Rによる重回帰分析

心理データ解析演習

猪原

07／11／07

発表内容

- **回帰分析とは何か**
- **重回帰分析とは何か**
- **Rとは何か**
- **Rで重回帰分析を実行**

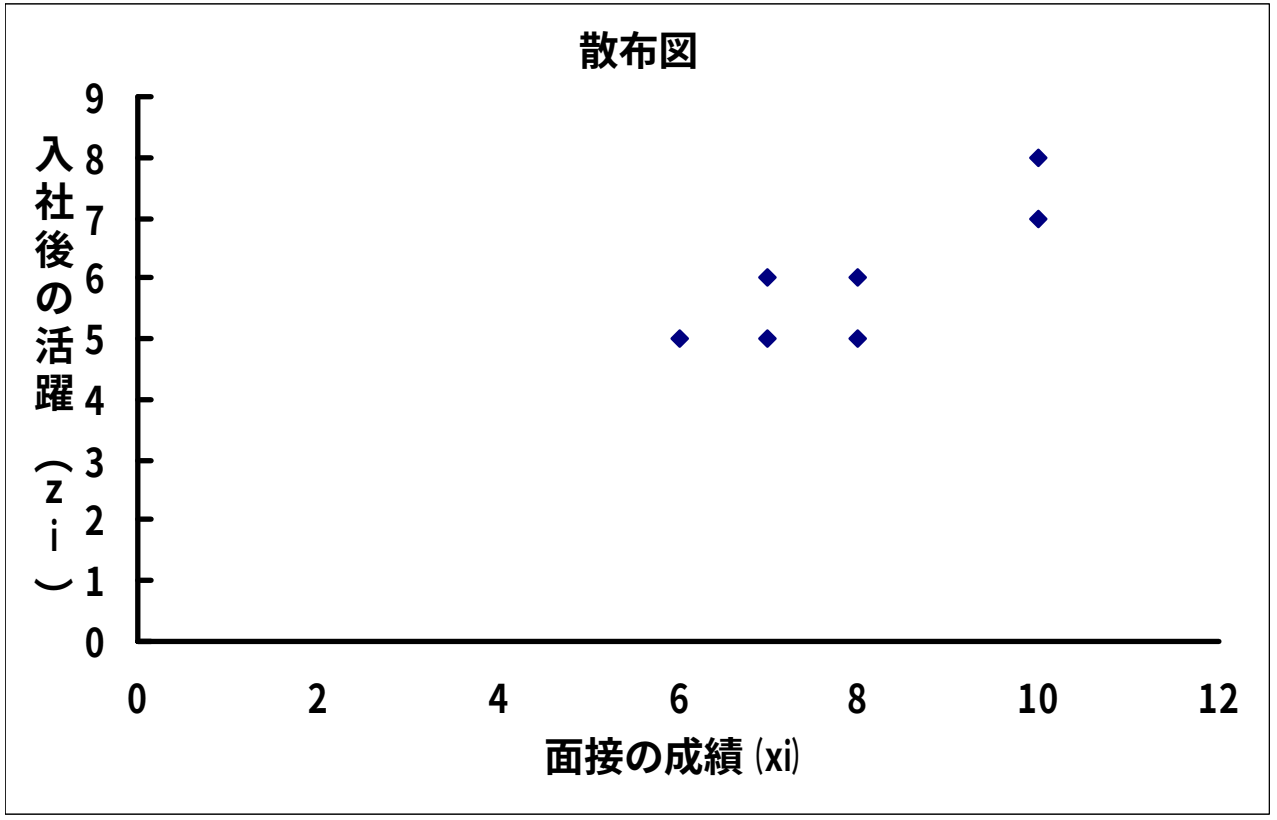
回帰分析とは　1

- 例えば，入社試験として面接を行う会社があったとする。面接の成績と入社後の活躍とはどのような関係があるか知りたい（データ出典：大村, 2006）。
- どうすればよい？

姓	入社後の活躍（zi）	面接の成績（xi）
山中	8	10
田口	7	10
中田	6	8
山口	6	7
中山	5	8
山田	5	7
田中	5	6

回帰分析とは　2

- 以下のような**散布図**を描いてみたり，**相関係数**を算出することで，面接の成績と入社後の活躍の関係がわかる。
- r = .85 （強い正の相関。面接の成績の良い人は入社後も活躍する傾向が強い）

