医学統計セミナー2019
量的データの解析(3)
相関分析と回帰分析

臨床研究センター
谷岡健資

2019年度 医学統計セミナー

開催日	講義内容
5月21日（火）	記述統計・グラフ表示・統計ソフトウェア
6月18日（火）	単群・2群比較のための諸種の検定
7月25日（木）	分散分析と多重比較
8月22日（木）	**相関分析と回帰分析**
9月26日（木）	クロス集計表の解釈と諸種の検定
10月17日（木）	ロジスティック回帰分析
11月14日（木）	生存曲線の推定と比較
12月　3日（火）	Cox比例ハザードモデル
1月23日（木）	感度・特異度・ROC曲線とその比較
2月27日（木）	繰り返し測定の分散分析、混合効果モデル

目次

1.相関係数について

□相関係数とは

ある画像データから計算した特徴量とある検査値の間に関係があるか否かを調べたいとします

被験者識別番号	画像データの値	ある検査値
1	21.4	6.5
2	18.3	4.5
3	12.9	2.8
4	25.4	7.1
5	23.1	5.9
6	15.6	3.4

＊扱うデータは同一被験者に対して2時点の値が観測されているような状況

相関係数は変量間の関係を調べるための指標です

1.相関係数について

□相関係数とは

ある画像データから計算した特徴量とある検査値の間に**関係**があるか否かを調べたいとします

被験者識別番号	画像データの値	ある検査値
1	21.4	6.5
2	18.3	4.5
3	12.9	2.8
4	25.4	7.1
5	23.1	5.9
6	15.6	3.4

質問

画像データから計算した特徴量とある検査値の間の関係とありますがどのような関係を知りたいのでしょうか？

1.相関係数について

□相関係数とは

ある画像データから計算した特徴量とある検査値の間に**関係**があるか否かを調べたいとします

質問
画像データから計算した特徴量とある検査値の間の関係とありますがどのような関係を知りたいのでしょうか？

⇒次のうち，いずれか一方が成り立っていると考えられるとき，
2変量（画像データの値と検査値の値）は相関関係があると言います

関係1:
画像データの値が高い場合，検査値の値も高くなる傾向にある
＝検査値の値が高い場合，画像データの値も高くなる傾向にある

関係2:
画像データの値が高い場合，検査値の値は低くなる傾向にある
＝検査値の値が高い場合，画像データの値も低くなる傾向にある

1.相関係数について

□相関係数とは

相関関係の種類について

関係1が成り立つとき，**正の相関がある**という

関係1
　画像データの値が高い場合，検査値の値も高くなる傾向にある
＝検査値の値が高い場合，画像データの値も高くなる傾向にある

関係2が成り立つとき，**負の相関がある**という

関係2:
　画像データの値が高い場合，検査値の値は低くなる傾向にある
＝検査値の値が高い場合，画像データの値も低くなる傾向にある

1.相関係数について

□標本の相関係数

相関係数（第1回の復習）

2変量間の関係を「-1から1」の範囲で測定する指標

2変量間に線形関係が仮定されたもとで，
相関係数の値が1に近い場合は正の相関関係があるといい，
相関係数の値が-1に近い場合は負の相関関係があるという．

正の相関関係
　画像データの値が高い場合，検査値の値も高くなる傾向にある
＝検査値の値が高い場合，画像データの値も高くなる傾向にある

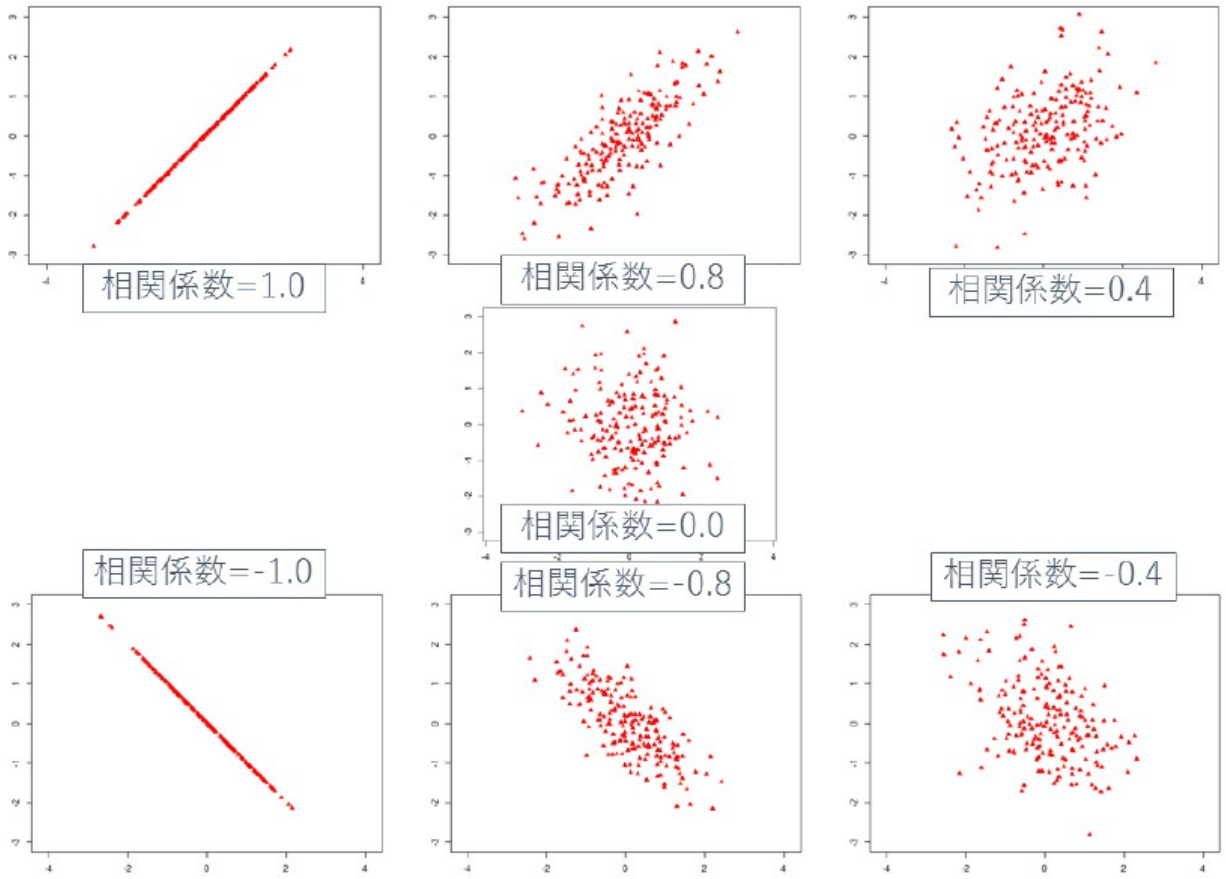
負の相関関係
　画像データの値が高い場合，検査値の値は低くなる傾向にある
＝検査値の値が高い場合，画像データの値も低くなる傾向にある

