回帰分析

第 13 講 - 変数間の関係を推測する

村田 昇

講義概要

- 回帰分析
- 回帰係数の推定
  - 点推定
  - 区間推定
- 回帰係数の検定
  - 係数の有意性
- 決定係数

回帰分析

回帰分析

- データのある変量をその他の変量を用いて説明・予測するモデル (**回帰モデル**) を構築するための分析法
- 変量の分類
  - 説明する側 : **説明変数** (または独立変数，共変量など)
  - 説明される側 : **目的変数** (または被説明，従属，応答変数など)
- 説明変数・目的変数ともに複数個あってもよい
  - 目的変数は通常は 1 つ (複数の場合は個別に回帰モデルを構築)
  - 説明変数が 1 つの場合を **単回帰**, 2 つ以上の場合を **重回帰**
  - この講義では単回帰のみ扱う

回帰モデル

- 説明変数 : X
- 目的変数 : Y
- Y を X で説明する関係式として一次関数を考える

$$Y = \alpha + \beta X \quad \text{(線形回帰モデル)}$$

  - α: **定数項**
  - β: X の **回帰係数**
- **注意** : 非線形な関係への対応
  - 適切な変数変換 (二乗, 対数など) を施して線形な関係に変換
  - 弱い非線形性を線形で近似

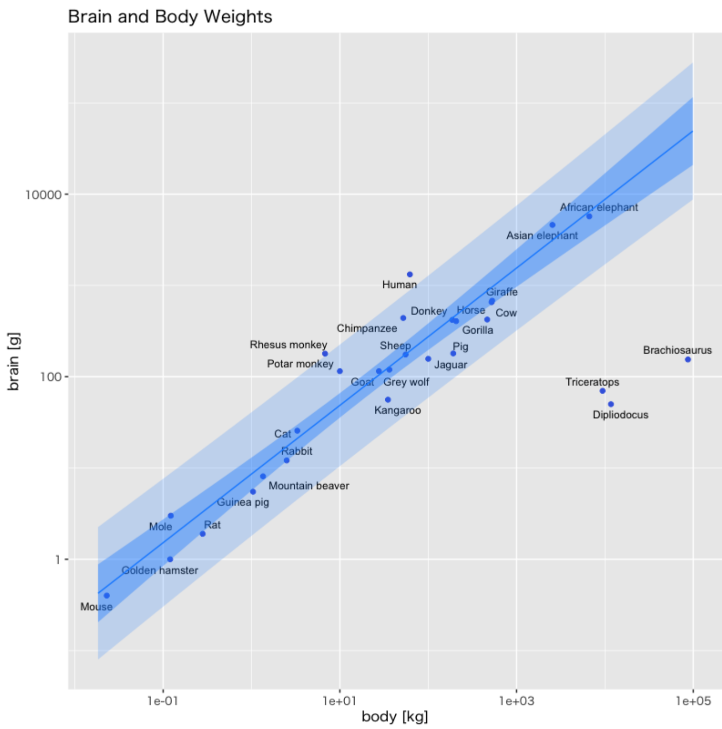

Figure 1: 脳の重さと体重の関係

回帰係数の点推定

回帰係数の点推定

- n 個の説明変数と目的変数の組 (X, Y) を観測

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

- 回帰モデル：データには観測誤差が含まれる

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

  - $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$: **誤差項** または **攪乱項**
- 線形回帰モデルのパラメータ α, β を推定

分析における仮定

- 説明変数 $X_1, \ldots, X_n$ は確率変数ではなく **確定値**
- 説明変数は一定値ではない
  ($X_1 = \cdots = X_n$ ではない)
- 誤差項 $\epsilon_1, \ldots, \epsilon_n$ は独立同分布な確率変数列
- 誤差項は 平均 0 分散 σ^2