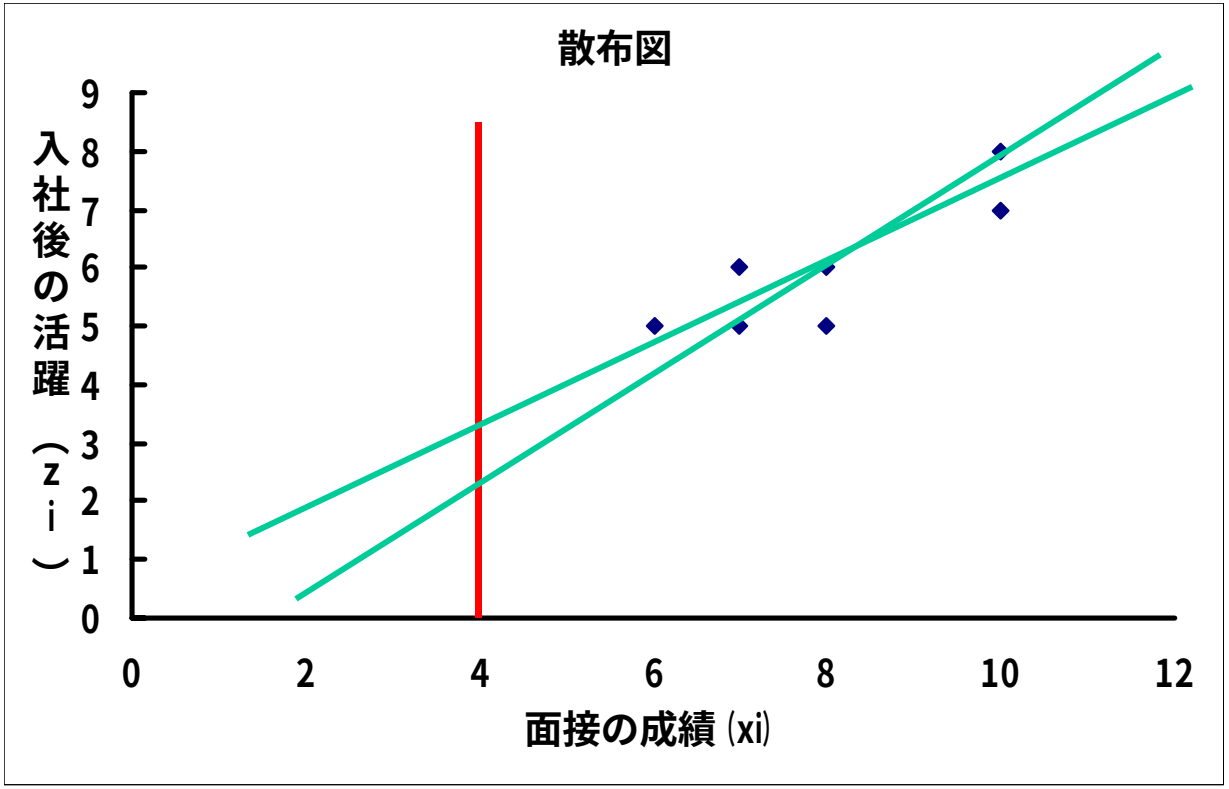


回帰分析とは　3

- 関係が分かったところで，次は面接の成績から入社後の活躍を予測して，採用するかどうか考えたいとする。
- 例えば，面接の成績が4点の人が現われたとする。
- この人には一体どれくらいの活躍が見込めるか，知りたい。
- どうすればよい？

