統計解析

古谷知之

本授業の目的と方法

- 統計基礎（単回帰分析）までの内容を理解していることを前提に、より応用的な統計モデリングについて理解を深める。
- 主に計量経済学・計量政治学・生物統計学などで用いられるような手法やデータを扱う。数式を多用するが、自身で理解するようつとめること。
- 統計ソフトRを用いて統計モデルの実装方法を学ぶ。

成績評価

- 以下の方法により授業の成績を評価する（100点満点）

- 出席課題（39点）
  - SFC-SFS上で提出。授業の理解度を評価。

- 最終レポート課題（61点）
  - データ分析に関する課題を出題する。

授業概要

* 履修者の状況に応じて変更される場合がありますが、全体としては以下のような授業構成となります。
* 講義の中でR演習を行うこともあります。

第1回　ガイダンス・単回帰分析
第2回　重回帰分析(1)
第3回　重回帰分析(2)
第4回　一般化線形回帰モデル(1)
第5回　一般化線形回帰モデル(2)
第6回　一般化線形回帰モデル(3)
第7回　一般化線形回帰モデル(4)
第8回　一般化線形回帰モデル(5)
第9回　一般化線形回帰モデル(6)
第10回　一般化線形混合モデル
第11回　状態空間モデル
第12回　R演習(1)
第13回　R演習(2)
第14回　R演習(3)

統計モデルの種類

	主な推定方法	データ分布	回帰係数
線形回帰モデル （単回帰・重回帰など）	最小二乗法	正規分布	一変数に一つ
一般化線形モデル	最尤推定法	正規分布以外の分布も可能	一変数に一つ
一般化線形混合モデル	最尤推定法	正規分布以外の分布も可能	変数の個体差に応じて推定可能
階層ベイズモデル	ベイズ推定	正規分布以外の分布も可能	変数の個体差に応じて推定可能

本授業で扱う統計モデル

- 線形回帰モデル
  - 単回帰モデル、重回帰モデル
- 一般化線形回帰モデル
  - 離散：ポアソン回帰モデル、二項反応モデル（ロジスティック回帰モデル、プロビット回帰モデル、補対数対数モデル）、負の二項分布モデル、ゼロ過剰ポアソン回帰モデル、ゼロ過剰負の二項分布モデル
  - 連続：ガンマ回帰モデル、ベータ回帰モデル、指数-ガウス回帰モデル
  - スパース：Lasso回帰モデル、Ridge回帰モデル
- 一般化線形混合モデル
  - マルチレベルモデル
- 状態空間モデル

授業内容

- 単回帰分析の手順と考え方
- 最小二乗法
- 不偏推定量
- 残差の性質
- 決定係数
- 回帰係数の統計的検定と信頼区間
- 予測値と予測誤差
- 単回帰分析のまとめ方

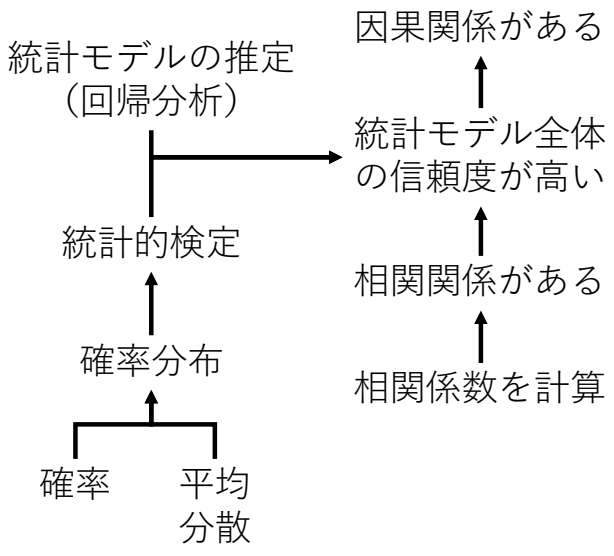
単回帰分析で検討すべきこと

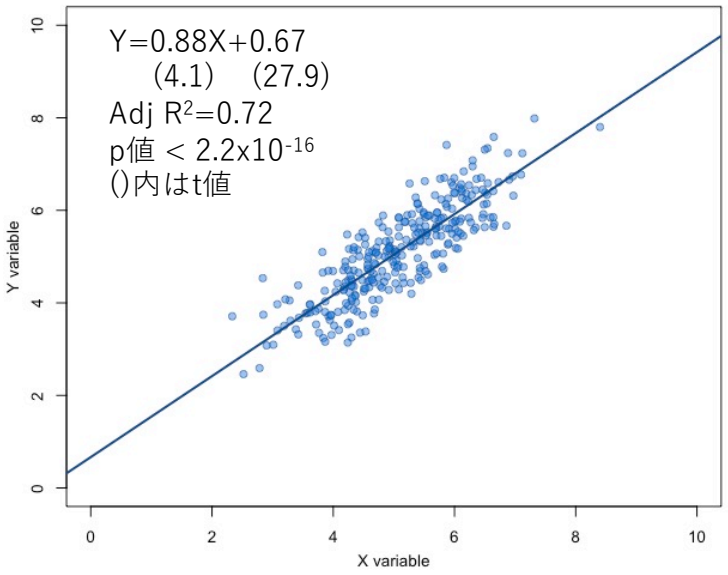
- 回帰係数の値：偏回帰係数
- 回帰係数の統計的有意性：t検定
- 回帰係数の信頼度：信頼区間
- 予測への適用可能性：決定係数
- 外れ値の検出：残差解析

回帰分析の主な手順

- 因果関係を証明したいテーマを設定する
- 因果関係を示すのに適した従属変数と独立変数を用意する
- 回帰係数、回帰係数のt値とp値、決定係数、自由度修正済み決定係数などの推定量・統計量を計算する
- 回帰係数のt検定を行い、採用する従属変数を選定する
- 推定結果が事前に想定した仮説とあわない（回帰係数が統計的に有意でない、決定係数の値が小さい、など）場合には、他の変数を用意するなどして分析をやり直す
- 決定係数が大きく、回帰係数がいずれも統計的に有意な結果が得られるた場合には、最終的に推定結果を解釈する

回帰分析により因果関係を示す手順





回帰分析

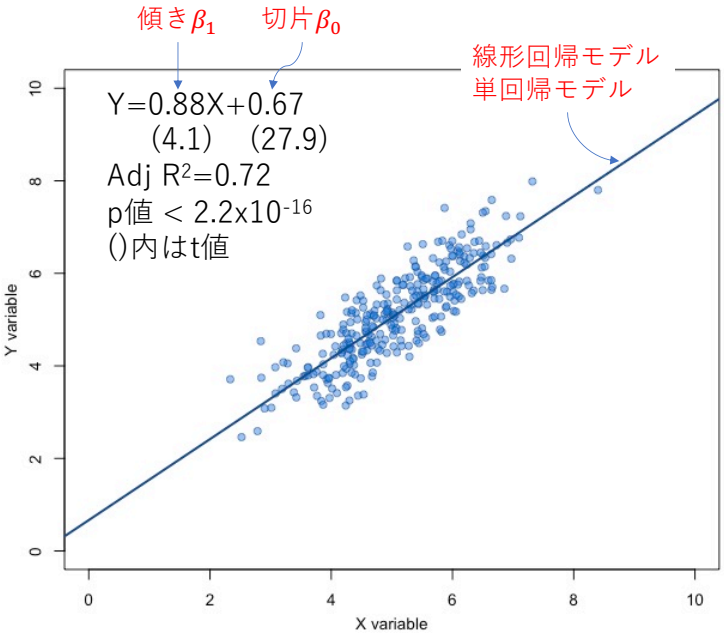
- 相関関係を計算するだけでは、因果関係を説明したことにはならない
- 因果関係を説明するには、原因と考えられる変数を説明変数（独立変数）により、結果と考えられる変数を被説明変数（従属変数）を説明（予測）する統計モデルを推定する
- 線形回帰：独立変数と従属変数との関係が線形関数で表せる
- 単回帰分析：一つの独立変数を含む回帰分析
- 重回帰分析：複数の独立変数を含む回帰分析

