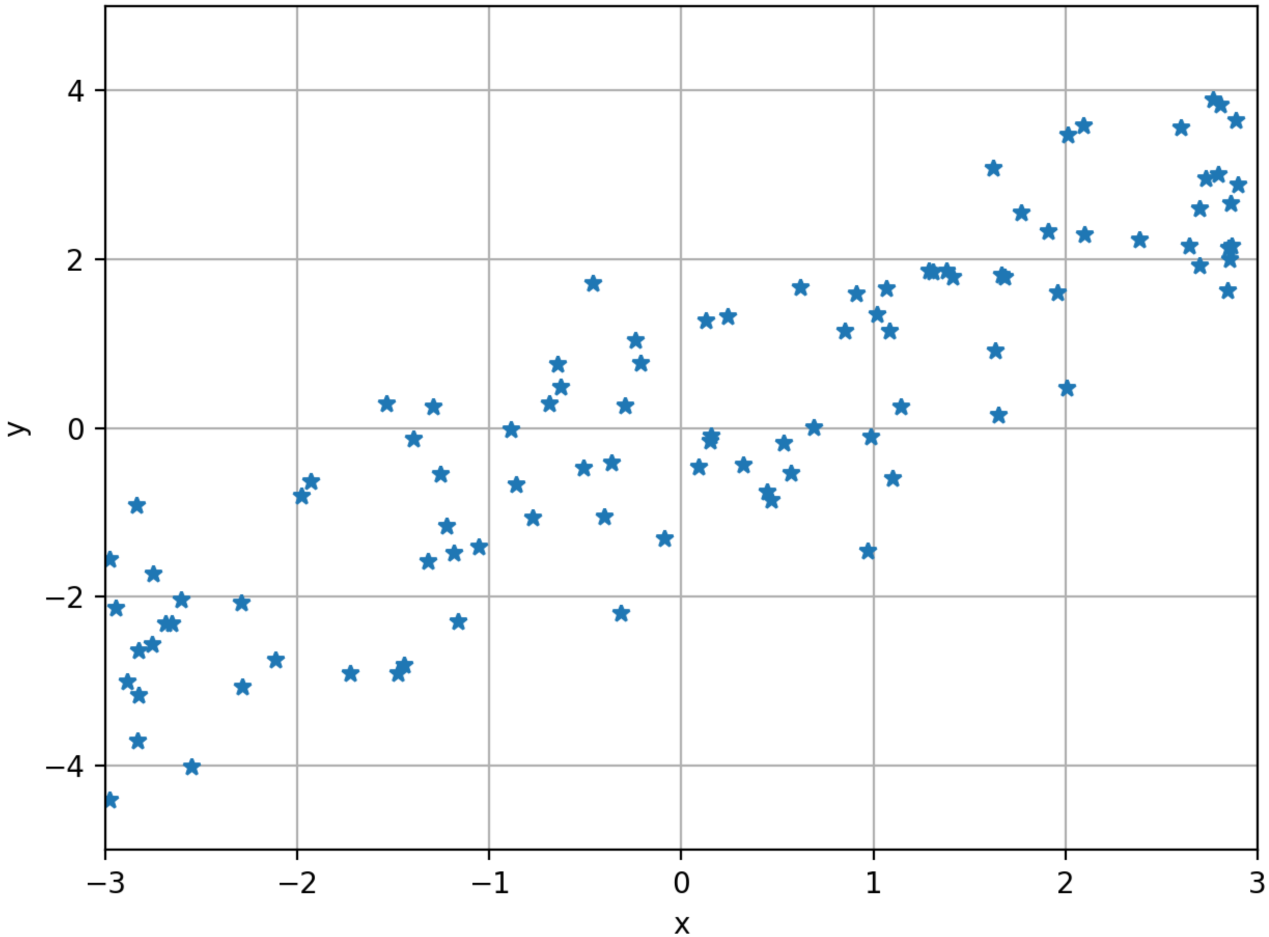
データ解析特論
第４回

システム情報系
福地　一斗
fukuchi@cs.tsukuba.ac.jp

線形回帰（最小二乗法）

回帰分析

- 目的変数yを説明変数xの関数として表したい
  - モデル化したい
  - 予測に使いたい



線形モデル（Fixed design）の推定問題

線形モデルのパラメトリック推定問題

- 既知母数：
  - **説明変数** $x_1, \dots, x_n \in \mathbb{R}^d$
- 未知母数：
  - **回帰係数** $b_0, b_1, \dots, b_d \in \mathbb{R}$ （特にb_0は**切片**）
- **誤差**（未観測）：$\epsilon_1, \dots, \epsilon_n \in \mathbb{R}$

