回帰分析の多変量バージョン
重回帰分析

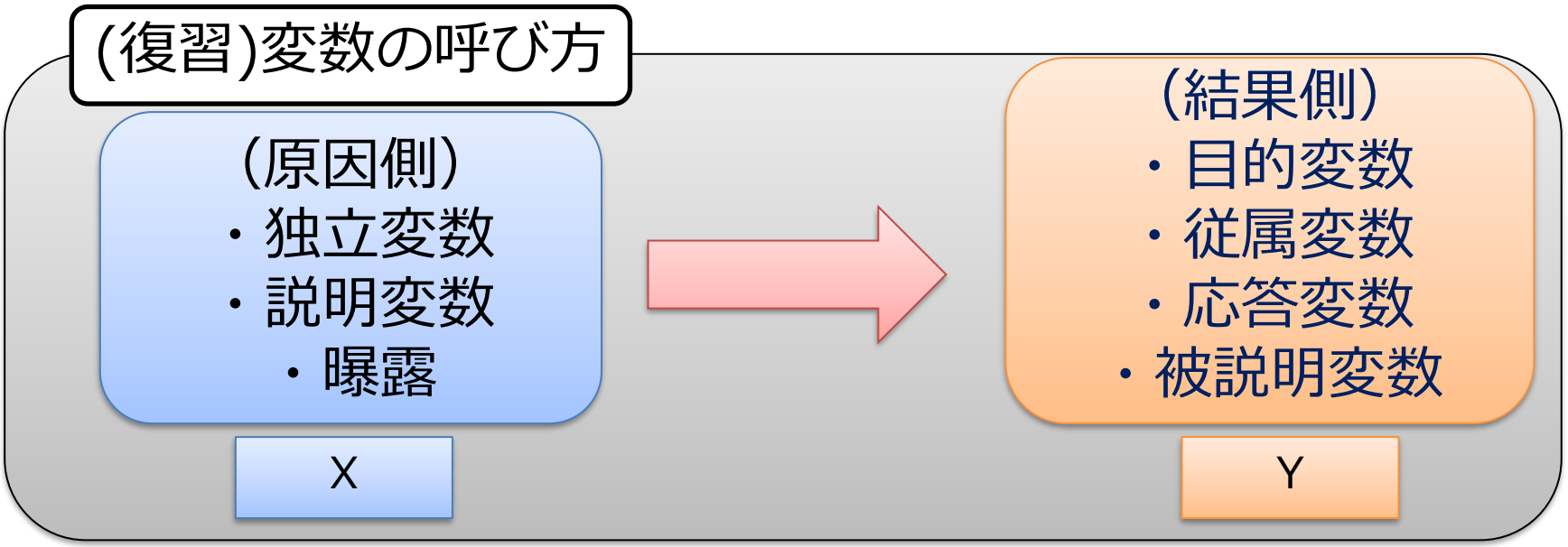
(復習)回帰分析

- 量的変数と量的変数の関係を直線関係にモデル化
- 母集団における，目的変数yと、独立変数（説明変数）xの関連性を、（1）式(回帰式)のようにモデル化する
- 誤差は互いに独立に分散が等しい正規分布に従うことを仮定している

$$y = a + bx + \varepsilon \qquad (1)$$

a: 切片, b: 回帰係数, ε: 誤差

- 母集団における真の関連性はわからないので，xとyのデータを取ってきて，aとbの推定値$\hat{a}$と$\hat{b}$を求めるのが基本的な流れ



2種類の回帰係数

- 非標準化偏回帰係数
  - 非標準化偏回帰係数の値はその変数が1単位変化した時，目的変数が何単位変化するかを表す
  - 式に含まれる変数の単位が変わると値が変わる
  - 変数の値そのものに意味がある場合に使いやすい(長さ，重さ，金額)
- 標準化偏回帰係数
  - 目的変数，説明変数のうち量的変数を平均0，分散1に標準化したデータで推定した偏回帰係数
  - 標準化偏回帰係数の値は説明変数が1標準偏差変化した時，目的変数が何標準偏差変化するかを表す
  - **単回帰分析の場合**はピアソンの積率相関係数と一致する
  - 式に含まれる変数の単位に依存しない
    →複数の説明変数がある時(多変量解析)の比較に用いられる

回帰分析の切片からわかること，解釈上の注意

- 回帰式で説明変数に0(ゼロ)を代入したときの目的変数の値
- 説明変数が0にならないような変数の場合は意味がない
- 理論上説明変数が0を取りうる場合でも分析対象に説明変数の値が0となるような対象が含まれない場合には意味のない値が出ることもある
  - 例: 22歳から65歳の大卒者の年収と年齢のデータを回帰分析した結果，以下のような結果が得られたとする

