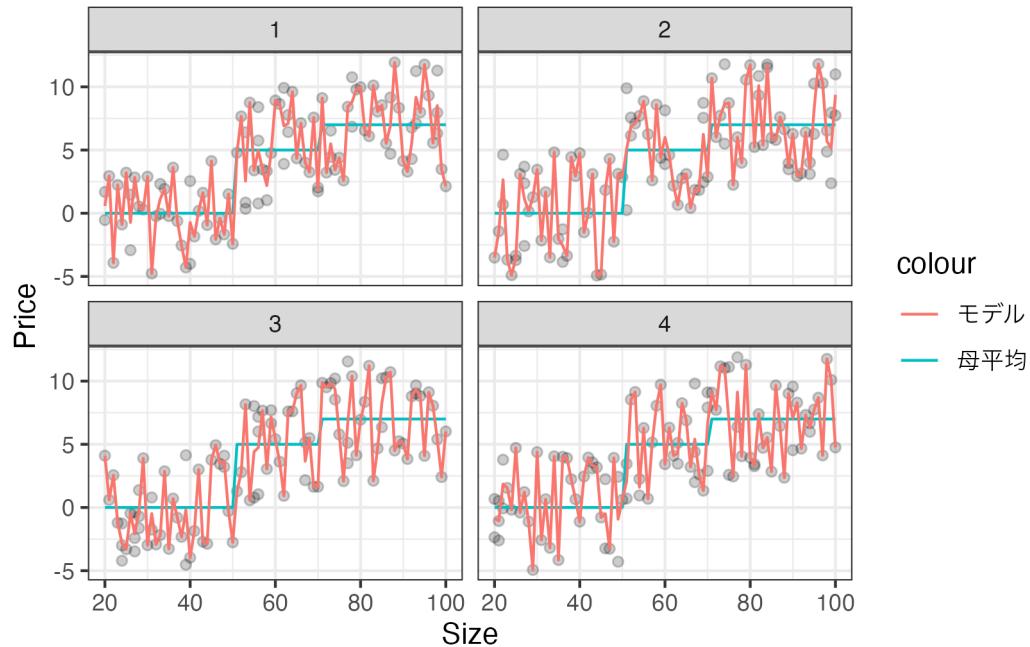
4.4 現実的な方法: 単純平均法

- ・単純な平均取引価格を予測値とする



4.5 現実的な方法: 丸暗記法

- 各 X について、データ上の平均値を予測値とする



4.6 複雑なモデルは望ましいのか?

- 丸暗記モデルは非常に複雑なモデルを生み出し、望ましいように思える
 - 現実は複雑
 - 推定に用いたデータとの矛盾はない
- * 予測値とデータ上の平均値は、必ず一致する
- 一般に予測性能は極めて悪い
 - 少数の事例のみで計算された平均値は、 X 以外の決定要因の偏りを反映してしまう