1.相関係数について

□標本の相関係数

$$-1 \leq \text{相関係数} \leq 1$$

－１:負の相関関係　　０:相関がない　　１:正の相関関係



1.相関係数について

□標本の相関係数

第1回目の質問（再掲）

下記の文章は正しいか？

「相関係数の値が1に近ければ，正の相関関係がある」

1.相関係数について

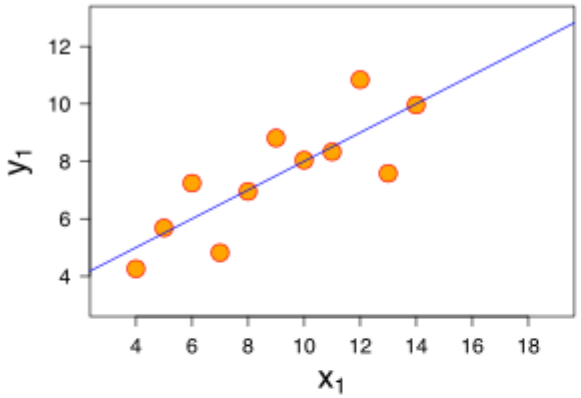
□標本の相関係数

「相関係数の値が1に近ければ，正の相関関係がある」⇒×

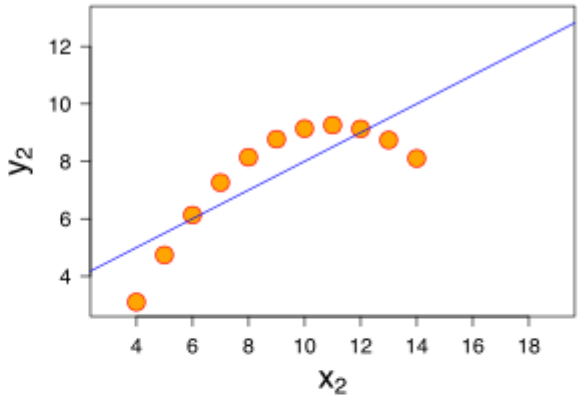
アンスコムの例

4つの図では全て横軸の平均9(分散11)：縦軸の平均7.5(分散4.12)：相関係数0.81

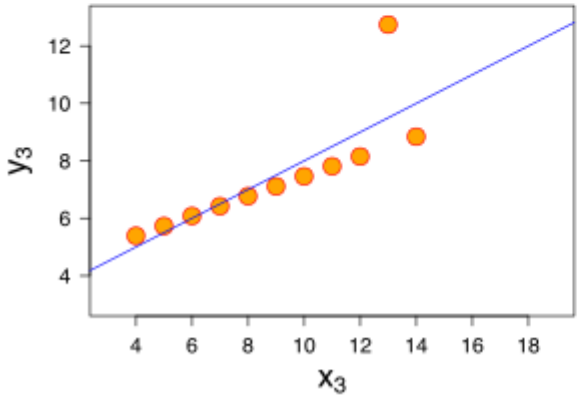
線形関係



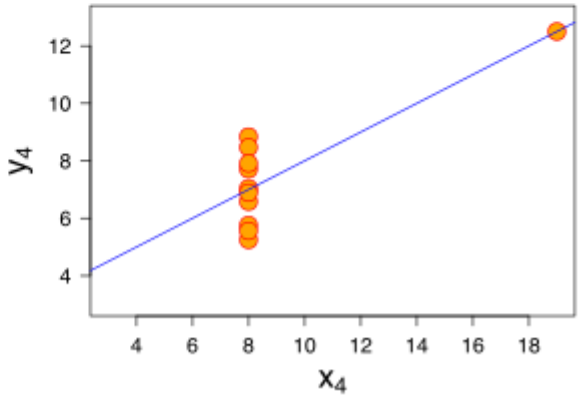
非線形関係



線形関係＋外れ値



線形関係でない＋外れ値



Wikipediaより転載

1.相関係数について

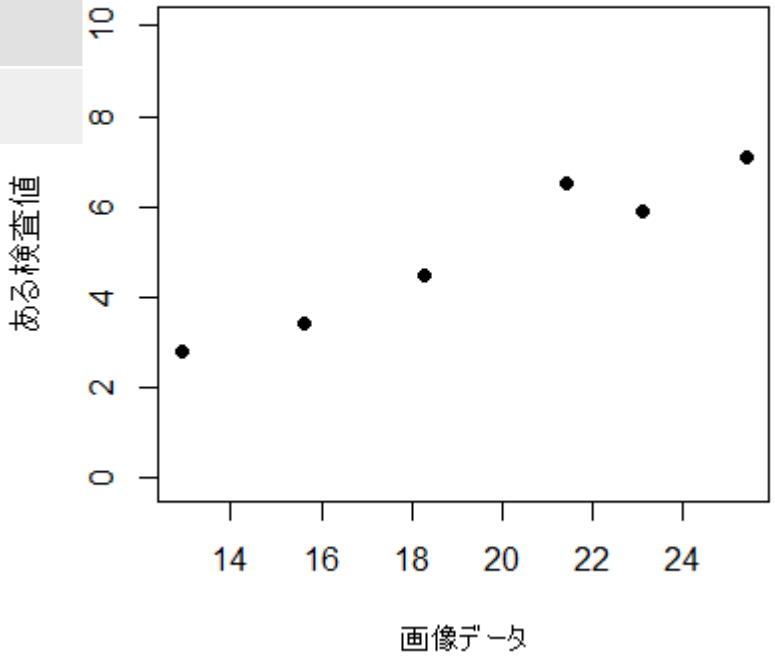
◘母集団での相関分析

ある画像データの値とある検査値に相関関係はあるか？

被験者識別番号	画像データの値	ある検査値
1	21.4	6.5
2	18.3	4.5
3	12.9	2.8
4	25.4	7.1
5	23.1	5.9
6	15.6	3.4

相関係数：0.97

散布図



ある画像データの値とある検査値間の
散布図より線形関係はありそう

相関係数も0.97で非常に1に近いことから
2変量間に正の相関があるといってもいいのでは？

1.相関係数について

□母集団での相関分析

下記の結果が主張できたこと

6名の被験者の「画像データの値」と「ある検査値」の正の線型関係が仮定でき，
相関係数が0.97である

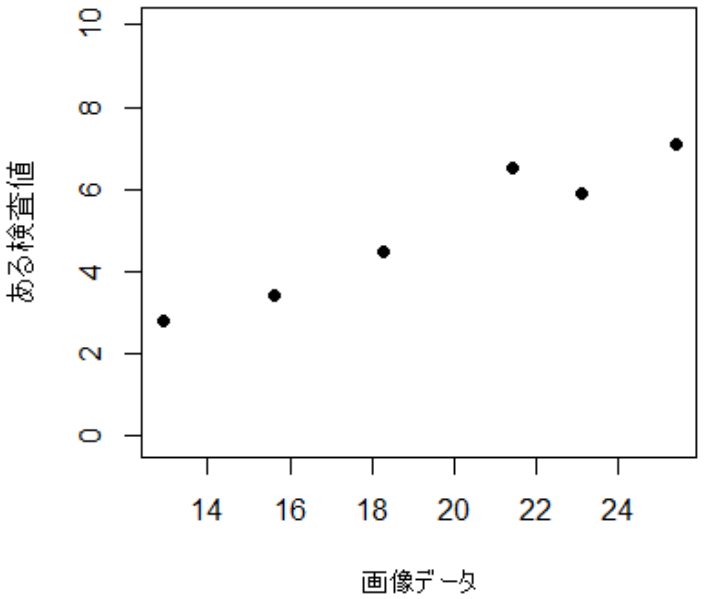
そもそも，
どのような目的のためにこのようなデータを
とるか考えなければならない

例）リサーチクエッションの例

ある検査値を測定するためには侵襲性が高い
または費用が非常にかかるため，
画像データから計算できる値で代替できないか

相関係数：0.97

散布図



1.相関係数について

□母集団での相関分析

下記の結果が主張できたこと

6名の被験者の「画像データの値」と「ある検査値」の正の線型関係が仮定でき，
相関係数が0.97である

主張したいこと

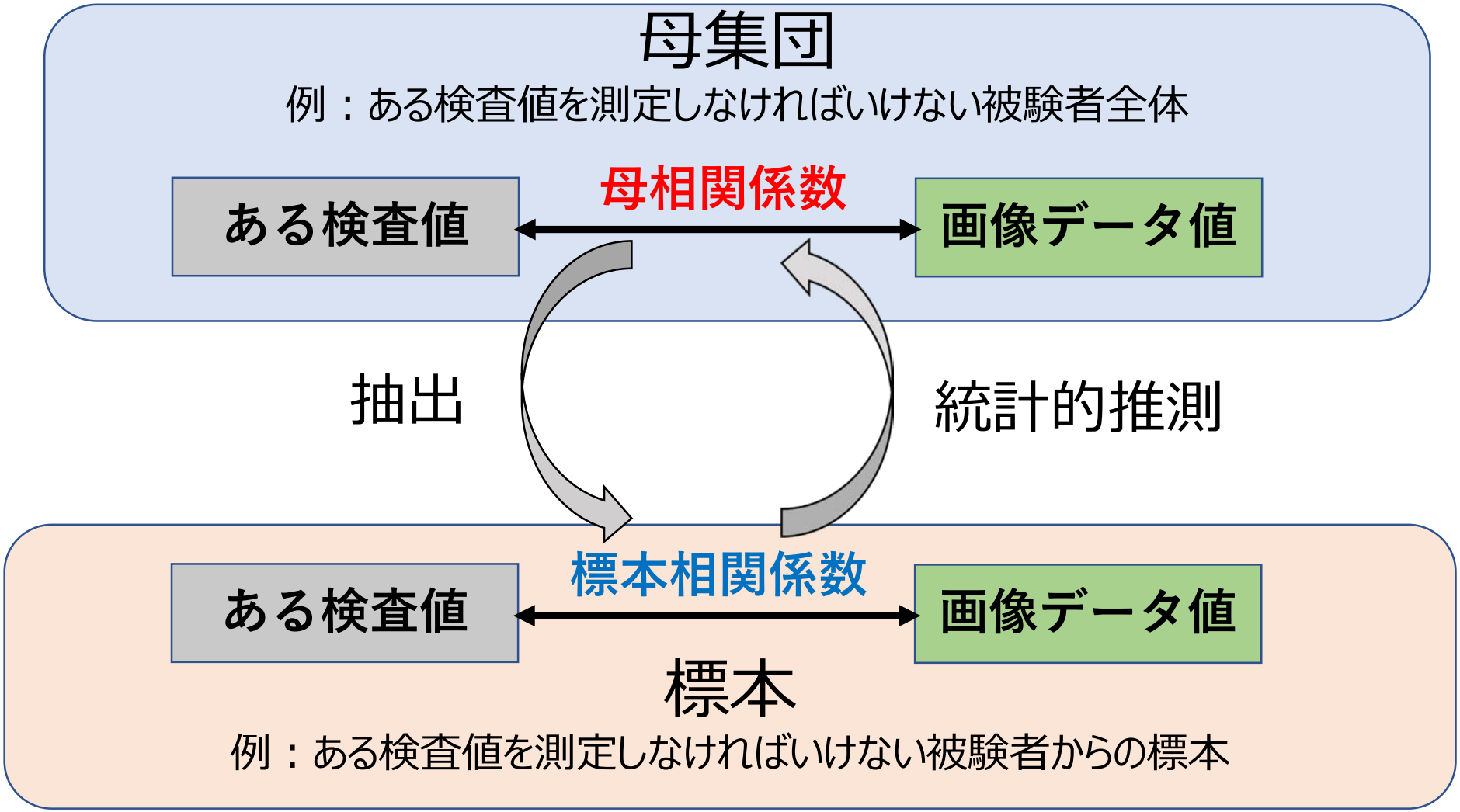
ある検査値をうける必要がある被験者全体（母集団）に対して
「ある検査値」と「画像データの値」に相関がある

すなわち，
「標本で計算される相関係数」から「母集団における相関係数」がどうであるかを
推測する必要がある！

1.相関係数について

◘母集団での相関分析

母集団の相関係数について推測する場合標本の相関係数から推測を実施する



1.相関係数について

□母集団での相関分析

母集団の相関係数（母相関係数）について推測する場合，
標本の相関係数から推測を実施する

平均値の検定と同様，下記を実行することができる.

[1] 母相関係数に関する検定

母相関係数がどのような値か否かを検討するための検定

[2] 母相関係数の95%信頼区間

母相関係数が95%の確率で属する区間を推測する方法

1.相関係数について

□母集団での相関分析

[1] 母相関係数に関する検定

母相関係数がどのような値か否かを検討するための検定

母相関係数の検定の目的

母集団での変量間の相関係数である母相関係数が0でないことを立証すること.