従属変数と独立変数

- 従属変数y
  - 説明される変数（結果）
  - 被説明変数、応答変数、などともいう
- 独立変数x
  - 説明する変数（原因）
  - 説明変数、予測変数、などともいう
- これら以外にも、ダミー変数などが用いられる
- 従属変数と独立変数は、分析する人の想定や仮定によって選択される
  - 何らかの「理論的根拠」や「合理的」な判断に基づいて選ばれる

単回帰分析

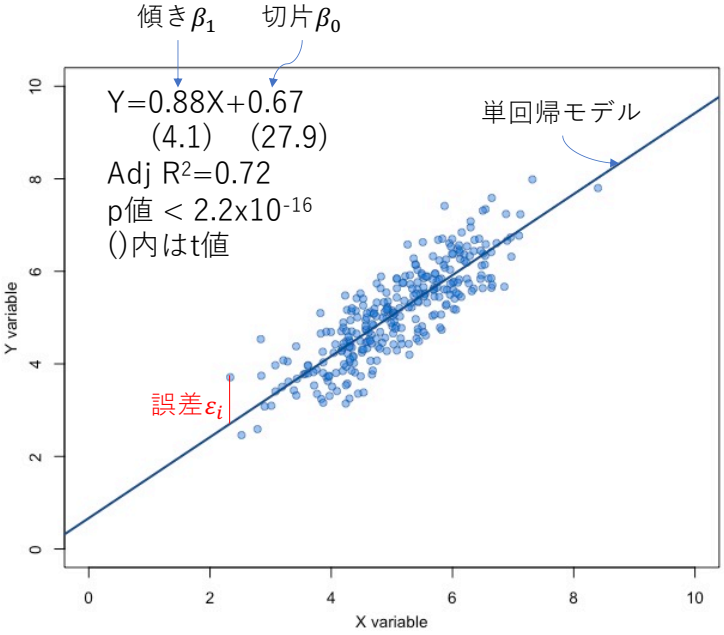
- ２つの変数x, yについて、
  $$y = \beta_0 + \beta_1 x$$
  のような線形の式形状で表される統計モデルを単回帰モデルという（左図の青直線のこと）
- ここで、単回帰モデルに用いられるβ_0とβ_1は回帰係数という（β_0は切片、β_1は傾きを意味する）
- 従属変数、独立変数ともに正規分布するデータを扱う



単回帰分析

- 実際には、単回帰モデル式
  $$y = \beta_0 + \beta_1 x$$
  と各データ(x_i, y_i)との間には、誤差εと呼ばれる乖離がある

- データ(x_i, y_i)について
  $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
  となる

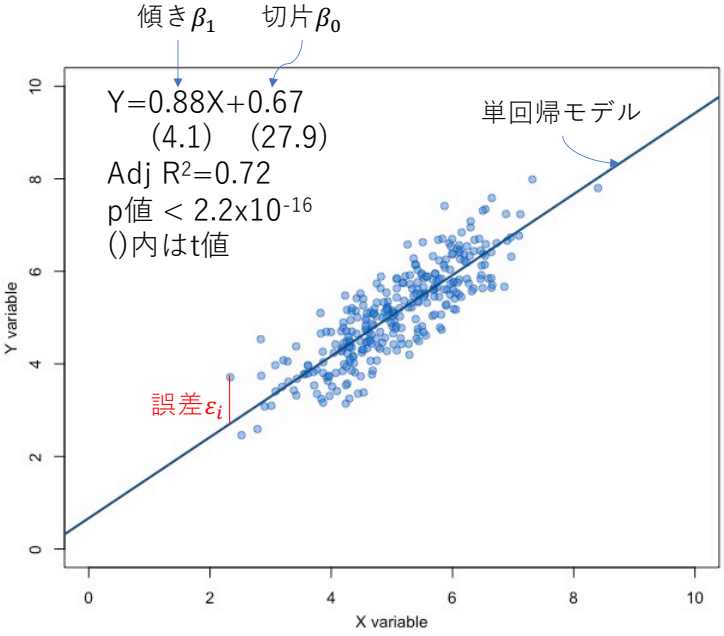


単回帰分析

- 誤差が大きいと、モデルの信頼性は低い（現実を十分に説明できていない）
- つまり誤差が小さいモデルを得ることが、現実のデータの状況を説明しうるモデルだということができる
- データ(x_i, y_i)の誤差ε_i

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

- 「誤差が小さい」とは各データの誤差ε_iの総和$\sum_{i=1}^{n} \varepsilon_i$が小さいということ



単回帰分析

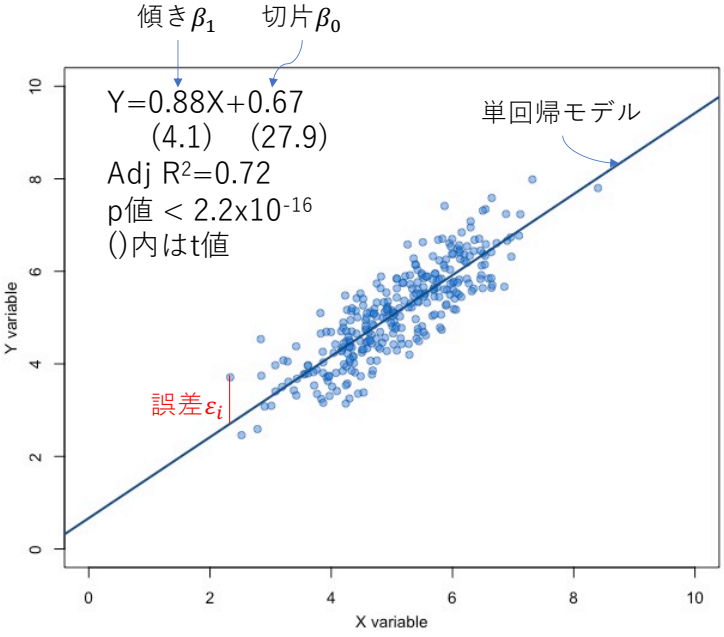
- データ(x_i, y_i)の誤差ε_i

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

- 誤差ε_iの総和$\sum_{i=1}^{n} \varepsilon_i$が最小となるような回帰係数$\beta_0, \beta_1$の組み合わせを求める

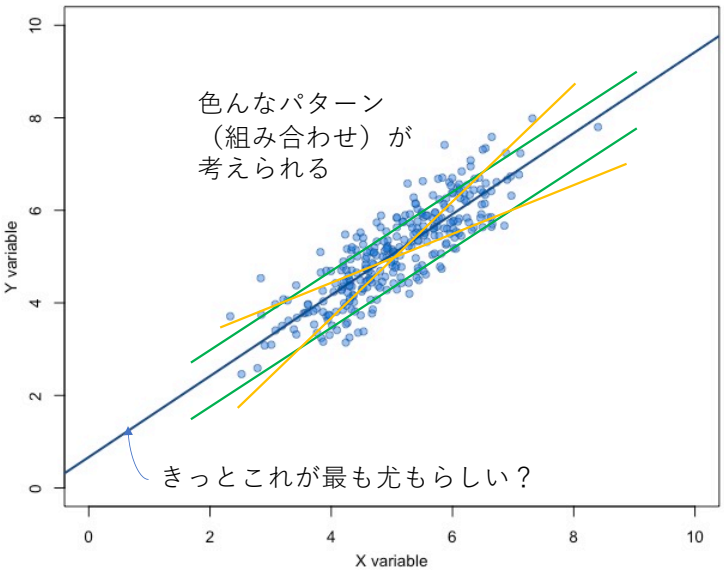
基本的な考え方

- 誤差の確率関数に関する最適解を求める



回帰係数の推計方法

- 様々な回帰係数の組み合わせやパターンが考えられるが、最も尤（もっと）もらしい組み合わせを１つだけ示すほうが、理解しやすい
- その組み合わせとは、回帰係数の分布の平均$(\overline{\beta_0}, \overline{\beta_1})$
- 回帰係数は確率分布する



