



データ解析実習

2012.11.12 月曜、4・5限

担当教員：高木英至

1

■ 本日の授業

- ▶ レポートの受け取り
- ▶ 回帰分析
 - ▶ 単回帰分析
 - ▶ 重回帰分析
- ▶ 時間があれば実習
 - ▶ ないかも



▶ 2

■ 回帰分析

- ▶ 単に相関をとることとは異なる
- ▶ 「**独立変数**」から「**従属変数**」を「説明」ないし「予測」する

$$x_1, x_2, \dots, x_p \rightarrow y$$

独立変数 説明変数	→	従属変数 目的変数
--------------	---	--------------

- ▶ 回帰分析
 - ▶ 単回帰分析：独立変数が1個
 - ▶ 重回帰分析：独立変数が2個以上

▶ 3

回帰モデル

$$x_1, x_2, \dots, x_p \rightarrow y$$

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon$$

- ▶ 線形回帰モデル

$$y = a + b_1 x_1 + b_2 x_2 + \dots + b_p x_p + \varepsilon$$

- ▶ 線形単回帰モデル

$$y = a + bx + \varepsilon$$

▶ 4

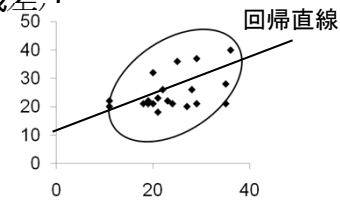
線形単回帰モデル

▶ n : 観測数 $i=1, \dots, n$ とすると、

$$y_i = a + bx_i + \varepsilon_i$$

a : 定数項 b : 回帰係数

ε_i : 誤差 (残差)



▶ 5

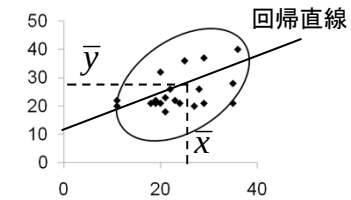
「最小二乗法」で係数の推定値を求めると

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

$$\hat{y}_i = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} + \frac{S_{xy}}{S_{xx}} x_i$$

$$\hat{y}_i - \bar{y} = \frac{S_{xy}}{S_{xx}} (x_i - \bar{x})$$

つまり、回帰直線は平均 $(\bar{x}, \bar{y})$ を通る。



▶ 6

決定係数

$$r_{xy} = r_{\hat{y}y} = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{y}) / (n-1)}{\sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1} \cdot \frac{\sum (\hat{y}_i - \bar{y})^2}{n-1}}}$$

$$= \sqrt{\frac{\sum (\hat{y}_i - \bar{y})^2 / (n-1)}{\sum (y_i - \bar{y})^2 / (n-1)}}$$

$$\text{決定係数 } r_{xy}^2 = \frac{\sum (\hat{y}_i - \bar{y})^2 / (n-1)}{\sum (y_i - \bar{y})^2 / (n-1)}$$

y の分散のうち、 x によって説明される割合

▶ 7

分散の分割

従属変数 y の分散は、予測値の分散と残差の分散の和になる

$$s_y^2 = s_{\hat{y}}^2 + s_{\varepsilon}^2 \quad (3.22), p.61$$

決定係数を書き換えると、

$$\text{決定係数 } r_{xy}^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

s_y^2 は、 r^2 と $(1-r^2)$ の比率で分割される。

$$s_{\hat{y}}^2 = s_y^2 r^2 \quad (3.23)$$

$$s_{\varepsilon}^2 = s_y^2 (1-r^2) \quad (3.24)$$

▶ 8

■重回帰分析：基本的な考え方

▶ 単回帰モデル

$$y_i = a + bx_i + \varepsilon_i$$

▶ 重回帰モデル

$$y_i = a + b_1x_{1i} + b_2x_{2i} + \cdots + b_px_{pi} + \varepsilon_i$$

- ▶ 基本的には単回帰と同じ
- ▶ a ：定数項
- ▶ b_j ：偏回帰係数
 - ▶ x_j 以外の変数の影響を統計的に除去している

▶ 9

重回帰モデル

$$y_i = \hat{a} + \hat{b}_1x_{1i} + \hat{b}_2x_{2i} + \cdots + \hat{b}_px_{pi} + \varepsilon_i$$