＝母集団の変量間に**相関関係が存在する**ということを立証すること

帰無仮説：「ある検査値」と「画像データの値」の母相関係数 ＝ 0

対立仮説：「ある検査値」と「画像データの値」の母相関係数 ≠ 0

1.相関係数について

◘母集団での相関分析

例題（丹後, 1993）
ある健診センターの男性受診者53名のデータを用いて，
Hb, Na, UA, Alb, T-Cholの5項目間の相関係数を計算した
どの項目間で母集団において関係があるか

	Hb	Na	UA	Alb
Na	0.015 (43)			
UA	**-0.533** (35)	0.227 (36)		
Alb	**0.360** (37)	0.086 (38)	**-0.459** (30)	
T-Chol	0.009 (43)	0.160 (44)	-0.245 (36)	0.317 (38)

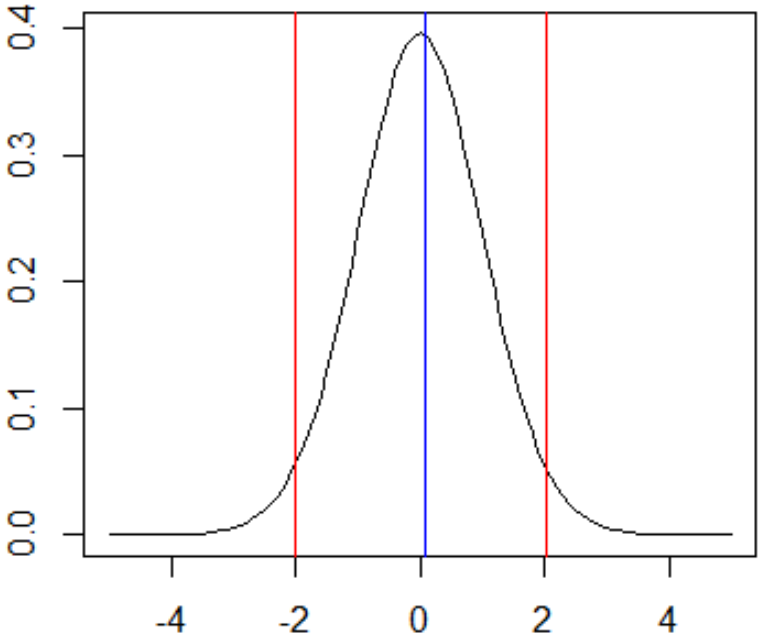
1.相関係数について

◘母集団での相関分析

例題（丹後, 1993）
ある健診センターの男性受診者53名のデータを用いて，
Hb, Na, UA, Alb, T-Cholの5項目間の相関係数を計算した
どの項目間で母集団において関係があるか

	Hb	Na	UA	Alb
Na	0.015 (43)			
UA	**-0.533** (35)	0.227 (36)		
Alb	**0.360** (37)	0.086 (38)	**-0.459** (30)	
T-Chol	0.009 (43)	0.160 (44)	-0.245 (36)	0.317 (38)

赤：棄却限界域　青：検定統計量実現値



帰無仮説の下での母相関係数の確率分布

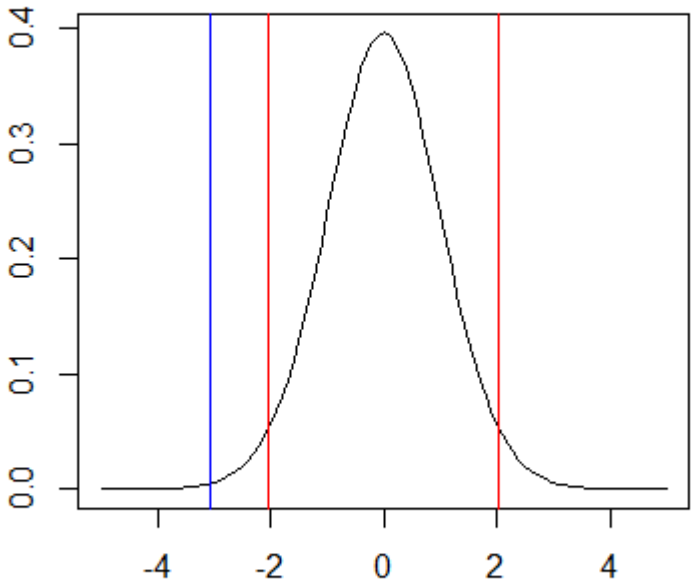
1.相関係数について

◘母集団での相関分析

例題（丹後, 1993）
ある健診センターの男性受診者53名のデータを用いて，
Hb, Na, UA, Alb, T-Cholの5項目間の相関係数を計算した
どの項目間で母集団において関係があるか

赤：棄却限界域　青：検定統計量実現値

	Hb	Na	UA	Alb
Na	0.015 (43)			
UA	**-0.533** (35)	0.227 (36)		
Alb	**0.360** (37)	0.086 (38)	**-0.459** (30)	
T-Chol	0.009 (43)	0.160 (44)	-0.245 (36)	0.317 (38)



帰無仮説の下での母相関係数の確率分布

1.相関係数について

□母集団での相関分析

[2] 母相関係数の95%信頼区間

母相関係数が95%の確率で属する区間を推測する方法

Case1: 95%信頼区間が0と重複していない場合

⇒ 母相関係数が0となる可能性は低いことから関係性はあると考える.

Case2: 95%信頼区間が0と重複している場合

⇒ 母相関係数が0となる可能性を否定できないことから関係性があると強く主張できない

1.相関係数について

◘母集団での相関分析

例題（丹後, 1993）
ある健診センターの男性受診者53名のデータを用いて，
Hb, Na, UA, Alb, T-Cholの5項目間の相関係数を計算した
どの項目間で母集団において関係があるか

Case 1

	Hb	Na	UA	Alb
Na	0.015 (43)			
UA	**-0.533** (35)	0.227 (36)		
Alb	**0.360** (37)	0.086 (38)	**-0.459** (30)	
T-Chol	0.009 (43)	0.160 (44)	-0.245 (36)	0.317 (38)

HbとUAの母相関係数の95%信頼区間

1.相関係数について

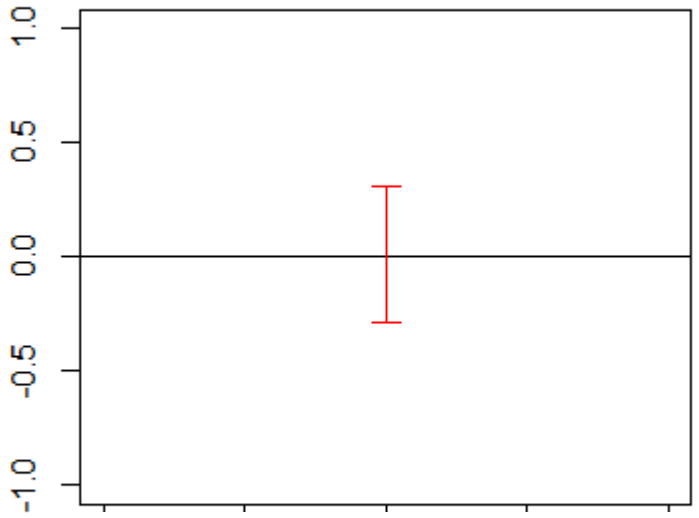
◘母集団での相関分析

例題（丹後, 1993）
ある健診センターの男性受診者53名のデータを用いて，
Hb, Na, UA, Alb, T-Cholの5項目間の相関係数を計算した
どの項目間で母集団において関係があるか

Case 2

	Hb	Na	UA	Alb
Na	0.015 (43)			
UA	**-0.533** (35)	0.227 (36)		
Alb	**0.360** (37)	0.086 (38)	**-0.459** (30)	
T-Chol	0.009 (43)	0.160 (44)	-0.245 (36)	0.317 (38)

HbとNaの母相関係数の95%信頼区間



2.回帰分析について

□単回帰分析の導入

「目的変数」と「説明変数」が与えられた際に，説明変数から目的変数の値を予測するための分析方法

例題（丹後, 1993）
17名のカドミウム作業者の血中B_2-MGと血中カドミウムCd値の関係データ

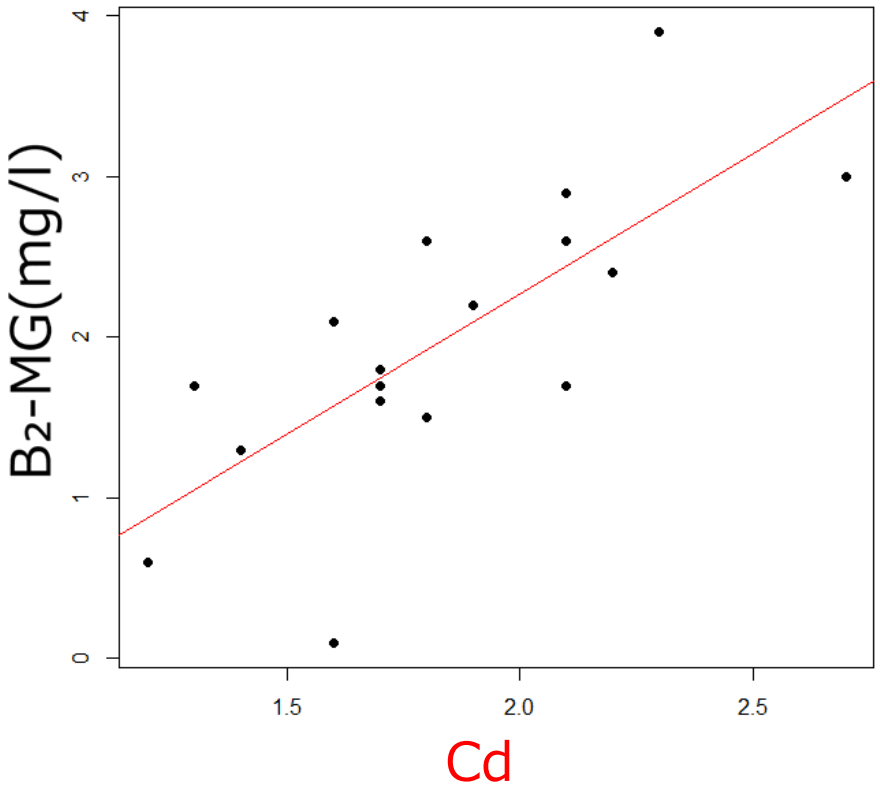
	説明変数	目的変数
対象	血中カドミウムCd	B_2-MG(mg/l)
1	1.3	1.7
2	2.1	2.6
3	2.7	3.0
4	1.6	0.1
…	…	…
17	1.2	0.6