4.7 補論: 丸暗記が有効な場合

- X 以外に Y の決定要因がなければ、丸暗記は有効
 - 例: 判断予測、コンピュータの動作予測
- 人間行動については、 X 以外の無数の決定要因が存在

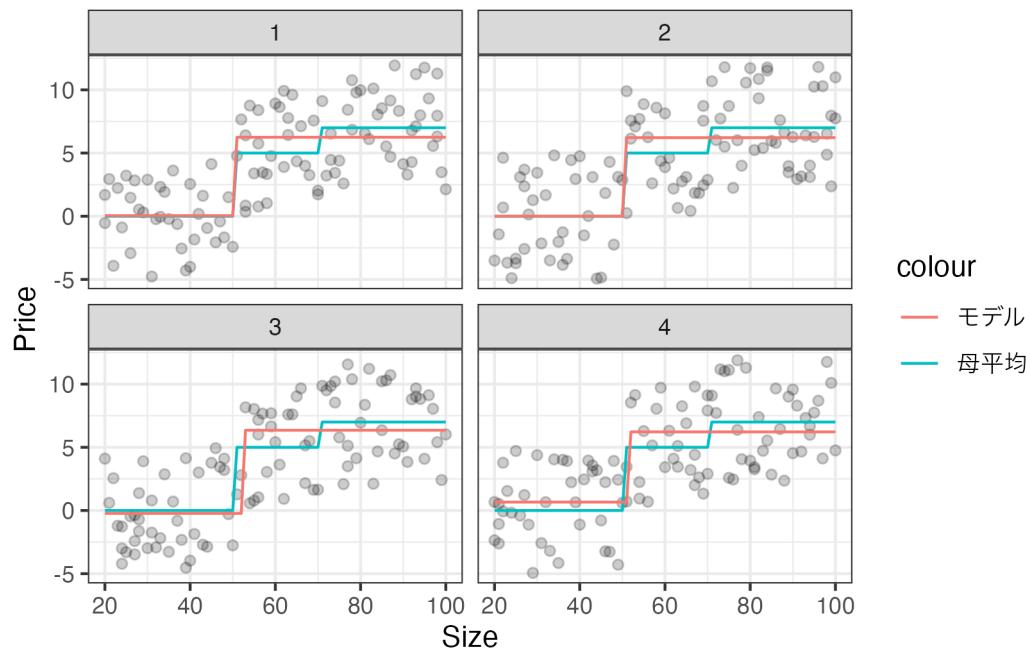
– 例: “双子”でも違う人生を歩む

– 丸暗記に向いていない

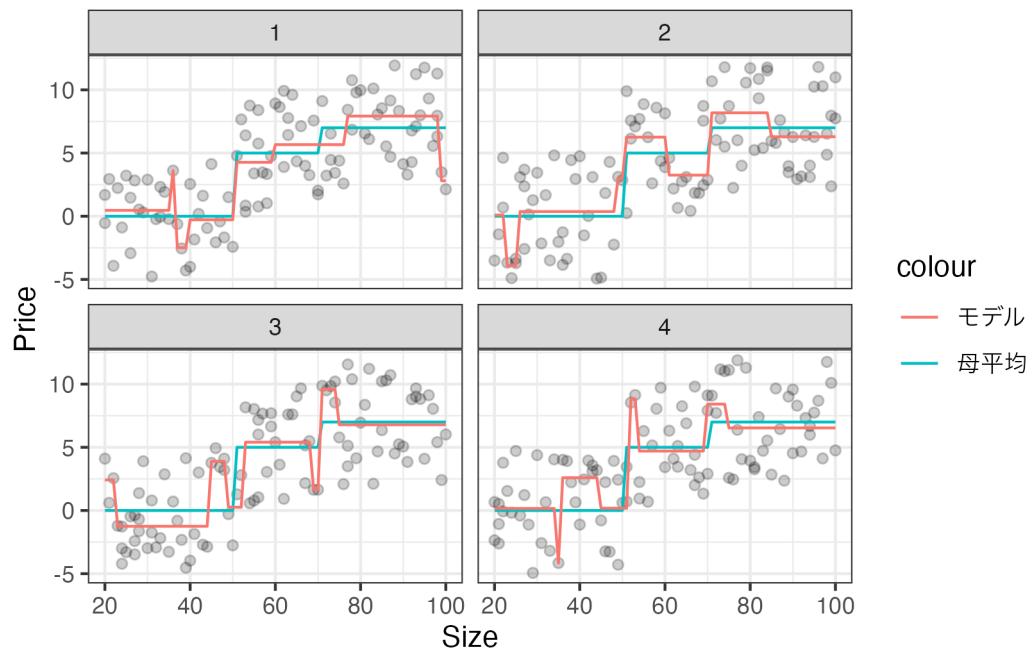
5 中庸なモデルの推定

- 多くの応用で平均値は単純すぎ、丸暗記は複雑すぎる
