統計解析ソフトRを用いたデータ解析
-回帰分析偏-

下川敏雄
和歌山県立医科大学 医学部／附属病院臨床研究センター

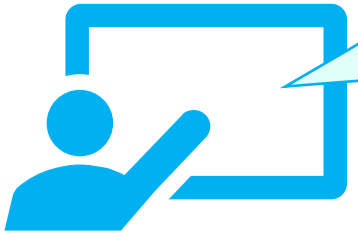
シェーマ

内容
重回帰分析(その1)：重回帰分析の概要
重回帰分析(その2)：Rによる重回帰分析の実践
発展的な回帰分析：ロジスティック回帰分析

重回帰分析(その1)：重回帰分析の概要

回帰分析とは

いま，たばこと肺がんの関係を明らかにするために，各地域のたばこの総販売個数と肺がんの罹患者数を調査した．



相関分析の結果，たばこの総販売個数と肺がんの罹患者数には正の相関関係があった．したがって，**たばこと肺がんには関連性がある**!!

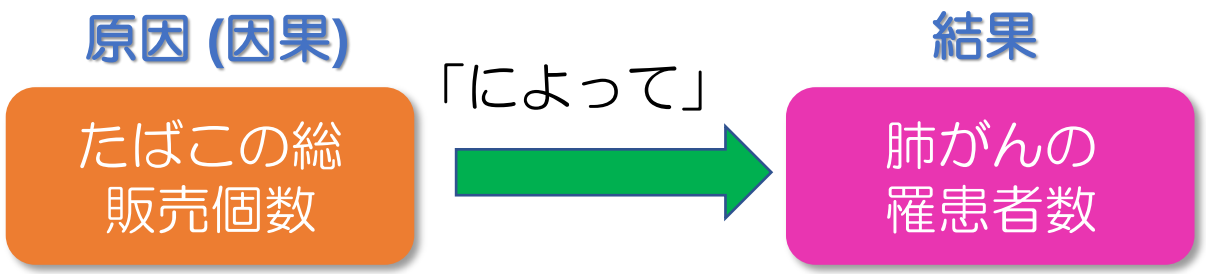
相関分析では，たばこの総販売個数が多い地域では肺がんの罹患者が多い(肺がんが多い地域ではたばこの総販売個数が多い)という関係がわかる．

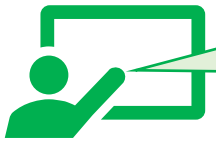
たばこの総販売個数 ⟷ 肺がんの罹患者数

このような調査では，「たばこが肺がんの原因である」ことを明らかにするために，**「たばこの総販売個数が高い(喫煙者が多い)ことで肺がんの罹患者数が高くなる」**ことを知りたいのでは？

回帰分析を用いる

仮説：たばこの総販売個数が高い(喫煙者が多い)ことで肺がんの罹患者数が多くなる





明確に原因と結果が明らかならば，「原因から結果を予測することができる」はずである.

原因から結果を予測するための予測式を作る統計的な方法が回帰分析である.

回帰分析の目的は，

(1) 原因から結果を予測するための予測式(回帰式／回帰モデル)を作成する [予測の観点] .

(2) 複数の原因がある場合(重回帰分析)には，複数の原因がどのように結果に影響するか(すなわち，要因構造)を明らかにする [要因構造探索の観点].

