重回帰分析の応用: 階層的重回帰分析

- 投入する説明変数を順次追加していき，偏回帰係数の変化をみる方法
  - 2変量(単回帰分析)→多変量(重回帰分析)
  - 興味のない説明変数(制御変数)だけを投入したモデル→制御変数+興味のある説明変数を投入したモデル
- 階層的重回帰分析によって，説明変数間の相互作用や目的変数に影響を与えるメカニズムを推測することができる
  - 階層的重回帰分析の発展形がパス解析や構造方程式モデリング

階層的重回帰分析の例

- 男性における食に関する主観的QOL(以下SDQOL)と年齢，就業状況，配偶者の有無の関係
  - 年齢，就業の有無だけを投入したモデル1では就業ありの人でSDQOLが高い
  - モデル1に配偶者の有無を追加すると就業の有無の影響はほぼ0になる
    →就業している人ほど配偶者がいる可能性が高い
    →配偶者がいる人ほどSDQOLが高い
    という関係が推測できる

	モデル1		モデル2	
	標準化係数	有意確率	標準化係数	有意確率
年齢	0.050	0.092	−0.096	0.003
就業の有無	0.081	0.007	−0.001	0.968
配偶者の有無			0.293	0.000

(復習)回帰式のあてはまりの評価(1)

- **決定係数**(R^2)

$$R^2 = 1 - \frac{\text{残差平方和}}{\text{目的変数の平方和}} = \frac{\text{回帰モデルの平方和}}{\text{目的変数の平方和}}$$

  - 平方和: ある値を二乗して合計したもの
    - ◆残差平方和:残差を二乗して合計したもの
    - ◆目的変数の平方和: 目的変数の値と目的変数の平均値の差を二乗して合計したもの
    - ◆回帰モデルの平方和: モデルによる予測値と予測値の平均値の差を二乗して合計したもの
  - 決定係数は0から1の値を取り，1に近いほど当てはまりがよいことを表している
    - ◆決定係数の値は目的変数のばらつきを回帰モデルによって説明できる割合として解釈できる
      →決定係数は**分散説明率**とも呼ばれる
    - ◆目的変数の予測，説明を目的にしている場合は決定係数が高いことが求められる
    - ◆説明変数の一部にのみ関心があって関心がある変数以外の条件を揃える(制御する)する目的で重回帰分析を使う場合は気にしなくてよい

回帰式のあてはまりの評価(2)

- 重回帰分析のときの注意点
  - 決定係数は説明変数が多くなると自動的に大きくなる

→説明変数の多さを考慮して比較するための指標が自由度調整済みR^2

$$\text{自由度調整済み}R^2 = 1 - \frac{\frac{\text{残差平方和}}{N-p-1}}{\frac{\text{目的変数の平方和}}{N-1}}$$

N: サンプルサイズ, p: 説明変数の数

分析結果のチェック：多重共線性(multicollinearity)の問題

- 説明変数として似たもの（相関が高い変数）を投入したときに生じる現象
  - どういう時に起こるか？：説明変数間の相関が高い場合，ある説明変数が他の説明変数で説明されてしまう(一次従属/線形従属)とき
  - 主な症状: 回帰係数などの推定値が出ない，推定された回帰係数がありえない値になる(符号が逆等)，標準誤差が大きくなる
- チェック方法
  - 説明変数間の相関係数をチェック
    →高い相関を持つ者同士が説明変数として投入されないように注意する
  - SPSSでは線形回帰の統計量の選択のところで「共線性の診断」にチェックを入れれば多重共線性の指標を出力できる
    - 分散拡大要因(Variance Inflation Factor: VIF)が5～10を超えていたら要注意
- 対処方法
  - 相関が高い説明変数を除去して1つだけにする
  - 相関が高い変数同士をまとめて合成変数として扱う(→主成分得点，因子得点，合計点等を使う)
  - サンプルサイズを増やす(解析段階では手遅れな事が多いが…)

交互作用項を入れるモデルの注意

- 交互作用項を構成する量的変数は中心化(centering)か標準化(standardize)しておく
  - 交互作用を構成する変数と交互作用項の相関が高くなり、多重共線性の問題が生じるのを防ぐため