- 決定木については、最大分割回数などを変更することで、単純平均と丸暗記法の間を移行できる

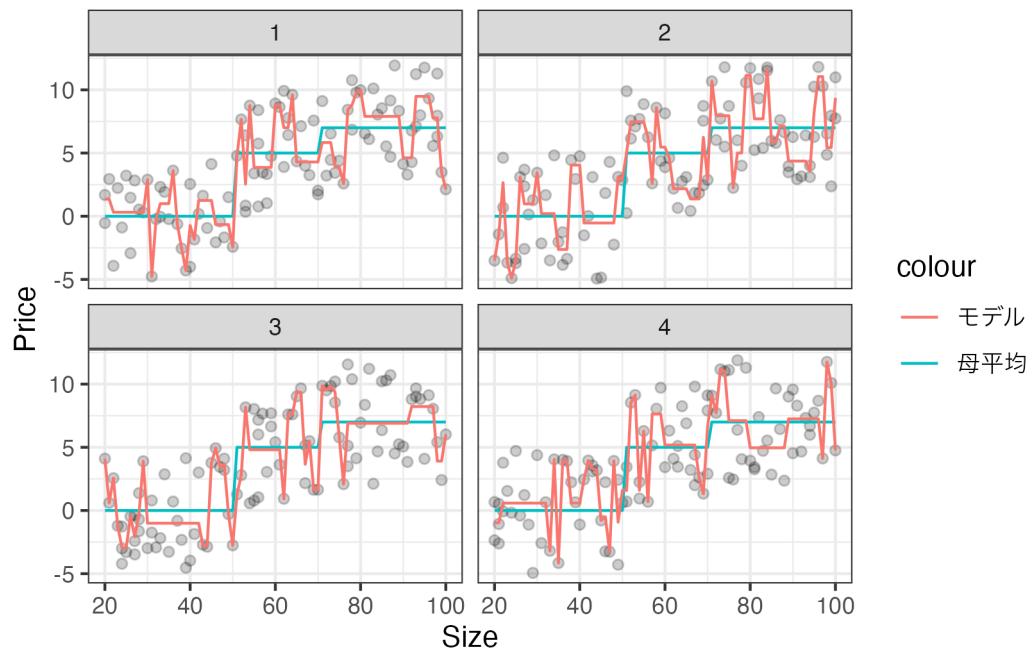
5.1 数値例: Depth 1



5.2 数値例: Depth 3

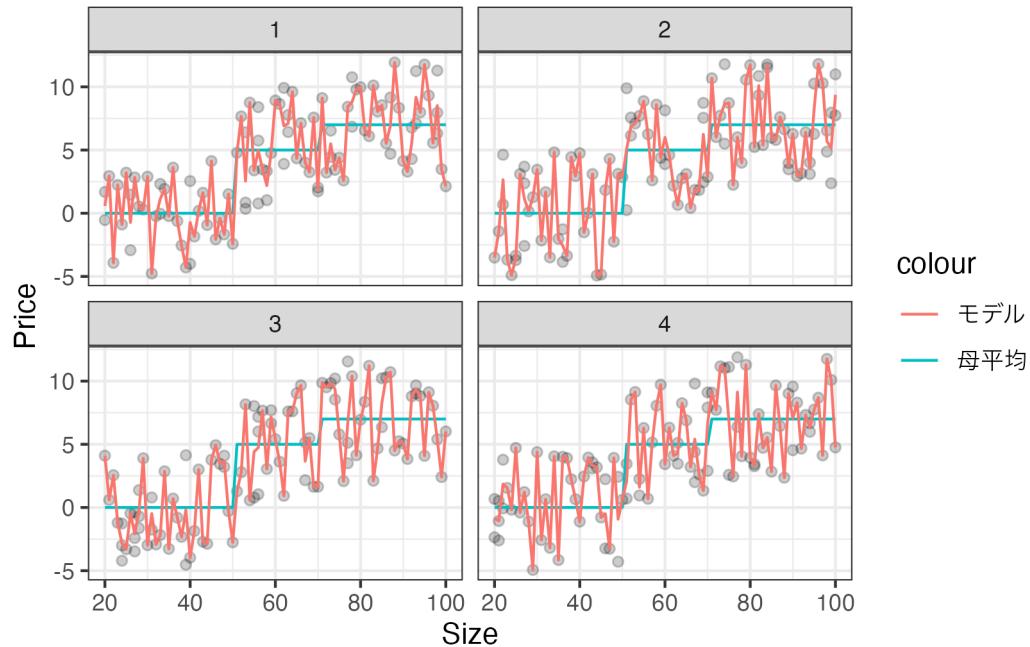


5.3 数値例: Depth 6



5.4 数値例: Depth 30

- 丸暗記モデルになってしまう



5.5 モデル選択/集計

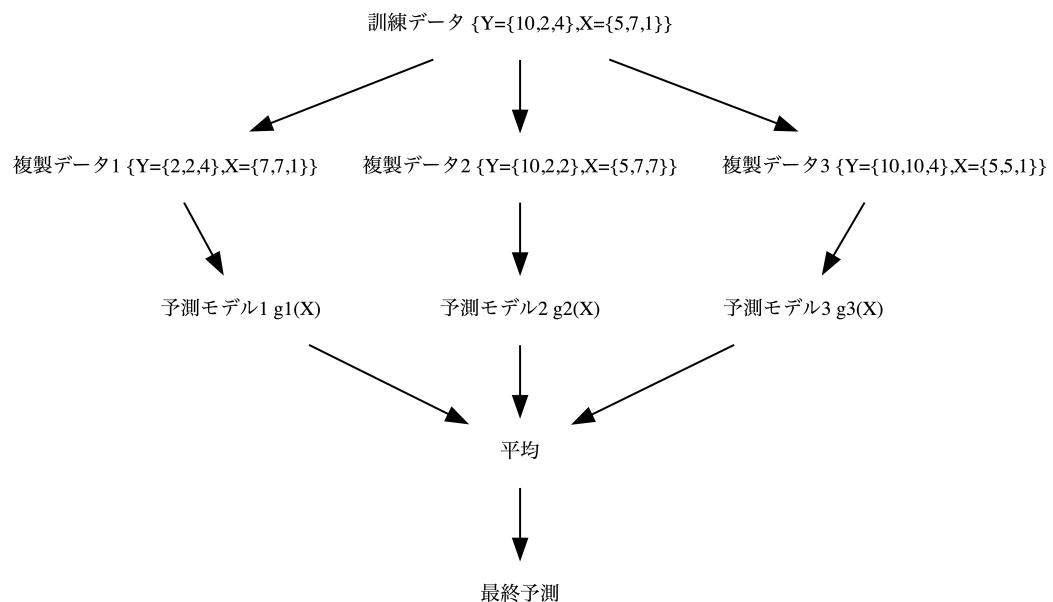
- 決定木の深さが異なれば、予測モデルは異なる
- 複数のモデルを”試作”し、予測性能を測定、最も性能が良いモデルを選ぶのは一つの方法
- 決定木については、複数のモデルを試作し、その予測結果を集計することがより有効な場合が多い
 - モデル集計/アンサンブル法などと呼ばれる

5.6 概要

- 複雑なモデルを推定すると、母平均から大きく乖離した事例（ハズレ値）の影響を強く受ける
 - 予測結果を安定するために、大量の予測モデルを作り、その平均値を最終予測とすることが有効（Bagging）
 - 一部の変数の使用を確率的に禁止することも併用できる（RandomForest）