回帰係数の推計方法

- 従属変数y_iが与えられたときに予測結果$\hat{y_i} = \overline{\beta_0} + \overline{\beta_1} x_i$が得られる事後確率$p(\hat{y_i}|y_i)$は、次式のように表せる

$$p(\hat{y_i}|y_i) = p(\overline{\beta_0} + \overline{\beta_1} x_i|y_i)$$

- 実際にはまだ回帰係数は得られていないので(β_0, β_1)と表す
- ベイズの定理を用いると、

尤度関数　　周辺確率

$$p(\beta_0 + \beta_1 x_i|y_i) = \frac{p(y_i|\beta_0 + \beta_1 x_i)p(\beta_0 + \beta_1 x_i)}{p(y_i)}$$

事後分布　　周辺確率

回帰係数の推計方法

- 事後分布を直接求めなくても、尤度関数を計算すればよい

尤度関数　　周辺確率

$$p(\beta_0 + \beta_1 x_i | y_i) = \frac{p(y_i | \beta_0 + \beta_1 x_i) p(\beta_0 + \beta_1 x_i)}{p(y_i)}$$

事後分布　　周辺確率

- 尤度関数はどうやって計算すればよいのか？

回帰係数の推計方法

- 尤度関数：予測結果$\hat{y_i}$が与えられた前提（条件）での実際のデータ（従属変数）y_iの当てはまり度合い

$$p(y_i | \hat{y_i}) = p(y_i | \beta_0 + \beta_1 x_i) = p(y_i | x_i, \beta_0, \beta_1)$$

- ところで、標本データとしてのy_iはどのようなデータか？

- yは正規分布する$y \sim N(\mu_y, \sigma_y^2)$
- 各標本$(y_1, y_2 \ldots, y_i, \ldots, y_n)$は互いに独立である

回帰係数の推計方法

- 従属変数y_iが正規分布する場合、尤度確率は正規分布に従う

$$p(y_i|x_i,\beta_0,\beta_1)=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left\{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1x_i)^2\right\}$$

- この尤度関数は y_iが平均$(\beta_0+\beta_1x_i)$、分散σ^2の正規分布に従うことを意味する

$$E(y_i|x_i,\beta_0,\beta_1)=\beta_0+\beta_1x_i$$

$$V(y_i|x_i,\beta_0,\beta_1)=\sigma^2$$

正規分布（ガウス分布）

- 平均μ、分散σ^2となる以下の確率密度関数に従う分布を正規分布という

$$N\left(\mu,\sigma^2\right)=f\left(x\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

$$E\left(Z\right)=\mu,V\left(Z\right)=\sigma^2$$

回帰係数の推計方法

- 従属変数y_iが正規分布する場合、尤度確率は正規分布に従う

$$p(y_i|x_i,\beta_0,\beta_1)=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left\{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1x_i)^2\right\}$$

- 回帰係数の組み合わせ(β_0,β_1)を得るために、全てのデータ$(y_1,\dots,y_n)$の尤度確率の積を計算する

$$\prod_{i=1}^{n}p(y_i|x_i,\beta_0,\beta_1)$$
$$=p(y_1|x_1,\beta_0,\beta_1)\times p(y_2|x_2,\beta_0,\beta_1)\times\cdots\times p(y_n|x_n,\beta_0,\beta_1)$$

回帰係数の推計方法

- 尤度確率の積は尤度関数と呼ばれる
- データが正規分布する場合、単回帰モデルの尤度関数は次式のように表すことができる

$$\prod_{i=1}^{n}p(y_i|x_i,\beta_0,\beta_1)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}exp\left\{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1x_i)^2\right\}$$

$$=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2\right\}$$

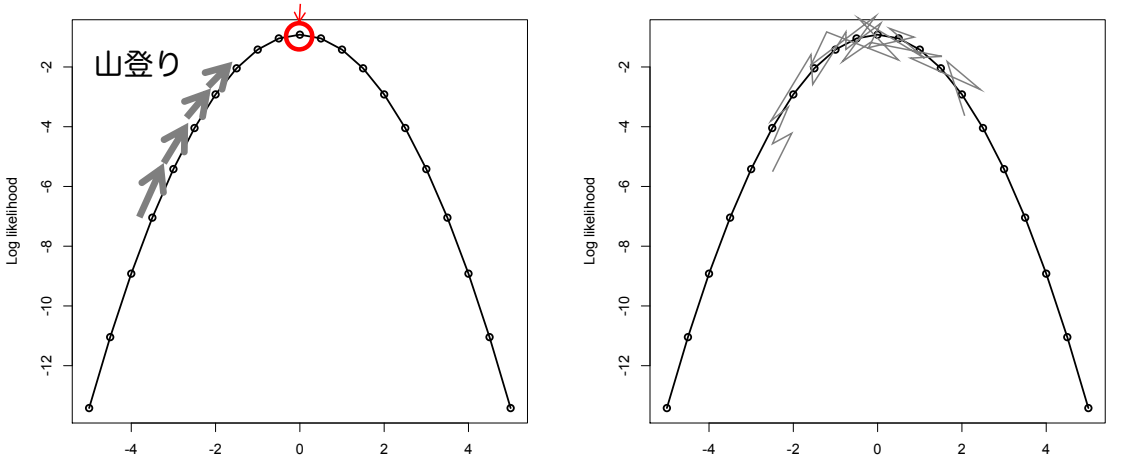
回帰係数の推計方法

- 尤度関数を直接数値計算で求めるのは困難なので

①尤度関数の対数（対数尤度関数）を使う（最尤推定法）

②シミュレーションで尤度関数を求める（ベイズ推定法）

①対数尤度関数が最大となる点を点推定　②シミュレーションで尤度関数を推定



最尤推定法による回帰係数の推計

- 対数尤度関数

$$\log\left[\prod_{i=1}^{n} p(y_i|x_i,\beta_0,\beta_1)\right]$$

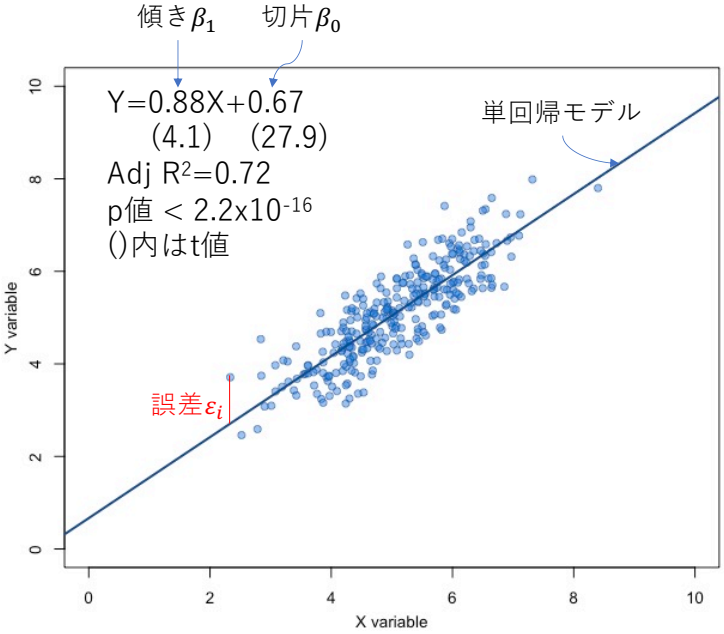
$$= \log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2\right\}\right]$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

- 上に凸な対数尤度関数を最大化することは、下に凸な $\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$ を最小化することと同じ

最小二乗法による回帰係数の推計

- $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$を最小化することの意味
- 誤差$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$の二乗和$\sum_{i=1}^{n} \varepsilon_i^2$を最小化することと同じ
- つまり誤差が最も小さい（最も尤もらしい）回帰モデル推定結果が得られるということになる