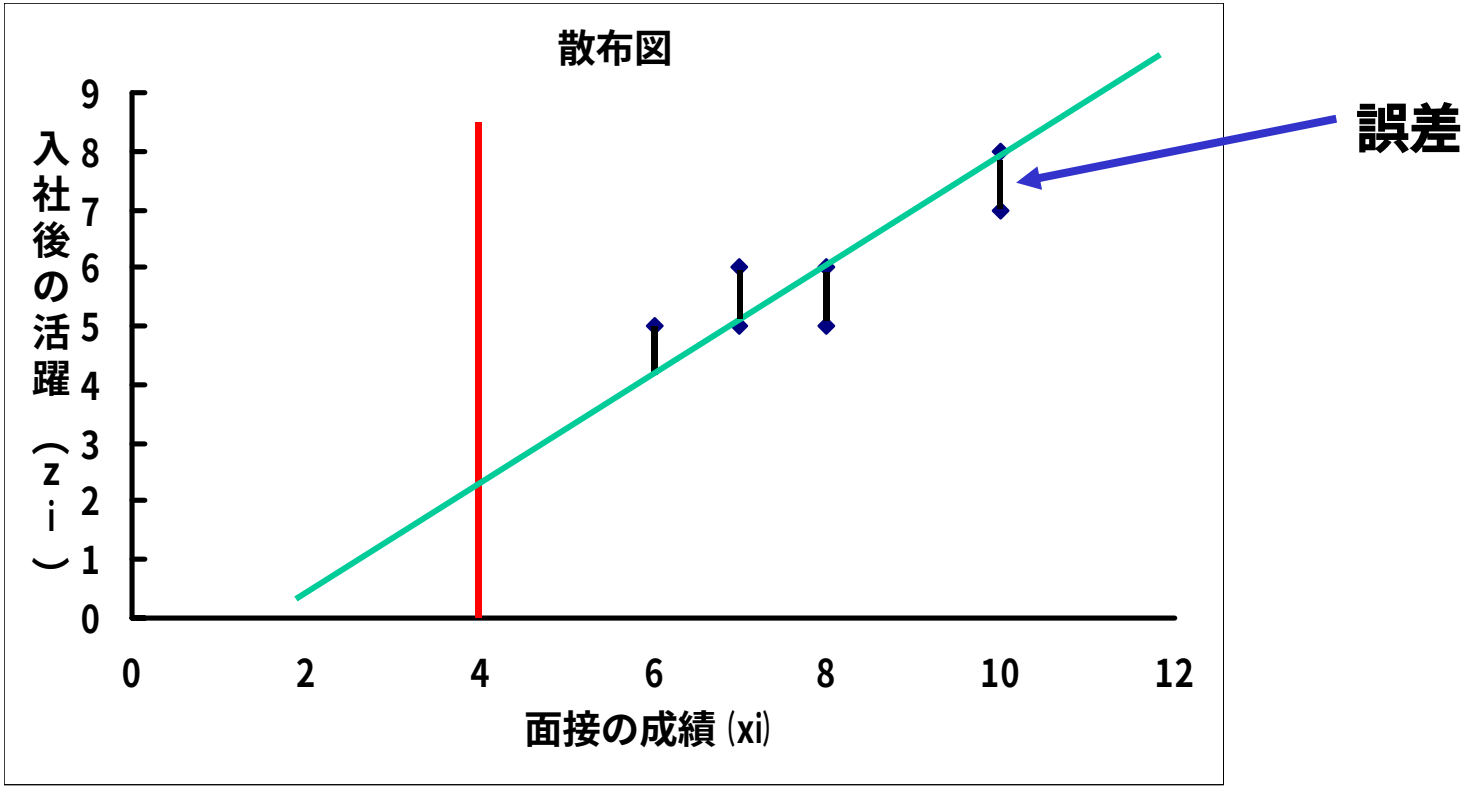
回帰分析とは　4

- 散布図を元に直線を引いてみると大体予測できる。
- でも直線の引き方は人によって異なる。最も予測の精度が良くなるように引きたい。



回帰分析とは　5

- このような場合に，直線からの誤差を，すべてのデータを考慮して最小にする直線を引くのが，**回帰分析**。
- 回帰分析によって引かれた直線を**回帰直線**，上のような回帰直線の引き方を，**最小二乗法**という。



回帰分析とは　6

- 回帰分析では，以下のようなことが分かる。
- **回帰直線**：$\hat{z} = a + bx$（回帰式）
  - x：説明変数（面接の成績）
  - z：基準変数（入社後の活躍）
  - $\hat{z}$：**予測値**
  - a：切片
  - b：**傾き**（**回帰係数**）
- **分散説明率**R^2：予測値 $\hat{z}$と xの相関係数の二乗を見ることで，データのばらつきのうち，回帰で説明できる割合がどれくらいかが分かる。**予測の精度**。