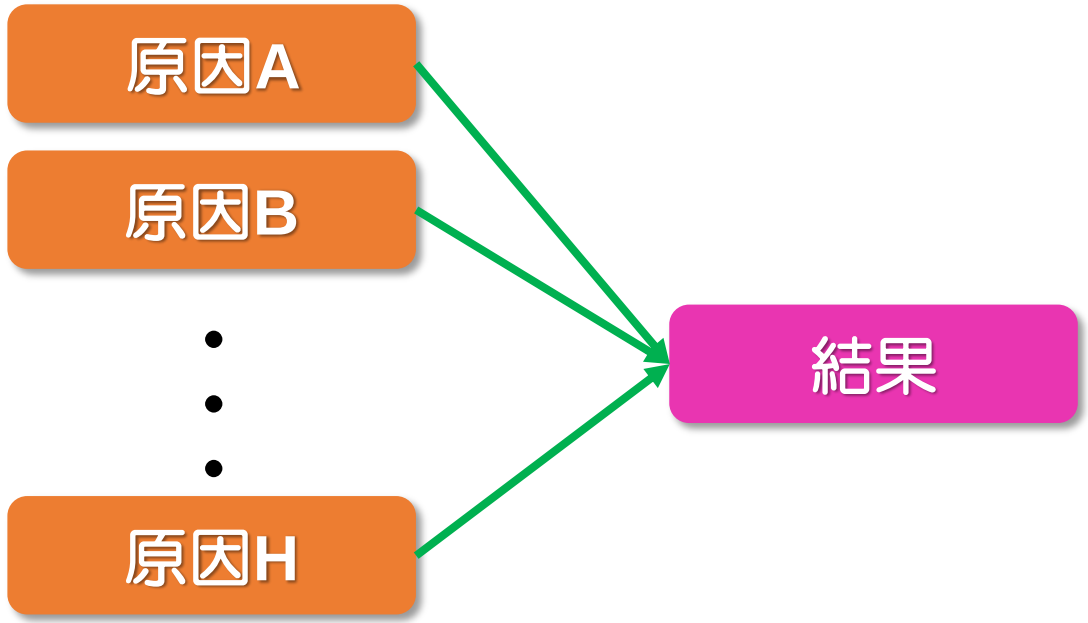
単回帰分析と重回帰分析

回帰分析は原因の数によって呼び方が異なる

単回帰分析

結果に対して，一つの原因のみの影響を分析する回帰分析を**単回帰分析**という．

重回帰分析



結果に対して，複数の原因の影響を分析する回帰分析を**重回帰分析**という．

回帰式(回帰モデル)とは：単回帰分析を例に

■ 回帰式の構造

結果(肺がんの罹患者数)の回帰(予測)式

結果(肺がんの罹患者数) ＝ 切片 ＋ 傾き × 原因(たばこの総販売個数) ＋ 誤差

これらをデータから計算(推定)できれば，原因から結果を予測することができる．
つまり，回帰分析では，切片と傾きを計算する．

■ 用語の定義

結果(肺がんの罹患者数) ＝ 切片 ＋ 傾き × 原因(たばこの総販売個数) ＋ 誤差

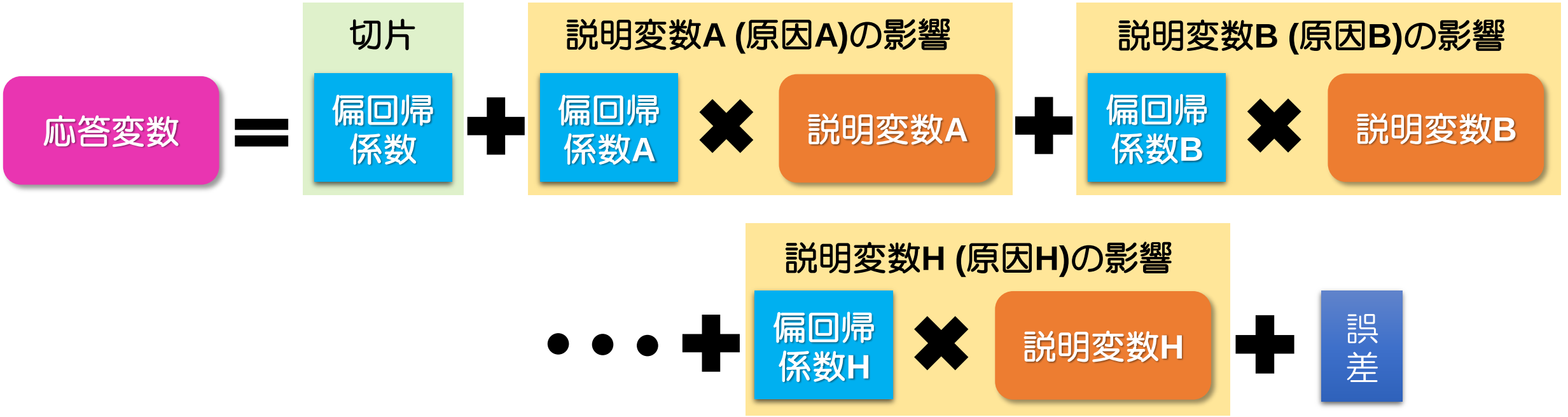
- 結果(肺がんの罹患者数) → 変数の名前
- 切片，傾き → 総称
- 原因(たばこの総販売個数) → 変数の名前
- 誤差

用語
(複数存在)

応答変数	回帰パラメータ	説明変数
従属変数	回帰係数	独立変数
目的変数		説明変数
出力変数		入力変数

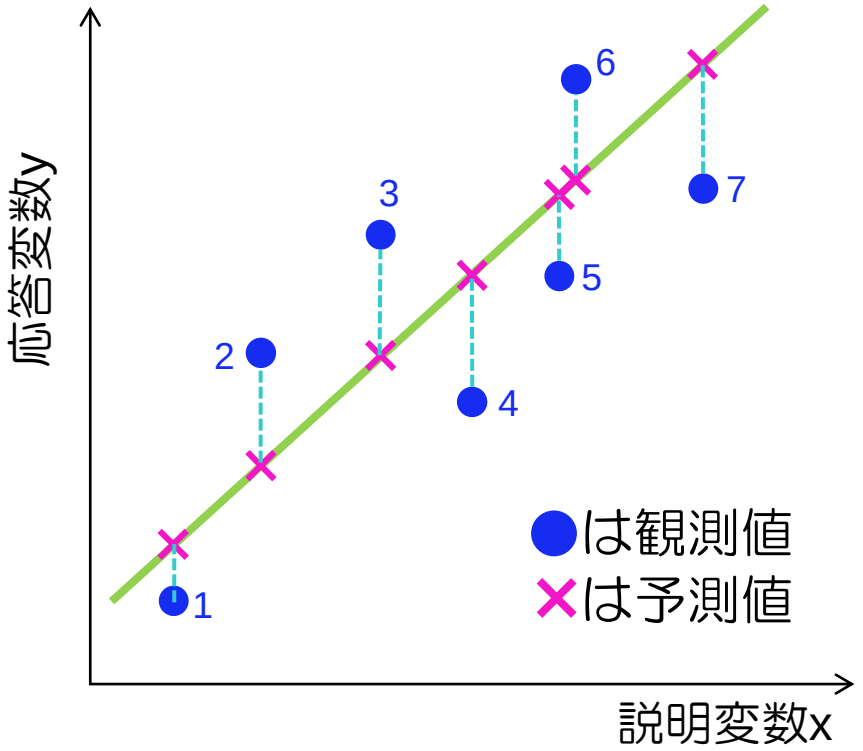
モデルで書くときは誤差(error)と呼び，データから推定された回帰式での誤差は，**残差(residual)**と呼ばれる

重回帰分析の場合の回帰式(回帰モデル)



- 重回帰分析では，回帰係数(回帰パラメータ)を偏回帰係数(偏回帰パラメータ)と呼ぶことがある.
- 重回帰分析の回帰式は，偏回帰係数×説明変数の総和で表される(線形結合).