係数[a]

モデル		標準化されていない係数		標準化係数			共線性の統計量	
		B	標準偏差誤差	ベータ	t 値	有意確率	許容度	VIF
1	(定数)	185.631	16.750		11.083	.000		
	age AGE OF RESPONDENT	-.597	.230	-.485	-2.594	.010	.033	29.868
	weight WEIGHT (kg)	-.307	.321	-.317	-.957	.339	.011	93.415
	sexd	-11.079	.668	-.604	-16.587	.000	.884	1.132
	agebweight	.009	.004	.705	2.007	.046	.010	105.218

a. 従属変数 height HEIGHT (cm) OBSERVED BY INTERVIEWER

係数[a]

モデル		標準化されていない係数		標準化係数			共線性の統計量	
		B	標準偏差誤差	ベータ	t 値	有意確率	許容度	VIF
1	(定数)	.781	.058		13.566	.000		
	Zweight Z 得点: WEIGHT (kg)	.336	.037	.336	9.091	.000	.856	1.168
	Zage Z 得点: AGE OF RESPONDENT	-.118	.035	-.118	-3.406	.001	.970	1.031
	sexd	-1.243	.075	-.604	-16.587	.000	.884	1.132
	zagebweight	.067	.033	.069	2.007	.046	.979	1.021

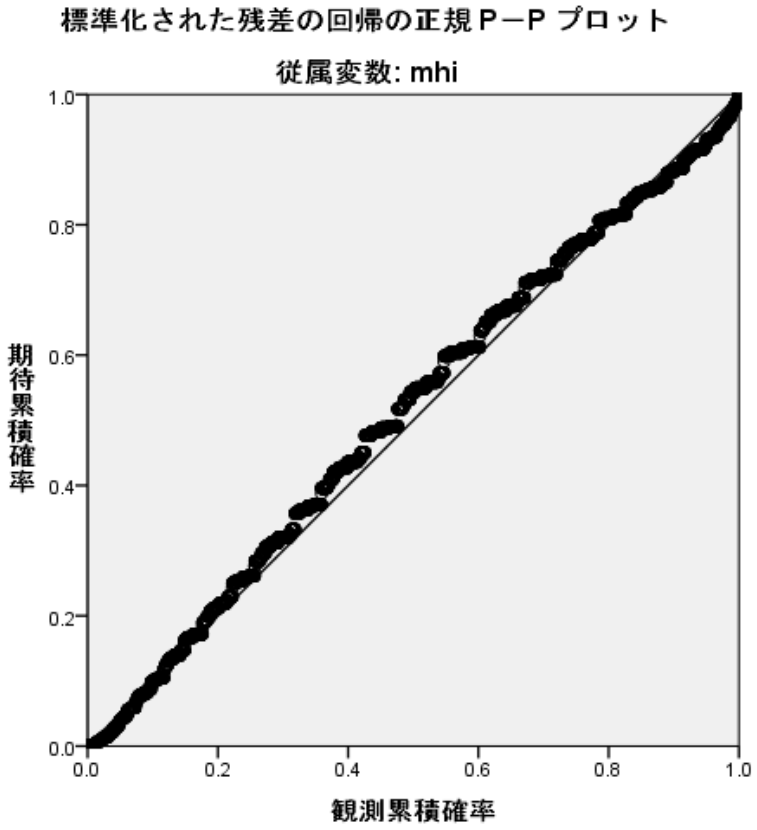
a. 従属変数 Zheight Z 得点: HEIGHT (cm) OBSERVED BY INTERVIEWER

分析結果のチェック: 回帰診断

- 残差に関する前提や，分析に与える影響が大きいレコードをグラフなどで確認する
  - 残差の正規性
    →正規確率プロット
  - 残差の等分散性
  - 残差の独立性
    →予測値に対する残差プロットでチェック
  - 外れ値・影響点のチェック
    →Cookの距離，てこ比などでチェック
    - Cookの距離はそのデータがパラメータに与える影響の大きさを表す．0.5以上で「影響が大きい」，1以上で「特に影響が大きい」とされる．
    - てこ比は説明変数が極端な値であるかの指標で，てこ比の平均の3倍(2倍を基準にすることもある)のてこ比を持つケースは注意が必要となる