    年収(万円)=240+0.5×年齢(歳)

    このとき，切片は240万円であるが，0歳児(年齢=0)の平均年収が240万円ということはできない
- 切片の値を解釈可能にするために，説明変数の値から説明変数の平均値を引いたものを説明変数として分析を行うこともある

回帰分析の回帰係数からわかること，解釈上の注意

- 回帰係数の値は説明変数の値が1大きくなると目的変数の値がいくつ大きく(小さく)なるかを表している
  - 回帰係数の値が0より大きい(正(プラス)の値)
    - 説明変数の値が大きくなると目的変数の値も大きくなる
      →二つの変数が**正の**相関関係にあることがわかる
  - 回帰係数の値が0より小さい(負(マイナス)の値)
    - 説明変数の値が大きくなると目的変数の値が小さくなる
      →二つの変数が**負の**相関関係にあることがわかる
- 非標準化回帰係数は目的変数，説明変数の単位によって値が変わる
  例: 22歳から65歳の大卒者の年収と年齢のデータを回帰分析した結果，以下のような結果が得られたとする

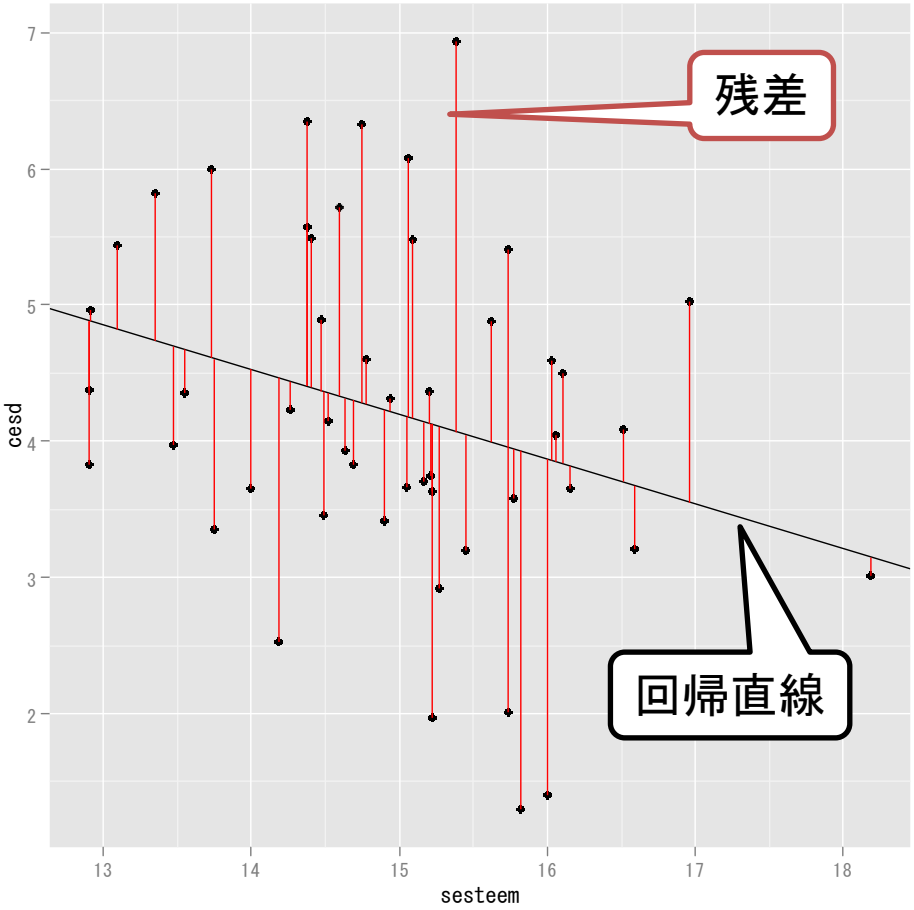
  年収(万円)=240+0.5×年齢(歳)

  データは同じまま**年収の単位を千円にする**と切片240万円=2400千円，回帰係数0.5万円/歳=5千円/歳だから回帰式は以下のようになる

  年収(千円)=2400+5×年齢(歳)
  - 目的変数，説明変数の単位に注意する
  - 回帰係数の大きさだけでは関連性の強さを判断したり，比較したりできない
    →標準化回帰係数を使う

散布図と回帰直線

- データから推定した回帰式をグラフにしたのが回帰直線(regression line)
- 回帰直線と散布図のそれぞれの点の差が残差(residual)