回帰係数の推定：単回帰分析を例に



推定された回帰式の予測値と観測値の差(残差)は

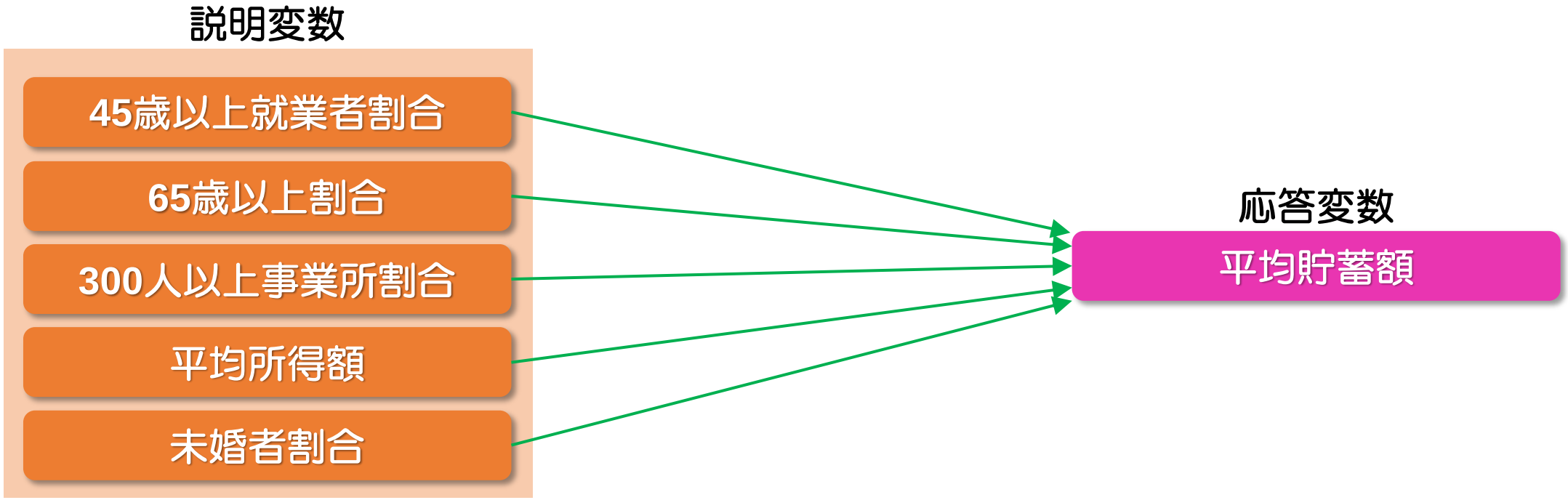
1番目の残差	＝	1番目の応答変数の値	－	1番目の予測値
2番目の残差	＝	2番目の応答変数の値	－	2番目の予測値
⋮		⋮		⋮
7番目の残差	＝	7番目の応答変数の値	－	7番目の予測値

残差が全体的に最小にするような回帰係数を計算する？ しかし，残差の総和は0になる.

残差の2乗和 (残差のバラつき)を最小にするような回帰係数を推定する．これを**最小2乗法**という.

重回帰分析の例示

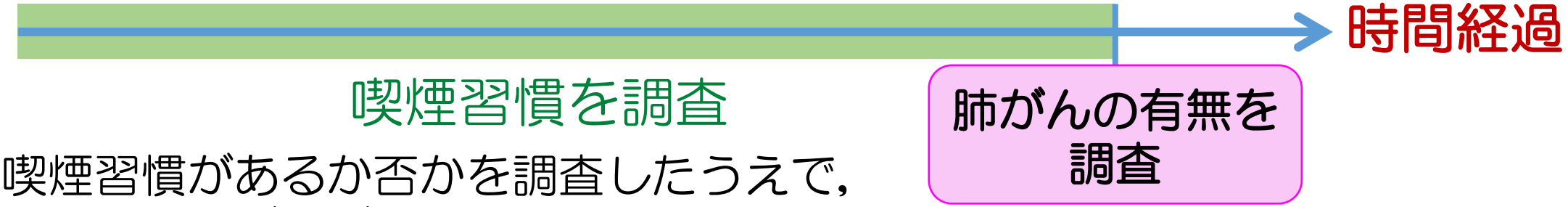
ある県では，県内の市区町村における就業状況と貯蓄額の関係を調査した．就業情報を説明変数，平均貯蓄額応答変数として，平均貯蓄額を推定するための重回帰分析を行う．



平均貯蓄額の推定値 ＝ **4007.9 — 84.2 × 45歳以上の就業者割合 ＋ 206.1 × 65歳以上の割合**
＋ 30776.9 × 300人以上事業割合 ＋ 4.8 × 平均所得額 — 418.9 × 未婚者の割合

