数理統計 補助資料
～予測のための回帰モデル～

2024年度2学期:　月曜1限, 水曜3限
担当教員:　石垣　司

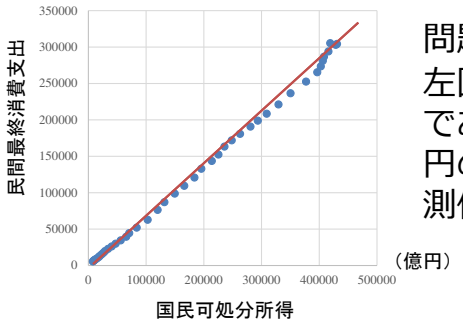
復習: 回帰分析と予測

回帰式 $y = \hat{b}_0 + \hat{b}_1 x$ を利用した予測

$\hat{b}_0$ と $\hat{b}_1$ はデータ $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ から最小2乗推定された係数

– $\hat{y}$: 説明変数 x に対する目的変数 y の予測値

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$



1955年度～1998年度(1968SNA)
（内閣府　国民経済計算年次推計）

問題
左図の回帰係数は $\hat{b}_0 \cong 0, \hat{b}_1 \cong 0.7$ である。国民可処分所得が500兆円のとき, 民間最終消費支出の予測値は？

重回帰分析の結果を用いた予測

顧客の1年間の購買金額の予測式

	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(>|t|) (p値)	
切片 (b_0)	106146	24196	4.39	0.000	***
年齢 (b_1)	841	382	2.21	0.028	*
家族人数 (b_2)	23170	2602	8.91	0.000	***
高齢者の有無 (b_3)	-1063	8202	-0.13	0.897	
子供の有無 (b_4)	7941	7633	1.04	0.299	
家からの時間(b_5)	-3208	598	-5.37	0.000	**
Adjusted R-squared (自由度調整済み決定係数)			0.106		

$$\hat{y} = 106146 + 841x_1 + 23170x_2 - 1063x_3 + 7941x_4 - 3208x_5$$

– **例: 夫婦と子供1人の核家族。登録者の年齢は45歳, 家から店舗までの所要時間は5分**

$\hat{y} = 106146 + 841 \times 45 + 23170 \times 3 - 1063 \times 0 + 7941 \times 1 - 3208 \times 5$
$= 205{,}402$円

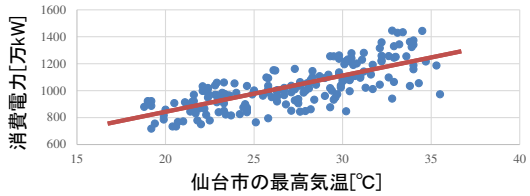
ここでは, 検定で有意とならなかった説明変数についても予測のために利用している。

目的変数 y の予測

1点の予測値を求めるためには回帰直線のみが必要

– **例: $x = 30$℃ のときの消費電力の予測値 $\hat{y} = 1092$ kW**

– **必要な仮定: 完全な多重共線性がない**



回帰直線のみが必要な場合, 確率的なモデルや議論は不要

一方, 需要予測等では予測の幅も重要な情報となる

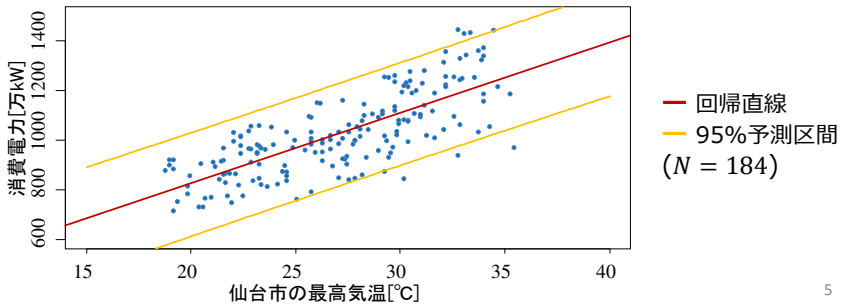
– **例:「明日の仙台市の最高気温は 30℃ の予報なので, 1092 kW 分の電力だけ準備すればよい」とはならない**

– **例: 欠品の確率が 1% 以下になるように在庫量を調整したい**

目的変数 y の予測区間とは？

予測区間

- 95%予測区間の意味：同じ母集団からサンプルサイズ1の標本抽出を100回繰り返したときに，そのうち約95個のデータがその中に入る区間
- いくつかの仮定に基づいて目的変数 y の分布を導出し，新しいデータが発生する区間を算出