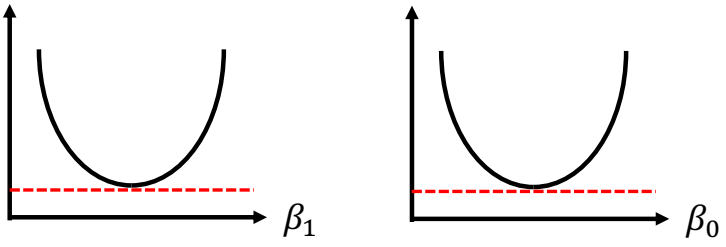
最小二乗法による回帰係数の推計

- $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$という関数について、未知の回帰係数$\beta_0$と$\beta_1$を計算して求めるにはどうしたらよいか？
- β_0とβ_1の関数がそれぞれ最小となるβ_0とβ_1を求めれば良い

β_1の関数: $\sum_{i=1}^{n}\{\beta_1^2 x_i^2 + 2\beta_1 x_i(\beta_0 - y_i) + (\beta_0 - y_i)^2\}$

β_0の関数: $\sum_{i=1}^{n}\{\beta_0^2 + 2\beta_0(\beta_1 x_i - y_i) + \beta_1^2 x_i^2 - 2\beta_1 x_i y_i + y_i^2\}$

- β_1に関する関数とβ_0に関する関数はともに下に凸な関数
  - グラフの傾きが 0 になる点のβ_1とβ_0を求めれば良い



最小二乗法による回帰係数の推計

- β_1の関数: $\sum_{i=1}^{n}\{\beta_1^2 x_i^2 + 2\beta_1 x_i(\beta_0 - y_i) + (\beta_0 - y_i)^2\}$
- β_1について偏微分すると0になることから、

$$2\beta_1 \sum_{i=1}^{n} x_i^2 + 2\beta_0 \sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i = 0$$

- β_0の関数: $\sum_{i=1}^{n}\{\beta_0^2 + 2\beta_0(\beta_1 x_i - y_i) + \beta_1^2 x_i^2 - 2\beta_1 x_i y_i + y_i^2\}$
- β_0について偏微分すると0になることから、

$$2n\beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} y_i = 0$$

- この２式からβ_1とβ_0を計算する

最小二乗法による回帰係数の推計

- β_1とβ_0を計算して得られた解を不偏推定量という→後述
- 不偏推定量それぞれ$\hat{\beta}_1$および$\hat{\beta}_0$と表される
- まず$\hat{\beta}_1$は以下のように計算される

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

分母分子をともにn^2で割ると

$$= \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i}{n}\frac{\sum_{i=1}^{n} y_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}$$

最小二乗法による回帰係数の推計

$$\hat{\beta}_1 = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2}$$

ここで分母と分子を式変形することで、以下のように表せる

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x}$$

最小二乗法による回帰係数の推計

$\hat{\beta}_1$の分子は、

$$\frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - 2\bar{x}\bar{y} + \bar{x}\bar{y}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y}\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \bar{x})(y_i - \bar{y})\}$$

最小二乗法による回帰係数の推計

$\hat{\beta}_1$の分母は、

$$\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - 2\bar{x}^2 + \bar{x}^2$$

$$= \frac{\sum_{i=1}^{n} x_i^2}{n} - 2\frac{\sum_{i=1}^{n} x_i}{n}\bar{x} + \bar{x}^2$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + n\bar{x}^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

最小二乗法による回帰係数の推計

- 次に$\hat{\beta}_0$は以下のように計算される

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \bar{y} - \hat{\beta}_1 \bar{x}$$

既に得られた$\hat{\beta}_1$と$\bar{x}$および$\bar{y}$を代入することで得られる

最小二乗法による回帰係数の推計

- 不偏推定量$\hat{\beta}_1$と$\hat{\beta}_0$をまとめると、以下のようになる

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

単回帰（線形回帰）モデルの特徴

- 独立変数xと従属変数yがともに正規分布するとき、回帰モデルの誤差項εと回帰係数β_0とβ_1も、ともに正規分布する
- 変数x, yと回帰係数β_0, β_1が与えられたとき、誤差項εは平均0, 分散σ^2の正規分布に従う（誤差項の分散は未知）
  - 誤差は従属変数の予測値と実績値との差分なので、理想的には差分＝０が望ましいが、少なくとも予測値と実績値の平均は一致するはず
  - ただし分散は異なるかもしれない
- また尤度関数は平均$\beta_0 + \beta_1 x_i$, 分散σ^2の正規分布となる
- 回帰係数も確率分布（正規分布）するが、最も説明しやすい値はその平均値と一致する（はず）
- 推計された回帰係数は、真のパラメータと一致する

不偏推定量

- 推定量$\hat{\beta}_j$の期待値が真のパラメータと一致するときに不偏推定量という
- 最小二乗法による推定量が不偏推定量となるかを示す

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \cdots$$

$$= \beta_1 + \frac{n\sum_{i=1}^{n} x_i \varepsilon_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \varepsilon_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

不偏推定量

- 両辺に期待値をとると

$$E\left[\hat{\beta}_1\right] = E\left[\beta_1 + \frac{n\sum_{i=1}^{n} x_i \varepsilon_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \varepsilon_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\right]$$

$$= E[\beta_1] + E\left[\frac{n\sum_{i=1}^{n} x_i \varepsilon_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \varepsilon_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\right]$$

$E(\varepsilon_i|x_i) = 0$

$$= \beta_1 + E\left[\frac{n\sum_{i=1}^{n} x_i E(\varepsilon_i|x_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} E(\varepsilon_i|x_i)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\right] = \beta_1$$

となるので、 $E\left[\hat{\beta}_1\right] = \beta_1$となる。同様に$E\left[\hat{\beta}_0\right] = \beta_0$となる。

一致推定量

- 推定量β_jについて、標本数を非常に大きく（無限大に）したとき、真の値と一致する推定量を一致推定量という
- 最小二乗法による推定量が一致推定量となるかを示す

$$\hat{\beta}_1 = \beta_1 + \frac{n\sum_{i=1}^{n} x_i\varepsilon_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \varepsilon_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

- データは互いに独立であることから大数の法則が成り立つ。

$$\frac{\sum_{i=1}^{n} x_i}{n} \to E[x_i], \frac{\sum_{i=1}^{n} \varepsilon_i}{n} \to E[\varepsilon_i], \frac{\sum_{i=1}^{n} x_i\varepsilon_i}{n} \to E[x_i\varepsilon_i], \frac{\sum_{i=1}^{n} x_i^2}{n} \to E\left[x_i^2\right]$$

一致推定量

- この性質を用いると、

$$\hat{\beta}_1 = \beta_1 + \frac{n\sum_{i=1}^{n} x_i\varepsilon_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \varepsilon_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \beta_1 + \frac{E[x_i\varepsilon_i] - E[x_i]E[\varepsilon_i]}{E\left[x_i^2\right] - (E[x_i])^2} = \beta_1$$

となる。ここで、 $E[x_i\varepsilon_i] = 0$および$E[\varepsilon_i] = 0$である。

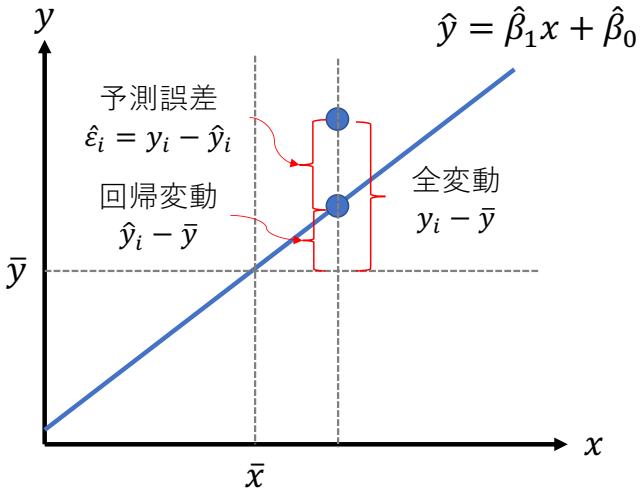
線形回帰モデルの留意点

- 線形回帰モデルは、分析において非常に多くの仮定を積み上げていることに注意しなくてはならない

- データは正規分布する
- データは互いに独立である
- 標本数が一定以上の大きさで確保できる
- 回帰係数は全てのデータに対して同一である
- 誤差項は互いに独立で同一の正規分布に従う