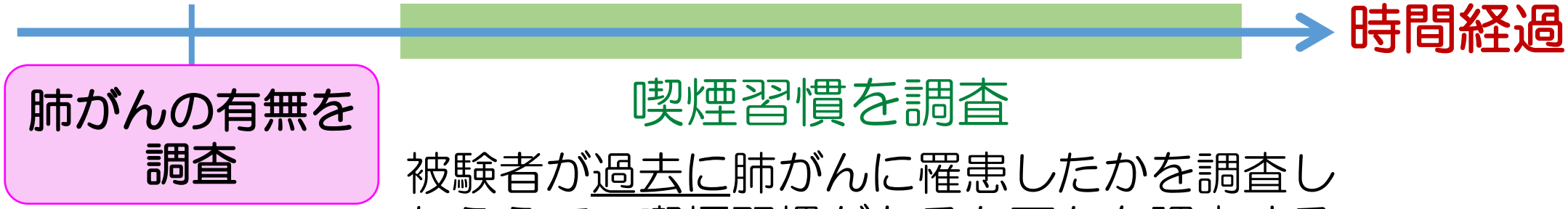
因果推論(回帰分析)を分析するうえにおいて重要な点

■ 正しい調査



喫煙習慣があるか否かを調査したうえで，
その被験者が肺がんに罹患したか調査する．

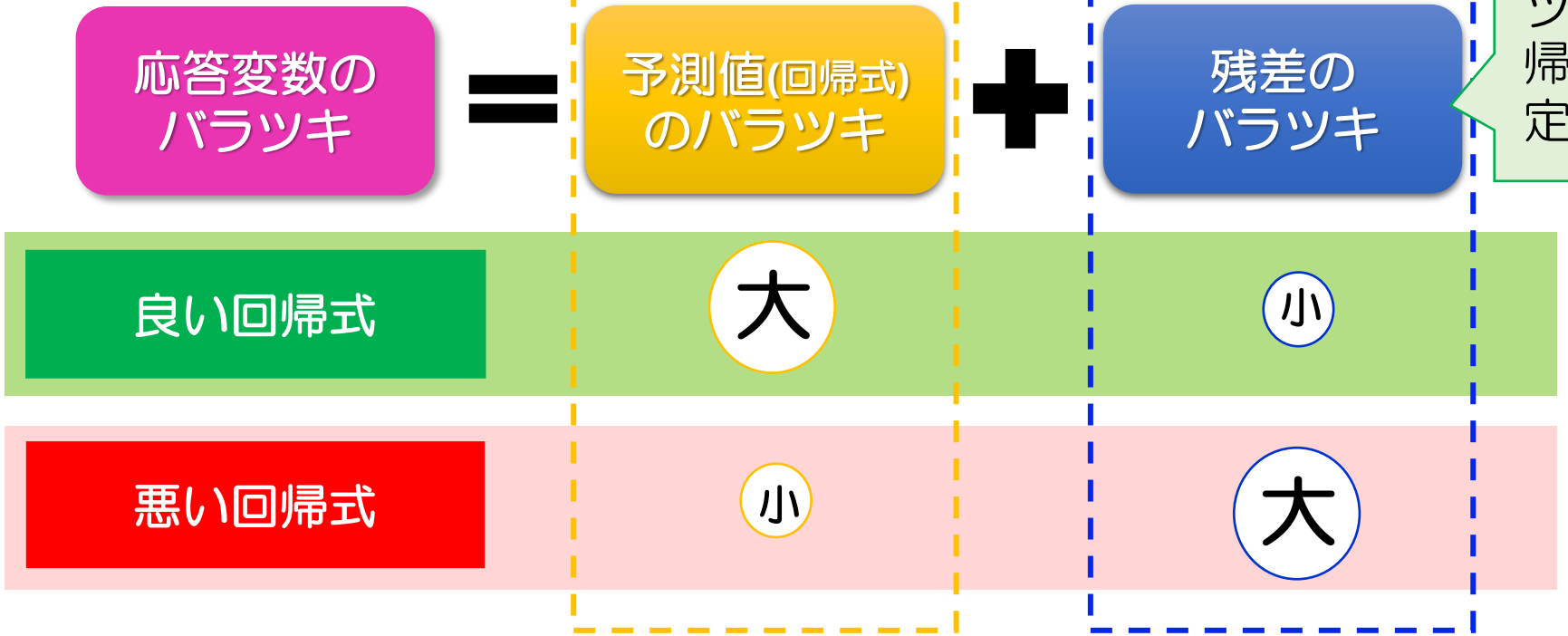
■ 誤った調査



被験者が過去に肺がんに罹患したかを調査したうえで，喫煙習慣があるか否かを調査する．

回帰分析における変動分解

回帰分析では，応答変数のバラツキは，予測値のバラツキと残差のバラツキに分けることができる．
これを回帰分析における変動分解という．

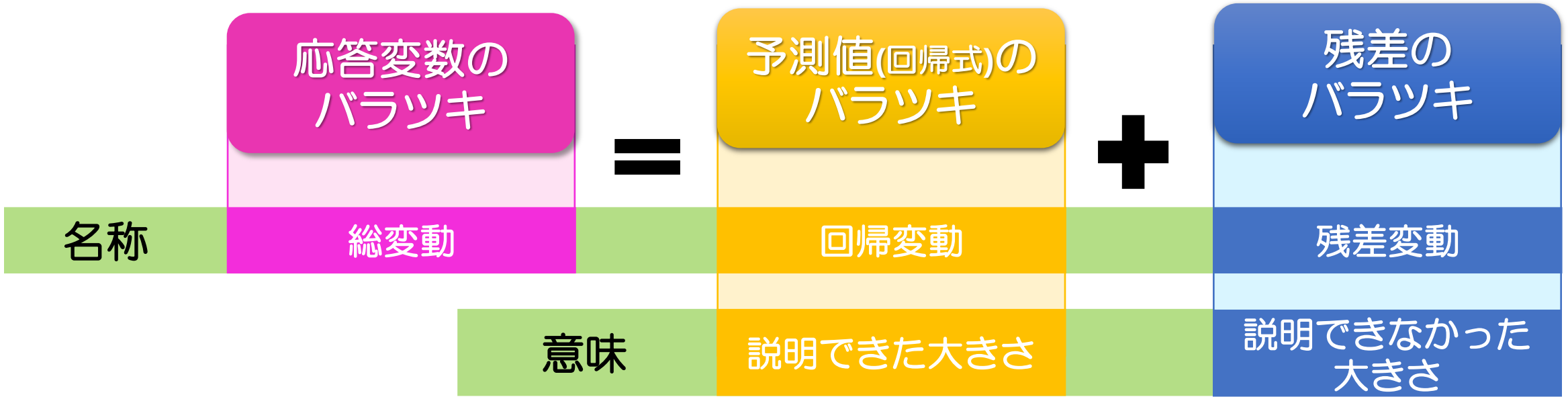


回帰分析では，誤差のバラツキを最小にするように回帰係数(回帰パラメータ)を推定している．

予測値のバラツキの大きさは，回帰式による応答変数の予測力の**高さ**を表す．

誤差のバラツキの大きさは，回帰式による応答変数の予測力の**低さ**を表す．

寄与率



寄与率とは，応答変数のバラツキの何割(何パーセント)を予測値のバラツキが示しているかを表したものである．いいかえれば，回帰式が応答変数をどの程度説明できたかを表した値である．したがって，寄与率は0.0 (0%) ～ 1.0 (100%)の範囲をとる．

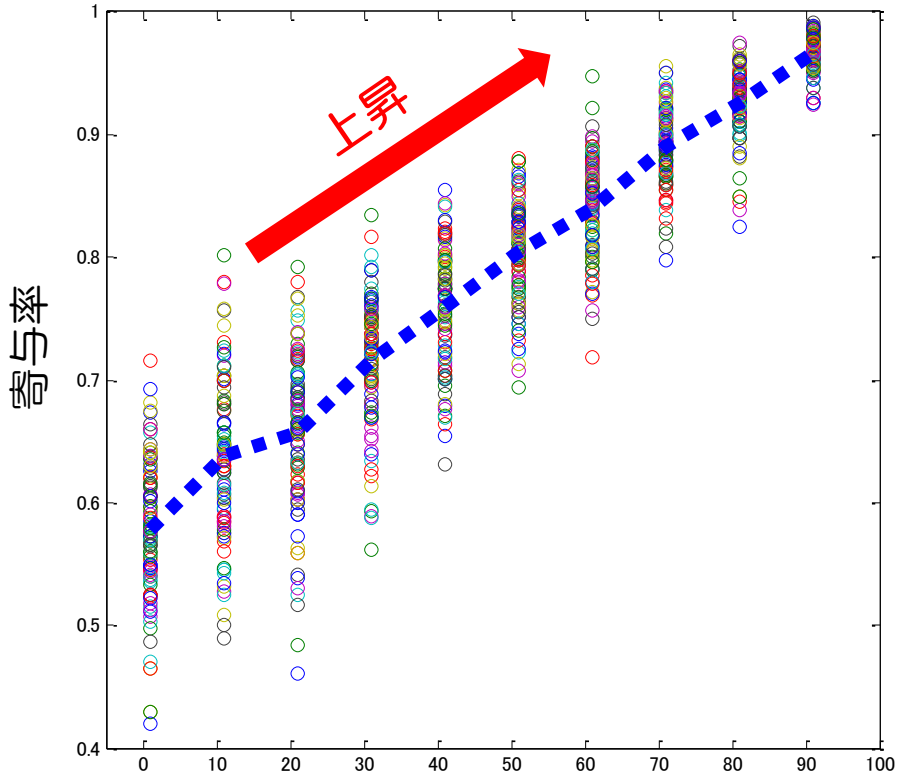
$$\text{寄与率} = \frac{\text{回帰変動}}{\text{総変動}} = 1 - \frac{\text{残差変動}}{\text{総変動}}$$