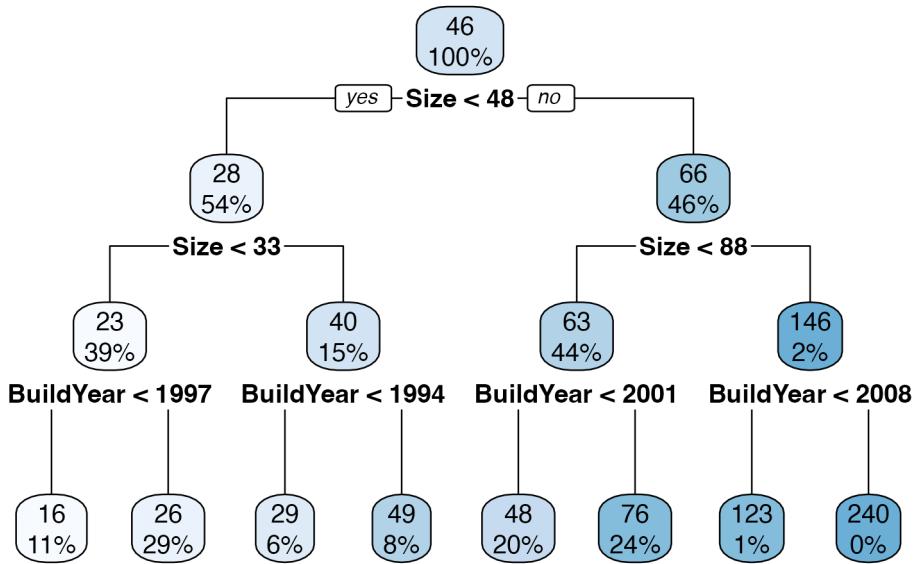
5.7 イメージ: Bagging

- Bootstrap 法を用いて、予測モデルを大量に作る (500-5000 個程度)