決定係数

- 推定された回帰モデル全体の当てはまりの良さを示す
- 予測したいデータy_i
- データの平均$(\bar{x}, \bar{y})$
- 推定した回帰式$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$を用いて推定されたデータ$\hat{y}_i$
- 予測誤差$\hat{\varepsilon}_i = y_i - \hat{y}_i$
- 回帰変動$\hat{y}_i - \bar{y}$
- 全変動$y_i - \bar{y}$



決定係数

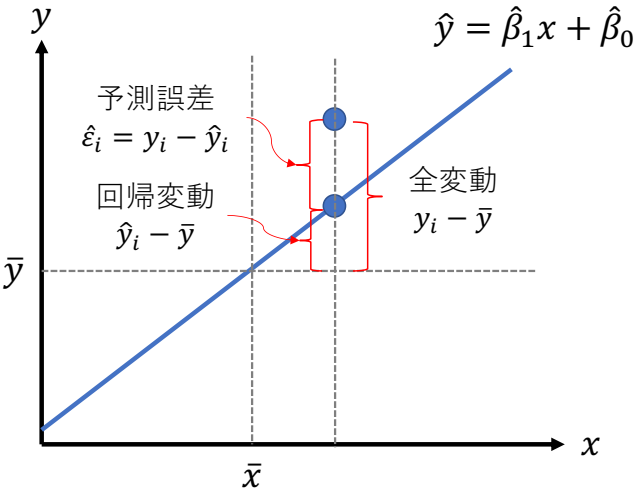
- 回帰変動$\hat{y}_i - \bar{y}$の平方和を全変動$y_i - \bar{y}$の平方和で割ったものを決定係数R^2という

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

- これは、予測誤差$\hat{\varepsilon}_i = y_i - \hat{y}_i$の平方和を全変動$y_i - \bar{y}$の平方和で割ったものを1から引いたものと等しい

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

- 決定係数は0-1の値をとり、1に近いほどモデルの説明力は高い



自由度修正済み決定係数

- 説明変数の数を増やすほど決定係数は 1 に近づく
- 説明変数が多い場合、そうした性質を補正した自由度修正済み決定係数$\mathrm{Adj.}\,R^2$が用いられる

$$\mathrm{Adj.}\,R^2 = 1 - \frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

- ここで、 nは標本数、 kは説明変数の数である

最小二乗法による回帰係数の推計

- 単回帰モデルの回帰係数$\hat{\beta}_1$と$\hat{\beta}_0$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- 回帰係数は不偏推定量であり、一致推定量である
- 回帰係数は正規分布する確率分布の平均値でもある

回帰係数の統計量

- 線形回帰分析では誤差ε_iが正規分布に従うという強い仮定をおいている。誤差項の分散は未知であるが、標本毎に同じであるとも仮定している。すなわち、

$$\varepsilon_i \sim N(0, \sigma^2)$$

- この仮定のもとでは、回帰係数の不偏推定量が以下のように正規分布するという性質を持つ

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

- 不偏推定量は回帰係数の分布の平均値と考えることもできる。このとき、一致推定量の性質とも合致する

残差の性質

- 推定された回帰モデルの残差$e_i = y_i - \hat{y}_i$（実績値と予測値との差分）は、回帰分析で説明がつかない部分を意味する
- 残差e_iは互いに独立に（各i毎に）平均0、分散σ^2の正規分布に従う。$e_i \sim N(0, \sigma^2)$
- 残差平方和S_eは次のように求められる

$$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

- これの自由度は$n-2$なので、回帰の残差分散s_e^2は

$$s_e^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}$$

となる

残差の性質

- 更に残差$e_i = y_i - \hat{y}_i$は、以下の 2 つの性質を持つ

- 残差の和は0になる

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

- 残差と従属変数とは直交する

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) x_i = 0$$

残差の性質

- 誤差εが正規分布に従う仮定$\varepsilon \sim N(0, \sigma^2 I)$の下で、残差$e$も正規分布に従う

$$e \sim N\left(0, \sigma^2 (I - X(X^T X)^{-1} X^T)\right)$$

$$X^T = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}$$

- ここで、 $H = X(X^T X)^{-1} X^T$はハット行列と呼ばれる

残差の性質

- 残差を標準偏差で基準化した残差e_i/σの平方和は、自由度$n - k - 1$（単回帰分析の場合は$n - 2$）のχ^2分布に従う

$$\sum_{i=1}^{n} \left(\frac{e_i}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2(n - k)$$

回帰係数に関する検定

- 回帰係数が正規分布するという性質を標準化すると、以下のような性質が得られる

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim N(0,1)$$

回帰係数に関する検定

- 残差の性質から以下の分布は自由度$n-2$のt分布に従う

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}} \Big/ \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2 (n-2)}$$
$$= \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sum_{i=1}^{n} e_i^2 / (n-2) \sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t(n-2)$$

- ここで残差の制約条件が$\sum_{i=1}^{n} e_i = 0$と$\sum_{i=1}^{n} e_i x_i = 0$の２つあることから、自由度は2下がっている

回帰係数に関する検定

- この統計量を用いて、回帰係数に意味があるのかを検証できる
- 「回帰係数は従属変数の説明に寄与していない」という帰無仮説を棄却できれば、「回帰係数に意味がないとは言えない」ということになる→「回帰係数に意味がある」ことを直接照明するのではなく、「回帰係数に意味がない」ことを否定する

- このとき帰無仮説は以下のようになる

$$H_0: \beta_1 = 0$$

回帰係数に関する検定

- 誤差の分散σ^2をその普遍推定量$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} e_i^2/(n-2)\right)^2$で置き換えると、以下のような関係性が導ける

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sum_{i=1}^{n} e_i^2/(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

回帰係数に関する検定

- 同様に、$H_0: \beta_0 = 0$に対する仮説検定は、

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} e_i^2 / n(n-2) \sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim t(n-2)$$

についてt検定を行えば良い

回帰係数は統計的に信頼できるのか？

- 回帰係数$\hat{\beta}$が平均μ、分散σ^2の正規分布$N(\mu, \sigma^2)$に従う確率変数であるとき、$\mu - 1.96\sigma^2 < \hat{\beta} < \mu + 1.96\sigma^2$が成立する確率は95%である。これを95%信頼区間という（復習）。

- ただしこのことが成立するのは、標本数nが非常に大きいとき

- 十分に大きな標本数nが得られないとき、この信頼区間はどのようにして得ることができるのか？
- また実際には、回帰係数の平均も分散も事前にはわからない

回帰係数の信頼区間

- （平均値としての）回帰係数の統計的信頼度を知りたい
- 正規分布に基づく統計的信頼区間を用いるには、回帰係数の平均と分散を知っていなくてはならない（正規分布の限界）
- しかし平均が事前には分からないので分散も分からない
  →正規分布による信頼区間をそのまま用いることができない
- シミュレーション（ベイズ推定）を用いて正規分布の平均と分散を推計することができるが、ここでは古典的な統計学のマナーに従って、正規分布に近い確率分布（t分布）を用いる

回帰係数の信頼区間

- 回帰分析の独立変数の数がk個（単回帰分析のとき$k=1$）、標本数がn個のとき独立変数xの偏回帰係数$\hat{\beta}$について、$\frac{\beta-\hat{\beta}}{SE}$は自由度$\mathrm{n}-k-1$の$t$分布に従う
- ここでSEは標準誤差(standard error)であり、不偏分散s_x^2（標本分散σ_x^2）を用いて計算される

$$\mathrm{SE}=\sqrt{s_x^2/n}$$

$$s_x^2=\frac{n-1}{n}\sigma_x^2$$

回帰係数の信頼区間

- 偏回帰係数$\hat{\beta}$の95%信頼区間は以下のようになる

$$\left[\hat{\beta} - t_{n-k-1,0.025}\sqrt{s_x^2/n}\,, \hat{\beta} + t_{n-k-1,0.025}\sqrt{s_x^2/n}\right]$$