寄与率は，決定係数とも呼ばれる．

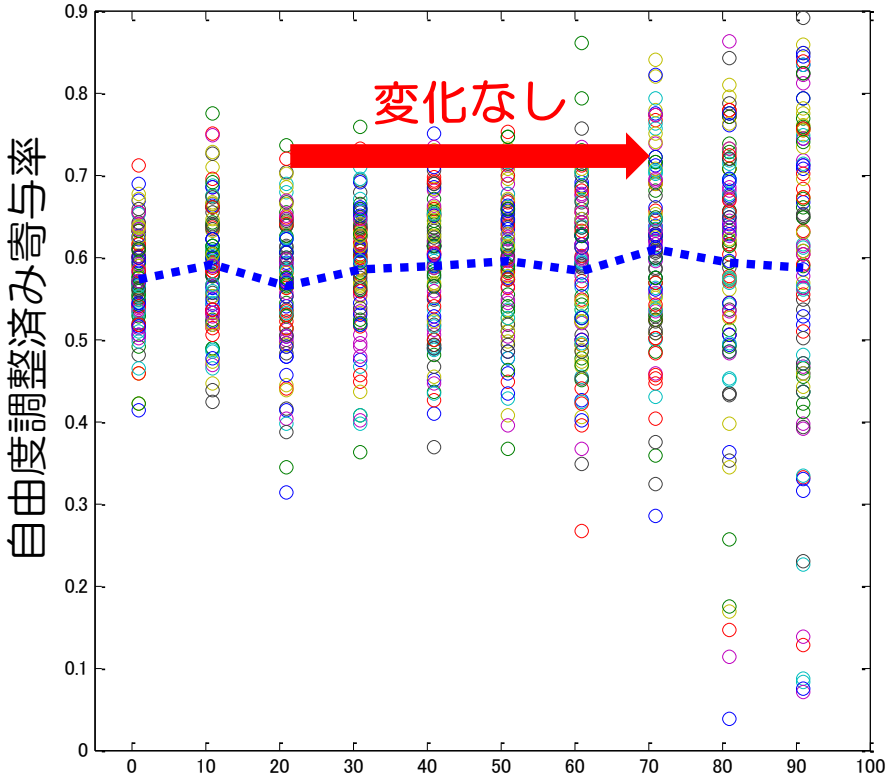
寄与率と自由度調整済み寄与率

説明変数の数が増えると，回帰変動が過大評価され，残差変動が過小評価される．この問題を調整した寄与率は，自由度調整済み寄与率と呼ばれる．

応答変数に影響しない説明変数の数を増加したときの寄与率及び自由度調整済み寄与率の推移(各説明変数の数に対して100回のシミュレートを実施している．青色の点線は平均値の推移を表している)



(a) 寄与率



(b) 自由度調整済み寄与率

寄与率は，応答変数に影響しない説明変数を増やせば増やすほど増加するので，適合度評価にはならない．そのため，自由度調整済み寄与率を用いなければならない．

先ほどの例示における重相関係数・寄与率・自由度調整済み寄与率

総変動 = 409,582,868

回帰変動 = 225,878,281

残差変動 = 18,3704,587

$$\text{寄与率} = \frac{225{,}878{,}281}{409{,}582{,}868} = 0.551\ (55.1\%)$$

$$\text{自由度調整済み寄与率} = 1 - \frac{18{,}3704{,}587 \div (46-5-1)}{409{,}582{,}868 \div (46-1)} = 0.495\ (49.5\%)$$