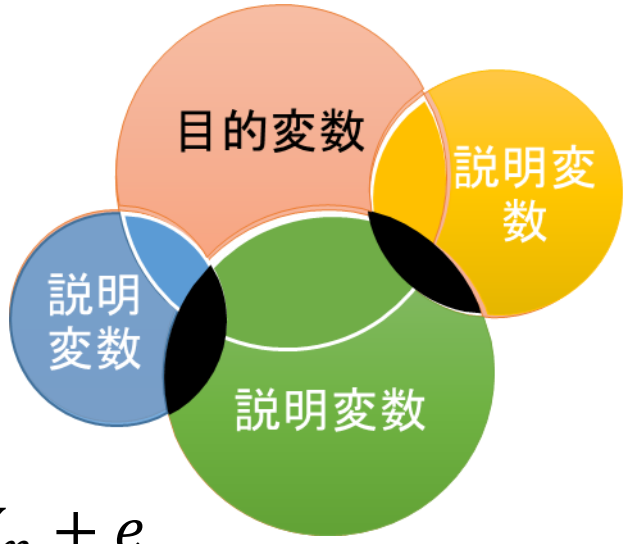


- どのように回帰式のパラメータを推定するか?
  - 残差の2乗の和（残差平方和sum of squares）が最も小さくなるようにパラメータを決める(=回帰直線と散布図のそれぞれの距離が一番小さくなるように直線を引く)

→最小二乗法(ordinal least squares estimation)

重回帰分析(multiple regression analysis)

- 単回帰分析の拡張版
  - 説明変数が複数
  - 多変量解析の基本形
- 重回帰分析のモデル式
  - 誤差の分布の仮定は単回帰分析と同じ
    →独立性，正規性，等分散性を仮定



$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

Y: 目的変数 Xi: 説明変数
a: 切片 bi: 偏回帰係数 e: 誤差

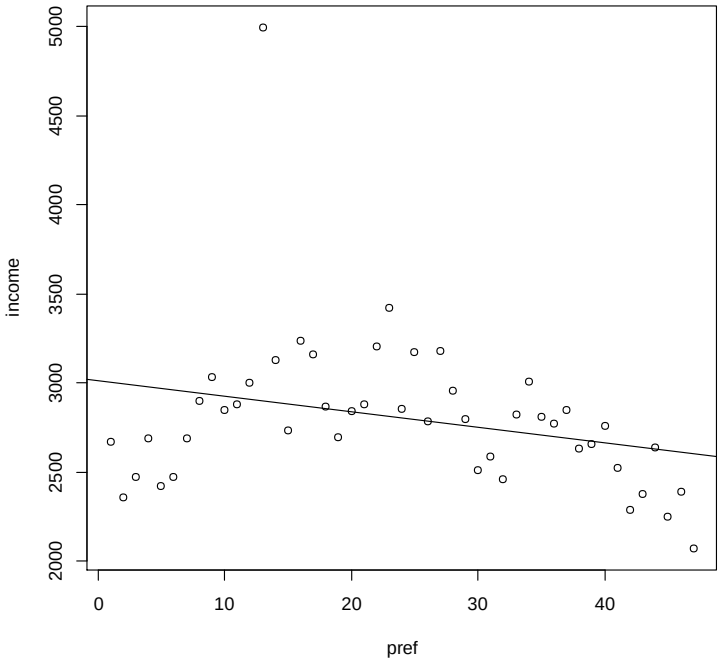
- 重回帰分析の考え方
  - 注目する変数以外の変数の値を固定して注目する変数を変化させたときの目的変数との関連をみる
  - 偏回帰係数は他の変数の影響を取り除いた，その**変数独自の影響力**を表している

回帰分析の説明変数として質的変数を使うには？

- (重)回帰分析では説明変数も目的変数も量的変数として扱われる
  - 名義変数を回帰分析の説明変数として使うと解釈ができない
  - 例:

都道府県を北海道＝1，青森＝2，…，沖縄＝47というように番号をつけてデータ化し，各都道府県民の平均所得を回帰分析で分析
→右の散布図．南に行くほど所得が下がるという解釈もできなくはないが，**都道府県を量的データとして扱うより，質的データとして扱って，平均を比較する方が妥当**

県番号	都道府県名	一人あたり所得(千円)
1	北海道	2,670
2	青森県	2,361
3	岩手県	2,472
	⋮	
45	宮崎県	2,251
46	鹿児島県	2,392
47	沖縄県	2,070