最小二乗法

- 係数 α, β の回帰式で説明できない目的変数の変動

$$e_i(\alpha, \beta) = Y_i - (\alpha + \beta X_i) \quad (i = 1, \dots, n)$$

- 方針

  回帰モデルの当てはまりがよい
  $\Leftrightarrow$ $e_1(\alpha, \beta), \dots, e_n(\alpha, \beta)$ の絶対値が小さい

- 評価基準

  $e_1(\alpha, \beta), \dots, e_n(\alpha, \beta)$ の平方和 (**残差平方和**) を最小にするように α, β を決定

$$S(\alpha, \beta) = \sum_{i=1}^{n} e_i(\alpha, \beta)^2 = \sum_{i=1}^{n} \{Y_i - (\alpha + \beta X_i)\}^2$$

- **最小二乗推定量**

  $S(\alpha, \beta)$ を最小にするパラメータの組 $(\hat{\alpha}, \hat{\beta})$

- 最小二乗推定量の解

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

  ただし

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

回帰分析の計算

- 関数 `stats::lm()` : 線形モデルを当てはめる

```r
lm(formula, data, subset, na.action, ...)
#' formula: 式 (目的変数 ~ 説明変数)
#' data: データフレーム
#' subset: 対象とする部分データ
#' na.action: 欠損値の扱い
#' ...: 他のオプション. 詳細は '?lm' を参照
```

実習

練習問題

- 回帰分析におけるモデルの推定量の精度に関する確率シミュレーションを考えなさい.
- 東京の気象データを用いて，必要であれば適当な期間を抽出し，日射量から気温を説明する回帰モデルを構成しなさい.

回帰係数の区間推定

誤差項に関する仮定

- ϵ_i は正規分布に従う

- 上の仮定より $\hat{\alpha}, \hat{\beta}$ は **正規分布** に従う
- 点推定の平均と分散

$$
\mathbb{E}[\hat{\alpha}] = \alpha, \qquad \mathbb{E}[\hat{\beta}] = \beta,
$$
$$
\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}, \qquad \mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}
$$

- σ^2 が **既知なら** 正規分布を用いて信頼区間を構成

誤差分散の推定

- 一般に σ^2 は **既知でない** ためデータから推定
  - ϵ_i の平均は 0
  - σ^2 は ϵ_i の共通の分散
- 誤差と回帰式の関係

$$
\epsilon_i = Y_i - (\alpha + \beta X_i) \quad (i = 1, \ldots, n)
$$

- σ^2 の自然な推定量 (良いとは限らない)

$$
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2 \quad \text{ただし} \quad \hat{\epsilon}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i), \quad (i = 1, \ldots, n)
$$

- **残差** $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ の性質 (資料; 正規方程式)

$$
\sum \hat{\epsilon}_i = 0, \quad \sum \hat{\epsilon}_i X_i = 0.
$$

- 残差の二乗平均の性質 (標本分散と同様の計算)

$$
\mathbb{E}[\hat{\epsilon}_i^2] = \sigma^2 (n-2)/n \quad (i = 1, \ldots, n)
$$

- σ^2 の不偏推定量

$$
\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2.
$$

回帰係数の性質

- $\hat{\alpha}, \hat{\beta}$ の分散の推定量 (資料; Gauss-Markov の定理)

$$
\mathrm{s.e.}(\hat{\alpha})^2 = \frac{\hat{\sigma}^2 \sum_i X_i^2}{n \sum_i (X_i - \bar{X})^2}, \quad \mathrm{s.e.}(\hat{\beta})^2 = \frac{\hat{\sigma}^2}{\sum_i (X_i - \bar{X})^2}
$$