$$\hat{y}_i = \hat{a} + \hat{b}_1x_{1i} + \hat{b}_2x_{2i} + \cdots + \hat{b}_px_{pi}$$

$$\varepsilon_i = y_i - \hat{y}_i$$

▶ 10

最小二乗法（１）

$$y_i = \hat{a} + \hat{b}_1x_{1i} + \hat{b}_2x_{2i} + \cdots + \hat{b}_px_{pi} + \varepsilon_i$$

$$\hat{y}_i = \hat{a} + \hat{b}_1x_{1i} + \hat{b}_2x_{2i} + \cdots + \hat{b}_px_{pi}$$

$$\varepsilon_i = y_i - \hat{y}_i$$

$$\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_p \text{の求め方} = \text{最小二乗法}$$

下記の Q が最小になるように $\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_p$ を求める。

$$Q(\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_p) \equiv \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^n \{y_i - (\hat{a} + \hat{b}_1x_{1i} + \hat{b}_2x_{2i} + \cdots + \hat{b}_px_{pi})\}^2$$

Q が最小となるには、次を充たすことが必要。

$$\frac{\partial Q}{\partial \hat{a}} = 0, \frac{\partial Q}{\partial \hat{b}_1} = 0, \frac{\partial Q}{\partial \hat{b}_2} = 0, \dots, \frac{\partial Q}{\partial \hat{b}_p} = 0.$$

▶ 11

最小二乗法（２）

$$\frac{\partial Q}{\partial \hat{a}} = 0, \frac{\partial Q}{\partial \hat{b}_1} = 0, \frac{\partial Q}{\partial \hat{b}_2} = 0, \dots, \frac{\partial Q}{\partial \hat{b}_p} = 0.$$

上の条件を整理すると：

$$\hat{a} = \bar{y} - (\hat{b}_1\bar{x}_1 + \hat{b}_2\bar{x}_2 + \cdots + \hat{b}_p\bar{x}_p) \quad (1)$$

$$S_{ji} = \sum_{i=1}^n (x_{ji} - \bar{x}_j)(x_{ji} - \bar{x}_i), S_{yj} = \sum_{i=1}^n (y_i - \bar{y})(x_{ji} - \bar{x}_j) \quad \text{とおくと、}$$

$$S_{11}\hat{b}_1 + S_{12}\hat{b}_2 + \cdots + S_{1p}\hat{b}_p = S_{y1}$$

$$S_{21}\hat{b}_1 + S_{22}\hat{b}_2 + \cdots + S_{2p}\hat{b}_p = S_{y2} \quad (2)$$

.....

$$S_{p1}\hat{b}_1 + S_{p2}\hat{b}_2 + \cdots + S_{pp}\hat{b}_p = S_{yp}$$

(2)の p 元連立１次方程式を解いて $\hat{b}_1, \hat{b}_2, \dots, \hat{b}_p$ を求める。

さらに(1)から $\hat{a}$ を求める。

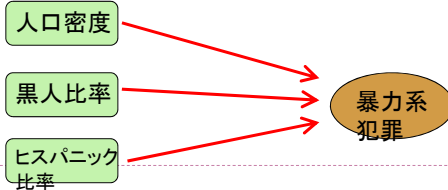
▶ 12

偏回帰係数の意味

$$\hat{y}_i = \hat{a} + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \cdots + \hat{b}_p x_{pi} \quad b_j \text{ は偏回帰係数}$$

$$\hat{y} = \hat{a} + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_p$$

1. x_1 が 1 単位増加 (減少) すれば、平均して、 $\hat{y}$ は $\hat{b}_1$ だけ増加 (減少) する
2. $\hat{b}_1$ は、 $x_2, \dots, x_p$ の影響を除いた、 $\hat{y}$ に対する x_1 の効果を表す



▶ 13

変数を標準化した場合

▶ 標準化 $Y_i = (y_i - \bar{y}) / s_y, X_{ji} = (x_{ji} - \bar{x}_j) / s_{x_j}$

$$y_i = \hat{a} + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \cdots + \hat{b}_p x_{pi} + \varepsilon_i \quad b_j \text{ は偏回帰係数}$$

$$Y_i = \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_p X_{pi} + \varepsilon_i \quad \beta_j \text{ は標準偏回帰係数}$$