または

$$\left[\hat{\beta} - t_{n-k-1,0.025}\,SE, \hat{\beta} + t_{n-k-1,0.025}\,SE\right]$$

- 標本数が多く有意水準5%のとき、$\hat{\beta} \pm 1.96SE$が目安となる

仮説検定

- 回帰係数が統計的に意味がある（有意である）ことを示すために、仮説検定と呼ばれる手法が用いられることがある

手順

- 回帰係数の分布に何らかの仮定を置く
- 回帰係数と回帰係数の分布との間に検定したい仮説を設定する
- 帰無仮説：回帰係数=0、対立仮説：回帰係数=0ではない
- 回帰係数を推計する
- 推計した回帰係数のp値を計算する
- p値が統計的有意水準α以下であれば帰無仮説を棄却する（有意水準αより大きければ回帰係数に統計的な意味はない）

仮説検定

- 一般に仮説検定は、母集団から抽出された標本について、標本の統計的性質が母集団の統計的性質と「異なる」ことを示すために用いられる

仮説検定の手順

- 検定したい統計量に何らかの確率分布を仮定する
- 「現実には起こりえない」とみなす統計的有意水準αを設定する帰無仮説と対立仮説を設定する
  - 帰無仮説H_0：棄却したい仮説、対立仮説H_1：帰無仮説と対立する仮説
- 仮定した確率分布に基づき統計量を推計する
- 帰無仮説H_0が正しい場合に標本が得られる確率p値を計算する
- p値が有意水準αより小さければ帰無仮説を棄却する

t分布

- 「帰無仮説H_0が正しい場合に標本が得られる確率p値」はどのようにして得られるか？
- 母集団から抽出された標本数nが大きく母分散σ^2が既知のときには正規分布を用いれば良い
- 標本数nが少なく母平均μが既知だが母分散σ^2が未知のときには、不偏分散s^2を用いることで、以下の統計量t値は自由度$\nu = n - 1$のt分布$t(\nu)$に従うと仮定する

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}}$$

- 不偏分散$s^2 = \frac{n-1}{n}\sigma^2$は母分散$\sigma^2$より少し小さい
- 不偏分散s^2は自由度$n-1$のχ^2分布に従う

t分布

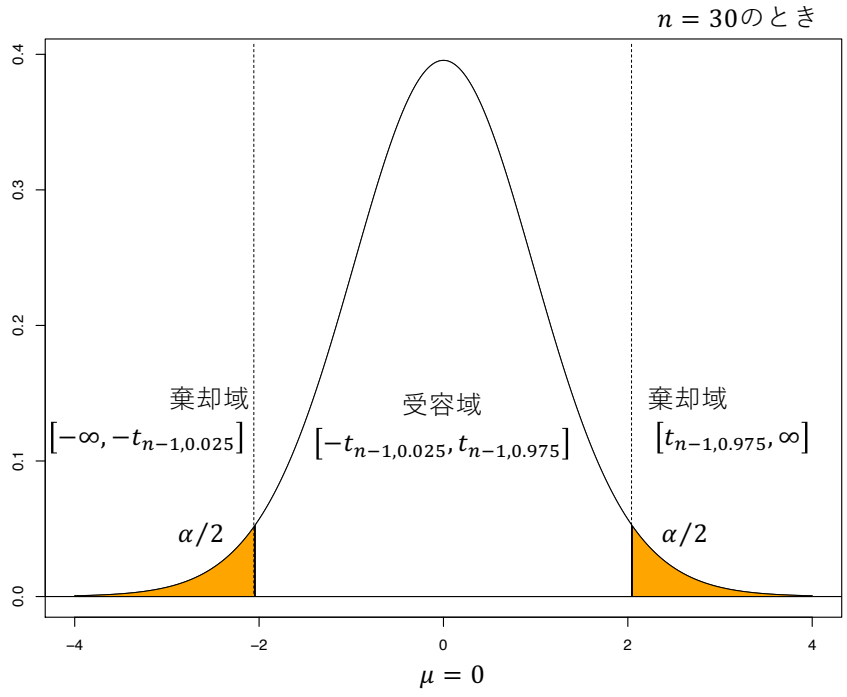
- 確率変数Xが自由度$\nu = n - 1$のt分布$t(\nu)$に従う$X \sim t(\nu)$とき、その確率関数$f(x|\nu)$は以下のように表される

$$f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{s}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- 一般に$\Gamma(z)$はガンマ関数と呼ばれ、 $\Gamma(z) = (z-1)!$
- 期待値$\mathrm{E}[t(\nu)] = 0$
- 分散$\mathrm{V}[t(\nu)] = \frac{\nu}{\nu-2}, \nu > 2$

t分布とp値

- t分布に従う確率変数は、受容域$[-t_{n-1,0.025}, t_{n-1,0.975}]$の中にデータの95%が収まる（$\alpha = 0.05$のとき）
- t値による検定ではαを事前に定めた上で$|t| > t_{n-1,\alpha/2}$かを判断する
- t値の絶対値が十分に大きいときp値は小さくなり、帰無仮説は棄却される
- 帰無仮説が正しいとき、 p値は大きい値をとる
- p値による検定では異常性をp値で確率的に示す



t分布とp値

- 実用的には、以下のように理解しておけば良い
- 標本数が非常に多い場合、t分布は正規分布に近づくので、95%信頼区間（$\alpha = 0.05$のとき）の絶対値が1.96より大きければ、t値が統計的に有意であると判断できる→「2 より大きければ良い」と理解しておく
- t分布は（所詮）ビッグデータやシミュレーションが無かった時代の産物なので、t分布の数式など覚えるには及ばない
- 最近の学術論文では、t値を用いた研究結果は学術論文として採用しないというものもある
- p値は様々な場面で出てくるので、理解しておくと良い

帰無仮説と対立仮説（回帰係数）

- 帰無仮説H_0：不偏推定量$\hat{\beta} = 0$（不偏推定量は存在しない）
- 対立仮説H_1：不偏推定量$\hat{\beta} \neq 0$

- 有意水準$\alpha = 0.05$とする
- 標本数nのデータから回帰係数βを推計する
- 帰無仮説H_0が正しいときにp値を計算する

$$t = \frac{\hat{\beta} - 0}{\sqrt{s^2/n}}$$

- p値が有意水準αより小さければ、「帰無仮説H_0は生じ得ないこと」と判断して帰無仮説H_0を棄却し、対立仮説H_1を受容する

帰無仮説と対立仮説（回帰係数）

- 帰無仮説H_0：不偏推定量$\hat{\beta} = 0$（不偏推定量は存在しない）
- 対立仮説H_1：不偏推定量$\hat{\beta} \neq 0$

- 有意水準$\alpha = 0.05$のとき、単回帰モデルの不偏推定量の95%信頼区間は以下のようになる
- ただし$t_{\alpha/2}(n-k-1)$は変数の数kと定数項から自由度$(n-k-1)$のt値を意味する。

$$\hat{\beta} - t_{\alpha/2}(n-k-1)\sqrt{s^2/n} \leq \beta \leq \hat{\beta} + t_{\alpha/2}(n-k-1)\sqrt{s^2/n}$$

有意水準αの設定と統計的判断

- 有意水準αは、分析者が分析の精度に応じて任意に設定できる
- この授業では$\alpha = 0.05$を多く使うが、それは信頼区間を95%に設定しておけば、そこから外れるような事象はほとんど起きないだろうと「勝手に」想定しているから
- 社会科学分野では便宜上$\alpha = 0.05$を多用するが、 $\alpha = 0.1$や$\alpha = 0.001$などであってもよい。
- 医療・薬学分野などでは、非常に小さい有意水準（$\alpha = 0.001$など）が採用されることがある
- 有意水準を小さくすれば、統計的判断が厳しくなり、例えば冤罪や偽陽性などの誤検出の可能性を少なくできる。他方、完全犯罪や偽陰性（検出失敗）を増やしてしまう。

