

重回帰分析

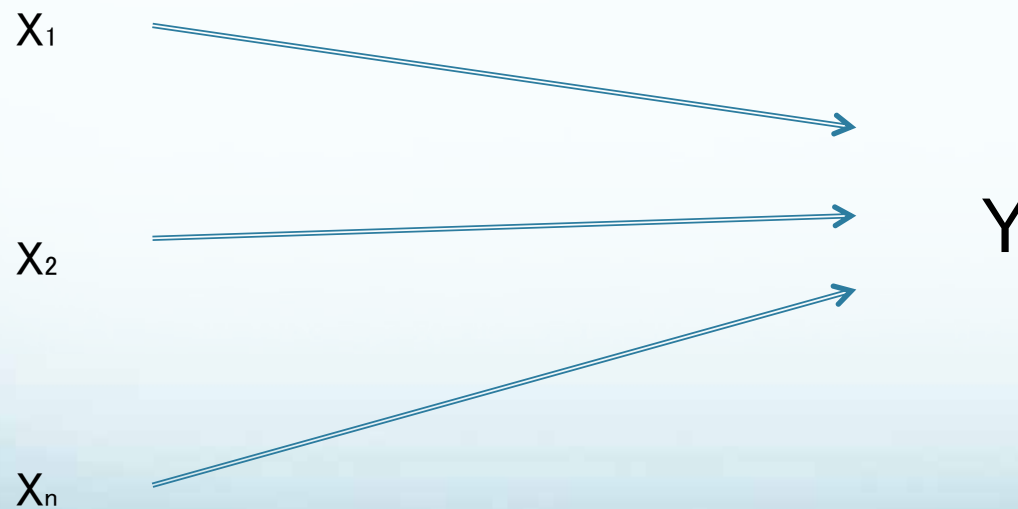
5/25/11
心理データ解析
M1 矢野裕理

今日のメニュー

- 回帰分析
- 単回帰分析
- 重回帰分析
- SPSSを使った重回帰分析

回帰分析とは？

- 独立変数 X_1, X_2, \dots, X_n によって従属変数 y を説明したり、予測するための統計的方法
 - イメージとしては…



回帰分析とは？

- 単回帰分析
 - 独立変数が1つの時
 - 予測する変数の説明変数が1つのとき
- 重回帰分析
 - 独立変数が2つ以上のとき
 - 予測する変数の説明変数が2つ以上のとき