定数項は消える。

$$\hat{Y}_i = (\hat{y}_i - \bar{y}) / s_y$$

$$= \frac{(\hat{a} + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \cdots + \hat{b}_p x_{pi}) - (\hat{a} + \hat{b}_1 \bar{x}_1 + \hat{b}_2 \bar{x}_2 + \cdots + \hat{b}_p \bar{x}_p)}{s_y}$$

$$= \frac{\sum_{j=1}^p \hat{b}_j (x_{ji} - \bar{x}_j)}{s_y} = \sum_{j=1}^p \hat{b}_j \cdot \frac{(x_{ji} - \bar{x}_j)}{s_y} = \sum_{j=1}^p \hat{b}_j \cdot \frac{s_{x_j}}{s_y} \cdot \frac{(x_{ji} - \bar{x}_j)}{s_{x_j}}$$

$$= \sum_{j=1}^p \hat{b}_j \cdot \frac{s_{x_j}}{s_y} \cdot X_{ji} \quad \text{つまり、} \hat{\beta}_j = \hat{b}_j \cdot \frac{s_{x_j}}{s_y}$$

▶ 14

標準偏回帰係数の意味

$$\hat{Y} = \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_p X_{pi} \quad \beta_j \text{ は標準偏回帰係数}$$

1. (標準化した値で) X_1 が 1 単位増加 (減少) すれば、平均して、 $\hat{Y}$ は $\hat{\beta}_1$ だけ増加 (減少) する
2. $\hat{\beta}_1$ は、 $X_2, \dots, X_p$ の影響を除いた、 $\hat{Y}$ に対する X_1 の効果を表す
3. $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ は値を相互に比較できる

▶ 15

平方和の分割

- 従属変数の平方和 (分散の分子) は、独立変数によって説明できる平方和と、誤差による平方和とに分割される (テキスト、p.250)。

$$SS_y = SS_{\hat{y}} + SS_{\varepsilon}$$

$SS_{\hat{y}}$: y の平方和

$SS_{\hat{y}}$: 独立変数から説明される平方和 (回帰モデルの平方和)

SS_{ε} : 誤差による平方和

▶ 16

重相関係数と決定係数

■ 重相関係数 R

$$\text{決定係数 } R^2 = r_{\hat{y}y}^2 = \frac{SS_{\hat{y}}}{SS_y} = 1 - \frac{SS_{\varepsilon}}{SS_y},$$

$$1 \geq R^2 \geq 0.$$

R^2 は、従属変数の平方和のうち、回帰モデル
(に導入した独立変数) によって説明される
比率を表す。

▶ 17

■ 重回帰分析における検定：ポイント

- ▶ よく使うのは2種類の検定
- ▶ 重回帰式の検定(pp.250-255)
 - ▶ F検定
- ▶ 偏回帰係数の検定
 - ▶ 独立変数の寄与の検定(F検定、pp.256-260)
 - ▶ 標準誤差を用いた検定(t検定、pp.260-262)
- ▶ 検定結果は、変数を標準化しても同じ
- ▶ 用語： $F(a, b, \alpha)$: 自由度 a, b で有意水準 α でのF値
(F分布表の値)

▶

重回帰式の検定(F検定)

- ・ 重相関係数の検定
- ・ 帰無仮説 $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$.
- ・ 独立変数は従属変数に説明に寄与していない。

分散分析表
(ANOVA)

$F_{p,p-1}(\beta_1, \beta_2, \dots, \beta_p) \rightarrow H_0$ を棄却する

変動要因	平方和	自由度	平均平方	F比
回帰モデル	$SS_{\hat{y}}$	p	$V_{\hat{y}} = \frac{SS_{\hat{y}}}{p}$	$F_0 = \frac{V_{\hat{y}}}{V_{\varepsilon}}$
誤差	SS_{ε}	$n-p-1$	$V_{\varepsilon} = \frac{SS_{\varepsilon}}{(n-p-1)}$	
全体	SS_y	$n-1$	$V_y = \frac{SS_y}{(n-1)}$	

