資料3-1

重回帰分析の概要

電力中央研究所 社会経済研究所

主任研究員 田中 拓朗

第8回送配電効率化・計画進捗確認WG

2025/2/17

電力中央研究所

アウトライン

- ◆ レベニューキャップ制度における重回帰分析の位置づけ
- ◆ 回帰分析とは
- ◆ 線形回帰モデルとは
- ◆ 回帰分析結果の表示

＜トピックス＞
- ◆ 決定係数 R^2
- ◆ 重回帰分析の注意点：多重共線性
- ◆ 重回帰モデルと中央値の比較

※本資料においては、重回帰分析の概要について理解することを目的としているため、統計学や計量経済学の厳密な議論には踏み込まない

レベニューキャップ制度における重回帰分析の位置づけ

- ✓ レベニューキャップ制度の「OPEX査定」「CAPEX査定」では、**重回帰分析が採用**されている
- ✓ 重回帰の**決定係数が低い場合**は、**中央値を用いた査定方法を採用**することとなっている

【収入上限の算定プロセス】

収入上限の算定プロセス

プロセス①
見積費用の査定

見積費用（申請） → 見積費用（査定）

＋

プロセス②
前規制期間の実績を踏まえた各種調整

- 予実差の調整
- 制御不能費用の調整
- 前期効率化実績の反映
- インセンティブ

→ 収入上限

見積費用の査定にあたっては大きく、①OPEX（人件費・委託費等）、②CAPEX（設備投資関連費用）に区分

費用区分	査定方法
事業経費（OPEX）	OPEX査定
設備投資関連費用（CAPEX）	CAPEX査定
外生的な費用	制御不能費用

OPEX総額を対象に各社の過去実績を用いて推計式を設定することとし、その設定においては統計手法として**重回帰分析を採用**

設備投資額を「投資量」と「投資単価」に分類

設備投資額 ＝ 投資量 × 単価

【第1規制期間におけるCAPEX統計査定の全体方針】

①送電・変電・配電の各設備における**物品費**及び**工事費**について、各社の過去実績単価を用いた推計式を設定することとし、その設定においては統計手法として**重回帰分析を採用**

②重回帰分析における説明変数については、定性的かつ定量的（**決定係数**や変数間の**多重共線性**）な観点から、適切な説明変数をそれぞれ設定。

③重回帰分析の結果、決定係数が低い費用については、以下のとおり**中央値**を用いた査定方法を採用。

- ✓様々な特殊要因によって、単価が高額となる案件については、統計的に対象案件を検出して個別査定を実施（高額案件の申請に当たっては、社内での検討プロセスを求める）。
- ✓高額案件以外については、それらの中央値単価を用いて査定を実施（必要最小限のグルーピングを行って、複数の中央値単価を設定することも検討）。

出典：電力・ガス取引監視等委員会「託送料金制度（レベニューキャップ制度）中間とりまとめ、詳細参考資料」2021年11月

【参考】第1規制期間におけるCAPEX統計査定の全体方針

第1規制期間におけるCAPEX統計査定の全体方針（ローカル・配電系統）

- CAPEX査定においては、各社の実情を踏まえつつコスト効率化を促すものとするため、効率的な事業者における実績単価を用いた統計的な査定を行う。

一般送配電事業者による見積 → 統計査定

CAPEX（見積） → 全社の効率性を反映した推計単価 → トップランナー的補正 → 効率的な事業者の状況を反映した推計単価

全社の効率性を反映した推計費用の統計的な算出

① 送電・変電・配電の各設備における物品費及び工事費について、**各社の過去実績単価を用いた推計式を設定**することとし、その設定においては**統計手法として重回帰分析**を採用。