iidは仮定しない

- 標本として**目的変数** $Y_1, . . . , Y_n \in \mathbb{R}$を観測：

$$Y_i = b_0 + b_1 x_{i1} + \cdots + b_d x_{id} + \epsilon_i$$

目的

- 既知母数$x_1, \dots, x_n$と標本$Y_1, \dots, Y_n$を元に$b_0, \dots, b_d$を推定

Fixed design：説明変数は既知母数であり固定（確率変数ではない）
⇔Random design (この講義では扱わない)：
説明変数は標本の一部でありランダム（確率変数）

線形モデルの行列表現

- 目的変数の線形モデル：

$$
\begin{aligned}
Y_i &= b_0 + b_1 x_{i1} + \cdots + b_d x_{id} + \epsilon_i \\
&= (1 \quad x_{i1} \quad \cdots \quad x_{id})^\top (b_0 \quad \cdots \quad b_d) + \epsilon_i
\end{aligned}
$$

- 線形モデルの行列表現 $Y = Xb + \epsilon$

$$
\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_d \end{pmatrix}}_{b} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\epsilon}
$$

線形モデルのパラメトリック推定問題（行列表記）

- 既知母数$X \in \mathbb{R}^{n\times(d+1)}$，未知母数$b \in \mathbb{R}^{d+1}$
- 誤差 $\epsilon \in \mathbb{R}^n$
- 標本$Y \in \mathbb{R}^n$: $Y = Xb + \epsilon$
- 目的 : XとYを元にbを推定

最小二乗推定量

- **回帰係数の推定値** $\hat{b} \in \mathbb{R}^{d+1}$
- **目的変数**Y_i**の**$\hat{b}$**を使った推定値**：$\hat{Y}_i = x_i^\top \hat{b}$
- **残差**：$Y_i - \hat{Y}_i$（$\hat{b}$を回帰係数とする線形モデルで説明できない”残り”）

（最小二乗推定量における）bの推定方針：$\hat{b}$とbが近ければ残差が小さい

- ⇨残差の二乗和を最小化する回帰係数$\hat{b}$を推定値とする
- ２乗誤差関数

$$L(\hat{b}) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2 = \|Y - X\hat{b}\|^2$$

- Lは $\hat{b}$ についての凸関数 ⇔ 最適解は以下の一次最適性条件を満たす

$$X^\top X b = X^\top Y$$

- $X^\top X$が正則ならば

$$\hat{b} = (X^\top X)^{-1} X^\top Y$$

最良線形不偏推定量（BLUE）

最小二乗推定量は良いのか？

- **線形不偏推定量**：Yの線型結合で表現できるbの不偏推定量
  - 行列$C \in \mathbb{R}^{(d+1)\times n}$について

$$\hat{b} = CY \quad \mathbb{E}\left[\hat{b}\right] = b$$

  - 最小二乗推定量も線形不偏推定量（$C = (X^\top X)^{-1} X^\top$）

- **最良線形不偏推定量（Best Linear Unbiased Estimator, BLUE）**
- =線形不偏推定量のうち全ての係数$\lambda \in \mathbb{R}^{d+1}$において

$$\mathbb{E}\left[\left(\lambda^\top (\hat{b} - b)\right)^2\right]$$

$\hat{b}$とbの間の誤差
λによって測り方が異なる

が最小

- ＝**誤差最小**の線形不変推定量

ガウスマルコフの定理

仮定

- Xがフルランク（$X^\top X$が正則）
- 誤差ϵ_iは以下を満たすと仮定（ガウスマルコフ条件）
  - 不偏性：$\mathbb{E}[\epsilon_i] = 0$
  - 等分散性：$\mathrm{Var}[\epsilon_i] = \sigma^2 < \infty$
  - 無相関性：$\mathrm{Cov}\left[\epsilon_i, \epsilon_j\right] = 0 \quad i \neq j$