  - $\mathrm{s.e.}(\hat{\alpha}), \mathrm{s.e.}(\hat{\beta})$ は **標準誤差** と呼ばれる
- 以下は $\hat{\beta}$ と独立で自由度 $n-2$ の χ^2 分布に従う

$$
\frac{(n-2)\mathrm{s.e.}(\hat{\beta})^2}{\mathrm{Var}(\hat{\beta})}
$$

回帰係数の区間推定

- 以下の確率変数は自由度 $n-2$ の t 分布に従う

$$\frac{\hat{\beta}-\beta}{\mathrm{s.e.}(\hat{\beta})} = \frac{(\hat{\beta}-\beta)/\sqrt{\mathrm{Var}(\hat{\beta})}}{\sqrt{(n-2)\mathrm{s.e.}(\hat{\beta})^2/(n-2)\mathrm{Var}(\hat{\beta})}}$$

- $\gamma \in (0,1)$ に対する β の $1-\gamma$ 信頼区間

$$\left[\hat{\beta} - t_{1-\gamma/2}(n-2) \cdot \mathrm{s.e.}(\hat{\beta}),\ \hat{\beta} + t_{1-\gamma/2}(n-2) \cdot \mathrm{s.e.}(\hat{\beta})\right]$$

区間推定の計算

- 関数 `confint()`：係数の信頼区間を求める

```
confint(object, parm, level = 0.95, ...)
#' object: 関数 lm で推定したモデル
#' parm: 区間推定をするパラメタ．指定しなければ全て
#' level: 信頼係数
#' ...: 他のオプション．詳細は '?confint' を参照
```

- 関数 `predict()`：予測値の信頼区間を求める

```
predict(object, newdata, se.fit = FALSE, scale = NULL, df = Inf,
    interval = c("none", "confidence", "prediction"),
    level = 0.95, type = c("response", "terms"),
    terms = NULL, na.action = na.pass,
    pred.var = res.var/weights, weights = 1,
    rankdeficient = c("warnif", "simple", "non-estim", "NA", "NAwarn"),
    tol = 1e-6, verbose = FALSE,
    ...)
#' object: 関数 lm で推定したモデル
#' newdata: 予測値を計算する説明変数
#' interval: 何も付けない (none)・信頼区間 (confidence)・予測区間 (prediction)
#' level: 信頼係数 (既定値は 0.95)
#' ...: 他のオプション．詳細は '?predict.lm' を参照
```

- 関数 `broom::augment()` によるデータの情報 (tidyverse)

```
augment(
  x,
  data = model.frame(x),
  newdata = NULL,
  se_fit = FALSE,
  interval = c("none", "confidence", "prediction"),
  conf.level = 0.95,
  ...
)
#' x: 関数 lm で推定したモデル
#' newdata: data と異なる説明変数であてはめ・予測を行う
#' se_fit: 標準誤差を付けるか否か
#' interval: 信頼区間 (confidence)・予測区間 (prediction) を付ける
#' 詳細は '?broom::augment.lm' を参照
```

実習

練習問題

- 前問で作成した回帰モデルについて区間推定を行いなさい．

回帰係数の有意性検定

回帰係数の有意性

- 説明変数 X が目的変数 Y を説明・予測するのに本当に役立っているかを検証

$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_1 : \beta \neq 0$$

- β の **有意性の検定**
  帰無仮説 H_0 が有意水準 γ で棄却されるとき，β は有意水準 γ で **有意である**

回帰係数の有意性検定

- 帰無仮説 H_0 が正しければ以下の統計量は自由度 n–2 の t 分布に従う

$$t = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})}$$

- 対立仮説 H_1 が正しければ, $\hat{\beta}$ は 0 でない値 β に近い値を取ることが期待されるため，$|t|$ は 0 から離れた値を取る

- 棄却域による検定
  有意水準を $\gamma \in (0, 1)$ とし，$\hat{\beta}$ の t **値** が以下の場合には帰無仮説を棄却

$$|t| > t_{1-\gamma/2}(n-2)$$

- p 値による検定
  以下で定義される $\hat{\beta}$ の p **値** が γ 未満の場合に帰無仮説を棄却

$$(p\ \text{値}) = 2\int_{|t|}^{\infty} f(x)dx$$

係数の検定

- 関数 `stat::summary()` : 情報の要約 (base R)

```
summary(object)
#' object: 関数 lm() で推定したモデル
#' 関数の出力 (リスト名 $"名前" で参照可能)
#' coefficients : 係数と t 統計量
#' fstatistics : F統計量 (モデルの評価)
#' 詳細は '?summary.lm' を参照
```