発表内容

- 回帰分析とは何か
- 重回帰分析とは何か
- Rとは何か
- Rで重回帰分析を実行

重回帰分析とは　1

- 回帰分析は，「入社後の活躍」と「面接の成績」といった2変数（説明変数と基準変数）間の関係について扱う分析だった。
- 重回帰分析は，多変量解析の一つとされるように，3変数以上（n個の説明変数と1個の基準変数）の間の関係について分析を行う手法である。
- 2変数間の回帰分析を「単回帰分析」と呼んで区別する。

重回帰分析とは　2

- 今度は，入社試験が面接と学科の2つになった。この2つの成績から，入社後の活躍を予測したい。
- どうすればよい？

姓	入社後の活躍 (zi)	面接の成績 (xi)	学科の成績 (yi)
山中	8	10	6
田口	7	10	9
中田	6	8	8
山口	6	7	6
中山	5	8	9
山田	5	7	5
田中	5	6	6

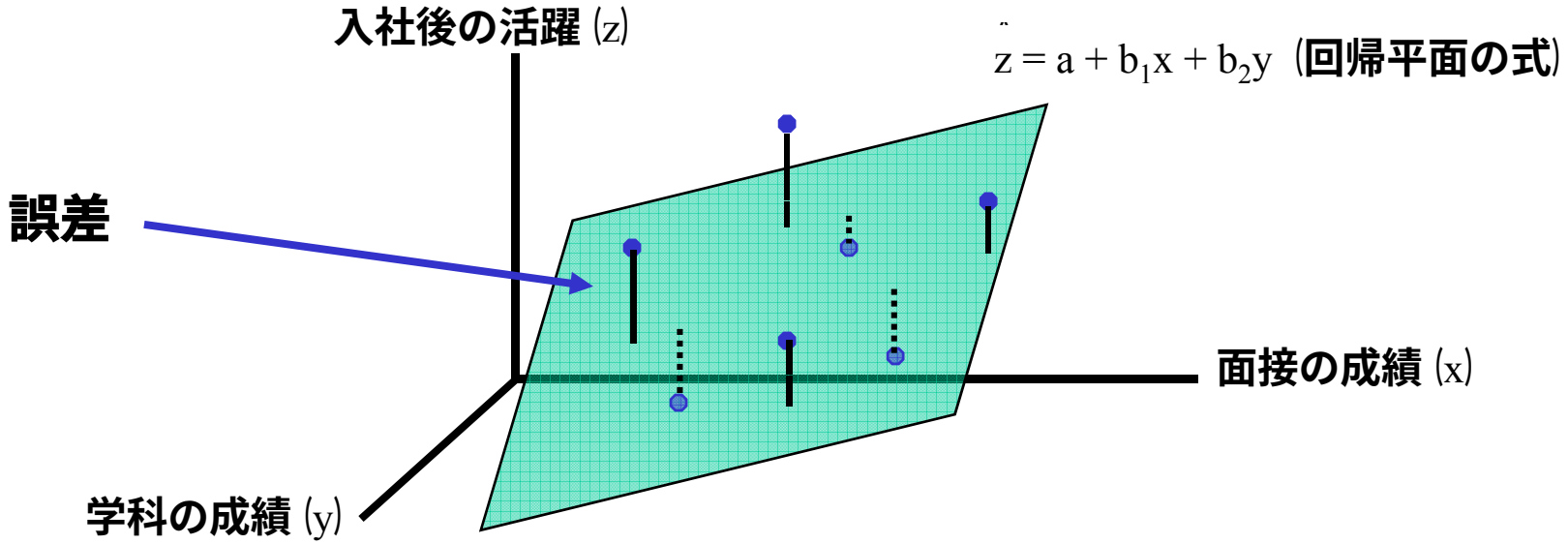
重回帰分析とは　3

- 面接の成績 (xi) と学科の成績 (yi) を合計して，単回帰分析を行えば良いのでは。
- しかし，「面接が10点，学科が6点」の人と，「面接が6点，学科が10点」の人の区別がつかなくなる。もし学科が入社後の活躍を全然予測しない説明変数だったら，面接のみの単回帰分析よりも精度を落とすことになる。
- それぞれの変数の予測力に応じたウェイトをかけて，予測の精度が最大になるようにしたい。