単回帰分析とは？

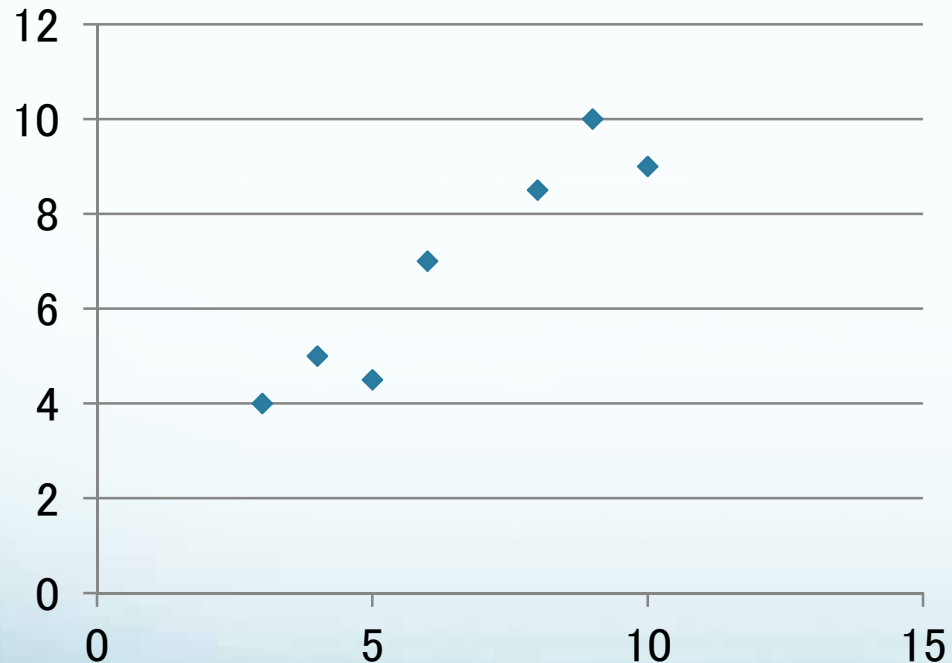
例題：ある化粧品会社が、一日の平均睡眠時間と美肌度を調べるとする。

以下の表はアンケートの一部を抜粋したものとする。

実験参加者	一日の平均睡眠時間	美肌度
1	8	8.5
2	3	4
3	5	4.5
4	4	5
5	10	9
6	6	7
7	9	10

単回帰分析とは？

一日の平均睡眠時間と 美肌度の関係



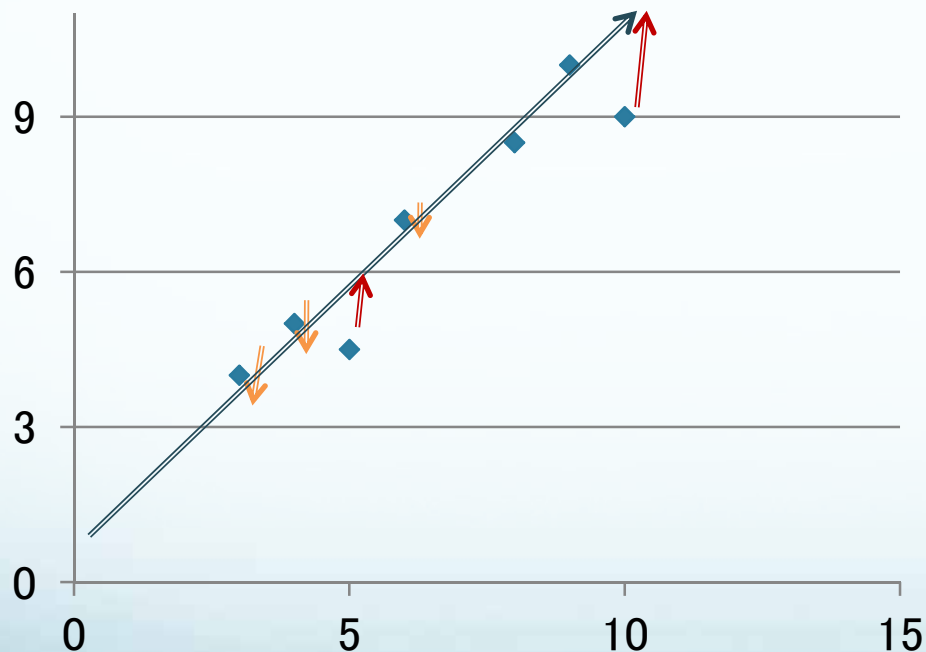
・左記のような散布図を描いてみると、一日の平均睡眠時間と美肌度の関係には強い正の相関がある。

($r=.94$)

・独立変数(一日の平均睡眠時間)から従属変数(美肌度)を予測するのが回帰分析の特徴である。

単回帰分析とは？

一日の平均睡眠時間と 美肌度の関係

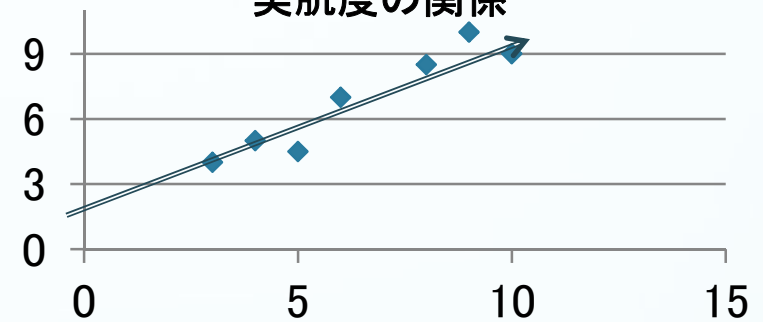


- ・では、一日に12時間寝る人は美肌度はいくつだといえるか。
- ・この際に直線からの誤差(オレンジの矢印)をすべてのデータを含めて最小にした直線が、**回帰直線**
- ・その回帰直線の弾き方を**最小二乗法**という
- ・回帰式は $y=a+bx$ で表すことができる
 - a:切片
 - b:傾き(回帰係数)
 - x:従属変数(美肌度)
 - y:独立変数(睡眠時間)
 - \hat{y} :予測値

