基礎統計

第12回講義資料

本日の講義内容

- 第9章：回帰分析

第9章：回帰分析

回帰分析

（目的）二変量の定量的な構造（モデル）を求める

- モデル：YをXで定量的に説明するもの
  回帰方程式，回帰関数　と呼ばれる
  - X：従属変数，被説明変数，内生変数
  - Y：独立変数，説明変数，外生変数
- YをX で説明しようとしている．
  二変量間の関係があるかどうかだけを調べる相関分析とは異なる．

回帰方程式

- 説明変数 X と 被説明変数 Y を考える。
- Y を X によって系統的に変化する部分 y と それ以外のばらつきの部分 u に分けて考える

$$Y = y + u$$

- 回帰方程式（回帰関数） $y = \beta_1 + \beta_2 x$ （線形関数の場合）
- 母回帰方程式 $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i, \quad i = 1,2,\cdots,n$

(X_i, Y_i) はi番目の観測値, u_i は誤差項（攪乱項）

母（偏）回帰係数 β_1, β_2 をデータから求めたい
（未知パラメータ）

回帰モデルの標準的な仮定

- 仮定１　X_i は確率変数でなく，すでに確定した値をとる.
- 仮定２　u_i は確率変数で期待値が0.　すなわち

$$E[\varepsilon_i] = 0, \quad i = 1,2,\cdots,n.$$

- 仮定３　異なった誤差項は無相関.

$$Cov[\varepsilon_i, \varepsilon_j] = E[\varepsilon_i \varepsilon_j] = 0$$

- 仮定４　誤差項の分散が一定で　σ^2 すなわち

$$V[\varepsilon_i] = E[\varepsilon_i^2] = \sigma^2, \quad i = 1,2,\cdots,n.$$

回帰係数の推定（最小二乗法）

- X で説明できない部分 ε_iを最小にするような回帰係数を求める

$$\varepsilon_i = Y_i - (\beta_1 + \beta_2 X_i)$$

- 具体的には，残差の二乗和 S を最小にする係数を求めることになる

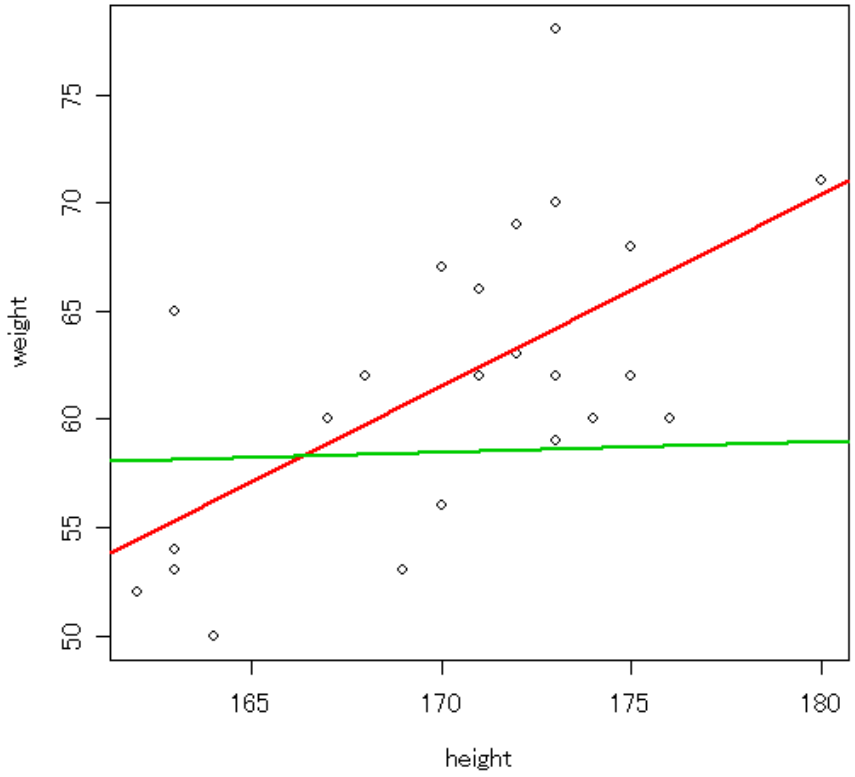
$$S = \sum \varepsilon_i^2 = \sum \{Y_i - (\beta_1 + \beta_2 X_i)\}^2$$