2.回帰分析について

□単回帰分析の導入

例題（丹後, 1993）
17名のカドミウム作業者の血中B_2-MGと血中カドミウムCd値の関係データ

右図のように，「Cd」と「B_2-MG」の関係が線形関係であることから，直線を用いて
「Cd」の値から「B_2-MG」の値を予測できそうということが分かる



2.回帰分析について

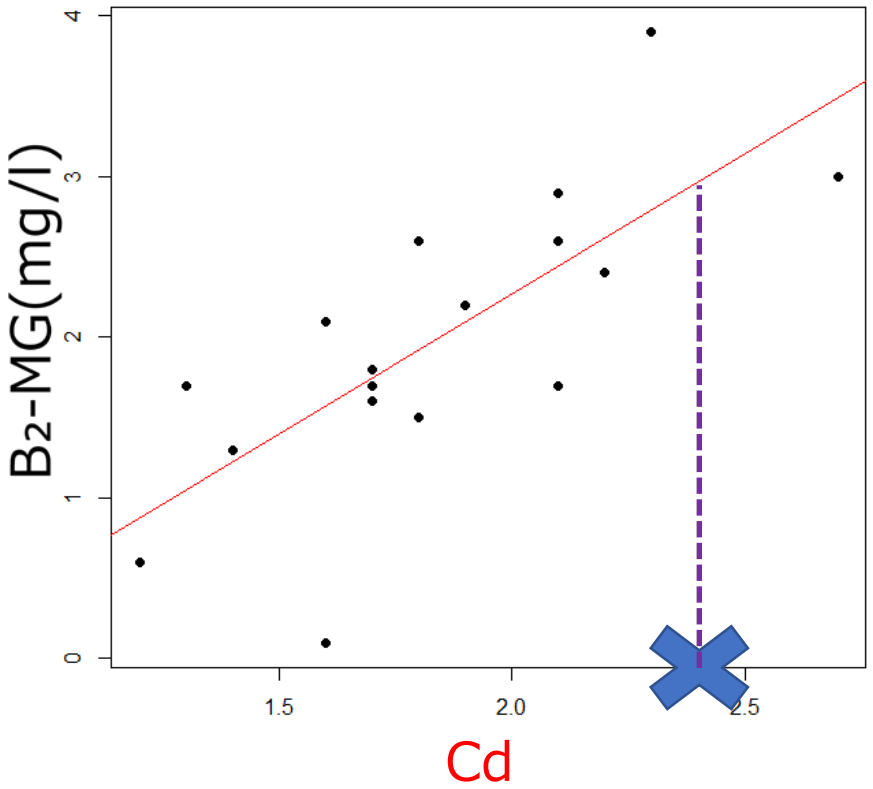
□単回帰分析の導入

例題（丹後, 1993）
17名のカドミウム作業者の血中B_2-MGと血中カドミウムCd値の関係データ

右図のように，「Cd」と「B_2-MG」の関係が線形関係であることから，直線を用いて
「Cd」の値から「B_2-MG」の値を予測できそうということが分かる

：新しい被験者のCdの値

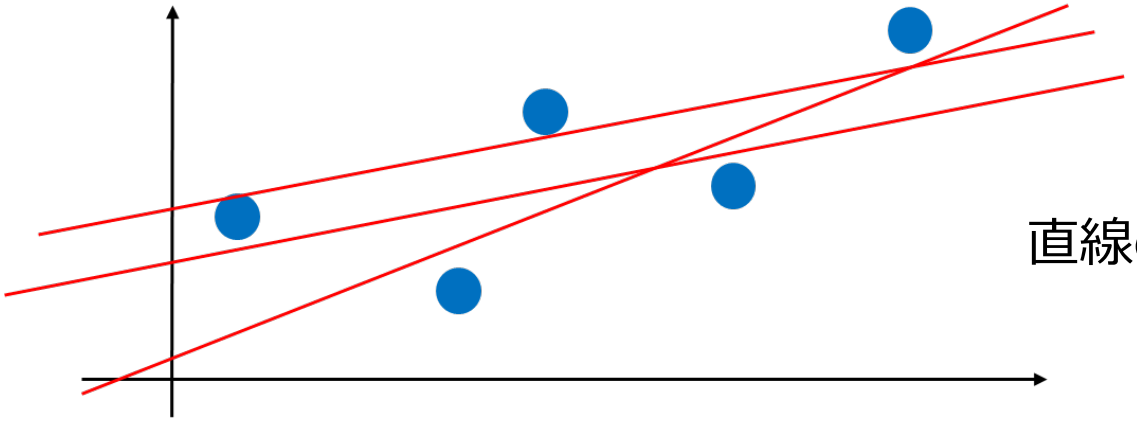
⇒この値から直線を用いて
B_2-MGを予測



2.回帰分析について

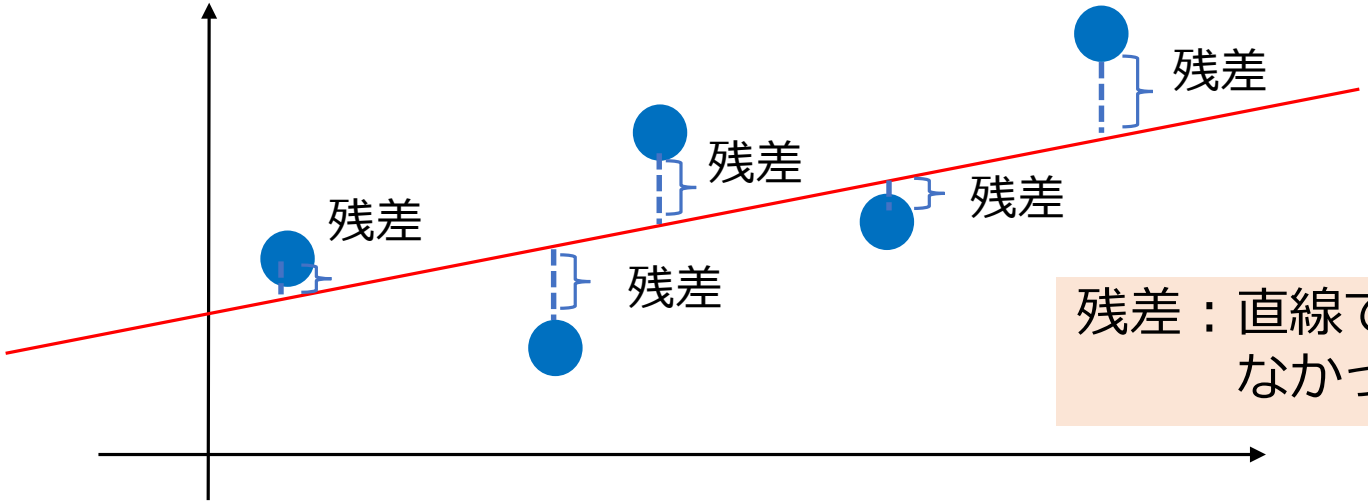
□単回帰分析の導入

どのようにデータから予測に適した直線を予測するのか？



直線の引き方はたくさんあるが…

⇒下記のようにデータと直線の遠さを表現し，残差が最小となるような直線を求める

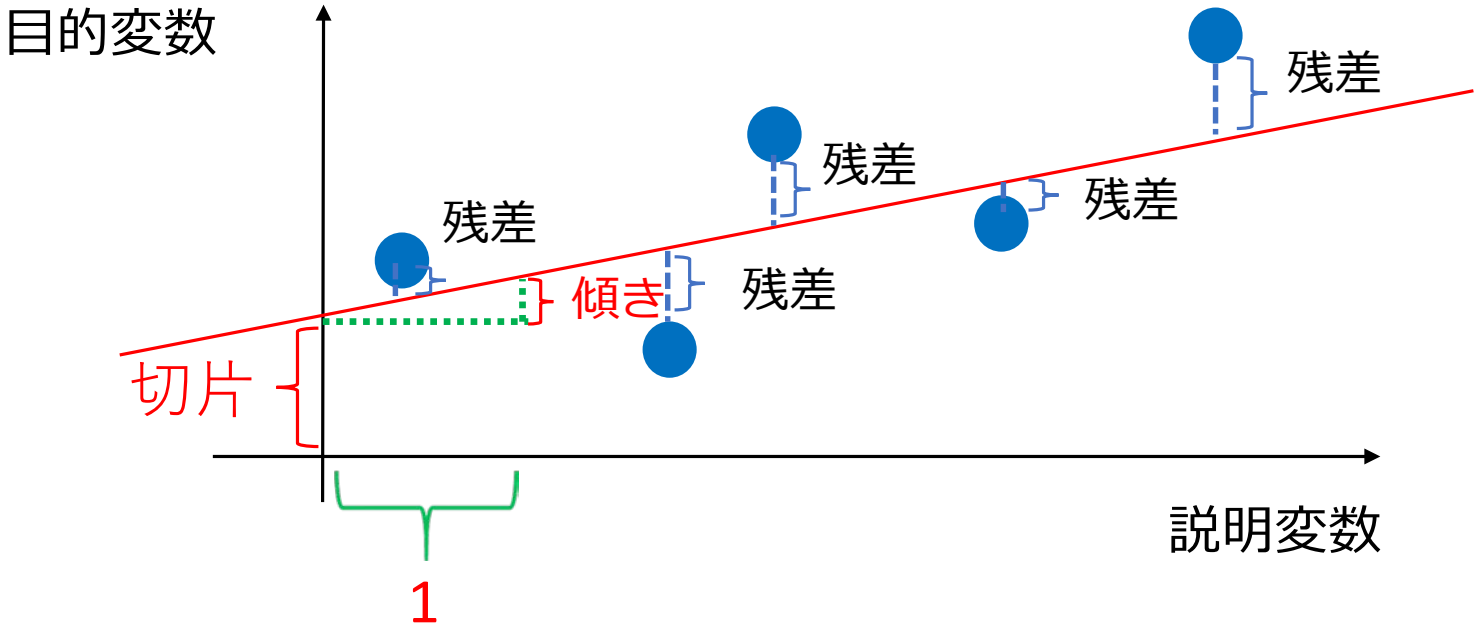


残差：直線でデータを説明できなかった部分

2.回帰分析について

□単回帰分析の導入

直線を求めるということは？



⇒切片と傾きを求めることに等しい