標本サイズ – 説明変数の数 - 1

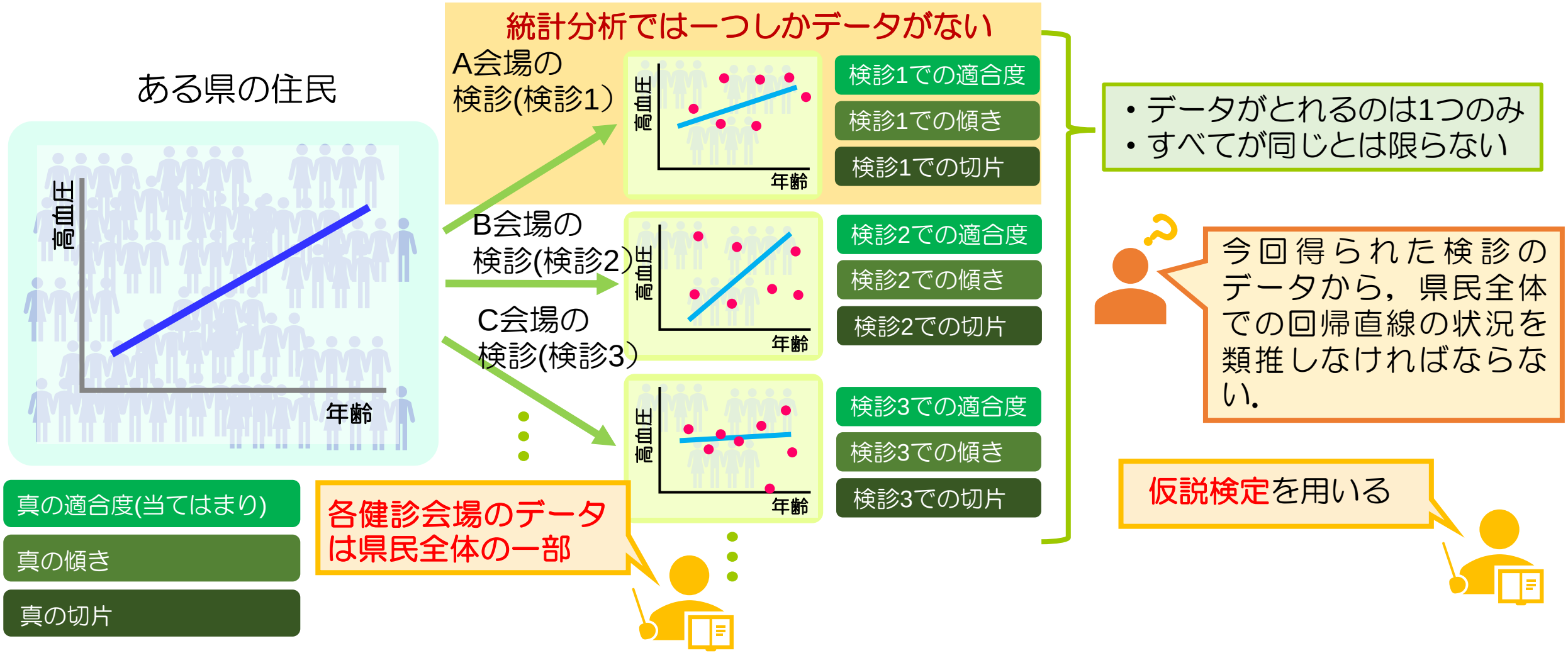
標本サイズ - 1

すなわち，推定された重回帰式(重回帰モデル)は応答変数(平均貯蓄額)の49.5%を説明できる.

重相関係数=0.7426である．$0.7426^2 = 0.551$となり，寄与率と一致する

回帰分析における諸種の検定

仮想例：いま，ある県において実施された特定検診のデータを用いて，年齢と高血圧の関係が調査された．



仮説検定とは何か：傾きに対する有意性検定を例に考える

二つのシナリオ

逆仮説

仮説が背反になっている

帰無仮説H_0

真の回帰式における傾きは0である

シナリオのパターンは1種類

帰無仮説H_0のシナリオが正しいとしたもとで，今回のデータがどれぐらいの確率で得られるかを考える．

帰無仮説H_0のシナリオが正しいと仮定したときに，今回のデータから推定されるような傾きが得られる確率

0.035 (3.5%)

これを**p値(有意確率)**という

対立仮説H_1

真の回帰式における傾きは0でない

シナリオのパターンは無限大

仮説検定とは，帰無仮説H_0が正しいと仮定したもとで，今回のデータが得られる確率を計算したもとで評価する方法である．

p値がどれぐらい小さいと帰無仮説H_0が誤っているといえるかを決定するカットオフ値が必要である．このカットオフ値を**有意水準α**といい，一般的には**0.05 (5%)**とされている．

仮説検定による判定の方法

p値が有意水準α(一般的にはα=0.05)未満

p値 (有意確率)

例示: p値=0.035

p値が有意水準α(一般的にはα=0.05)以上

例示の場合

帰無仮説H_0が間違っている (棄却)

有意である (棄却される)

結論

(真の)傾きは0でない

有意であるということは，その説明変数は，応答変数に影響している.

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (棄却されない)

結論

(真の)傾きは0でないという根拠なし

仮説検定において0であるとは言えないので注意

F検定(回帰の分散分析)