この時**最小二乗推定量が最良線形不偏推定量**

- 最小二乗推定量は誤差最小の意味で**一番良い**推定量

非線形回帰問題

- 既知母数：$x_1, \ldots, x_n \in \mathbb{R}^d$
- 未知母数：$\theta \in \mathbb{R}^p$
- 標本として**目的変数** $Y_1, . . . , Y_n \in \mathbb{R}$を観測：
$$Y_i = f_\theta(x_i) + \epsilon_i$$
  - f_θはθでパラメータ化された非線形関数$f_\theta: \mathbb{R}^d \to \mathbb{R}$

目的

- $Y_1, \ldots, Y_n$を元にθを推定

非線形回帰の線形回帰への帰着

- f_θが非線形関数$\phi: \mathbb{R}^d \to \mathbb{R}^p$と$\theta \in \mathbb{R}^p$の線形和で表せるとする

$$f_\theta(x) = \theta_0 + \phi_1(x)\theta_1 + \cdots + \phi_p(x)\theta_p$$

- Y_iは$\phi(x_i)$を説明変数とする線形モデルに従う

$$Y_i = \theta_0 + \phi_1(x)\theta_1 + \cdots + \phi_p(x)\theta_p + \epsilon_i$$

- $\phi(x_i)$を説明変数，Y_iを目的変数とした最小二乗法によってθを推定可能

例：

- 単一の説明変数の多項式モデル

$$f_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_m x^M$$

  - $\phi(x) = (1, x, x^2, \dots, x^M)$と置くと$f_\theta(x) = \phi(x)^\top \theta$

線形回帰に関わる統計量と検定

線形モデルの再現度合いと誤差分散

線形回帰では母集団における線形モデルを仮定
- ⇒本当に仮定して問題ないか？

考え方：線誤差項が”大き過ぎる”場合は線形モデルがあっていない

誤差項の分散の推定
- ガウスーマルコフ条件を仮定
- **残差平方和**

$$S = \left\|Y - \hat{Y}\right\|^2 = \left\|Y - X\hat{b}\right\|^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

  - $\mathbb{E}[S] = \nu_e \sigma^2$
- **残差自由度**：$\nu_e = n - \mathrm{rank}(X)$ （多くの場合 $\mathrm{rank}(X) = 1 + d$）
- **誤差分散の不偏推定量**

$$\hat{\sigma}^2 = \frac{S}{\nu_e}$$

重相関係数と決定係数

誤差分散からスケールの影響を取り除く

- **重相関係数**（Multiple Correlation Coefficient）：
  目的変数Yと推定値$\hat{Y}$の相関係数

$$R = \frac{n^{-1}\sum_{i=1}^{n}\left(Y_i - \hat{\mathbb{E}}[Y]\right)\left(\hat{Y}_i - \hat{\mathbb{E}}[\hat{Y}]\right)}{\hat{\mathbb{V}}'[Y]^{1/2}\hat{\mathbb{V}}'[\hat{Y}]^{1/2}}$$

$\hat{\mathbb{E}}$ ：標本平均
$\hat{\mathbb{V}}'$：標本分散
$\hat{\mathbb{V}}$ ：不変分散

- **決定係数**：重相関係数の二乗

$$R^2 = 1 - \frac{n^{-1}S}{\hat{\mathbb{V}}'[Y]}$$

  - $1-$（Yの分散に対する誤差分散の占める割合）

- **自由度調整済み決定係数**
  - $1-$（Yの不偏分散に対する誤差分散の占める割合）

$$\tilde{R}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\mathbb{V}}[Y]}$$

1に近いほど回帰の当てはまりが良い

回帰係数に関する検定

- ❑ 未知母数：$b_0, b_1, \ldots, b_d \in \mathbb{R}$
- ❑ 誤差項はiidかつ正規分布に従う $\epsilon_i \sim N(0, \sigma^2)$
- ❑ 標本として**目的変数** $Y_1, . . , Y_n \in \mathbb{R}$を観測：

$$Y_i = b_0 + b_1 x_{i1} + \cdots + b_d x_{id} + \epsilon_i$$

目標

1. 特定（k番目）の説明変数が目的変数に影響を与えるのか検証
   - ❑ 回帰係数に対するT検定

2. 説明変数が目的変数に影響を与えるのか検証
   - ❑ 回帰モデルについてのF検定

回帰係数についてのT検定

- 未知母数：$b_0, b_1, \ldots, b_d \in \mathbb{R}$
- 誤差項はiidかつ正規分布に従う $\epsilon_i \sim N(0, \sigma^2)$
- 標本として**目的変数** $Y_1, \ldots, Y_n \in \mathbb{R}$を観測：