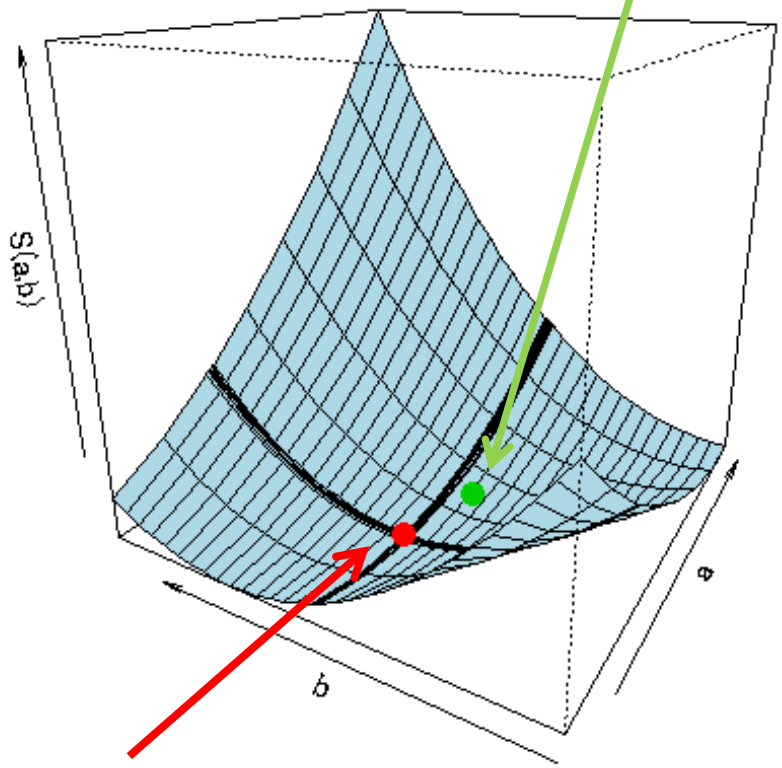
- Sを β_1, β_2 でそれぞれ偏微分して 0とおいた式を解くことにより標本回帰係数が得られる

$$\hat{\beta}_2 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \qquad \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

$$S(a,b)=\sum\{Y_i-(a+bX_i)\}^2$$

当てはまりの
よくない回帰
直線ほどSが
大きくなる





もっとも当てはまりの良
い回帰直線に対するS
（もっとも小さい）

回帰残差

- 回帰残差

$$\begin{aligned} \widehat{e_i} &= Y_i - \widehat{Y_i} \\ &= Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_i \quad (i = 1, 2, \cdots, n) \end{aligned}$$

- 推定値の標準誤差

$$s$$

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{e_i^2}$$

誤差項の分散
の推定値

回帰方程式の当てはまりと決定係数

- 決定係数（当てはまりの良さをはかる基準）

$$\eta^2 \equiv 1 - \frac{\sum \hat{e}_i^2}{\sum (Y_i - \overline{Y})^2} = \frac{\sum (\hat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2}$$

- 標本回帰係数（最良線型不偏推定量）
  - 不偏推定量
  - 分散

$$E(\hat{\beta}_1) = \beta_1, \quad E(\hat{\beta}_2) = \beta_2$$

$$V(\hat{\beta}_1) = \frac{\sigma^2 \sum X_i^2}{n \sum (X_i - \overline{X})^2}, \quad V(\hat{\beta}_2) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$$

最良線型不偏推定量

線型で不偏な推定のうち分散が最小となる推定量

偏回帰係数の統計的推測

- 偏回帰係数の標本分布

$$\widehat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(X_i - \overline{X})^2}\right)$$