帰無仮説H_0：(真の)回帰式に意味はない

対立仮説H_1： (真の)回帰式に意味はある

帰無仮説H_0が正しいと仮定したとき，**今回のデータにおける適合度はどれぐらいの確率(p値)で得られるのか**について計算する

p値

p値≧有意水準α　比較　p値＜有意水準α

帰無仮説H_0が受容

回帰式(モデル)に意味あるとはいえない

帰無仮説H_0が棄却

回帰式(モデル)に意味ある

傾きに対する有意性検定

帰無仮説H_0：(真の)傾きは0である

対立仮説H_1： (真の)傾きは0でない

帰無仮説H_0が正しいと仮定したとき，**今回のデータにおける傾きはどれぐらいの確率(p値)で得られるのか**について計算する

p値

p値≧有意水準α　比較　p値＜有意水準α

帰無仮説H_0が受容

傾きが0でないとはいえない

帰無仮説H_0が棄却

傾きが0でない

切片に対する有意性検定

帰無仮説H_0：(真の)切片は0である

対立仮説H_1： (真の)切片は0でない

帰無仮説H_0が正しいと仮定したとき，**今回のデータにおける切片はどれぐらいの確率(p値)で得られるのか**について計算する

p値

p値≧有意水準α　比較　p値＜有意水準α

帰無仮説H_0が受容

切片が0でないとはいえない

帰無仮説H_0が棄却

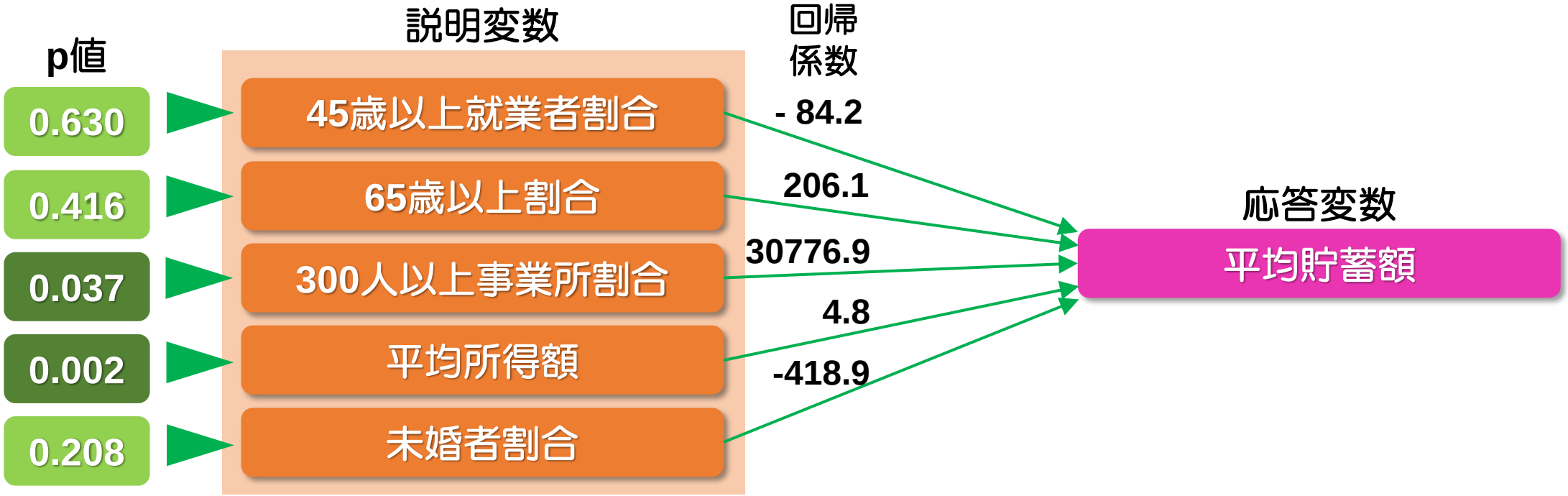
切片が0でない

「比較」の部分のロジックな説明

帰無仮説H_0が正しいと仮定したときに，今回の結果(適合度，傾き，切片)が得られる確率をp値(有意確率)という．このとき，どれぐらいの確率未満であれば，帰無仮説H_0が誤っている(逆仮説である対立仮説H_1が正しい)と判定するのかの基準が有意水準α(通常はα=0.05)である．

重回帰分析の例示：回帰係数に対する有意性検定を付与

ある県では，県内の市区町村における就業状況と貯蓄額の関係を調査した．就業情報を説明変数，平均貯蓄額応答変数として，平均貯蓄額を推定するための重回帰分析を行う．



平均貯蓄額の推定値 ＝ 4007.9 — 84.2 × 45歳以上の就業者割合 ＋ 206.1 × 65歳以上の割合
＋ 30776.9 × 300人以上事業割合 ＋ 4.8 × 平均所得額 — 418.9 × 未婚者の割合

ただし，有意な説明変数は２個(300人以上事業所割合，平均所得額)のみである

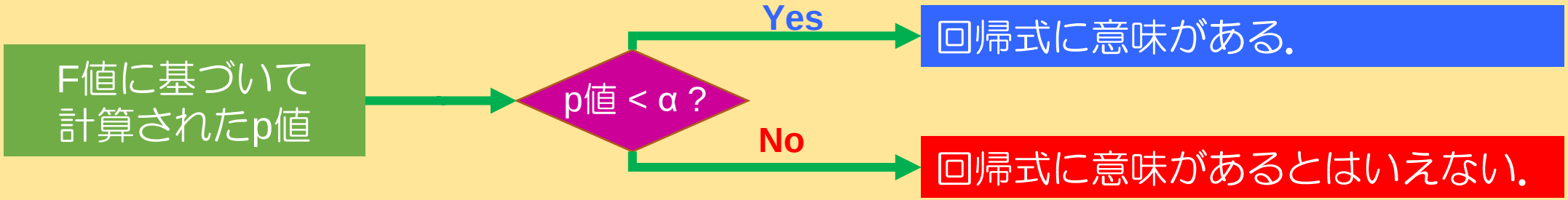
回帰分析における分散分析表

推定された回帰式(回帰モデル)に意味があるかどうかを判断するために用いられるのが分散分析表である.

変動	平方和	自由度	分散	F値
回帰	回帰変動	P	回帰変動 ÷ P	回帰の分散 ÷ 残差の分散
残差	残差変動	N – P - 1	残差変動 ÷ N – P - 1	
合計	総変動	N - 1		

Pは説明変数の数，Nは標本サイズ(サンプル数)を表している.

F値に基づいてp値というものが計算される．それに基づいて，次のように解釈する.



回帰分析における分散分析表

分散分析では，仮説検定という統計学的な判断が行われる．仮説検定は，手法毎に帰無仮説H_0，対立仮説H_1が立てられる．今回の場合には，

H_0：回帰式(回帰モデル)に意味がない，　　H_1：回帰式(回帰モデル)に意味がある

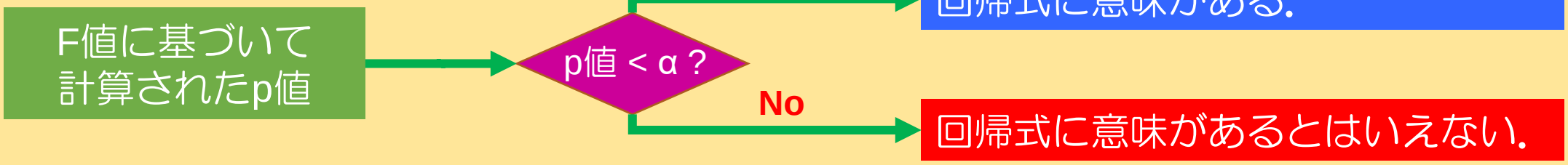
上記の仮説でもわかるように，対立仮説H_1が本来言いたい仮説になっている．

仮説検定では，**「帰無仮説H_0が正しいと仮定したときに，得られたデータにおいて，その仮定がどれぐらい確かだと言えるのか」を確率的な概念のもとで評価される．**

このとき，確率的な概念として用いられるものがp値である．したがって，p値(**ExcelではP-値**)が非常に小さな値をとるということは，**「帰無仮説H_0が正しいと仮定したものの，その仮定は誤っている（確かでない）」**と判断される．ちなみに，p値を計算するための数字(今回の場合はF値)を**検定統計量**という．

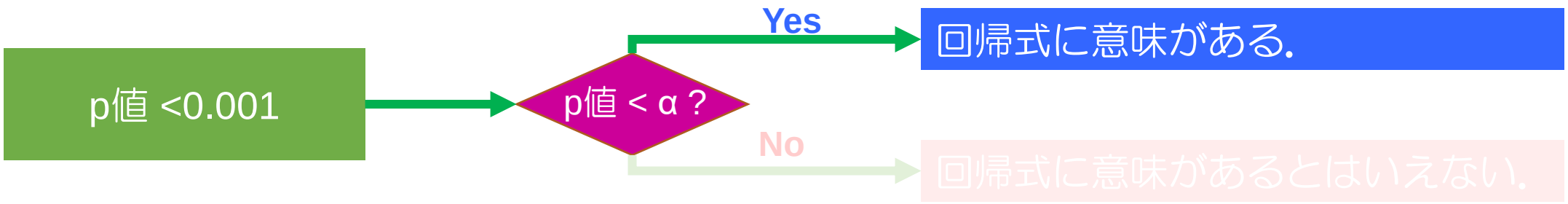
p値が「非常に小さい」と判断するしきい値になるものが有意水準αである．一般的には，有意水準αには0.05が採用される．