傾きは**回帰係数**とよばれ，説明変数が目的変数にどれほど影響を与えているのかを把握するために用いられる．

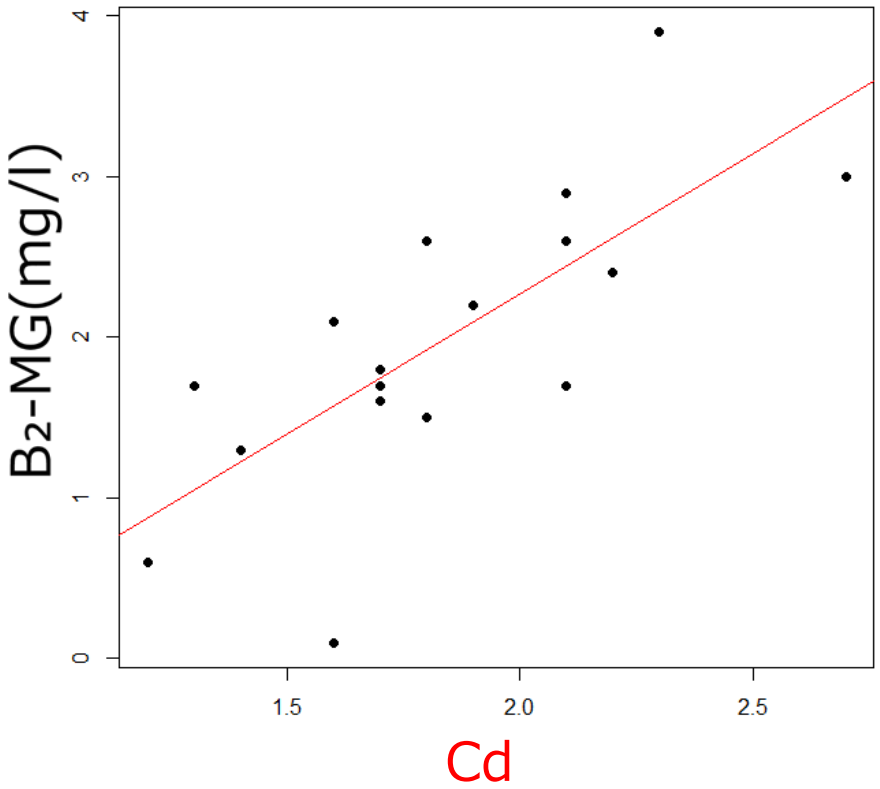
2.回帰分析について

□単回帰分析の導入

例題（丹後, 1993）
17名のカドミウム作業者の血中B_2-MGと血中カドミウムCd値の関係データ

切片	回帰係数
-1.23	1.75

この結果から云える回帰係数の解釈は？
⇒傾きの定義から考えてみてください



2.回帰分析について

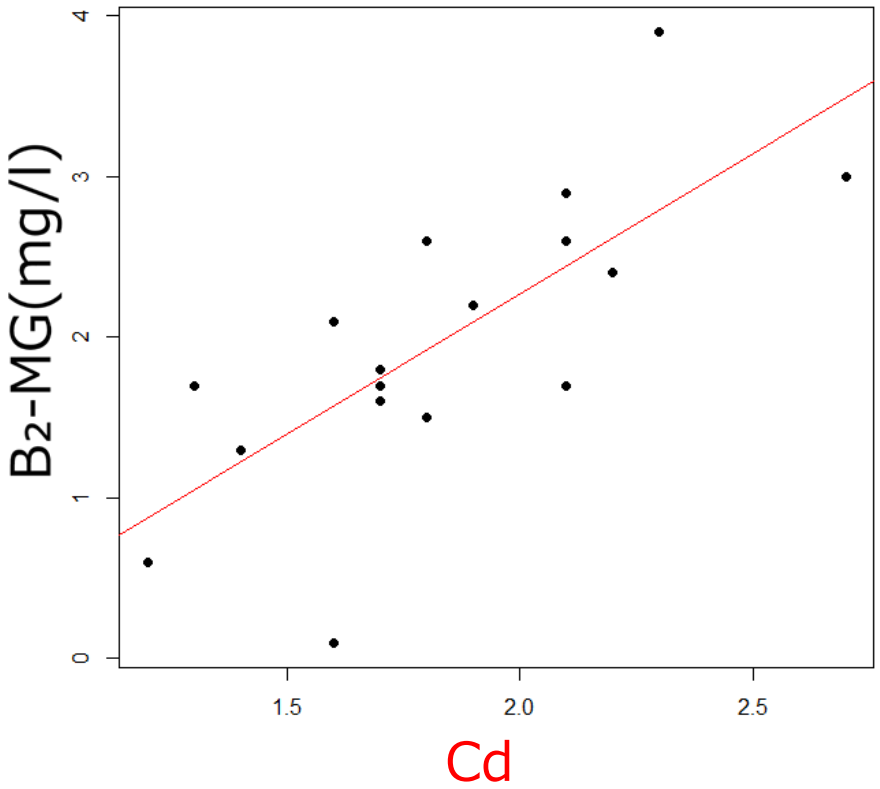
□単回帰分析の導入

例題（丹後, 1993）
17名のカドミウム作業者の血中B_2-MGと血中カドミウムCd値の関係データ

切片	回帰係数
-1.23	1.75

この結果から云える回帰係数の解釈は？

⇒　血中カドミウムCd値が1上昇すると，血中B_2-MGが1.75上昇する傾向にある.



2.回帰分析について

□回帰係数の検定

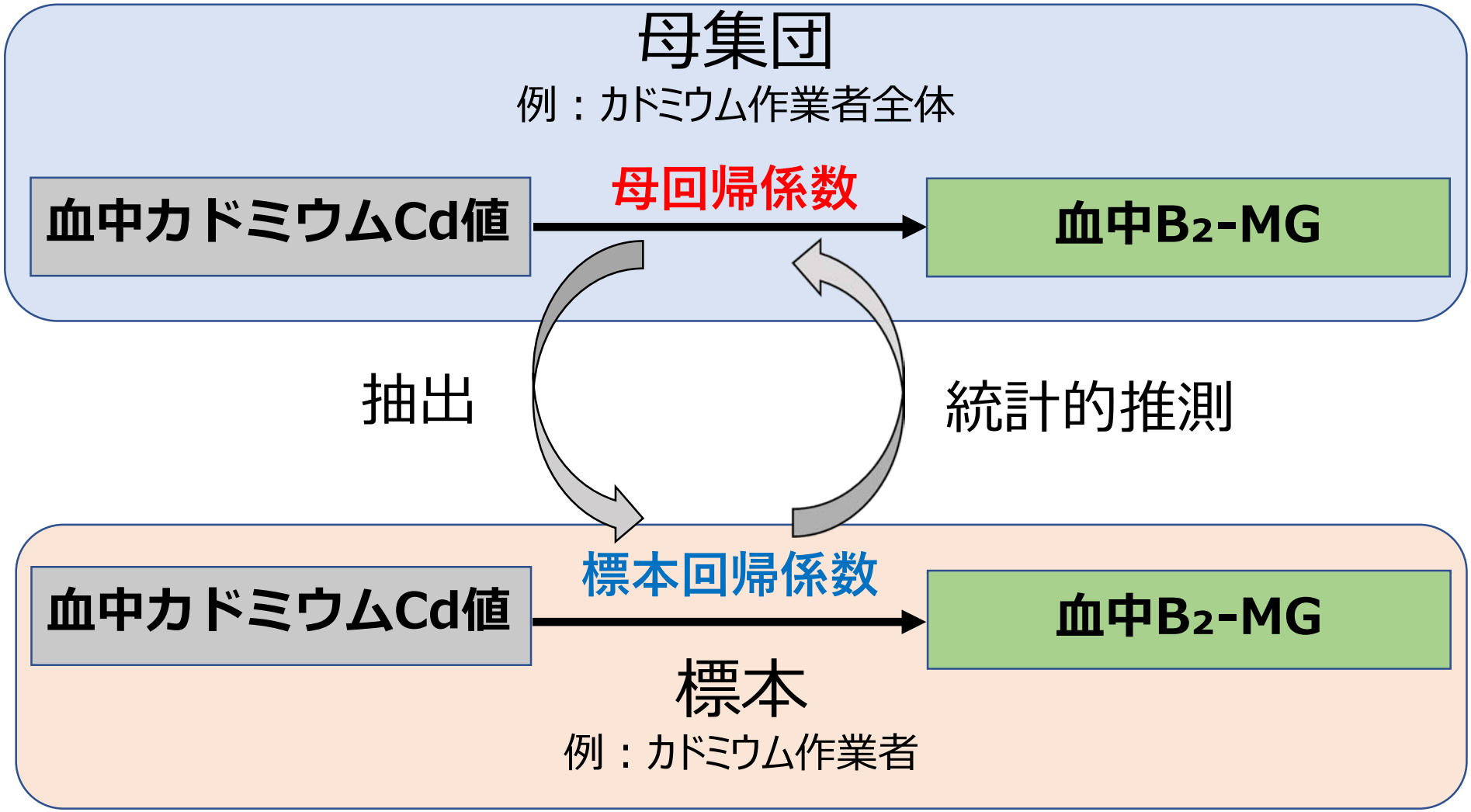
例題（丹後, 1993）
17名のカドミウム作業者の血中B_2-MGと血中カドミウムCd値の関係データ

回帰係数の結果から主張できることは
17名のカドミウム作業者において，血中カドミウムCd値が1上昇すると血中B_2-MGが1.75上昇する傾向にある

あくまでも17名のカドミウム作業者に対する結果であり，
多くの場合はカドミウム作業者全体に対して
血中カドミウムCd値が血中B_2-MGに影響を与えるか否かに興味がある場合がほとんどである.