② 重回帰分析における説明変数については、定性的かつ定量的（決定係数や変数間の多重共線性）な観点から、適切な説明変数をそれぞれ設定。

③ 重回帰分析の結果、決定係数が低い費用については、以下のとおり中央値を用いた査定方法を採用。
- ✓ 様々な特殊要因によって、**単価が高額となる案件**については、統計的に対象案件を検出して**個別査定を実施**（高額案件の申請に当たっては、社内での検討プロセスを求める）。
- ✓ 高額案件以外については、それらの**中央値単価を用いて査定を実施**（必要最小限のグルーピングを行って、複数の中央値単価を設定することも検討）。

トップランナー的補正水準の設定

④ 重回帰分析により算出した各社の推計単価と、各社の実績単価を比較して効率性スコアを算出。

④' 全社の実績単価における中央値と、各社の実績単価を比較して効率性スコアを算出。

⑤ トップランナー的補正における効率性スコアの水準は、**上位3位**を基本とする。

⑥ 第1規制期間においては、スモールスタートの観点も踏まえ、トップランナー的補正を行った推計単価に対して、**過去実績も反映（70%）。**

189

出典：電力・ガス取引監視等委員会「託送料金制度（レベニューキャップ制度）中間とりまとめ、詳細参考資料」2021年11月, s.189

回帰分析とは

✓ 回帰分析とは、**ある変数Y（目的変数）**の値を、**他の変数（説明変数）**を用いて**説明・予測**するためのデータ分析手法

【本資料での定義】
- **ある変数Y**は、目的変数や被説明変数と呼ばれているが、本資料では**目的変数**で統一する
- **他の変数**は、説明変数や独立変数と呼ばれているが、本資料では**説明変数**で統一する

◆ 回帰分析の活用例

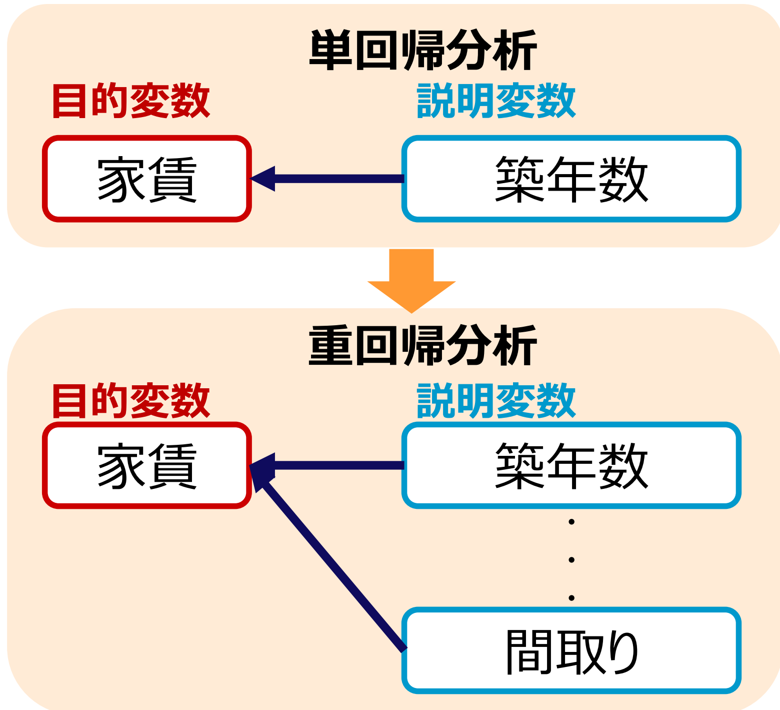
- 家賃（ある変数Y）は、どのような要因（他の変数）によって決定されるのかを知りたい
- 省エネ機器の購入（ある変数Y）は、省エネ補助金（他の変数）によって増加するのか知りたい
- 顧客満足度（ある変数Y）は、サービス品質（他の変数）にどの程度影響を受けているかを知りたい