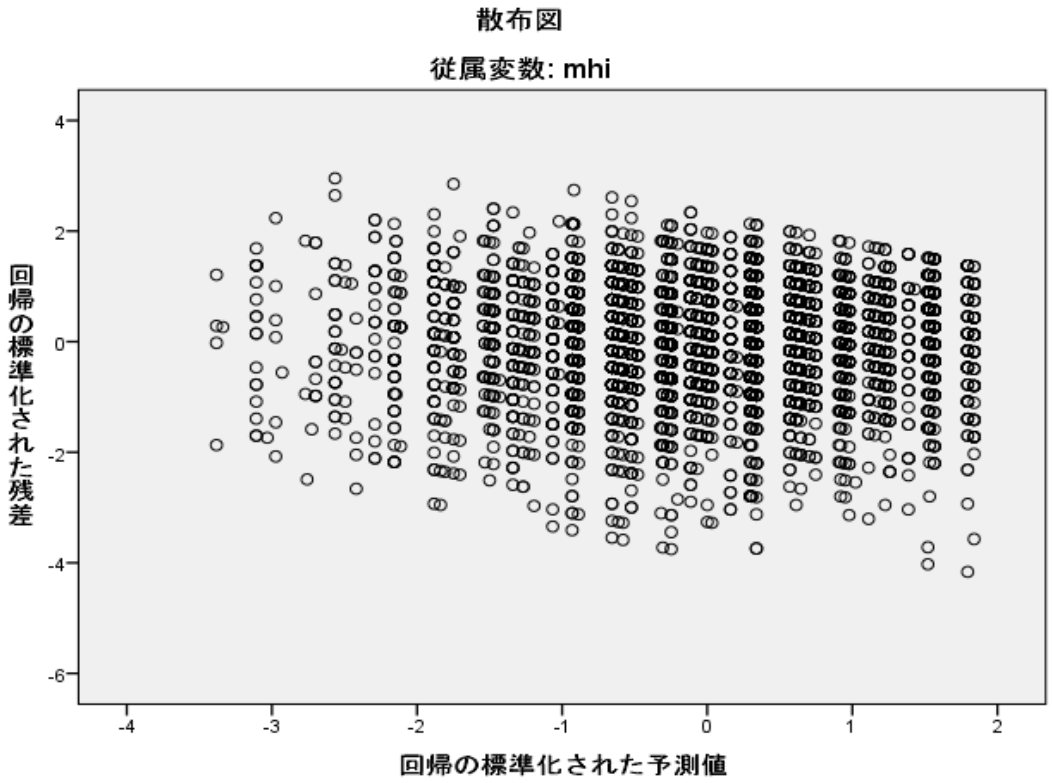
正規確率プロット

- 標準化した残差に対して，標準化した残差が正規分布に従うと仮定した時の期待値をプロットしたもの
- 残差の正規性を確認できる
  - プロットが直線状に並んでいればOK