2.回帰分析について

◘回帰係数の検定



2.回帰分析について

□回帰係数の検定

すなわち，標本の回帰係数から母集団の回帰係数の値を推測する

平均値の検定と同様，下記のように推定と検定を実施することができる

[1] 母回帰係数の検定
母集団で説明変数が目的変数に影響を与えているか否かを調べる方法

[2] 母回帰係数の区間推定
母集団で説明変数が目的変数に与えている影響を区間で推測する方法

2.回帰分析について

◘回帰係数の検定・95%信頼区間

[1] 母偏回帰係数に関する検定

母集団で説明変数が目的変数に影響を与えているか否かを調べるための検定

母偏回帰係数の検定の目的

母集団での説明変数の回帰係数が0でないということを立証すること

＝母集団で説明変数が**目的変数に影響を与えている**ということを立証すること

帰無仮説：「血中カドミウムCd値」の母回帰係数 ＝ 0

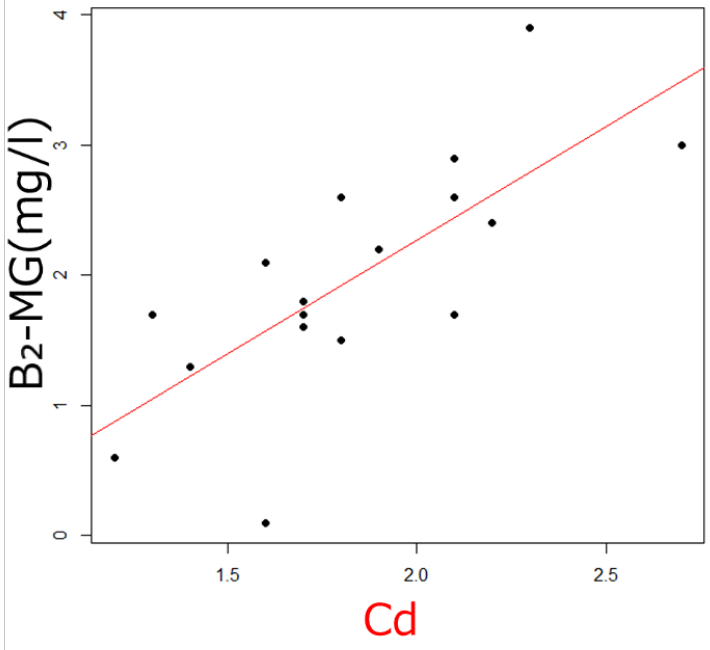
対立仮説：「血中カドミウムCd値」の母回帰係数≠ 0

2.回帰分析について

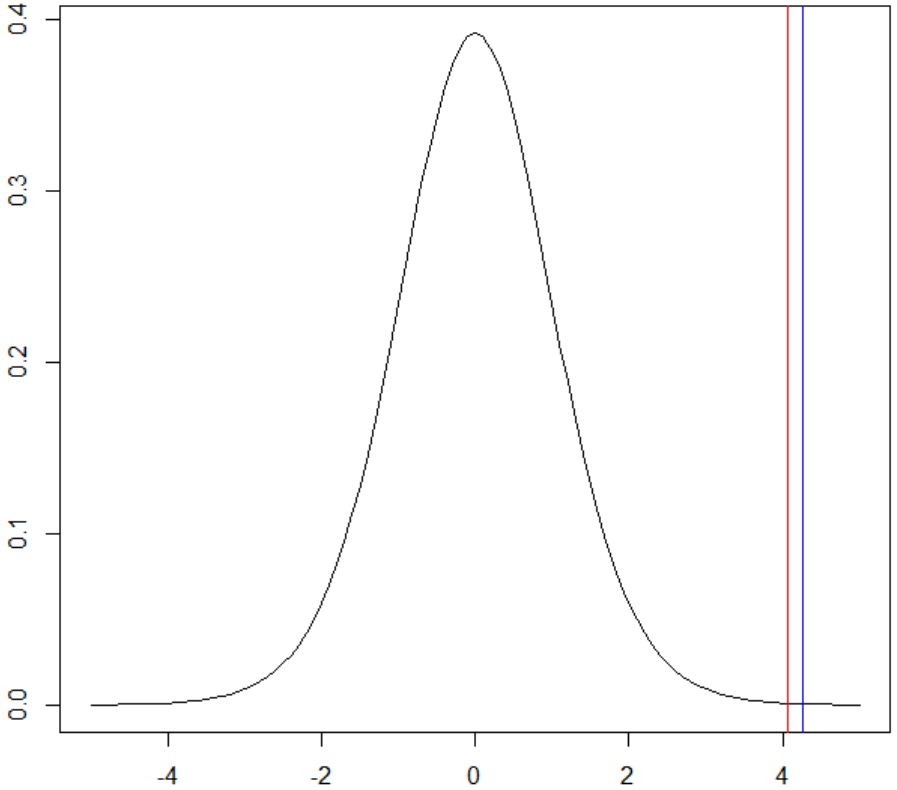
◘回帰係数の検定・95%信頼区間

丹後(1993)の例では下記より帰無仮説は棄却されることから
母集団でCdはB_2-MGに影響を与えていると考えられる

下記の標本の回帰係数から
母集団の回帰係数が0かどうかを
検定する



赤：棄却限界域　青：検定統計量実現値



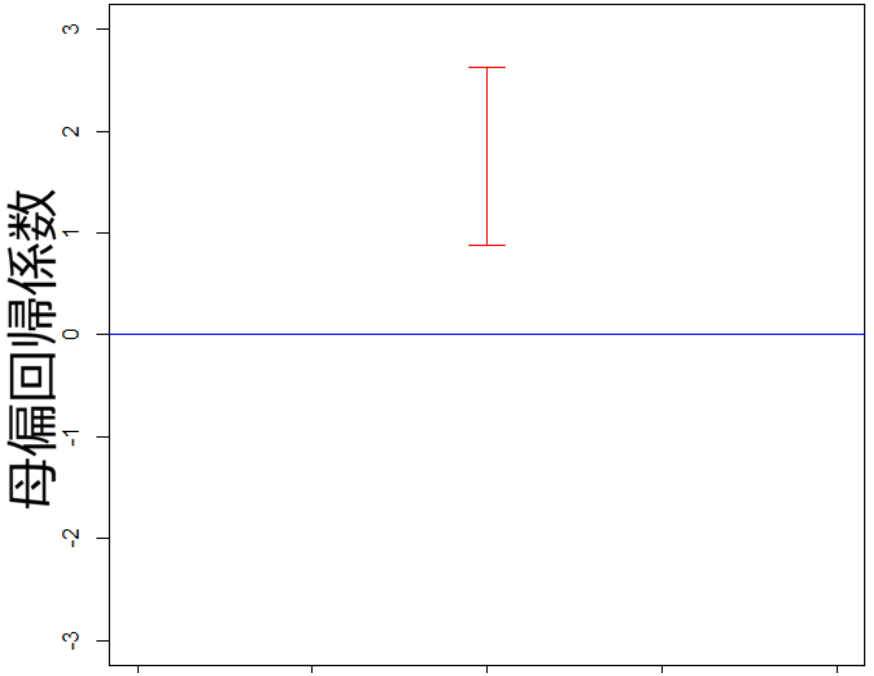
帰無仮説の下での母回帰係数

2.回帰分析について

□回帰係数の区間推定

[2] 母回帰係数の区間推定
母集団で説明変数が目的変数に与えている影響を区間で推測する方法

丹後(1993)のケースでは
母回帰係数の95%信頼区間が
0をまたいでいないことから,
血中カドミウムCd値が
血中B_2-MGに影響を与えている
と解釈することができる