◆ 説明変数が1つ：単回帰分析

目的変数の**予測精度を向上**させたい

他の説明変数の影響を考慮することで、**各説明変数の純粋な効果を明らかに**したい

◆ 説明変数を複数個扱う：重回帰分析



線形回帰モデル

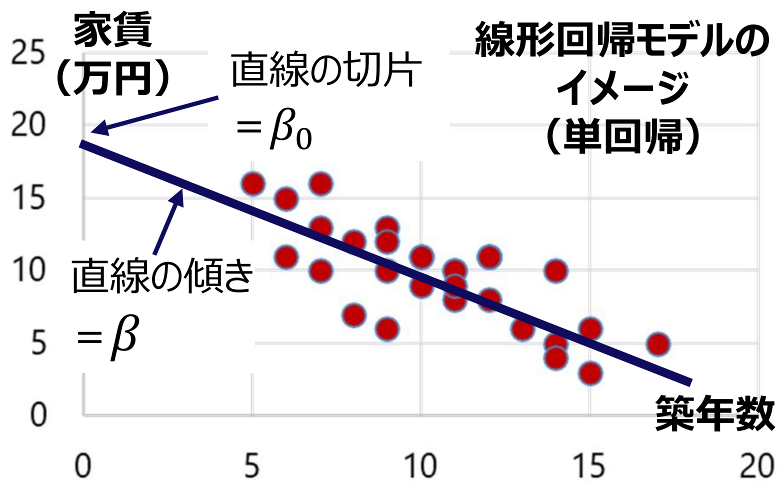
✓ **目的変数**と**説明変数**の関係性について、以下のような**線形関係**を仮定して分析するモデルを、**線形回帰モデル**という

◆ 線形回帰モデルの例（家賃を説明・予測するモデル）

Y:目的変数　**X:説明変数**　**ϵ:誤差項**

$$\text{家賃} = \beta_0 + \beta_{\text{築年数}} \times \text{築年数} + \beta_{\text{駅からの距離}} \times \text{駅からの距離} + \beta_{\text{延床面積}} \times \text{延床面積} + \epsilon$$

β_0 :定数項　**$\beta_1, \beta_2, \dots, \beta_K$:回帰係数**



- **ϵ:誤差項**
  - ✓ 目的変数Yを説明する観測不可能な<u>その他の要因</u>
- **β_0:定数項**
  - ✓ 説明変数の値がすべて0の場合のYの基準値と解釈される
    - ただし、説明変数の中には、例えば延床面積のように、0とすることが非現実的なものもあるため、基本的には特段の解釈は求めず、数学的な計算上のY基準値として扱われる
- **$\beta_1, \beta_2, \dots, \beta_K$:回帰係数**
  - ✓ 説明変数X_kの値が１単位増えた時に、目的変数が平均的にどの程度変化するのかを表すパラメータ

回帰分析結果の表示※

$$家賃 = \beta_0 + \beta_{築年数} \times 築年数 + \beta_{駅からの距離} \times 駅からの距離 + \beta_{延床面積} \times 延床面積 + \epsilon$$

回帰係数（β）
■説明変数が1単位変化した時、**目的変数が平均的に何単位変化するのか**を表す

標準誤差
■回帰係数の推定結果の**統計的な精度**
□ 値が小さいほど、推定結果の統計的な精度が高い

P値
■各説明変数が目的変数に与える影響について、**統計的に有意かどうかを判断**するために利用する値

□P値が小さいときに、説明変数が目的変数に影響すると評価

✓ 厳密には、「影響しないという仮説を棄却する」ことを意味している

✓ 各説明変数の効果の大きさ自体を表すものではない点に注意

説明変数	回帰係数	標準誤差	P値
切片項	16.581	2.265	0.000
駅からの距離	-0.324	0.083	0.008
築年数	-0.191	0.073	0.040
延べ床面積※※	-0.036	0.073	0.639
観測数	10		
決定係数R^2	0.874		
自由度修正済み決定係数	0.811		