予測値に対する残差プロット

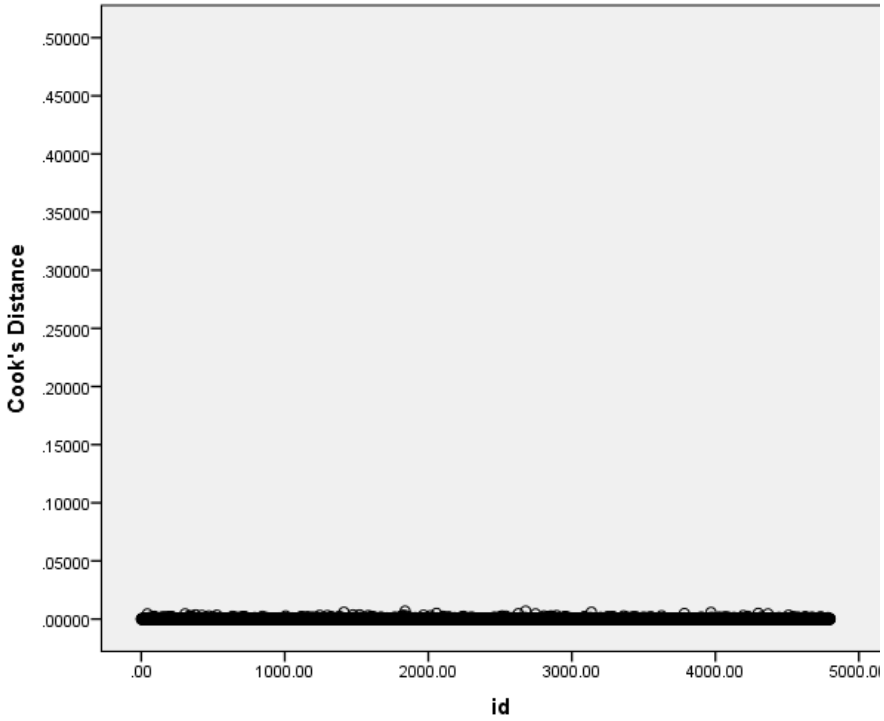
- 予測値に対して，残差をプロットしたもの
- 残差の等分散性，独立性を確認できる
  - 予測値の区間で残差が0を中心に均一に散らばっていればOK
  - 区間によってばらつきが違う場合は等分散性が成り立っていないことを疑う
  - 残差のばらつきかたに傾向がある場合は独立性が成り立っていないことを疑う

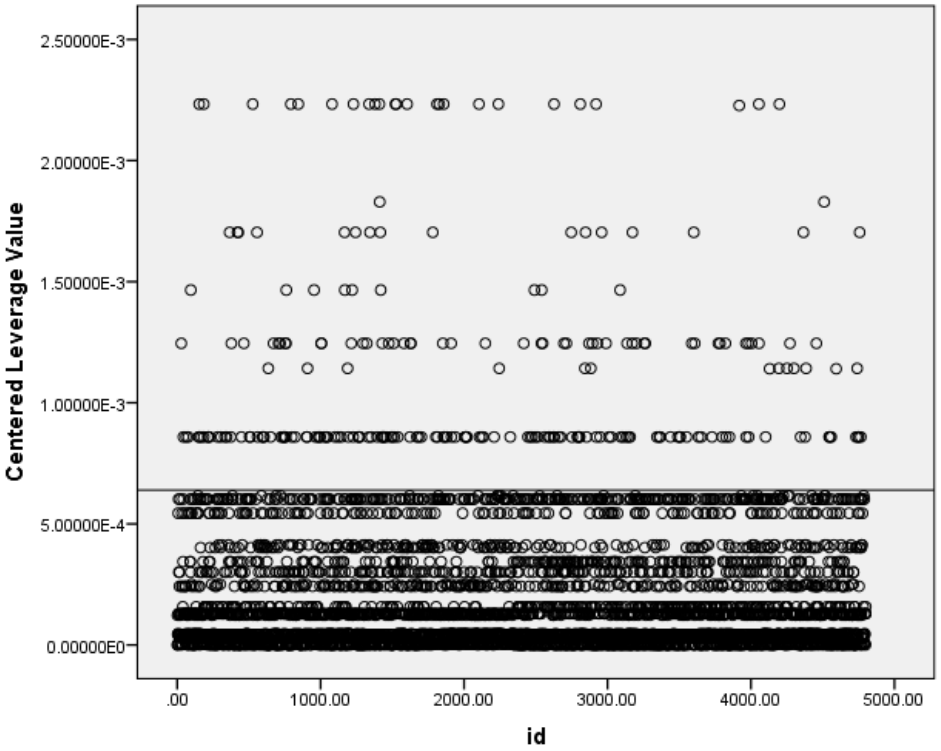


Cookの距離とてこ比のプロット

・Cookの距離
0.5を超えるものはないので，問題のある点はなさそう

・てこ比
平均の3倍を超える点が散見される．





残差の仮定を満たさない場合や影響点・外れ値への対処方法(1)