所得＝3014.9－8.81×県番号

出典： 県民経済計算（平成13年度 - 平成26年度）（93SNA、平成17年基準計数）

回帰分析の説明変数として質的変数を使う方法ーダミー変数

- ダミー変数
  - 質的変数のカテゴリを0と1の組み合わせで表現した変数
  - ダミー変数の値がすべて0のカテゴリ(=**参照カテゴリ**)と比べてどのくらい目的変数に影響しているかを表すことができる
    →ダミー変数の偏回帰係数は参照カテゴリと該当するカテゴリの**平均の差**と解釈できる
  - カテゴリ数ー1個のダミー変数を作れば全てのカテゴリを表現できる

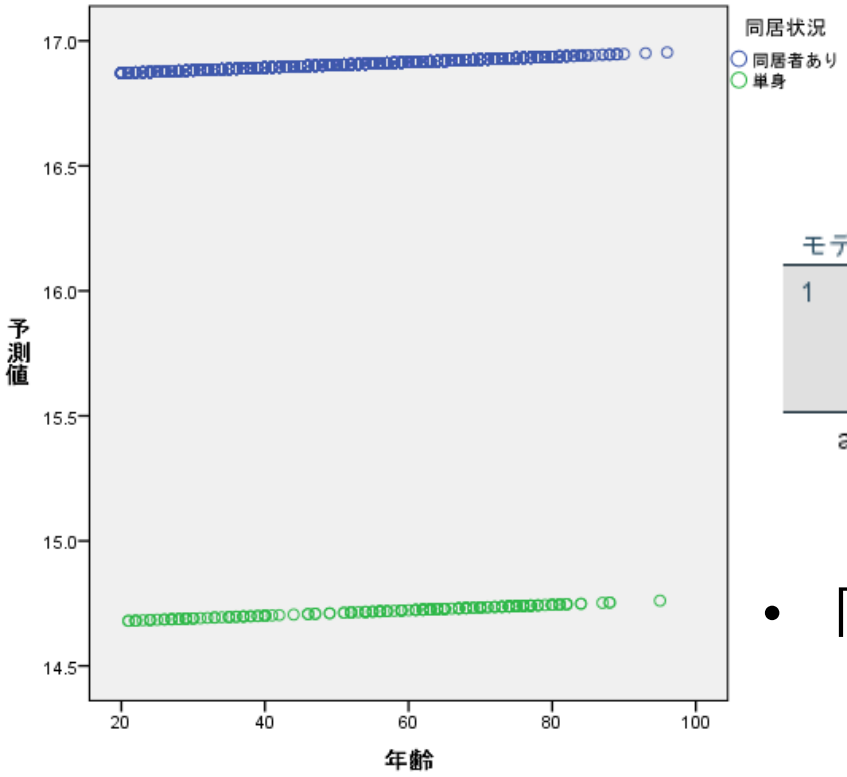
D	性別
0	男
1	女

男性が参照カテゴリ

D1	D2	D3	学歴
1	0	0	大卒以上
0	1	0	専門・短大卒
0	0	1	高卒
0	0	0	中卒

中卒が参照カテゴリ

ダミー変数の偏回帰係数の解釈



係数[a]

モデル		非標準化係数 B	非標準化係数 標準誤差	標準化係数 ベータ	t値	有意確率
1	(定数)	16.849	.171		98.486	.000
	同居状況	-2.192	.190	-.209	-11.550	.000
	年齢	.001	.003	.006	.357	.721

a. 従属変数 食に関する主観的QOL

- 同居状況がダミー変数
  - 0が同居，1が単身
  - 係数の表では同居状況の係数が-2.192
    →**単身の人は同居の人に比べて2.192点食に関する主観的QOL得点が低い**

2種類の回帰係数ー非標準化偏回帰係数

- 解釈の仕方は単回帰分析の回帰係数の解釈とほぼ同じ
  - 違いは**他の説明変数の値を0に固定したときの**，説明変数と目的変数の関連を表しているという点のみ
- 非標準化偏回帰係数(unstandardized partial regression coefficient)
  - 注目する変数以外の説明変数を0に固定した上で，注目する変数が1単位変化した時，目的変数が何単位変化するかを表す
    - 説明変数が**量的変数なら傾き,質的変数なら参照カテゴリとの平均の差**を表す
  - 式に含まれる変数の単位が変わると値が変わる
  - 変数の値そのものに意味がある場合に使いやすい(長さ，重さ，金額)