決定係数R^2
■ 回帰モデルの**当てはまり具合を評価する指標**であり、当てはまり度が高いほど、1に近づく
■ 次スライド参照

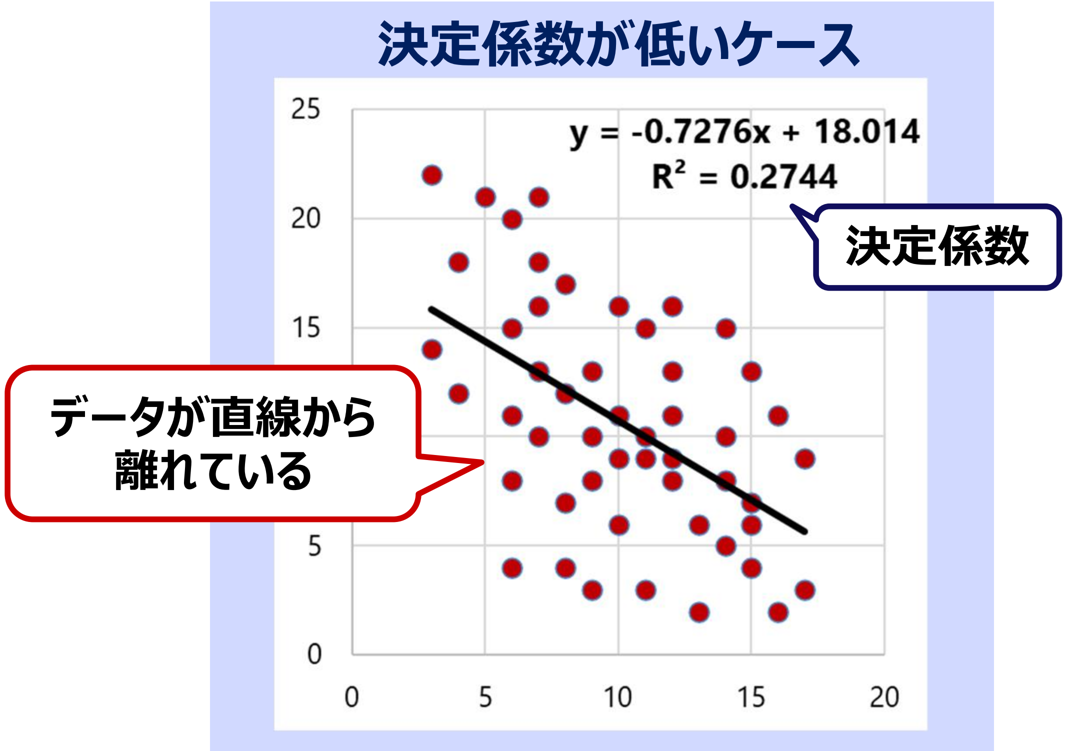
※本推定結果は、実データではなく、ダミーデータを用いた結果

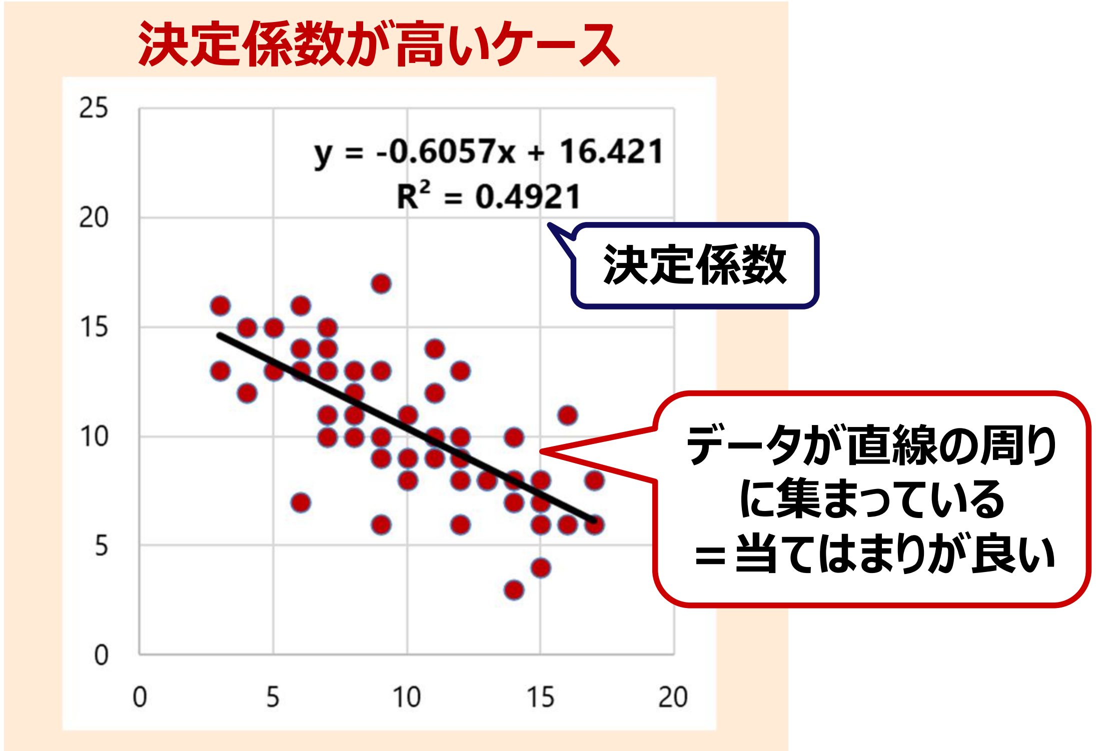
※※回帰係数の符号がマイナスとなっており、直感と反するように思われるが、p値が大きいため、その効果は統計的には有意でないと判断される

決定係数 R^2

- ✓ 決定係数R^2は、線形回帰モデルのあてはまりを評価する代表的な指標の1つ
- ✓ 1に近いほど、回帰モデルの当てはまりが良いと判断する※