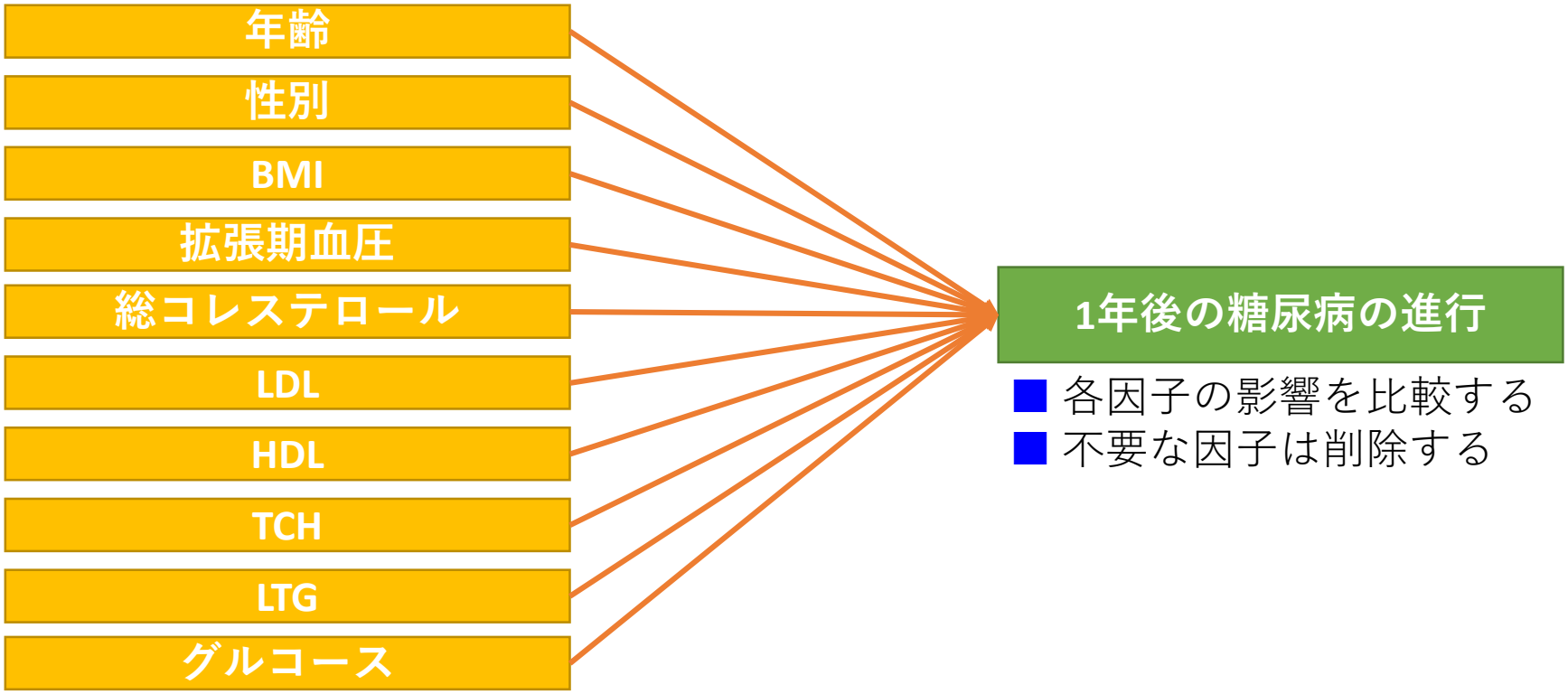
3.重回帰分析について

糖尿病患者442名に対して，10変数(ベースライン値)として，

年齢　　性別　　BMI 拡張期血圧　総コレステロール
LDL　HDL TCH　LTG　　グルコース

がとられている．応答Yは，ベースライン時点から1年後の糖尿病の進行程度を数値的に表したものである(Efron et al., 2004).

どの因子が，1年後の進行程度に影響を及ぼしていると考えてよいか？



- 各因子の影響を比較する
- 不要な因子は削除する

3.重回帰分析について

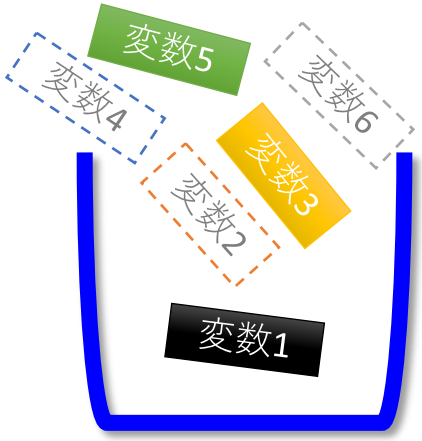
JMPの手順

STEP.1 「モデルのあてはめ」を選択し，説明変数，応答変数をそれぞれ「モデル効果の構成」「Y」に追加する．

STEP.2 「手法」を「ステップワイズ法」に変更(変数選択のため)

統計モデル

多変量解析では，たくさんの説明変数(共変量)を入れるほど情報がたくさんになるので，良い統計モデルになるということ？



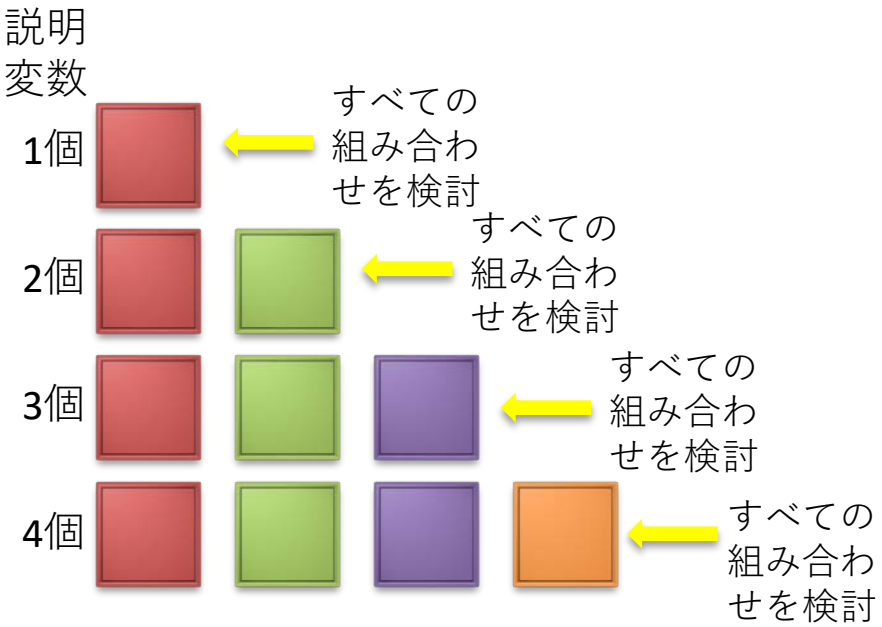
必ずしもそうではない．説明変数が多いほど不必要な変数は単なるノイズでしかないわけだし，また，多重共線性の問題などがある．そのため，必要なものだけで統計モデルをつくることが推奨されることが多い．そのための方法が変数選択である．

3.重回帰分析について

STEP.1 「停止ルール」で変数選択に用いる測度を設定し，「方向」で変数選択のアルゴリズムを選択する．
そして，「実行」ボタンを押す

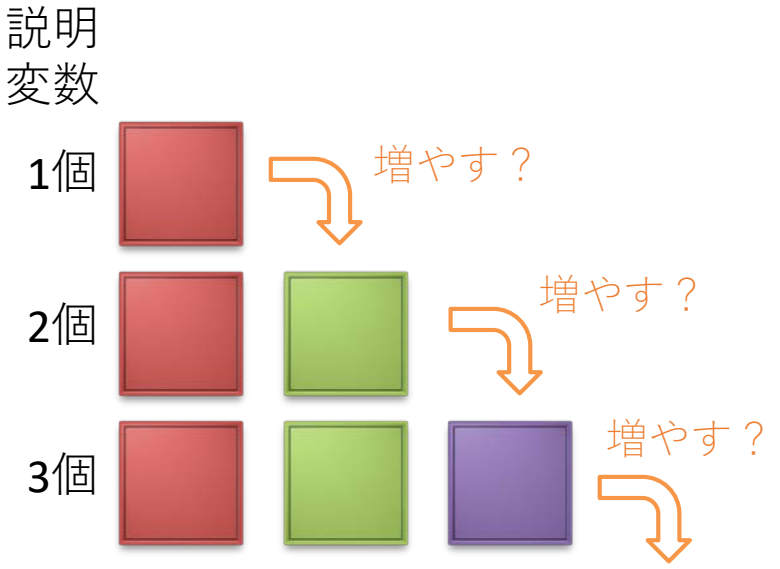
注意：変数減少法の場合は，「停止ルール」横の「すべて選択」を押してから実行ボタンを押す．

総当り法



説明変数が1個,2個,...のそれぞれのパターンを計算しその中から最適なものを選ぶ

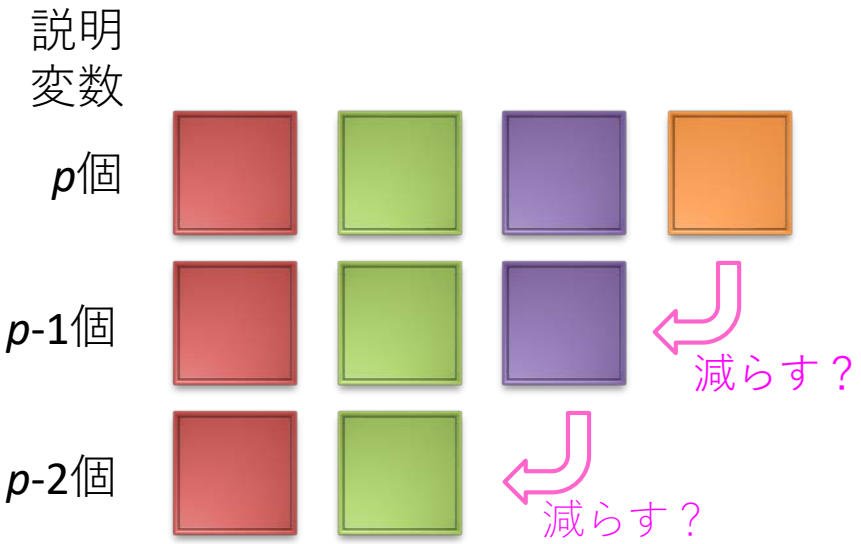
変数増加法
(前進ステップワイズ法)



説明変数が1個の場合からスタートして，変数を追加したほうが良ければ増やし，そうでなければ変数の追加をしない．

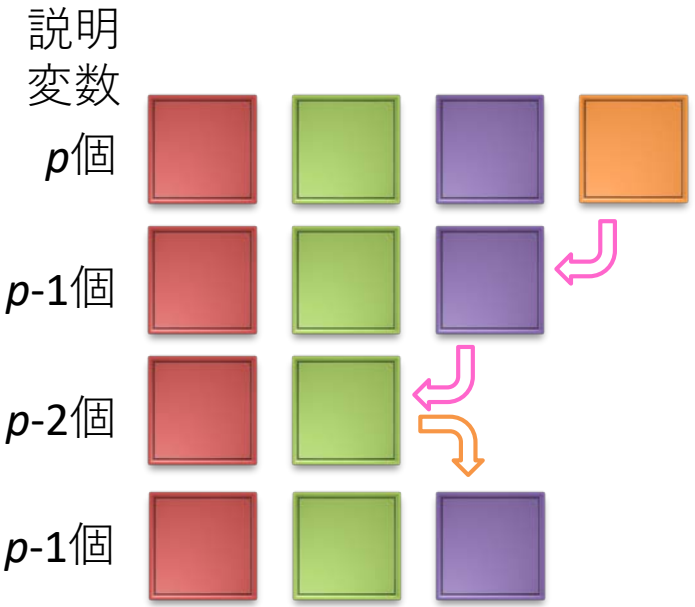
3.重回帰分析について

変数減少法
(後退ステップワイズ法)



全ての説明変数からスタートして，変数を減らしても影響がなければ減らし，そうでなければ変数の削除をしない．

変数増減法
(ステップワイズ法)



変数減少法からスタートするが，変数増化法では変数が削除された場合にその変数を考慮しなかったが，変数増減法では変数の削除と削除した変数の追加の両方を検討しながら各ステップを進める．

3.重回帰分析について

情報量規準に基づく方法
　モデルの当てはまりの良さを情報量規準(BIC, AIC等)を用いて評価する

検定に基づく方法
　変数を追加(or 削除)した場合とそうでない場合のモデルを検定を用いて比較する.