単回帰分析とは？

最小二乗法を使い、 y の x への回帰直線を得る
この直線の x に既存の x の値を代入した時の \hat{y} の値を y の理論値という。

一日の平均睡眠時間と美肌度の関係



実験参加者	一日の平均睡眠時間(x)	美肌度(y)	得点の理論値(\hat{y})	残差($y - \hat{y}$)
1	8	8.5	$8a + b$	$8.5 - (8a + b)$
2	3	4	$3a + b$	$4 - (3a + b)$
3	5	4.5	$5a + b$	$4.5 - (5a + b)$
4	4	5	$4a + b$	$5 - (4a + b)$
5	10	9	$10a + b$	$9 - (10a + b)$
6	6	7	$6a + b$	$7 - (6a + b)$
7	9	10	$9a + b$	$10 - (9a + b)$

単回帰分析とは？

残差の二乗和の最小値を求める(最小二乗法)

$$L=[8.5-(8a+b)]^2+[4-(3a+b)]^2+[4.5-(5a+b)]^2+[5-(4a+b)]^2+[9-(10+b)]^2+[7-(6a+b)]^2+[10-(9a+b)]^2$$

- 単回帰分析では、以下の式で求まる。

$$a = \frac{\text{xとyの偏差積和}}{\text{xの偏差平方和}} = \frac{S_{xy}}{S_{xx}}$$

$$b = \text{yの平均} - \frac{\text{xとyの偏差積和}}{\text{xの偏差平方和}} \times \text{xの平均} = \bar{y} - \frac{S_{xy}}{S_{xx}} \times \bar{x}$$

決定係数とは？

- 決定係数
 - 理論値と実測値が全く違うようでは意味がない
 - 決定係数
 - 単回帰モデルの予測の精度を表す指標
 - $S^2(\hat{y}) / S^2(y) = r(x,y)^2$
 - r^2 : 分散説明率
 - y の観測値の変動(平均との偏差) = 回帰によって説明される変動(理論値の偏差) + 回帰によって説明されない変動(残差)

単回帰分析とは？

- 単回帰分析
 - 美肌度(y)が変化するの**は一日の平均睡眠時間(x)が原因だと考える**
 - **美肌度(y)は一日の平均睡眠時間(x)に従属しており(従属変数)**
 - **一日の平均睡眠時間(x)は独立の個々の値を取ることができる。(独立変数)**
 - では、値段の違う化粧品の影響を知るには？
 - 説明変数が2つ以上→重回帰分析

重回帰分析とは？

- 2変数間の回帰分析は「単回帰分析」
- 「重回帰分析」は3変数以上の間の関係について分析を行う手法である。
- 予測値(\hat{y}) = $a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3$
 - \hat{y} : 予測値、 a : 切片、 b : 偏回帰係数、 X_1 : 独立変数

重回帰分析とは？

美肌度(y)を一日の平均睡眠時間(x_1)、一ヶ月に化粧品にかける金額(x_2)、一週間に野菜を食べる回数(x_3)から予測する

実験参加者	一日の平均睡眠時間(x_1)	化粧品にかける金額(x_2)	一週間に野菜を食べる回数(x_3)	美肌度(y)
1	8	4000	4	8.5
2	3	3000	3	4
3	5	1000	3	4.5
4	4	3500	4	5
5	10	1500	2	9
6	6	3000	4	7
7	9	6000	6	10

重回帰分析とは？

最小二乗法を用いて、従属変数の実測値と予測値との誤差 ($y-\hat{y}$) の二乗和が最小になるような回帰定数、偏回帰係数を求める

実験参加者	美肌度 (y)	美肌度の理論値	残差
1	8.5	$a+(8b_1+4000b_2+4b_3)$	$8.5-a+8b_1+4000b_2+4b_3$
2	4	$a+(3b_1+3000b_2+3b_3)$	$4-a+3b_1+3000b_2+3b_3$
3	4.5	$a+(5b_1+1000b_2+3b_3)$	$4.5-a+5b_1+1000b_2+3b_3$
4	5	$a+(4b_1+3500b_2+4b_3)$	$5-a+4b_1+3500b_2+4b_3$
5	9	$a+(10b_1+1500b_2+2b_3)$	$9-a+10b_1+1500b_2+2b_3$
6	7	$a+(6b_1+3000b_2+4b_3)$	$7-a+6b_1+3000b_2+4b_3$
7	10	$a+(9b_1+6000b_2+6b_3)$	$10-a+9b_1+6000b_2+6b_3$