- 標準偏差の推定量

$$s.e.(\widehat{\beta}_2) = \frac{s.e.}{\sqrt{\sum(X_i - \overline{X})^2}}$$

$$s.e. = \sqrt{s^2} = \sqrt{\sum e_i^2/(n-2)}$$

- 検定統計量

$$t_2 = \frac{\widehat{\beta}_2 - \beta_2}{s.e.(\widehat{\beta}_2)} \sim t(n-2)$$

自由度 n-2 のt分布

重回帰モデル

$$Y_i = \underbrace{\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}}_{\text{回帰方程式}} + \varepsilon_i, \quad i = 1,2,\cdots,n$$

- 説明変数と誤差項について
  - 仮定１－４は単回帰分析で示したものと同じ

    説明変数は確定値，誤差項の期待値は0，誤差項は互いに無相関，誤差項の分散は等しい(分散均一性)

  - 仮定５：説明変数間に多重共線性がない

$$\alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \cdots + \alpha_k X_{ki} = 0, \quad i = 1,2,\cdots,n$$

を満たす$\alpha_1, \alpha_2, \cdots, \alpha_k$ は $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ のときのみ

重回帰方程式の推定

- 最小二乗法：　単回帰分析のときと同じ

$$S = \sum \varepsilon_i^2 = \sum \{Y_i - (\beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})\}^2$$

Sを $\beta_1, \cdots, \beta_k$ で偏微分して0とおいて求める

$$\frac{\partial S}{\partial \beta_1} = 0, \frac{\partial S}{\partial \beta_2} = 0, \cdots, \frac{\partial S}{\partial \beta_k} = 0$$

標本回帰方程式　$Y = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$

回帰値　$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$

残差　$\hat{e}_i = Y_i - \hat{Y}_i$

誤差項の分散　$s^2 = \dfrac{\sum e_i^2}{n-k}$

自由度調整済み決定係数　$\overline{\eta}^2 = 1 - \dfrac{\sum e_i^2 / (n-k)}{\sum (Y_i - \overline{Y})^2 / (n-1)}$

第10章:分散分析

分散分析

- 第10章：分散分析
  - 一元配置分散分析
    - 一元配置モデル
    - 分散分析
  - 二元配置分散分析
    - 二元配置モデル
    - ダミー変数による回帰モデルによる表示
    - 分散分析
    - 交互作用がない場合の主効果の検定
    - 主効果の検定

実験データの分析

工学,医学,農学など、様々な分野において行われる実験は

- いくつかの対象や条件の比較を目的としており、
- 同一条件下で実験を繰り返しても結果が同じにならない

という意味で、データに実験誤差が含まれる、という特質をもつ.

実験データの分析：母集団平均の比較

- 2標本問題（2つの母集団平均の比較）
  - 母平均の差の検定
    - 母分散が既知
    - 母分散が未知（等しい）
    - 母分散が未知（等しくない）
- K標本問題（3つ以上の母集団平均の比較）
  - **分散分析**
    - 一元配置分散分析
    - 二元配置分散分析
    - 多元配置分散分析

平均値の差の検定

- 2つのグループで結果に差があるかどうかを検定する
- ＜対照実験＞
  20匹のラットを10匹ずつ2群に分け、一方にはふつうの食餌を与え、他方には血中の赤血球数を減らすと考えられている薬を混入した食餌を与えた場合の、血液1mm^{3}中の赤血球数を調べた結果が次の表である。

表1　ラット20匹の血液1立方ミリメートル中の赤血球の数（単位：100万個）										
投薬群	7.97	7.66	7.59	8.44	8.05	8.08	8.35	7.77	7.98	8.15
対照群	8.06	8.27	8.45	8.05	8.51	8.14	8.09	8.15	8.16	8.42