$$Y_i = b_0 + b_1 x_{i1} + \cdots + b_d x_{id} + \epsilon_i$$

目標

1．特定（k番目）の説明変数が目的変数に影響を与えるのか検証

仮説

- 帰無仮説 H_0：$b_k = 0$（k番目の説明変数は目的変数に影響しない）
- 対立仮説 H_1：$b_k \neq 0$（k番目の説明変数は目的変数に影響する）

- 帰無仮説H_0が与えられた有意水準αで棄却できるか調査
  - 棄却できればH_1を支持

回帰係数についてのT統計量

- T統計量

最小二乗法で推定された回帰係数

$$T_k = \frac{\hat{b}_k}{\hat{\sigma}\sqrt{[(X^\top X)^{-1}]_{kk}}}$$

$\hat{b}_k$の標準偏差の推定値

- 帰無仮説H_0のもとでT_kは自由度ν_eのt分布に従う

- 自由度ν_eのt分布と有意水準αから求めた両側検定の閾値：$\tau_{\nu_e,\alpha}$

$$|T_k| > \tau_{\nu_e,\alpha}$$

であればH_0を棄却 = H_1を支持

  - つまりk番目の説明変数は目的変数に影響する

- $|T_k| \leq \tau_{\nu_e,\alpha}$であれば$H_0$を棄却できない
  - k番目の説明変数は目的変数に影響すると断定する根拠が乏しい

回帰係数についてのF検定

- 未知母数：$b_0, b_1, \ldots, b_d \in \mathbb{R}$
- 誤差項はiidかつ正規分布に従う $\epsilon_i \sim N(0, \sigma^2)$
- 標本として**目的変数** $Y_1, \ldots, Y_n \in \mathbb{R}$を観測：

$$Y_i = b_0 + b_1 x_{i1} + \cdots + b_d x_{id} + \epsilon_i$$

目標

2. 説明変数が目的変数に影響を与えるのか検証

仮説

- 帰無仮説 H_0：$b_1 = \cdots = b_d = 0$
- 対立仮説 H_1：$\neg(b_1 = \cdots = b_d = 0)$

- 帰無仮説H_0が与えられた有意水準αで棄却できるか調査
  - 棄却できればH_1を支持

回帰モデルについてのF統計量

- F統計量

$$F = \frac{\frac{S_0 - S}{\mathrm{rank}(X) - 1}}{\hat{\sigma}^2} = \frac{\frac{S_0 - S}{\mathrm{rank}(X) - 1}}{\frac{S}{\nu_e}} = \frac{\frac{S_0 - S}{\mathrm{rank}(X) - 1}}{\frac{S}{n - \mathrm{rank}(X)}}$$

  - 残差平方和 $S = \left\| Y - X\hat{b} \right\|^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$
  - $b_1 = \cdots = b_d = 0$の残差平方和 $S_0 = \left\| Y - 1\hat{b}_0 \right\|^2 = \sum_{i=1}^{n} \left(Y_i - \hat{b}_0 \right)^2$
- 帰無仮説H_0のもとでFは自由度$(\mathrm{rank}(X) - 1, n - \mathrm{rank}(X))$の$F$分布に従う

- 自由度$(\mathrm{rank}(X) - 1, n - \mathrm{rank}(X))$の$F$分布と有意水準$\alpha$から求めた片側検定の閾値：$\phi_{\mathrm{rank}(X)-1, n-\mathrm{rank}(X), \alpha}$

$$F > \phi_{\mathrm{rank}(X)-1, n-\mathrm{rank}(X), \alpha}$$

であればH_0を棄却 = H_1を支持
  - つまり説明変数は目的変数に影響する
- $F \leq \phi_{\mathrm{rank}(X)-1, n-\mathrm{rank}(X), \alpha}$であれば$H_0$を棄却できない
  - 説明変数は目的変数に影響すると断定する根拠が乏しい