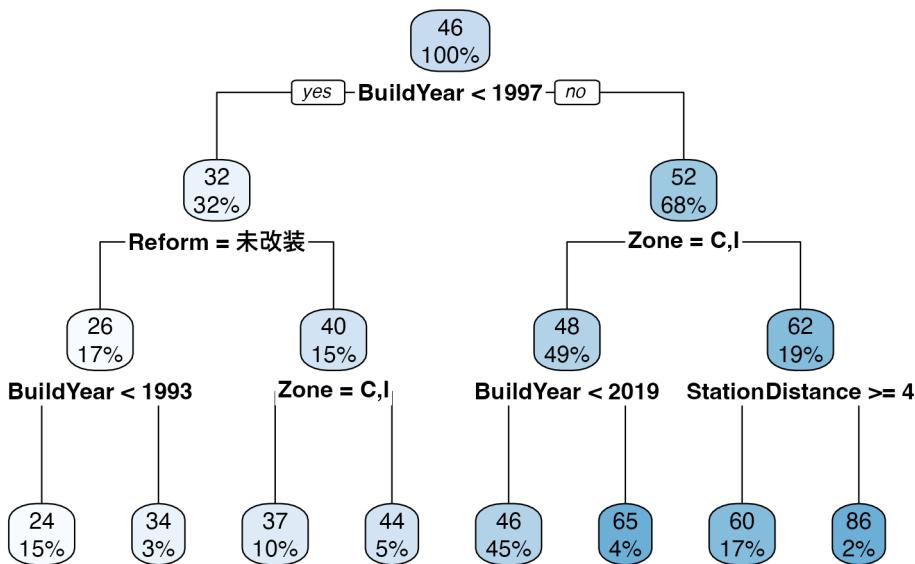
5.8 イメージ: RandomForest

- Tree を実際のデータに当てはめると



5.9 イメージ: RandomForest

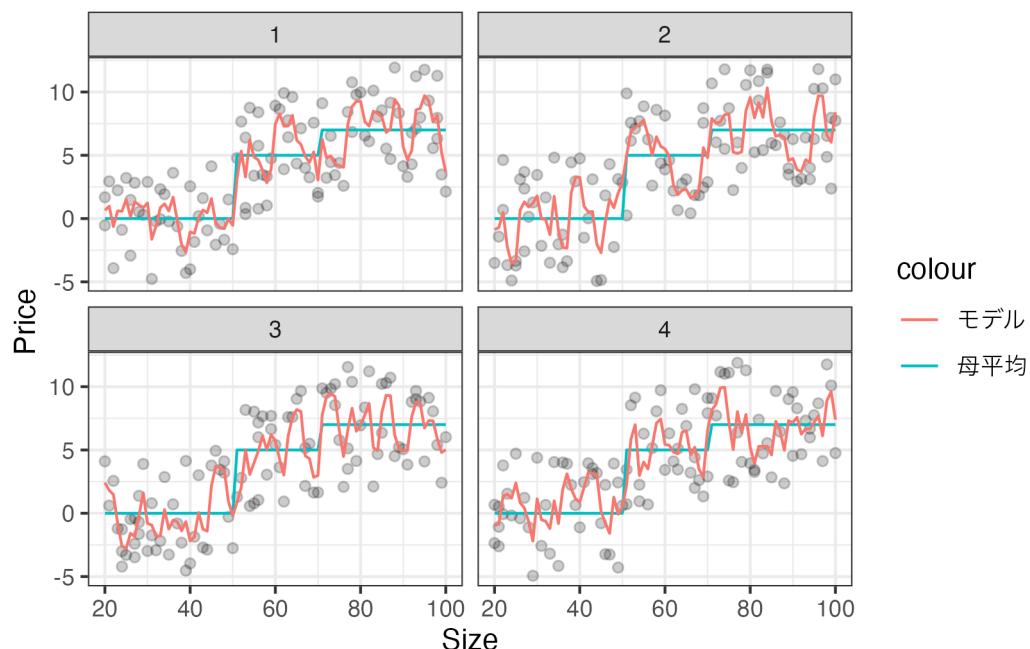
- 一部の予測モデルから、支配的な予測変数である Size を除外



5.10 RandomForest の利点

- Bootstrap は、“ハズレ値”を含まないデータとそれを用いたモデルも生成される
 - ハズレ値の影響を軽減できる
- 変数の一部を確率的に使用できなくすることで、モデル間で用いる変数の多様性が促進される
 - より多くの変数の情報が活用できる
- 注: 実際にはモデル単位ではなく、サンプル分割ごとに、(元々の変数数の square root 個) ランダムに使用禁止する

5.11 数値例: Random Forest



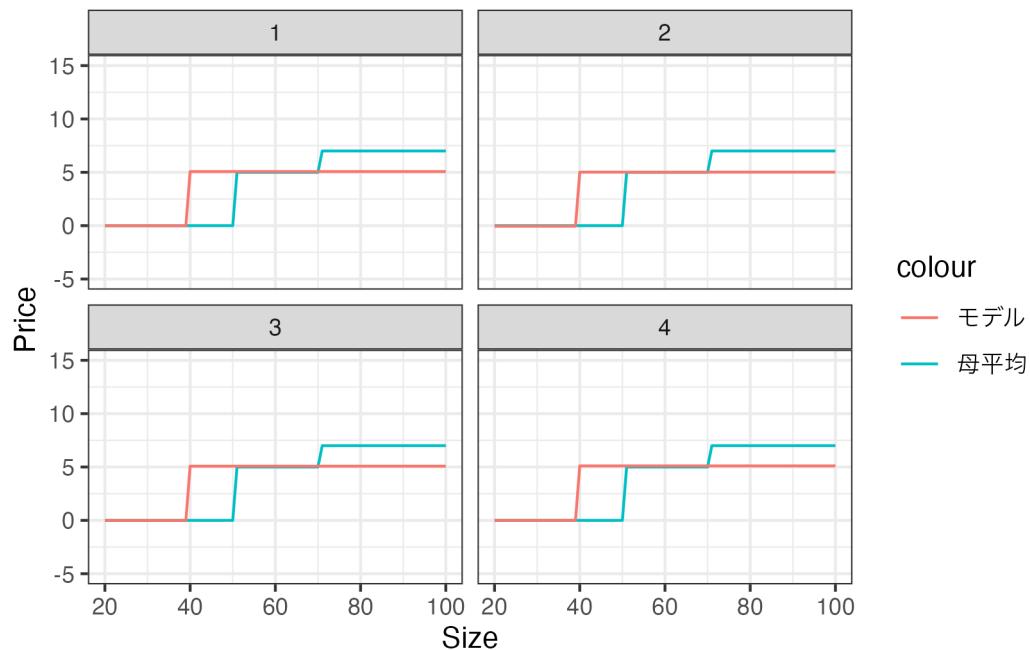
5.12 Big data は優れた予測モデルを保証するか?