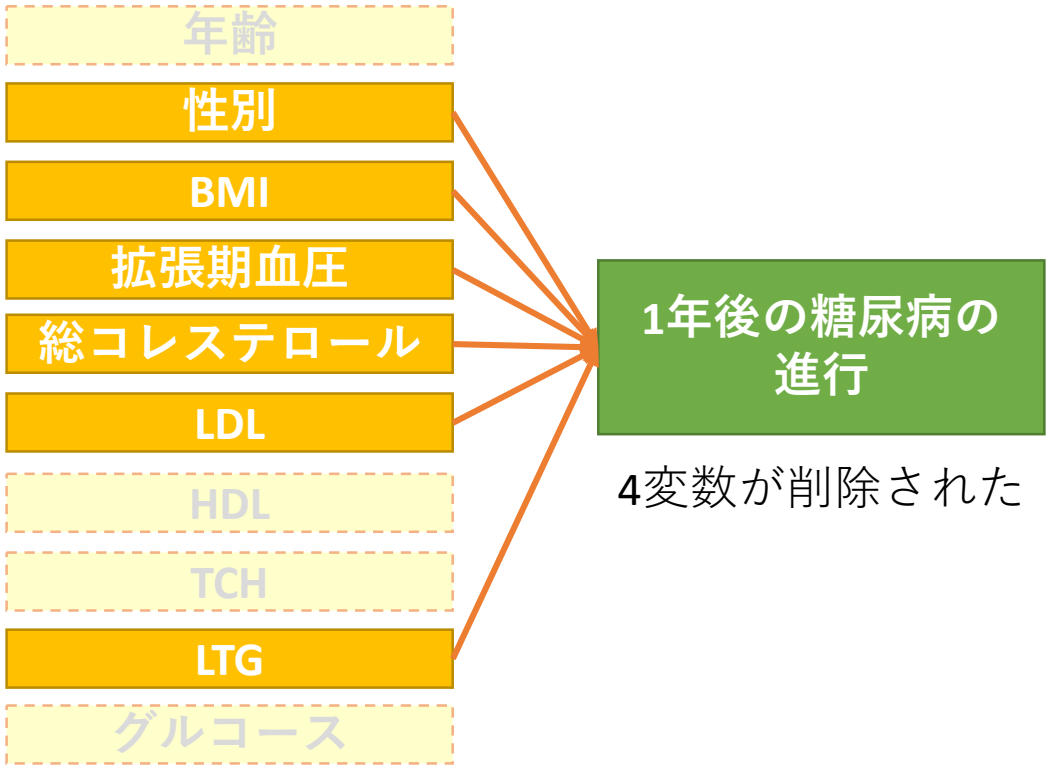
交差検証(確認)法(Cross-validation)に基づく方法
　モデルの予測度の高さをコンピュータを使って計算して評価する.

JMPのOutput

事例では(変数減少法, BIC)

現在の推定値

ロック	追加	パラメータ	推定値
☑	☑	切片	-346.15314
☐	☐	年齢	0
☐	☑	性別{1-2}	10.7955055
☐	☑	BMI	5.71110674
☐	☑	血圧	1.12655255
☐	☑	総コレステロール	-1.0428764
☐	☑	LDL	0.84327695
☐	☐	HDL	0
☐	☐	TCH	0
☐	☑	LTG	73.3065264
☐	☐	グルコース	0



4変数が削除された

3.重回帰分析について

STEP.1 「モデルの作成」→「OK」ボタンを押す

重回帰分析では,

(応答のバラツキ) ＝ (予測値のバラツキ)＋(誤差のバラツキ)

モデルの当てはまりの良さ　　モデルの当てはまりの悪さ

に分けることができる．これを**変動分解**という．

寄与率(R2値)とは,

(予測値のバラツキ)／(応答のバラツキ)

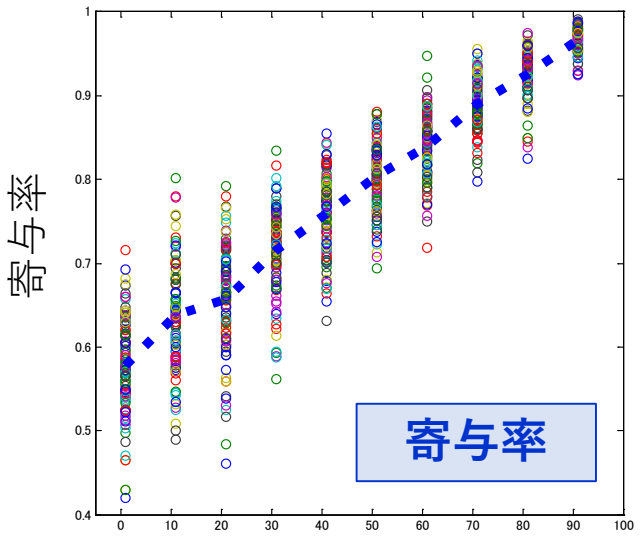
で，モデルがどれぐらいの
割合で，応答のバラツキを表しているか(説明できているか)を表す統計量である．

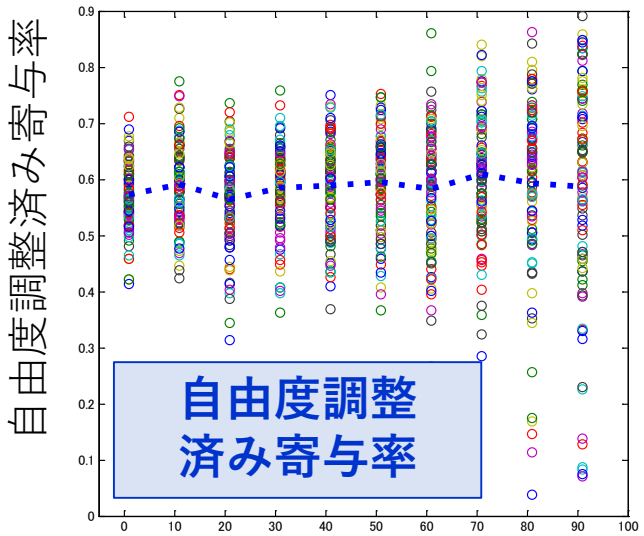
JMPのOutput

あてはめの要約	
R2乗	0.514884
自由度調整R2乗	0.508193
誤差の標準偏差(RMSE)	54.06454
Yの平均	152.1335
オブザベーション(または重みの合計)	442

JMPでは，R2乗が寄与率を表している．この事例では，全体の0.515の割合(51.5%)の説明能力があることを表している．

3.重回帰分析について





寄与率では，応答変数に影響しない説明変数であっても，沢山投入すれば，寄与率はそれにつれて上昇する．

JMPのOutput

あてはめの要約

R2乗	0.514884
自由度調整R2乗	0.508193
誤差の標準偏差(RMSE)	54.06454
Yの平均	152.1335
オブザベーション(または重みの合計)	442

そのため，通常は，自由度調整済み寄与率(自由度調整R2乗)を用いて評価する．
その結果，推定された回帰モデルは，全体の0.508(50.8%)を説明する能力があることを示している．

3.重回帰分析について

JMPのOutput

分散分析

要因	自由度	平方和	平均平方	F値
モデル	6	1349515.1	224919	76.9487
誤差	435	1271494.0	2923	p値(Prob>F)
全体(修正済み)	441	2621009.1		<.0001*

重回帰分析では,

(応答のバラツキ) ＝ (予測値のバラツキ)＋(誤差のバラツキ)

に分けることができる.

(応答のバラツキ)：全体の平方和, (予測値のバラツキ)：モデルの平方和
(誤差のバラツキ)：誤差の平方和

平均平方：「平方和／自由度」により計算されたもの
F値：(モデルの平均平方)／(誤差の平均平方)により計算されたもの

- F値とは, (自由度で調整されたモデルのバラツキ)／(自由度で調整された誤差のバラツキ)を表している.
- いいかえれば, 「当てはまりの良さ／当てはまりの悪さ」を検定したものが, 回帰の分散分析であり, 回帰モデルに意味があるか否かを検討することができる.

事例では, p<0.001なので, 回帰モデルに意味があることが示されている.

3.重回帰分析について

JMPのOutput

パラメータ推定値

項	推定値	標準誤差	t値	p値(Prob>|t|)
切片	-346.1531	25.77073	-13.43	<.0001*
性別[1]	10.795506	2.852819	3.78	0.0002*
BMI	5.7111067	0.707262	8.07	<.0001*
血圧	1.1265526	0.215843	5.22	<.0001*
総コレステロール	-1.042876	0.220751	-4.72	<.0001*
LDL	0.843277	0.229754	3.67	0.0003*
LTG	73.306526	7.308256	10.03	<.0001*

帰無仮説H_0「回帰係数が0である」に対して，対立仮説H_1「回帰係数が0でない」を検定したものが回帰係数の検定である．

モデル選択において，残った変数においても，この検定で有意にならない場合もあるので注意

応答(糖尿病の進行程度)に対する予測式

$$-346.153 + 10.796 \times (\text{性別}) + 5.711 \times (\text{BMI}) + 1.127 \times (\text{血圧}) - 1.043 \times (\text{総コレステロール}) + 0.843 \times (\text{LDL}) + 73.307 \times (\text{LTG})$$

LTGが最も影響が強いといってよいか？

そうとはいえない．そもそも，単位が異なる．

6.おわりに

ご清聴ありがとうございました