両側検定と片側検定

- 帰無仮説H_0：$\mu = \mu_0$のとき対立仮説には以下の３つがあり得る

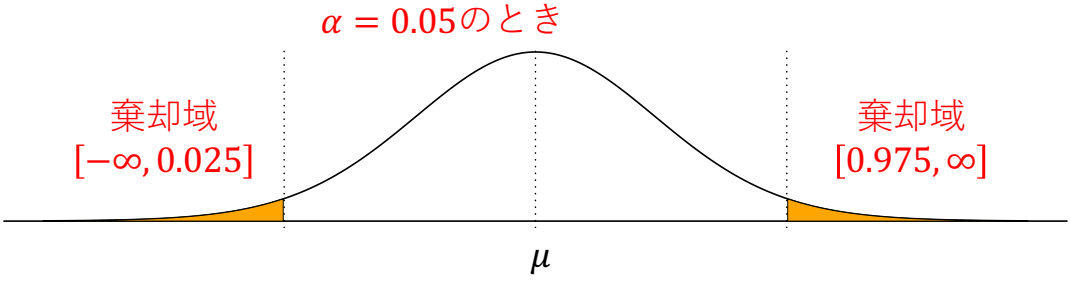
- 対立仮説H_{1a}：$\mu \neq \mu_0$
- 対立仮説H_{1b}：$\mu < \mu_0$
- 対立仮説H_{1c}：$\mu > \mu_0$

- 対立仮説の性質に応じて、両側検定（H_{1a}）か片側検定（H_{1b}、H_{1c}）かが決まる
- 片側検定を行う特段の理由がなければ両側検定を行う

両側検定の棄却域

- 帰無仮説H_0：$\mu = \mu_0$
- 対立仮説H_{1a}：$\mu \neq \mu_0$
- 有意水準α

のとき、棄却域は$[-\infty, \alpha/2]$及び$[1-(\alpha/2), \infty]$となる



残差解析と外れ値の検出

- 残差についての性質を調べることで、回帰分析用いられたデータが外れ値なのかどうかを判断する材料を提供できる。主に以下の手法が用いられる

- 残差プロット
- 残差の正規Q-Qプロット
- S-Lプロット
- 梃子（てこ）比とクックの距離

レバレッジ（梃子比）　　梃子＝てこ

- ハット行列Hの第i対角要素をレバレッジ（梃子比） h_{ii}という
- 梃子比は以下の性質を持つ

$$\frac{1}{n} < h_{ii} < 1$$

- 従属変数の個数をk個とすると、以下の性質が成り立つ

$$\sum_{i=1}^{n} h_{ii} = k + 1$$

レバレッジとスチューデント化残差

- レバレッジ（梃子比）を用いて、残差の分散とスチューデント化残差を以下のように得ることができる
- 残差e_iの分散

$$Var(e_i) = \sigma^2(1 - h_{ii})$$

- スチューデント化残差

$$\frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

クックの距離

- １つのデータを除去して推計される予測値と、全データを用いて推計される予測値との差の平方和を、誤差分散の推定値で割ったもの

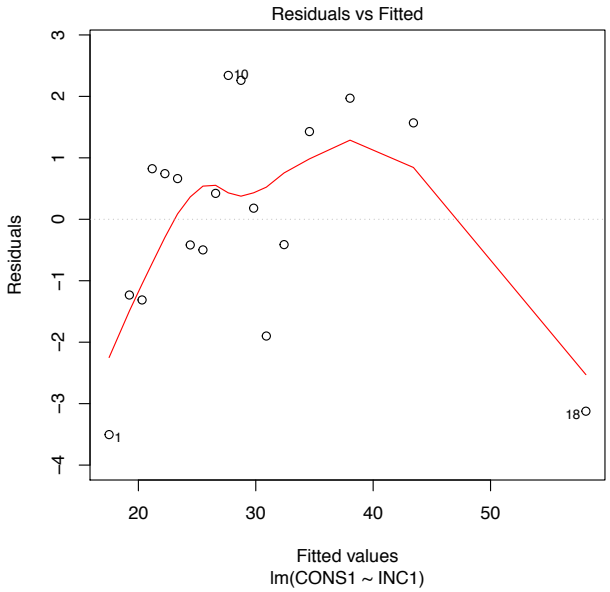
$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p \cdot s_e^2}$$

- ここで、$\hat{y}_j$は全データを用いて得られる予測値、$\hat{y}_{j(i)}$はデータiを除去して得られる予測値、pは説明変数の数（単回帰分析のとき$p = 2$）、s_e^2は残差分散の推定値である。

残差プロット

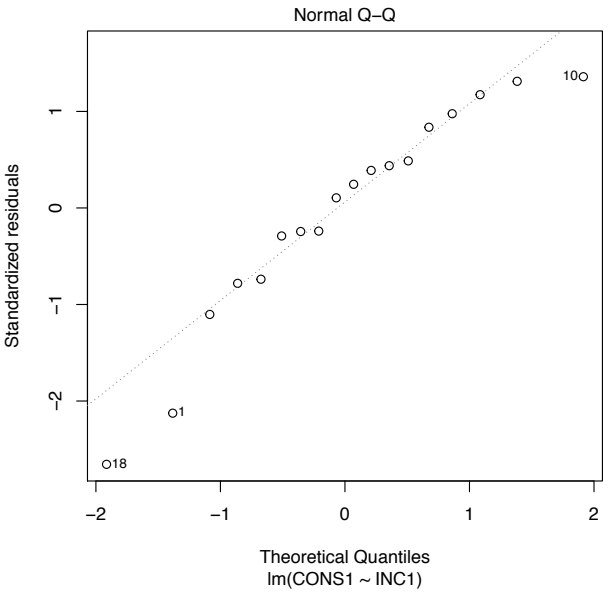
- 予測値$\hat{y}_i$を横軸、残差e_iを縦軸に描いた散布図
- 縦軸の絶対値が$2\hat{\sigma}$を超える残差が多く見られるようなら、データを採用するという仮定を疑うべき→そのデータは外れ値である可能性がある

- 右の結果の場合、$i = 1, 10, 18$が外れ値の可能性がある



残差の正規QQプロット

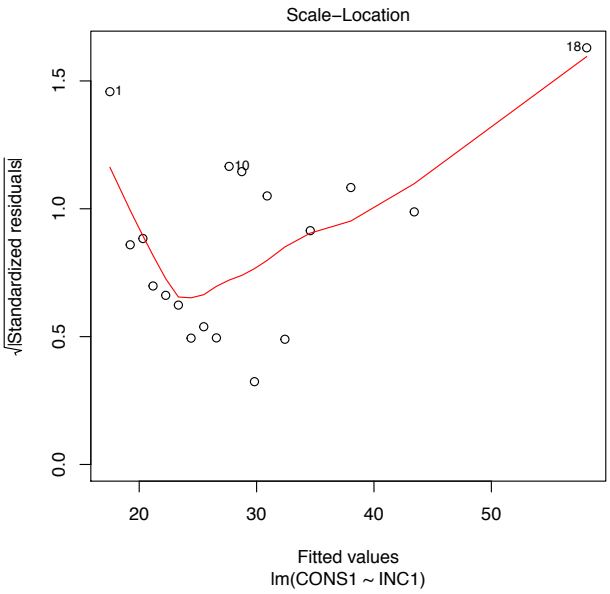
- スチューデント化残差に対する正規Q-Qプロット
- Q-Qプロットではデータ正規分布に従うとき45°線上にデータが乗ってくる性質がある。残差が正規分布に近ければ45°線上にプロットされる
- 縦軸で絶対値が 2 を超える残差が多い場合は、外れ値の可能性がある
- 右の結果の場合、$i = 1, 18$が外れ値の可能性がある