目的変数 y に関する分布 #1

問題設定

- 仮定 1, 2, 3, 4 を満たす
  - 仮定1：説明変数は確率変数ではなく定数である
  - 仮定2：説明変数間に多重共線性はない
  - 仮定3：誤差項 $\{e_1, \cdots, e_N\}$ は互いに独立である
  - 仮定4：誤差項 e は正規分布 $N(0, \sigma^2)$ に従う
- 最小2乗推定量 $\hat{\boldsymbol{b}}$ は過去データ $\{\boldsymbol{y}, X\}$ から既に計算済み
- 新しい目的変数が $y_0 \sim \text{i.i.d.} N(\boldsymbol{x}_0^T\boldsymbol{b}, \sigma^2)$ に従い観測される

  $y_0 = \boldsymbol{x}_0^T\boldsymbol{b} + e_0, e_0 \sim \text{i.i.d.} N(0, \sigma^2)$ と同じ意味であることに注意

  $\hat{\boldsymbol{b}}$ は過去データ $\{\boldsymbol{y}, X\}$ から推測されているため y_0 とは独立

  σ^2 は未知。誤差項の標本分散 $S^2 = \frac{1}{N-P-1}\sum_{i=1}^{N} e_i^2$ は観測可能

目的変数 y に関する分布 #2

予測値と観測値のズレの期待値

- 新しく観測される目的変数 y_0（※確率変数という点に注意）
- 最小2乗推定量を利用した予測値 $\hat{y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{b}}$
- y_0 と $\hat{y}_0$ のズレ $d = y_0 - \hat{y}_0$

check!

$$E[d] = 0, V[d] = \sigma^2\{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0\}$$

y_0 と $\hat{y}_0$ のズレの分布

$$\frac{y_0 - \hat{y}_0}{\sqrt{V[d]}} = \frac{y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{b}}}{\sqrt{\sigma^2\{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0\}}} \sim N(0,1)$$

σ^2 を誤差項の標本分散 S^2 で置き換え

$$\frac{y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{b}}}{\sqrt{S^2\{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0\}}} \sim t^{(N-P-1)}$$

目的変数 y の分布 #2

$$E[d] = E[y_0 - \hat{y}_0] = E[y_0] - E[\hat{y}_0] = E[y_0] - E\left[\boldsymbol{x}_0^T\hat{\boldsymbol{b}}\right]$$
$$= \boldsymbol{x}_0^T\boldsymbol{b} - \boldsymbol{x}_0^T\boldsymbol{b} = 0$$

$$V[d] = V[y_0 - \hat{y}_0] = V\left[y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{b}}\right]$$

1. $\boldsymbol{x}_0$ は定数という仮定
2. $\hat{\boldsymbol{b}}$ は過去のデータ $\{\boldsymbol{y}, X\}$ を使用して求めている
3. $y_0 \sim \text{i.i.d.} N(\boldsymbol{x}_0^T\boldsymbol{b}, \sigma^2)$ としてiidで$\{\boldsymbol{y}, X\}$とは別に観測される変数

1,2,3から，y_0 と $\boldsymbol{x}_0^T\hat{\boldsymbol{b}}$ は独立。

また、$V\left[\boldsymbol{x}_0^T\hat{\boldsymbol{b}}\right] = \boldsymbol{x}_0^T V[\hat{\boldsymbol{b}}]\boldsymbol{x}_0 = \sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0$ より

$$V[d] = V[y_0 - \hat{y}_0] = V[y_0] + V\left[\boldsymbol{x}_0^T\hat{\boldsymbol{b}}\right] = \sigma^2 + \sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0 =$$
$$= \sigma^2\{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0\}$$

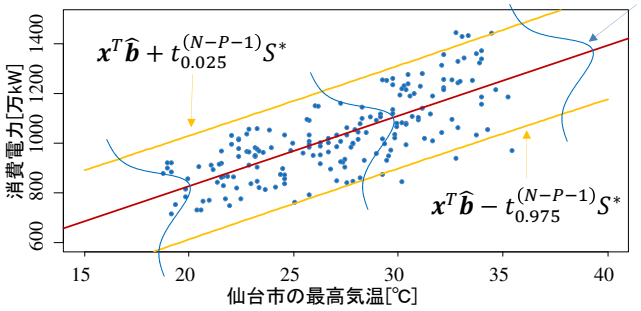
目的変数 y の予測区間

目的変数 y の95%予測区間

$$\boldsymbol{x}_0^T\hat{\boldsymbol{b}} - t_{0.975}^{(N-P-1)}S^* \leq y_0 \leq \boldsymbol{x}_0^T\hat{\boldsymbol{b}} + t_{0.025}^{(N-P-1)}S^*$$

$$S^* = \sqrt{S^2\{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0\}}$$

$t_\alpha^{(N)}$：**自由度 N の t 分布の上側確率 $100 \times \alpha$ % 点**



自由度182の t 分布
(ほぼ標準正規分布)
と同じ形状の分布

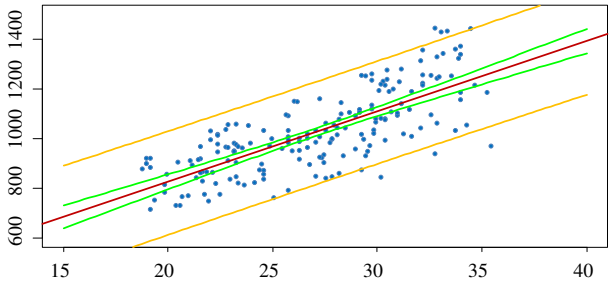
— 回帰直線
— 95%予測区間
($N = 184$)