- 関数 `broom::tidy()` : 回帰係数の情報 (tidyverse)

```
tidy(x, conf.int = FALSE, conf.level = 0.95, exponentiate = FALSE, ...)
#' x: 関数 lm() で推定したモデル
#' conf.int: 信頼区間を付けるか否か
#' conf.level: 信頼係数
#' 詳細は '?broom::tidy.lm' を参照
```

- 関数 `broom::glance()` : モデルの統計情報 (tidyverse)

```
glance(x, ...)
#' x: 関数 lm() で推定したモデル
#' F統計量は statistic/p.value の列
#' 詳細は '?broom::glance.lm' を参照
```

実習

練習問題

- 前問で作成した回帰モデルについて係数の検定を行いなさい.

決定係数

決定係数

- X が Y の変動をどの程度説明できるかを数量化
- **決定係数** (あるいは **寄与率**)

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- $\hat{Y}_i$ は**あてはめ値** または **予測値** と呼ばれる

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \quad (i = 1, \dots, n).$$

- 以下の等式が成立

$$\begin{aligned}
\hat{\epsilon}_i &= Y_i - \hat{Y}_i \quad (i = 1, \dots, n) \\
\sum_{i=1}^n \hat{\epsilon}_i &= 0, \\
\frac{1}{n} \sum_{i=1}^n Y_i &= \bar{Y}, \\
\frac{1}{n} \sum_{i=1}^n \hat{Y}_i &= \bar{Y}.
\end{aligned}$$

- 決定係数

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- R^2 の成分
  - R^2 の分子 : あてはめ値の (標本平均まわりでの) 変動
  - R^2 の分母 : 目的変数の (標本平均まわりでの) 変動
- R^2 の意味
  - 回帰式が目的変数の変動をどの位説明できるか評価
  - 大きいほど説明力が高いと解釈される

決定係数の別表現

- R^2 は以下のように書き直すことも可能
  - 目的変数の観測データとあてはめ値の相関の二乗

$$R^2 = \left\{ \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} \cdot \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \right\}^2$$

  - 説明変数と目的変数の観測データの間の相関の二乗

$$R^2 = \left\{ \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \right\}^2$$

自由度調整済み決定係数

- 不偏分散による R^2 の修正
  - 残差 ϵ_i と目的変数 Y_i の標本分散による表現

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

  - 不偏推定量で代替：**自由度調整済み決定係数** (または寄与率)

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

決定係数の計算

- 関数 `stat::summary()`：情報の要約 (base R)

```r
summary(object)
#' object: 関数 lm() で推定したモデル
#' 関数の出力 (リスト名 $"名前" で参照可能)
#' r.squareds : 決定係数
#' adj.r.squareds : 自由度調整済み決定係数
#' 詳細は '?summary.lm' を参照
```

- 関数 `broom::tidy()`：回帰係数の情報 (tidyverse)

```r
tidy(x, conf.int = FALSE, conf.level = 0.95, exponentiate = FALSE, ...)
#' x: 関数 lm() で推定したモデル
#' conf.int: 信頼区間を付けるか否か
#' conf.level: 信頼係数
#' 詳細は '?broom::tidy.lm' を参照
```

- 関数 `broom::glance()`：モデルの統計情報 (tidyverse)

```r
glance(x, ...)
#' x: 関数 lm() で推定したモデル
#' F統計量は statistic/p.value の列
#' 詳細は '?broom::glance.lm' を参照
```

演習

練習問題

- 前問で作成した回帰モデルについて決定係数を確認しなさい.
- 説明変数として降水量を用いた回帰モデルについて検討しなさい.