- 残差の正規性を満たさない場合
  - 正規分布に近づくように変数変換を行う
    - 対数変換，逆数変換など
    - 簡単に行えるが，回帰係数の解釈が難しくなる
  - 一般化線形モデルを使用する
    - ポアソン回帰，ガンマ回帰など
- 残差の等分散性を満たさない場合
  - 変数変換を行う
    - 正規性を満たさない場合と同様
  - 一般化線形モデル，一般化推定方程式，(一般化)線形混合モデルを使用する
    - 一般化推定方程式，(一般化)線形混合モデルは残差の分散共分散行列をモデル化することができるためより柔軟にモデル構築を行うことができる

残差の仮定を満たさない場合や影響点・外れ値への対処方法(2)

- 残差の独立性を満たさない場合
  - 一般化推定方程式，(一般化)線形混合モデルを使用する
    - これらのモデルは残差の分散共分散行列をモデル化することができるためより柔軟にモデル構築を行うことができる
- 外れ値・影響点がある場合
  - 外れ値や影響点を除外した結果と除外しない結果を比較する(感度分析)
  - 結果が概ね一致していれば同じ結論を導くことができるので，除外しない結果を元に議論すればよい
  - 結果が大きく異なる場合は，データ収集のプロセス，外れ値が理論的・臨床的にあり得る値か等を元に，どちらの結果のほうがもっともらしいか考察する

説明変数の選択方法

- 自分の仮説に従って選ぶのが原則
  - 先行研究のレビューが重要
  - 因果推論を行う際には非巡回有向グラフ(DAG; Directed Acyclic Graph)による視覚化と整理が変数選択に有効
  - 回帰係数が有意じゃない変数でも仮説モデルにあったら残す
    - 仮説が妥当なものであれば有意でない(関連がない)ことも重要な知見となる
- ステップワイズ法(変数増加法，変数減少法，変数増減法，変数減増法)は使わない
  - ステップワイズ法はP値を基準に変数を選択する方法
  - ステップワイズ法の欠点
    - 有意な変数が残って有意でない変数は残らない
      - 自分が注目していた変数や本当は関連がある変数が残らないこともある
      - 本当は関係ない変数なのに残ることも(αエラー)
    - オーバーフィッティングが起こる
      - 解析データへの当てはまりはいいが，別のデータには当てはまらないモデルになってしまう

Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, *5*(1), 32. https://doi.org/10.1186/s40537-018-0143-6

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour?: Stepwise modelling in ecology and behaviour. *Journal of Animal Ecology*, *75*(5), 1182–1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x

サンプルサイズを決める

- 推定の安定性ベースの決め方
  - EPV(Event per variable)を元に決める
  - EPV: 説明変数1つあたりのイベント数
  - EPVが10以上になるようにデータを集める
    例: 説明変数を10個使いたければ10×10=100ケースのイベントが必要となる．重回帰分析の場合全てのケースがイベントと考えればよいので，この場合は10×10＝100ケース集めればよいことになる．
- 検出力ベースの決め方
  - 有意にしたい回帰係数の大きさなどから決める
  - パワーアナリシス用のソフトで計算できる
    - G*Power: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
- 両方共計算して多い方を採用すればOK

参考文献

- 中山和弘．看護学のための多変量解析入門．医学書院, 2018.
- 酒井麻衣子．SPSS完全活用法ーデータの入力と加工(第4版)，東京図書，2016.
- 三輪哲・林雄亮(編著)．SPSSによる応用多変量解析，オーム社, 2014.
- 柳井晴夫・緒方裕光(編著)．SPSSによる統計データ解析―医学・看護学、生物学、心理学の例題による統計学入門．現代数学社，2006.
- 永田靖，吉田道弘．統計的多重比較法の基礎．サイエンティスト社，1997.
- 永田靖．多重比較法の実際．応用統計学，27(2), pp.93-108, 1998.

参考文献

- Smith, G. (2018). Step away from stepwise. Journal of Big Data, 5(1), 32. https://doi.org/10.1186/s40537-018-0143-6
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour?: Stepwise modelling in ecology and behaviour. Journal of Animal Ecology, 75(5), 1182–1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x