◆ 決定係数が高いケースでは、データが線形回帰モデルの直線の近くに集まっていることがわかる（＝当てはまりが良い）





※決定係数が1に近いほど当てはまりが良いということは言えても、「どの程度であれば、当てはまりが良いと判断できるか」ということに関する絶対的な閾値があるわけでない

【参考】決定係数 R^2の概要

◆ 決定係数R^2は、**目的変数Yの全体的な変動（TSS）**を**回帰モデルの変動（ESS）**でどれだけ説明できているのかを表し、1**に近いほど、回帰モデルの当てはまりが良いと判断**する

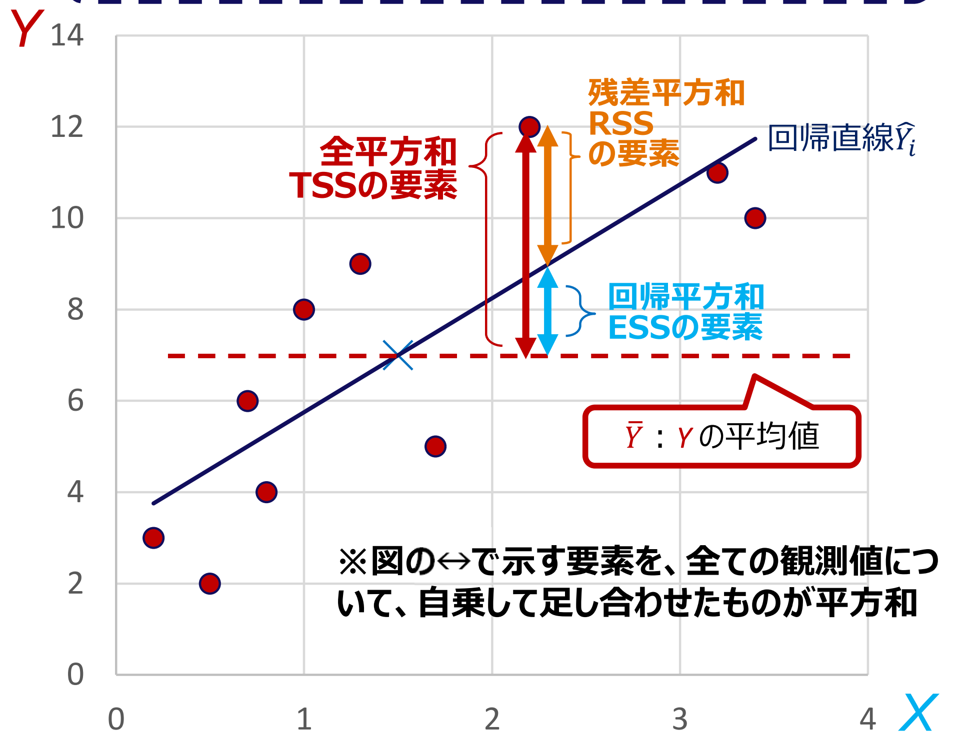
$$R^2 = \frac{\text{回帰平方和(ESS)}}{\text{全平方和(TSS)}},\ 0 \le R^2 \le 1$$