姓	入社後の活躍 (zi)	合計 (xi + yi)
山中	8	16
田口	7	19
中田	6	16
山口	6	13
中山	5	17
山田	5	12
田中	5	12

重回帰分析とは　4

- このような，2つ以上の説明変数を用いた予測値が最大の精度（分散説明率）を持つようにウェイトをかけるのが，**重回帰分析**。
- （最適）ウェイトとは，単回帰分析においては最小二乗法によって導かれた回帰直線の**傾き** b に当たる。説明変数が2つの場合は，データの散布図が3次元となり，以下のような**回帰平面**を最小二乗法で求め，平面の傾きを決定する b_1，b_2 を（最適）ウェイトとして算出することになる。
- 説明変数が3つ以上だと超回帰平面となり，視覚的に把握するのは困難になるが，手続きは同じ。

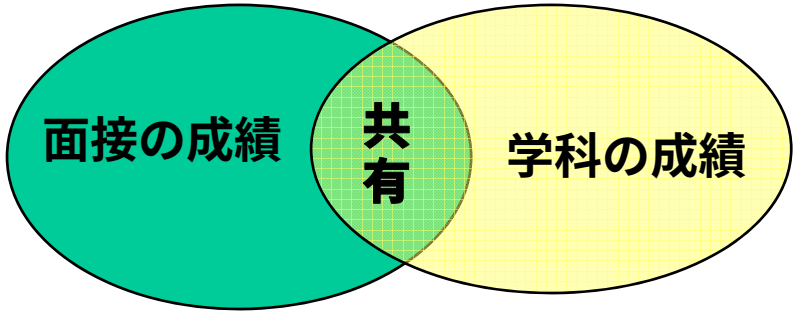


重回帰分析とは　5

- **重回帰分析**では，以下のようなことが分かる（説明変数2個の場合）。
- **回帰平面**：$\hat{z} = a + b_1 x + b_2 y$ （回帰式）
  - x, y : 説明変数（面接の成績，学科の成績）
  - z : 基準変数（入社後の活躍）
  - $\hat{z}$: 予測値
  - a : 切片
- **偏回帰係数**b_1, b_2 ：後述
- **分散説明率**R^2

重回帰分析とは　6

- 面接の成績と学科の成績が，それぞれどの程度入社後の活躍を予測するのが知りたい。
- ただし，面接の成績と学科の成績の間にも何らかの関係があると考えられるので，例えば，学科の成績にある程度の予測力が認められても，それは面接の成績が計測しているものの一部かもしれない。だとすれば，面接だけ行えば良いということになる。
- このため，2つの説明変数が共有している成分については除外して，それぞれの独自の成分だけを知る必要がある。



重回帰分析とは　7

- 実は，最小二乗法によって求めたそれぞれの説明変数の（最適）ウェイトである偏回帰係数は，そうした独自成分を反映した値であり，この疑問の答えとなる指標である。
- ただし，偏回帰係数の解釈には注意が必要である。
  - 偏回帰係数の値は，説明変数の単位に依存する。例えば，面接の成績を１０点満点，学科の成績が２００点満点，という場合には，そのまま偏回帰係数の比較ができない。こうした場合には，全ての変数の標準偏差（ばらつきの指標）を等しくした値である「標準偏回帰係数」によって比較が可能になる。
  - また，偏回帰係数の値は，もはや「説明変数１単位辺りの基準変数の変化量」を示していない。「ある説明変数から他の説明変数の影響を除いた残差変数１単位辺りの基準変数の変化量」であるため，他の説明変数によって強く影響される値であることに注意しなくてはならない。

発表内容

- 回帰分析とは何か
- 重回帰分析とは何か
- Rとは何か
- Rで重回帰分析を実行

Rとは　1

- **無料の統計ソフトの1つ。**
- **2007年現在，国内では重回帰分析をするのにSPSSを用いることが多いように（個人的には）思われるが，高価で個人所有するのが難しい。**
- **Rは無料で重回帰分析が実行可能。**
- **その他も様々な分析が可能で，「いつでも，どこでも，貧乏でも」が大きなメリットの1つ。**