2種類の回帰係数ー標準化偏回帰係数

- 解釈の仕方は単回帰分析の回帰係数の解釈とほぼ同じ
  - 違いは**他の説明変数の値を0に固定したときの**，説明変数と目的変数の関連を表しているという点のみ
- 標準化偏回帰係数(standardized partial regression coefficient)
  - 目的変数，説明変数のうち量的変数を平均0，分散(標準偏差)を1に標準化して推定した偏回帰係数
    - 説明変数が量的変数の場合，説明変数が1標準偏差変化した時，目的変数が何標準偏差変化するかを表す
    - 説明変数が質的変数の場合は参照カテゴリに比べて何標準偏差違うかを表す
    - 式に含まれる変数の単位が変わっても値が変化しない
  - 説明変数がすべて量的変数の場合は値の大きさで影響の大きさを比較できる
  - 量的変数と質的変数が混在するときは値の意味が違う

→関連の強さを比較するときは相関比（η^2）で

変数の影響力の比較

- 相関比（η^2）
  - モデル全体の平方和（SS_{total}，下表の「総和のタイプIII平方和」)と変数の平方和（SS_{var}，下表)の各変数のタイプIII平方和の比

被験者間効果の検定

従属変数: cesd CES-D

ソース	タイプ III 平方和	自由度	平均平方	F 値	有意確率	偏イータ 2 乗
修正モデル	654.775[a]	3	218.258	41.433	.000	.056
切片	4929.525	1	4929.525	935.805	.000	.308
sexd	23.271	1	23.271	4.418	.036	.002
disease	66.719	1	66.719	12.666	.000	.006
control	455.723	1	455.723	86.513	.000	.040
誤差	11056.872	2099	5.268			
総和	47670.000	2103				
修正総和	11711.647	2102				

a. R2 乗 = .056 (調整済み R2 乗 = .055)

- 例:
- SEXD（性別）のη^2:23.271/47670=0.0004882
- control(コントロール感)のη^2:455.723 /47670=0.00956

論文での結果の示し方の例

表 3　孤独感に関連する要因―接触相手・提供者の有無，重回帰分析―　$n = 412$

	β	P
基本属性		
母親の年齢	.10	*
経済的ゆとり	.01	ns
健康状態のよさ	−.13	**
内的作業モデル　安定尺度得点	−.50	***
母親意識　肯定的意識尺度得点	−.10	**
育児環境		
外出の困難感のなさ	−.13	**
接触対象者・サポート提供者あり		
夫・パートナー	−.03	ns
実父母	−.06	ns
ママ友達	−.10	**
友人	−.08	*
Adjusted R^2	.43	

†: $P<.1$, *: $P<.05$, **: $P<.01$, ***: $P<.001$, ns: not significant.

説明変数

標準化偏回帰係数，β(ベータ)と書かれていることが多い

偏回帰係数の検定結果
帰無仮説は通常「偏回帰係数=0」
帰無仮説が棄却されれば，有意な関連があると判断する

自由度調整済み決定係数

乳児の母親の孤独感の関連要因を検討した研究．目的変数は孤独感，母親の年齢などの属性，育児環境，周囲からのサポートの有無．

馬場 千恵, 村山 洋史, 田口 敦子, 村嶋 幸代, 乳児を持つ母親の孤独感と社会との関連について, 日本公衆衛生雑誌, 2013, 60 巻, 12 号, p. 727-737, https://doi.org/10.11236/jph.60.12_727

変数の相互作用の回帰モデルにおける表現と確認方法(1)

- 交絡・媒介
  - 注目する要因のみを投入するモデルと注目する要因と交絡要因(媒介要因)の候補を同時に投入するモデルを比較する
  - 注目する要因のみ投入するモデルで目的変数との関連があり，交絡要因(媒介要因)を同時に投入するモデルでは関連が消失または弱くなった場合に交絡または媒介が起こっていると考える
  - 交絡か媒介かは関連性の分析だけではわからない．理論的な考察も必要
- 抑制・歪曲
  - 注目する要因のみを投入するモデルと注目する要因と抑制変数(歪曲変数)の候補を同時に投入するモデルを比較する
  - 注目する要因のみ投入するモデルで目的変数との関連がなく，抑制変数(歪曲変数)も同時に投入するモデルでは関連が出現したら抑制が，関連性の符号が逆になあったら歪曲が起こった可能性があると考える

変数の相互作用の回帰モデルにおける表現と確認方法(2)

- 修飾
  - **交互作用項**を投入し，関連があれば修飾が起こっていると考える
  - 交互作用項
    - 注目する要因と効果を修飾すると考えられる要因の積をとったもの
- 考え方
  - 交互作用項を投入した回帰モデルは以下のように表現できる
  - $Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2$

これをX_1について整理すると

$Y = a + (\boldsymbol{b_1} + \boldsymbol{b_3X_2})\boldsymbol{X_1} + b_2X_2$

X_1とYの関連性がX_2の値によって変わると見ることができる
→X_2が修飾要因
b_3はどの程度修飾効果が大きいかを表すとみることができる