◆ 線形回帰モデルでは、全平方和（TSS）＝回帰平方和（ESS）＋残差平方和（RSS）の関係から、

$$R^2 = 1 - \frac{\text{残差平方和(RSS)}}{\text{全平方和(TSS)}}$$

が成立し、**残差 e_i ※ が全般的に小さい**と1に近づくことがわかる

※残差e_i ：●観測値（Y_i）と――回帰直線（$\hat{Y}_i$）との差
$e_i = Y_i - \hat{Y}_i$



- ■ **全平方和**（TSS, Total Sum of Squares）
  - ✓ $TSS = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$
- ■ **回帰平方和**（ESS, Explained Sum of Squares）
  - ✓ $ESS = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$
- ■ **残差平方和**（RSS, Residual Sum of Squares）
  - ✓ $RSS = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N}(e_i)^2$

回帰モデルのあてはまりを評価するその他の指標

✓ 決定係数 R^2以外にも、回帰モデルの当てはまりを評価する指標は多くある

- **自由度修正済み決定係数： *Adjusted* $\mathbf{R^2}$※**
  - 決定係数R^2は説明変数が多くなるほど高くなる性質があり、あてはまりの良さを過大に評価してしまうリスクがあるため、Adjusted R^2 では、説明変数が増えることをペナルティとして考慮する
  - **説明変数の数が異なるモデル同士を比較**する場合は、***Adjusted*** $\mathbf{R^2}$ を利用

※*Adjusted* $\mathbf{R^2}$

$$= 1 - \frac{\text{残差平方和(RSS)}}{\text{全平方和(TSS)}} \underbrace{\frac{N-1}{N-K-1}}_{\text{自由度修正部分}}$$

$$= 1 - \frac{\sum_{i=1}^{N}(Y_i - \widehat{Y_i})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \frac{N-1}{N-K-1}$$

Y_i: 観測値、$\widehat{Y}_i$:予測値、$\bar{Y}$:平均値
N: 観測数、K:定数項以外の説明変数の数

- 決定係数以外にも、回帰モデルを評価する指標として、以下のようなものがある
  - 平均自乗誤差（MSE：Mean Squared Error）
  - 平均絶対値誤差（MAE：Mean Absolute Error）
  - 平均絶対パーセント誤差（MAPE：Mean Absolute Percentage Error）
  - 赤池情報量規準（AIC：Akaike Information Criterion）
  - ベイズ情報量規準（BIC： Bayesian Information Criterion）など

重回帰分析の注意点：多重共線性

✓ 重回帰分析では、説明変数同士の相関が強いと多重共線性の問題を引き起こしうるため、説明変数間の相関に注意しながら、変数選択を行うことが重要である

- 多重共線性とは、重回帰分析において**説明変数同士が強い相関**を持っている状態
  - 家賃を説明・予測する重回帰モデルにおける例
    - 説明変数として、間取り（1LDKなど）と専有面積（㎡）を考える場合、両者ともに物件の大きさを表す変数であると考えられることから、これらの変数間に多重共線性が生じる可能性がある