- 薬効があると認められるか？

一元配置分散分析

- 因子：実験結果に影響を与えると考えられる要因　A
- 水準：因子に対して与える条件　$A_1, A_2, \cdots, A_a$

- 因子の数が1つの場合を一元配置、複数の場合を多元配置と呼ぶ

- 観測値 Y_{ij} ： それぞれの水準では平均だけが異なり、分散は一定であるとする. $N(\mu_i, \sigma^2)$ に従うと仮定する。
- モデル

$$Y_{ij} = \mu_i + \varepsilon_{ij}, i = 1,2,\ldots,a,\ j = 1,\ldots,n_i$$

n_i ：水準 A_i の繰り返しの数

一元配置モデル

- 観測値の総数 $n = \sum_{i=1}^{a} n_i$
- 一般平均 $\mu = \dfrac{\sum_{i=1}^{a} n_i \mu_i}{n}$ 水準 A_i の効果 $\delta_i = \mu_i - \mu, \quad \sum n_i \delta_i = 0$
- モデル

$$Y_{ij} = \mu + \delta_i + \varepsilon_{ij}, \; i = 1,2,\ldots,a, \; j = 1,\ldots,n_i$$

（共通の効果）＋（第i水準の効果）＋（それ以外の誤差）

分散分析

- 帰無仮説:
  すべての水準で平均が等しく水準による効果が0である.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a = \mu \quad \text{または} \quad H_0 : \delta_1 = \delta_2 = \cdots = \delta_a = 0$$

- $$F = \frac{S_a/(a-1)}{S_e/(n-a)}$$ 自由度$(a-1, n-a)$のF分布$F(a-1, n-a)$に従う

S_a/σ^2は自由度$\nu_a = a-1$のχ^2分布に従う

S_e/σ^2はS_aとは独立に自由度 $\nu_e = n-a$のχ^2分布に従う

比はF分布に従う

平方和の分解

- 級間平方和

$$S_a = S_t - S_e = \sum_{i=1}^{a} n_i (\overline{Y}_i - \overline{Y})^2,$$

- 全平方和と級内平方和

$$S_t = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2, \quad S_e = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$$

$$\overline{Y} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} Y_{ij} \Big/ n, \quad \overline{Y}_i = \sum_{j=1}^{n_i} Y_{ij} \Big/ n_i$$

分散分析表

要因	平方和	自由度	分散比	F値
級間	S_a	$\nu_a = a-1$	$V_a = S_a/\nu_a$	$F = V_a/V_e$
級内	S_e	$\nu_e = n-a$	$V_e = S_e/\nu_e$	
計	S_t	$\nu_t = n-1$		

事例:一元配置分散分析

- 次のデータを一元配置分散分析せよ
  反応温度が有効成分の生成量に影響を与えているかどうかを検定してみる。

表1　反応温度別の原料100g当たりの有効成分の生成量(g)

反応温度	40度	50度	60度
データ	20.0	24.6	24.9
	21.8	26.8	21.1
	21.1	27.2	25.3
	22.4	24.8	22.0
	21.1	25.4	
		27.6	

実行結果

- 分散分析表

分散分析表

変動要因	変動	自由度	分散	観測された分散比	P-値	F 境界値
グループ間	63.427	2	31.714	15.414	0.000484	3.885
グループ内	24.689	12	2.057			
合計	88.116	14				

実習2：製品中のある成分の含有量データ

反応温度	50度	55度	60度	65度
データ	77.4 78.2 78.1 77.8 77.9	78.3 78.2 78.4 77.3 79.1	79.2 79.3 79.1 78.2 79.3	78.9 78.8 78.1 78.1 78.9

繰り返しのある二元配置分散分析(1)

- 因子が2つ以上の場合,多元配置分散分析をおこなう
- 二元配置モデル

因子　A, B

水準　$A_1, A_2, \cdots, A_a, \quad B_1, B_2, \ldots, B_b$

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$$

互いに独立で正規分布に従うとする。

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n_{ij}$$

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

A_iの主効果　　B_jの主効果

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0$$

$$\sum_{j=1}^{b} (\alpha\beta)_{ij} = 0, i = 1,2,\cdots,a$$

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, j = 1,2,\cdots,b$$

ダミー変数を用いた回帰モデルによる表示

- ダミー変数を含む回帰モデルによって表現できる

$$D_{1l,ijk} = \begin{cases} 1 : l = i \\ 0 : l \neq i \end{cases} \qquad D_{2m,ijk} = \begin{cases} 1 : m = j \\ 0 : m \neq j \end{cases}$$