- **ここより詳しい紹介は田中哲平氏の解説をどうぞ。**
  - （フリー統計ソフトRを用いた分散分析　http://www.educ.kyoto-u.ac.jp/cogpsy/personal/Kusumi/datasem07/tanaka.pdf）

Rとは　2

- 使い方の基本としては，
  - R Console を立ち上げる。
  - 「ファイル」→「ディレクトリの変更」で好みの作業ディレクトリにする。
  - 「ファイル」→「新しいスクリプト」or「スクリプトを開く」でスクリプトエディタを呼び出す。
  - スクリプトに行いたい分析を書き込み，その部分を選択し，「Ctrl＋R」で実行してみる。
    - データの読み込み
  - 結果がConsoleに出てくる。
    - データの表示
  - 思い通りの結果が出てくれば，さらにスクリプトに分析を追加していく。エラーが出れば，スクリプトを修正する。

発表内容

- 回帰分析とは何か
- 重回帰分析とは何か
- Rとは何か
- Rで重回帰分析を実行

Rで重回帰分析を実行　1

- **論文で重回帰分析を用いる際，何を報告するのか。**
- **例）文章理解研究の場合**
  - Komeda & Kusumi (2006). The effect of a protagonist's emotional shift on situation model construction. *Memory & Cognition*. **34**, 1548-1556.
  - **登場人物の感情のシフトの有無（特に関心のある説明変数）が各文の読み時間（基準変数）に与える効果を，他のシフト，文字数，呈示位置（あまり関心のない説明変数）を除いた上で観察したい。**
- **相関行列**
- **標準偏回帰係数**
- **分散説明率**

Rで重回帰分析を実行　2

- まず，**相関行列**（全ての変数間の相関係数）を報告する。
- 後で出てくる偏回帰係数の解釈などに役立つ，基本情報。

Table 1
Bivariate Correlations Between Predictor Variables in Experiments 1 and 2

Variables	1	2	3	4	5	6
1. Number of characters	–					
2. Serial position	.17**	–				
3. Temporal shifts	−.07	−.04	–			
4. Causal shifts	.01	−.10	.25**	–		
5. Spatial shifts	.06	.14**	.61**	.15**	–	
6. Emotional shifts	.17**	.34**	−.17**	−.15**	−.05	–

**$p < .01$.　$N = 368$.

- → 入社試験のデータで実行してみましょう。
  - 相関係数の出力

Rで重回帰分析を実行　3

- 次に，各説明変数と基準変数との**標準偏回帰係数** (β)
- **文字数，呈示位置，時・因果・空間**のシフトを除外しても，**感情**のシフトが有意に読み時間を延長させることがわかる。

Table 2
Beta Weights From the Regression Analyses of Reading Times in Experiments 1 and 2

	Experiment 1		Experiment 2	
Variables	*β*	*t*	*β*	*t*
Number of characters	.420***	21.11	.429***	16.50
Serial position	−.193***	−7.07	−.214***	−5.76
Time	.088***	4.76	.104**	3.79
Causation	.068***	5.27	.065***	4.34
Space	−.019	−1.09	−.052*	−2.70
Emotion	.064**	3.09	.077**	2.83

*$p < .05$. **$p < .01$. ***$p < .001$. Two-tailed.

- また，Komeda & Kusumi(2006)ではそれほど重要ではないが，多くの研究では**分散説明率** (R^2) も報告される。
- → **入社試験のデータで実行してみましょう。**
  - 重回帰分析の結果の出力
  - 偏回帰係数の標準化

引用･参考文献

- **南風原朝和** (2002). **心理統計学の基礎　統合的理解のために**. **有斐閣**
- Komeda, H. & Kusumi, T. (2006). The effect of a protagonist's emotional shift on situation model construction. *Memory & Cognition*. **34**, 1548-1556.
- **大村平** (2006). **多変量解析のはなし　改訂版**. **日科技連出版社**
- **田中哲平** (2007). **フリー統計ソフトRを用いた分散分析** http://www.educ.kyoto-u.ac.jp/cogpsy/personal/Kusumi/datasem07/tanaka.pdf