- 多重共線性が引き起こす主な問題
  - 回帰係数の**標準誤差やP値（スライド7）が大きく**なり、統計的に有意でなくなる
  - データのわずかな変動や観測期間の変更などにより、**回帰係数が大きく変化**したり、理論から予想される符号や値から**大きく乖離する**ことがある

- 問題の深刻度は、分析に使用する**説明変数間の相関の程度に依存**
  - 説明変数間に相関があるからといって、それらの変数が**常に分析から除外されるわけではない**

- 多重共線性の判断
  - 分散拡大係数（**VIF**）を計算し、この数値が大きく、多重共線性が疑われる場合は、変数選択を再考する

$$VIF_j = \frac{1}{1 - R_j^2}$$

ここで、R_j^2は説明変数X_jを他の説明変数で説明する回帰モデルの決定係数

重回帰モデルと中央値の比較

✓ 重回帰モデルと中央値の選択においては、**重回帰モデルの決定係数だけでなく、中央値モデルの当てはまり**も評価し、相対的に評価するなどの総合的な判断が求められる

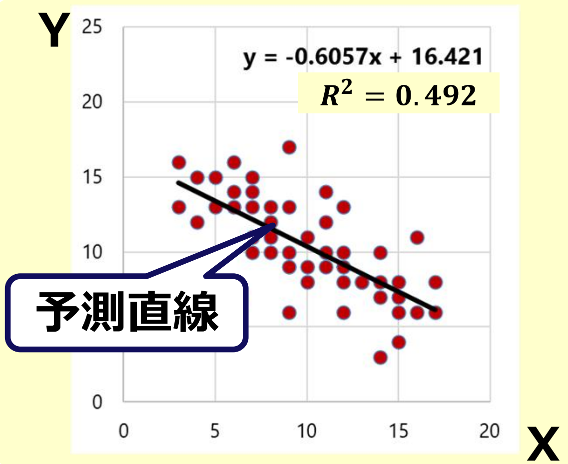
- **中央値**も、予測モデルの一つと捉えることが可能（仮に**中央値モデル**と呼ぶ※）
  - **中央値モデル**： 目的変数＝ **中央値** ＋ 誤差項
  - **重回帰モデル**： 目的変数＝ β_0 ＋ β_1 ×説明変数1 ＋ β_2 ×説明変数2 ＋ 誤差項

※中央値は、定数項のみの50%分位点（メディアン）回帰モデルとして表現可

- **重回帰モデル**と**中央値モデル**について、同じダミーデータ（スライド8の決定係数が高いケース）に対する当てはまりを、仮に、右枠の※※で示した重回帰の決定係数で評価すると、以下のようになる