重回帰分析とは？

- 標準偏回帰係数 (β)
 - 各データを標準化して、回帰直線を求める。その偏回帰係数のこと。
 - 測定単位の影響を受けない。
 - データを標準化しているので、切片は0になる
 - 数値の大きさを、その独立変数の従属変数への影響力の強さを比較できる。→重回帰分析では重要！
- 独立変数の原点や単位があいまいなとき(例: 怒りを測る5件法の尺度など)は標準偏回帰係数、明確なときは偏回帰係数をつかうとよい

重回帰分析とは？

- 多重共線性：複数の独立変数間に強い相関があると多重共線性が起こっている可能性がある
 - 重要な独立変数も有意ではなくなる。
 - 回帰係数の分散を増やし、回帰式を不安定にする。
- 解決策
 - 相関係数の高い変数のうち、少なくとも1つの独立変数を削除する
 - 独立変数をまとめる(独立変数に対して因子分析や主成分分析を行って合成する)
 - ただし、再度理論から仮定した因果モデルを考慮し直す必要も

SPSSで単回帰分析を行う

- データの読み込みをする
 - 「データドキュメントを開く」を選択
 - デスクトップからデータファイルを選択
 - (ファイルの種類をexcelの”xls”にする)
 - ⇒「開く」
 - ワークシートから「example」を選択
 - ⇒「OK」

SPSSで単回帰分析

- 分析メニュー → 回帰 → 線型
 - 「従属変数」に「美肌度」を指定
 - 「独立変数」に「平均睡眠時間」を指定
 - 「統計」ボタンを押し、「記述統計量」にチェックを入れる。「続行」を押す。
 - 各変数の平均値や標準偏差, 相関が出力される。
 - 「OK」を押す

SPSSで重回帰分析

- 分析メニュー → 回帰 → 線型
 - 「従属変数」に「美肌度」を指定
 - 「独立変数」に「化粧品にかける金額」「野菜を食べる回数」を指定
 - 「統計」ボタンを押し、「記述統計量」にチェックを入れる。「続行」を押し。
 - 各変数の平均値や標準偏差, 相互相関が出力される。
 - 「OK」を押し

例題

- 『多変量データ解析法』(足立, 2006)
 - Tシャツの売れ行きを、素材(の良さ)・値段・デザイン(の良さ)という3つの説明変数の大小によって予測する
- 「データドキュメントを開く」を選択
- デスクトップからデータファイルを選択
 - (ファイルの種類をexcelの”xls”にする)
 - ⇒「開く」
 - ワークシートから「足立(2006)より」を選択
 - ⇒「OK」

SPSSで重回帰分析

- 分析メニュー → 回帰 → 線型
 - 「従属変数」に「売れ行き」を指定
 - 「独立変数」に「素材」「値段」を指定
 - 「統計」ボタンを押し、チェックを入れる。「続行」を押し。
 - 「OK」を押し

SPSSで重回帰分析

- 各変数間の相互相関が出力される。
 - 上からPearsonの積率相関係数、有意確率、データ数。
 - 値段は売れ行きと有意な負の相関を示している。
 - 素材は売れ行きと有意な負の相関を示している。
 - 値段と素材は有意な正の相関を示している。
- 重相関係数(R)、重決定係数(R^2)、自由度調整済み(サンプルサイズの調整)の R^2 が出力される
- 回帰式全体の有意性の検定。0.01%水準で有意である。
 - この有意確率が重決定係数(R^2)の有意水準となる

SPSSで重回帰分析

- 標準化していない回帰係数(B)と標準偏回帰係数(β)、およびその有意確率が出力される。
- 「値段」が「売れ行き」に対して負の有意な影響力をもっていることがわかった。(「素材」は有意差がなかった)
- 許容度の目安は0.2以下で多重共線性が生じている可能性が高い
- $VIF > 10$ では多重共線性の可能性が高い
 - この場合は多重共線性はなかった

参考文献

- 足立浩平 (2006). 多変量データ解析法—心理・教育・社会系のための入門 ナカニシヤ出版
- 森敏昭・吉田寿夫 (1990). 心理学のためのデータ解析テクニカルブック 北大路書房
- 栗本達児(2005). 単回帰分析と重回帰分析
<http://kyoumu.educ.kyoto-u.ac.jp/cogpsy/personal/Kusumi/datasem05/kurimoto.pdf>
- 猪原敬介(2007).Rによる重回帰分析
<http://www.educ.kyoto-u.ac.jp/cogpsy/personal/Kusumi/datasem07/inohara.pdf>