注意：予測区間と信頼区間

回帰直線の95%信頼区間

- **同じ母集団から標本抽出を100回繰り返し，それぞれの標本から回帰直線を算出したときに，そのうち約95本の回帰直線が入る区間**

予測区間と信頼区間は意味が異なるので要注意



— 回帰直線
— 95%予測区間
— 95%信頼区間
($N = 184$)

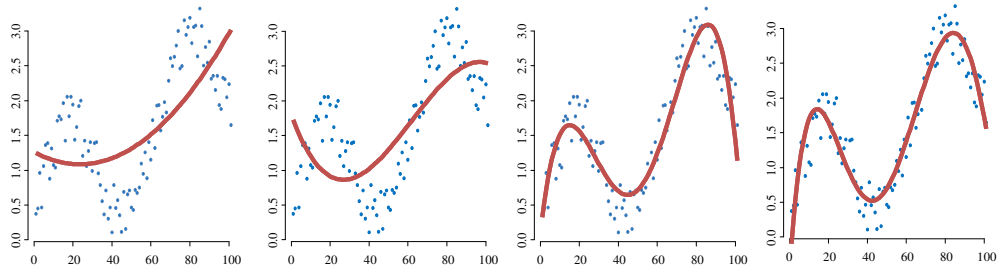
「予測」に関するこの先の学習の道筋 #1

多項式回帰：説明変数の多項式で表現される回帰モデル

- **M 次の多項式回帰モデル**

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 +, \cdots, +b_Mx^M$$

回帰係数は最小2乗法で推定可能
非線形構造をもつデータの傾向に合わせた関数で回帰できる



2次の多項式モデル　3次の多項式モデル　4次の多項式モデル　5次の多項式モデル

「予測」に関するこの先の学習の道筋 #2

両者は「計量分析」で学習済みとのこと。LASSOは2000年代，Deep Learningは2010年代の数理統計関連の技術的な一大ブレイクスルー

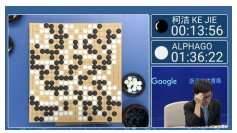
LASSO 回帰

- **スパース正則化を用いた回帰分析**
- **特徴選択（説明変数の選択）を過学習を抑制**

Deep Learning

- **高性能に入力（説明変数）が与えられた時に出力（目的変数）を返すのが機能。その出力を予測や判別に利用している**

Deep Learning により実現した機能



将棋
プロ棋士(2013)
名人(2017)

囲碁
元世界ランク1位(2016)
現役世界ランク1位(2017)

画像認識
人間の精度を超える
(2015)

https://pocketalk.jp/
自動翻訳

画像生成
「Tohoku University」
Mage.spaceにて作成

ChatGPT
(2022)

線形回帰モデルのまとめ

「実証」のための線形回帰モデル

- **実証：定量的に仮説や理論を検証し，現状の理解を促進**
- **伝統的な計量経済学モデルの基礎**

「予測」のための線形回帰モデル

- **予測：過去の構造・パターンを利用し定量的に未知を推計**
- Deep Learning **を含む統計・機械学習モデルの基礎**

実証と予測の両方で線形回帰モデルは中心的基盤

演習問題

ある製品の来月の受注数 y **は今月の営業コスト** x **と相関関係にあることが分かっている。今月の在庫量はゼロで，今月の営業コスト** $x_0 = 100$ **であったとき，来月の受注に関して欠品の確率が** 1% **以下となるためには何個の製品を生産すればよいか答えなさい。ただし，条件は下記である。**

- **過去**60**か月のデータを用いた線形回帰分析の結果を利用する**
- **このデータを発生する現象は仮定**1, 2, 3, 4**を満たしている**
- **自由度**30**以上の** t **分布の確率は標準正規分布の確率で代替できる**
- **標準正規分布の上側確率** $100 \times \alpha$ **% 点：** $z_{\alpha=0.01} = 2.23, z_{\alpha=0.005} = 2.58$
- **利用する回帰モデル：** $y = \hat{b}_0 + \hat{b}_1 x$

$$\sqrt{S^2\{1 + \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0\}} = 100,\ \hat{b}_0 = 200,\ \hat{b}_1 = 10,\ \boldsymbol{x}_0^T = [1 \quad x_0]$$