- 「Big data = サンプルサイズが多い」、のであれば、完璧な予測モデル $g(X) = Y$ を保証しない
 - 削減不可能な誤差 $Y - E[Y|X]$ が存在するため
- 理想的な予測モデル $g(X) = E[Y|X]$ は保証できるか?
 - 伝統的な方法 = 極めて強い仮定のもとでのみ可能

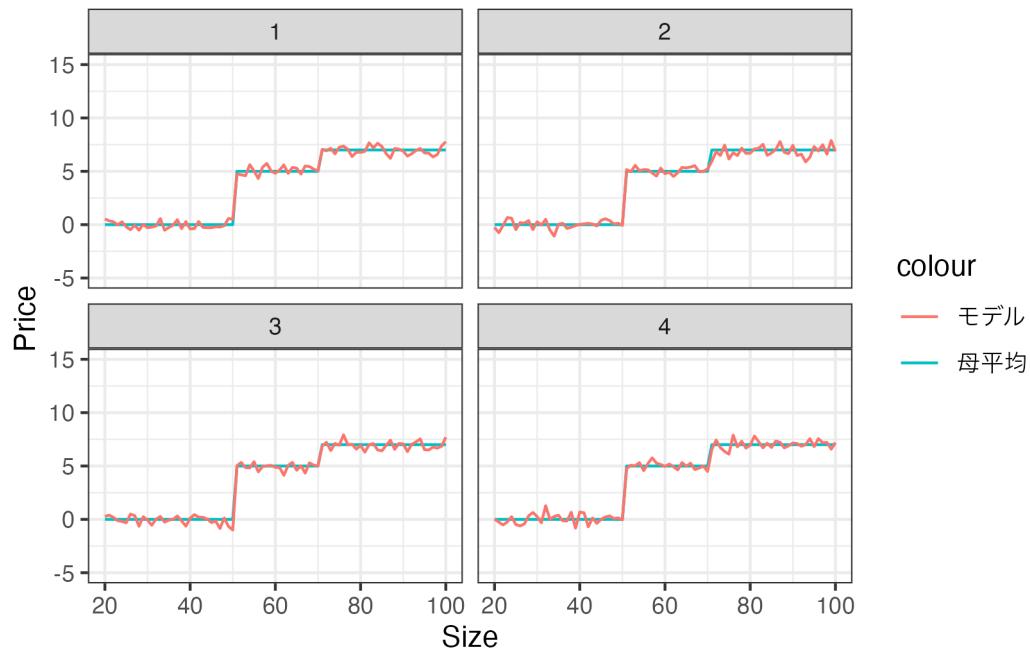
– 多くの機械学習 = 緩やかな仮定で可能

5.13 数値例: 伝統的アプローチ (5000 事例)

- 研究者が、「Size = 40 で分割する」と決定
 - 推定されるモデルを限定する



5.14 数値例: Random Forest with large sample (5000 事例)



5.15 まとめ

- 実際/理想/完璧な予測モデルの区別が重要
- 理想の予測モデル: $g(X) = E[Y|X]$
 - 個人差があれば、完璧な予測 $g(X) = Y$ は不可能
- 限られたデータから推定されたモデルは、理想の予測モデルとはならない $g(X) \neq E[Y|X]$
 - 機械学習を用いて、近づけることは可能
- 予測を使う前に、テストデータを用いた性能調査が必須

5.16 指導: 実務への応用

- 実務課題の解決に役立つ、予測課題を設定する必要がある
 - 予測できると嬉しい Y は何か?
 - * 組織内で蓄積しているデータと活用可能な X から、高い精度で予測できるそうか?

5.17 補論: (例) 推薦アルゴリズム構築

- サブスク制動画配信サービスにおける動画推薦アルゴリズム構築に役立つ、予測モデルを構築する
 - X = 過去の作品視聴履歴やいいね数
- 何を Y とすべきか?
 - ヒント.
 - * Y = 「その動画を視聴するかどうか」ではない
 - * 業務課題は、サブスク会員数の維持/拡大
- [Netflix Prize](#)

5.18 補論: 予測性能が低い = 役に立たない?

- 巨大なデータ & 機械学習を用いて、理想的な予測モデル $g(X) = E[Y|X]$ を推定してもなお完璧な予測モデル $g(X) = Y$ にはならない
 - 削減不可能な誤差 $Y - E[Y|X]$ が存在するため
- 予測性能は低いので役に立たない?
 - 母平均の特徴はある程度捉えているので、因果効果や格差の推定に有益 (後述)

5.19 Reference