※※ 中央値モデルは、重回帰と前提が異なる点もあるため、ここでは下記の式で決定係数を算出。自由度調整は行っていない点に注意

$$R^2 = \frac{回帰平方和(ESS)}{全平方和(TSS)}$$

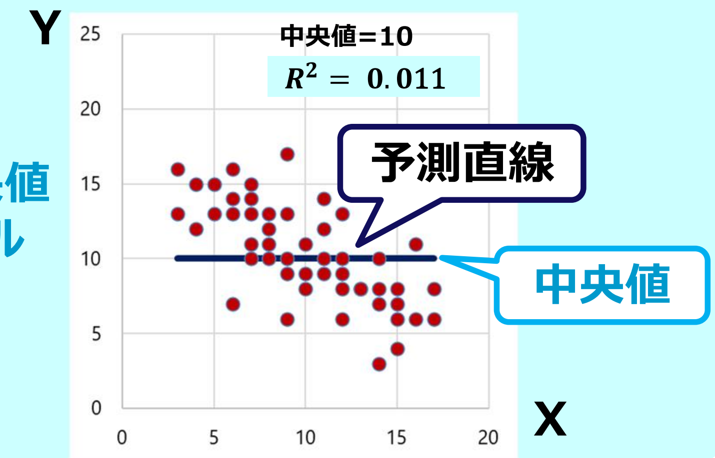


重回帰モデル＊

＊ 簡略化のため、説明変数が1つの単回帰モデルの例を表示

どちらの直線の方が、データに対する当てはまりが良いか？

中央値モデル



重回帰モデルと中央値の比較

✓ 実際のモデルの予測評価には、決定係数だけでなく、**平均自乗誤差**や**平均絶対値誤差、平均絶対パーセント誤差**なども用いられる

1. 平均自乗誤差（**MSE**：Mean Squared Error）

- 予測精度を、観測値と予測値の差の**自乗和で評価している点**が特徴
- 値が小さいほど、モデルの当てはまりが良いと判断

2. 平均絶対値誤差（**MAE**：Mean Absolute Error）

- 予測精度を、観測値と予測値の差の**絶対値和で評価している点**が、MSEと異なる
- 値が小さいほど、モデルの当てはまりが良いと判断

3. 平均絶対パーセント誤差（**MAPE**：Mean Absolute Percentage Error）

- 予測精度を、観測値と予測値の差を観測値で除した**割合の絶対値和で評価している点**が特徴
- 値が小さいほど、モデルの当てはまりが良いと判断

前スライドの推定モデルを、各指標で評価した結果

MSE

重回帰モデル	中央値モデル
○ 5.5	10.94

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$$

MAE

重回帰モデル	中央値モデル
○ 1.89	2.74

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \widehat{Y}_i\right|$$

MAPE

重回帰モデル	中央値モデル
○ 21.92	32.29

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| \times 100$$

○：各指標で当てはまりが良いと判断されたモデル

Y_i：観測値、$\widehat{Y}_i$：予測値、N：観測数

【参考】MSE、MAE、MAPEの特徴

1. 平均自乗誤差（MSE：Mean Squared Error）

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$$

定義上、メディアン重回帰モデルよりも**重回帰モデルを評価する傾向**がある

2. 平均絶対値誤差（MAE：Mean Absolute Error）

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \widehat{Y}_i|$$

定義上、重回帰モデルよりも**メディアン重回帰モデルを評価する傾向**がある

3. 平均絶対パーセント誤差（MAPE：Mean Absolute Percentage Error）

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| \times 100$$

定義上、上記1と2のような明らかな関係性はない

Y_i: 観測値、$\widehat{Y}_i$:予測値、N: 観測数
今回のケースの予測値には、重回帰モデルの結果や中央値が入る

■ **重回帰モデル**

✓ 残差の**自乗和を最小化**するモデル

$$\min_{\beta_0,\beta_1,\ldots,\beta_K}\sum_{i=1}^{N}(Y_i - (\beta_0+\beta_1 X_{i1} + \cdots + \beta_K X_{iK}))^2$$
$$= \min_{\beta_0,\beta_1,\ldots,\beta_K}\sum_{i=1}^{N}(e_i)^2$$

✓ 上記の目的関数は、**MSEの構成要素**となっている

■ **メディアン重回帰モデル**

✓ 残差の**絶対値和を最小化**するモデル

$$\min_{\beta_0,\beta_1,\ldots,\beta_K}\sum_{i=1}^{N}|Y_i - (\beta_0+\beta_1 X_{i1} + \cdots + \beta_K X_{iK})|$$
$$= \min_{\beta_0,\beta_1,\ldots,\beta_K}\Sigma_{i=1}^{N}|e_i|$$

✓ 上記の目的関数は、**MAEの構成要素**となっている

参考文献

- 「計量経済学　ミクロデータ分析へのいざない」 末石直也　日本評論社　2015
- 「計量経済学」 西山慶彦・新谷元嗣・川口大司・奥井亮　有斐閣　2021
- 「計量経済学」 浅野晳・中村二朗　有斐閣　2009
- 「新しい計量経済学　データで因果関係に迫る」 鹿野繁樹　日本評論社　2015
- 「Mostly Harmless Econometrics: An Empiricist's Companion」, Joshua D. Angrist, and Jörn-Steffen Pischke, Princeton University Press, 2009