▶ 19

偏回帰係数の検定

- ▶ 帰無仮説 $H_0: \beta_j = 0$.
 - ▶ 独立変数 x_j は従属変数 y の説明に寄与しない。
- ▶ F検定を使う方法
 - ▶ x_j を回帰式に導入することによる従属変数の説明の増加分が有意か否かを検定する
 - ▶ $F_j > F(1, n-p-1, \alpha)$ なら、帰無仮説 H_0 を棄却
- ▶ t検定を使う方法
 - ▶ t_j : β_j の推定値のt分布での値
 - ▶ $t_j > t$ 値(自由度 $n-p-1, \alpha$) $\rightarrow H_0$ を棄却
- ▶ 偏回帰係数の場合、F検定とt検定の結果は一致する

▶ 20

■重回帰分析：例

- ▶ 埼玉県内の市町村への流入者数（人口当たりの）は何によって説明されるか？
- ▶ 分析単位：市町村
- ▶ 従属変数(dependent variable)
 - ▶ 流入者数（人口千人当たりの）
- ▶ 独立変数（independent variables）
 - ▶ 平均所得
 - ▶ 金融業（従事者数／千人）
 - ▶ サービス業（従事者数／千人）
 - ▶ 公務（従事者数／千人）

主な結果：記述統計量

記述統計量

	平均値	標準偏差	N
流入	46.1638	9.3900	92
所得千円	6021.76	1516.39	92
金融業	12.9692	5.6147	92
サービス	122.3055	18.2727	92
公務	18.6792	5.7928	92

相関係数

	流入	所得千円	金融業	サービス	公務
Pearson の相関					
流入	1.000	.300	.276	.332	.131
所得千円	.300	1.000	.166	-.023	-.123
金融業	.276	.166	1.000	.735	-.018
サービス	.332	-.023	.735	1.000	.153
公務	.131	-.123	-.018	.153	1.000
有意確率 (片側)					
流入		.002	.004	.001	.107
所得千円	.002		.057	.418	.122
金融業	.004	.057		.000	.432
サービス	.001	.418	.000		.072
公務	.107	.122	.432	.072	
N					
流入	92	92	92	92	92
所得千円	92	92	92	92	92
金融業	92	92	92	92	92
サービス	92	92	92	92	92
公務	92	92	92	92	92

重相関係数(R)、決定係数(R²)

モデル集計

モデル	R	R2 乗	調整済み R2 乗	推定値の標準誤差
1	.468 ^a	.219	.183	8.4869

a. 予測値: (定数)、公務、金融業、所得千円、サービス。

回帰モデルにより
従属変数の22%が
説明された

回帰式の検定（分散分析、F 検定）

分散分析^b

モデル	平方和	自由度	平均平方	F 値	有意確率
1 回帰	1757.188	4	439.297	6.099	.000 ^a
残差	6266.418	87	72.028		
全体	8023.606	91			

a. 予測値: (定数)、公務、金融業、所得千円、サービス。

b. 従属変数: 流入

偏回帰係数の検定（分散分析）

係数^a

モデル		非標準化係数		標準化係数	t	有意確率
		B	標準誤差	ベータ		
1	(定数)	9.594	8.573		1.119	.266
	所得千円	2.022E-03	.001	.327	3.310	.001
	金融業	-4.413E-02	.246	-.026	-.179	.858
	サービス	.175	.075	.340	2.317	.023
	公務	.191	.159	.118	1.204	.232

a. 従属変数: 流入

$$\hat{y} = 9.594 + 0.002x_1 - 0.044x_2 + 0.175x_3 + 0.191x_4$$

$$\hat{Y} = 0.327X_1 - 0.026X_2 + 0.340X_3 + 0.118X_4$$



偏回帰係数の検定（分散分析）

係数^a

モデル		非標準化係数		標準化係数	t	有意確率
		B	標準誤差	ベータ		
1	(定数)	9.594	8.573		1.119	.266
	所得千円	2.022E-03	.001	.327	3.310	.001
	金融業	-4.413E-02	.246	-.026	-.179	.858
	サービス	.175	.075	.340	2.317	.023
	公務	.191	.159	.118	1.204	.232

a. 従属変数: 流入

従属変数に対する各独立変数の効果(偏回帰係数)はここに出ている

検定の帰無仮説は「その偏回帰係数が母集団ではゼロ」

検定量(t値)はここに出ている。帰無仮説が間違っているほど、t値は高い。

有意なのは、所得千円とサービスだけである



講義は おしまい