- 二元配置モデル（帰無仮説）

$$Y_{ijk} = \mu_{11} + \sum_{\ell=2}^{a} \gamma_i D_{1\ell,ijk} + \sum_{m=2}^{b} \delta_j D_{2m,ijk} + \varepsilon_{ijk}$$

対立仮説のもとでは

$$Y_{ijk} = \sum_{l=1}^{a} \sum_{m=1}^{b} \mu_{ij} Z_{lm,ijk} + \varepsilon_{ijk}$$

$$Z_{lm,ijk} = D_{1l,ijk} \cdot D_{2m,ijk}$$

二元配置分散分析

- 交互作用の有無　→　F検定によって確かめる
- 帰無仮説：交互作用は含まない

$$H_0 : \mu_{ij} = \mu_{11} + \gamma_i + \delta_j, i = 2,3,\mathrm{K}\ ,a,\ j = 2,3,\mathrm{K}\ ,b$$

- 検定統計量

$$F = \frac{V_{A\times B}}{V_e}, \quad V_e = \frac{S_e}{n-ab}, \quad V_{A\times B} = \frac{S_{A\times B}}{(a-1)(b-1)}$$

  - 交互作用を含むモデルの回帰残差の平方和 S_e
  - 交互作用を含まないモデルの回帰残差の平方和 S_0^*
  - 2つの平方和の差

$$S_{A\times B} = S_0^* - S_e$$

交互作用がない場合の主効果の検定

因子Bの主効果の検定

- 帰無仮説：因子Bの主効果がない

- 検定統計量

$$Y_{ijk} = \sum_{l=1}^{a} \gamma_l D_{1l,ijk} + \varepsilon_{ijk}$$

$$F_1 = \frac{V_B}{V_e^*}, \quad V_B = \frac{S_B}{b-1}, \quad V_e^* = \frac{S_0^*}{n-a-b+1}$$

F_1 自由度（b-1, n-a-b+1）のF分布に従う

$$S_B = S_0^{**} - S_0^*$$

$$S_0^{**} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{i\cdot})^2, \bar{Y}_{i\cdot} = \sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} Y_{ijk} \Big/ n_{i\cdot}, n_{i\cdot} = \sum_{j=1}^{b} n_{ij}$$

実際には…

$$F_2 = \frac{V_B}{V_e}$$

自由度（b-1, n-ab）のF分布に従う

交互作用がない場合の主効果の検定

因子Aの主効果の検定

- 帰無仮説：因子Aの主効果がない
- 検定統計量

モデル $Y_{ijk} = \sum_{m=1}^{b} \delta_m D_{2m,ijk} + \varepsilon_{ijk}$

$$F_1 = \frac{V_A}{V_e^*}, \quad V_A = \frac{S_A}{a-1}, \quad V_e^* = \frac{S_0^*}{n-a-b+1}$$

実際には…

$$F_2 = \frac{V_A}{V_e}$$

自由度（a-1, n-ab）のF分布に従う

F_1 自由度（a-1, n-a-b+1）のF分布に従う

$$S_A = S_0^{***} - S_0^*$$

$$S_0^{***} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{\cdot j})^2, \bar{Y}_{\cdot j} = \sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} Y_{ijk} \Big/ n_{\cdot j}, n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$$

繰り返しのある二元配置分散分析(2)

- 原料100g当たりの有効成分の生産量(g)のデータを用いて，二元配置分散分析をおこなう。

反応時間 / 反応温度	10分	30分	40分
50度	23.7 22.2 21.7	25.7 23.3 22.5	24.1 23.6 24.6 23.9
60度	24.0 23.6	24.5 24.1 25.4 24.9	25.8 24.4 25.6

(3)考察

- 観測値に関する統計情報
- 交互作用項を含むモデルの結果
- 交互作用項を含まないモデル
  - 両方の因子の主効果を含むモデル
  - 因子Aのみの主効果のモデル
  - 因子Bのみの主効果のモデル
- 要因別変動表と検定統計量
  - 交互作用の検定
  - 交互作用がない場合の因子の主効果の検定