分散分析

1 元配置分散分析（1-way ANOVA）

- 異なるM個の実験条件，n回の独立な試験を実施
- 未知母数：
  - 平均効果 $b_0 \in \mathbb{R}$
  - 各実験条件が計測結果に与える効果 $b_1, \dots, b_M \in \mathbb{R}$
- i番目の試験の実験条件 $m_i \in \{1, \dots, M\}$
- i番目の試験の計測結果（標本）

$$Y_i = b_0 + b_{m_i} + \epsilon_i$$

誤差$\epsilon_i \sim N(0, \sigma^2)$

目標:

- $m_1, \dots, m_n$と$Y_1, \dots, Y_n$を元に条件が結果に影響を与えるか検証

仮説

- 帰無仮説 H_0 ：$b_1 = \cdots = b_M = 0$
- 対立仮説 H_1 ：$\neg(b_1 = \cdots = b_M = 0)$

試験番号	1	…	i	…	n
実験条件	3	⋯	1	⋯	8
結果	8.21	⋯	9.01	⋯	5.85

線形モデルへの帰着

- ❑ m_iをもとに説明変数$x_i \in \mathbb{R}^M$を作成
  - ❑ $j = m_i$のときのみ$x_{ij} = 1$それ以外は$x_{ij} = 0$
- ❑ 線形モデル

$$Y_i = b_0 + b_1 x_{i1} + \cdots + b_M x_{iM} + \epsilon_i$$

は１元配置分散分析の計測結果のモデルと同じ

以下の仮説は上記の説明変数におけるF検定と同じ

- ❑ 帰無仮説 H_0 ： $b_1 = \cdots = b_M = 0$
- ❑ 対立仮設 H_1 ：$\neg(b_1 = \cdots = b_M = 0)$

１元配置分散分析の実行方法

1. 上記に従って説明変数x_iを構築
2. 線形回帰におけるF検定を実施
3. H_0を棄却 ⇒ 実験条件は観測結果に影響を与えている
   - ❑ H_0が棄却できない ⇒ 影響があると断定する根拠が乏しい

2元配置分散分析

- 2つの要因それぞれ異なるM, K個の水準，n回の独立な試験を実施
- 未知母数：
  - 平均効果 $a \in \mathbb{R}$
  - 要因1の水準mの効果 $b_m \in \mathbb{R}$
  - 要因2の水準kの効果 $c_k \in \mathbb{R}$
  - 要因1,2の水準m, kの交互作用 $d_{mk} \in \mathbb{R}$
- i番目の試験の要因1の水準 $m_i \in \{1, \ldots, M\}$
- i番目の試験の要因2の水準 $k_i \in \{1, \ldots, K\}$
- i番目の試験の計測結果（標本）

$$Y_i = a + b_{m_i} + c_{k_i} + d_{m_i k_i} + \epsilon_i$$

誤差$\epsilon_i \sim N(0, \sigma^2)$

目標:

- $m_1, \ldots, m_n$，$k_1, \ldots, k_n$，$Y_1, \ldots, Y_n$を元に単独要因と交互作用それぞれが結果に影響を与えるか検証

試験番号	1	...	i	...	n
要因1	3	$\cdots$	1	$\cdots$	8
要因2	7	$\cdots$	9	$\cdots$	2
結果	8.21	$\cdots$	9.01	$\cdots$	5.85

2 元配置分散分析

各要因，相互作用それぞれ別に影響があるか検証

- 要因1単独の影響

仮説

- 帰無仮説 H_0 ：$b_1 = \cdots = b_M = 0$
- 対立仮説 H_1 ：$\neg(b_1 = \cdots = b_M = 0)$

いずれも線形回帰の
F検定により検証できる

- 要因2単独の影響

仮説

- 帰無仮説 H_0 ：$c_1 = \cdots = c_K = 0$
- 対立仮説 H_1 ：$\neg(c_1 = \cdots = c_K = 0)$

- 相互作用の影響

仮説

- 帰無仮説 H_0 ：$d_{11} = \cdots = d_{MK} = 0$
- 対立仮説 H_1 ：$\neg(d_{11} = \cdots = d_{MK} = 0)$

まとめ　データ解析特論　第4回

- 線形回帰
  - 最小二乗法
  - ガウス-マルコフ定理
  - 多項式回帰
- 線形回帰に関する検定
  - 回帰係数のT検定
  - 回帰係数のF検定
- 分散分析
  - １元配置分散分析
  - ２元配置分散分析

演習課題

- Manabaから確認，提出
- 締め切り２週間後