S-Lプロット

- スチューデント化残差の絶対値の平方根を予測値に対して描いた散布図
- 縦軸が$\sqrt{2}$を超えるようなら、データが外れ値であることへの注意が必要

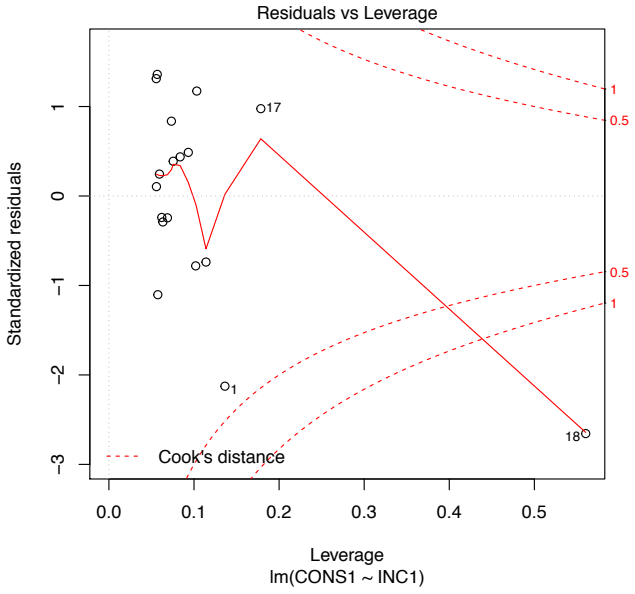
- 右の結果の場合、 $i = 1, 18$が外れ値の可能性がある



梃子比とクックの距離

- 梃子比h_{ii}が$2(k+1)/n$より小さければ注意が必要
- 実践的にはクックの距離D_iが $0.2 < D_i \leq 0.5$なら「要注意」、 $0.5 < D_i$なら当該データを「解析から除去」するのが望ましい

- 右の結果の場合、 $i = 18$を除去したほうがよく、 $i = 1, 17$は要注意データだと判断できる



分析テーマの例

- 経済理論：消費は所得に依存する
- 「所得が増えれば消費が増える」という単回帰モデルを構築
- 独立変数（説明変数）：所得x_i
- 従属変数（被説明変数）：消費y_i
- 分析単位：年収階級i
- モデル式

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- 期待される分析結果
  - $\beta_1 > 0$かつβ_1が5%水準で統計的に有意（t値が1.96以上）
  - 自由度修正済みR^2が1に近い

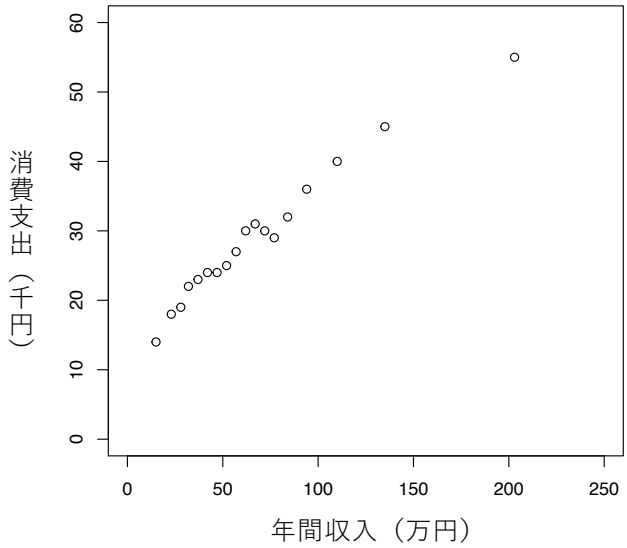
分析に用いるデータ

- 『家計調査』2017年9月、第２－６表「年間収入階級別１世帯当たり１か月間の収入と支出」
- データは政府統計ポータルサイトe-statから入手可能
- 「二人以上の世帯」
- 所得x_i ＝「年間収入(10万円)」、支出y_i ＝「消費支出(千円)」を用いる
- 「年間年収階級」の18階級を分析単位とする

散布図を描いてみる

- 散布図を見る限り、相関関係はありそう
- 年間収入が増加するに従い、消費支出も増加するように見える

- 理論に基づかなくても因果関係がある場合があるので、データを見ることが大事



単回帰分析の分析結果

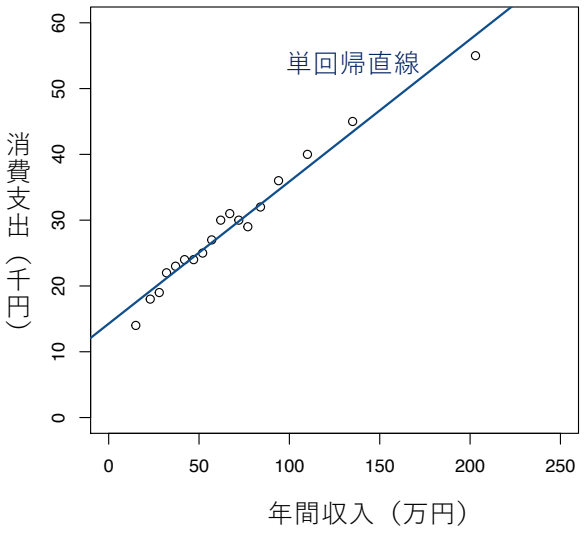
- 単回帰分析の結果は以下のようになった ()内はt値

$$\hat{y} = 14.3 + 0.216x$$

(18.56)　(23.02)

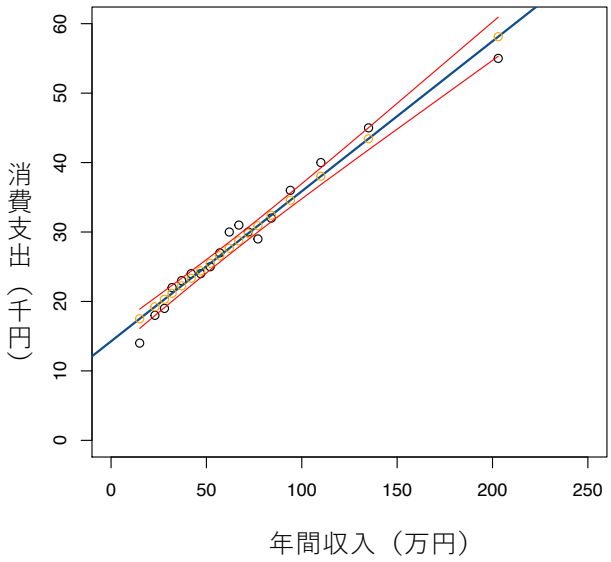
自由度修正済み$R^2 = 0.97$

- $\hat{\beta}_1 = 0.22 > 0$かつ$\hat{\beta}_1$のt値=23.02と5%水準で統計的に有意
- 自由度修正済み$R^2 = 0.97$と1に近い



予測値と予測誤差

- 単回帰式（$\hat{y} = 14.3 + 0.216x$）に年間収入を代入すると消費支出の予測値$\hat{y}$が得られる
- 消費支出の予測値$\hat{y}$と実績値yとの差が予測誤差ε
- 予測値$\hat{y}$をプロットすると左図のようになる（オレンジ色）
- 95%信頼区間は赤色の曲線で挟まれた区間



回帰係数の信頼区間

- 前述の推定結果では、年間収入に対する偏回帰係数の標準誤差$SE = 0.0094$が得られたことから、信頼区間は

$$[0.216 - 1.96\times0.0094, 0.216 + 1.96\times0.0094]$$

すなわち

$$[0.198, 0.234]$$

となる

単回帰分析の結果のまとめ方（例）

- 単回帰分析の結果は、以下のように整理すると分かりやすい
- 回帰係数、t値、自由度修正済みR^2、標本数の記述は必須
- 標準誤差と95%信頼区間を記述するとより親切

説明変数	回帰係数	t値	標準誤差	95%信頼区間
切片	14.26	18.56	0.769	[12.75, 15.77]
年間収入(万円)	0.216	23.02	0.00938	[0.198, 0.234]
自由度修正済みR^2 0.97 標本数18				