今回の場合の仮説検定(F検定による判断)



先ほどの例示における分散分析表の例

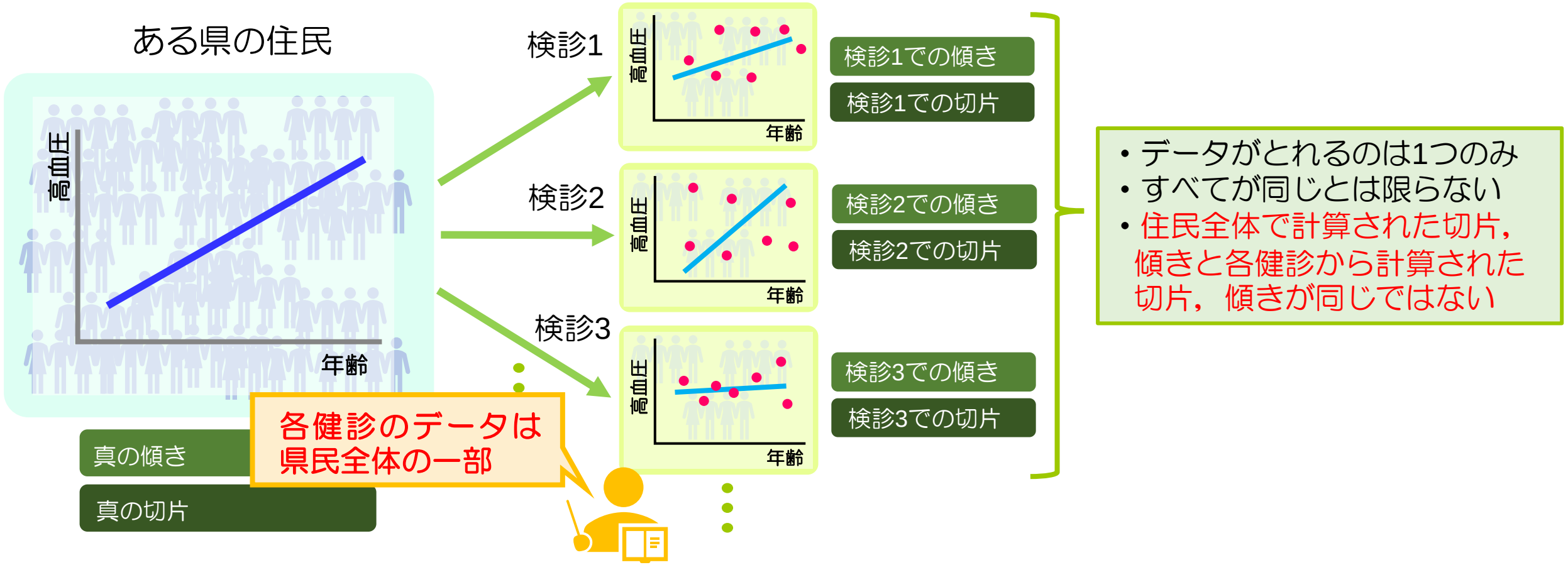
変動	平方和	自由度	分散	F値	p値
回帰	225,878,281	5	225,878,281 ÷ 5 = 45,175,656.2	45,175,656.2 ÷ 4,592,614.7 = 9.8366	<0.001
残差	183,704,588	46 – 5 – 1 = 40	183,704,588 ÷ 40 = 4,592,614.7		
合計	409,582,869	46 – 1 = 45			



すなわち，回帰モデルには意味があると判断される.

回帰分析における信頼区間とは？

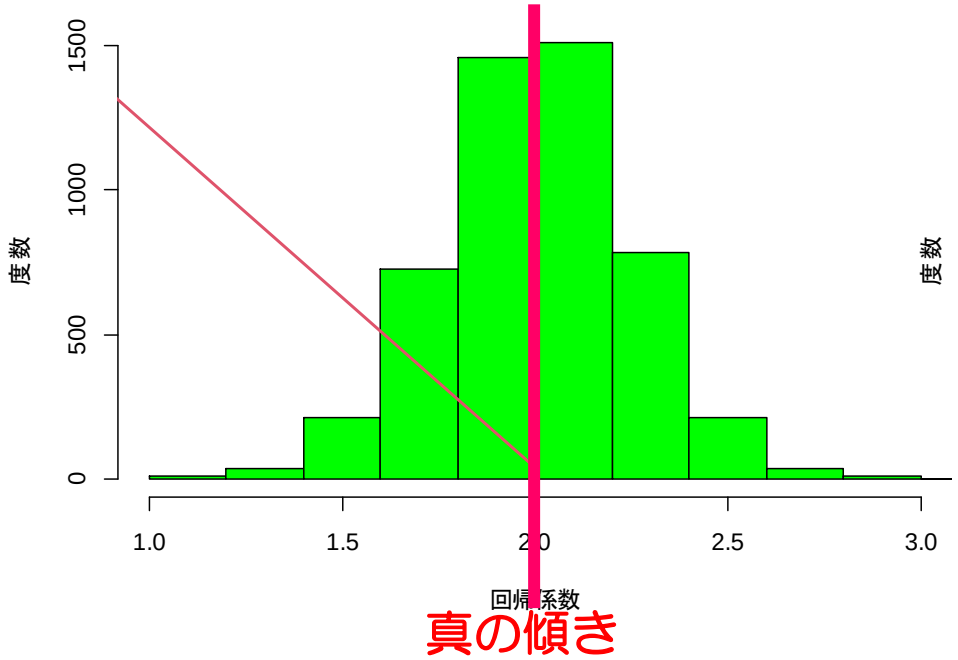
仮想例：いま，ある県において実施された特定検診のデータを用いて，年齢と高血圧の関係が調査された．



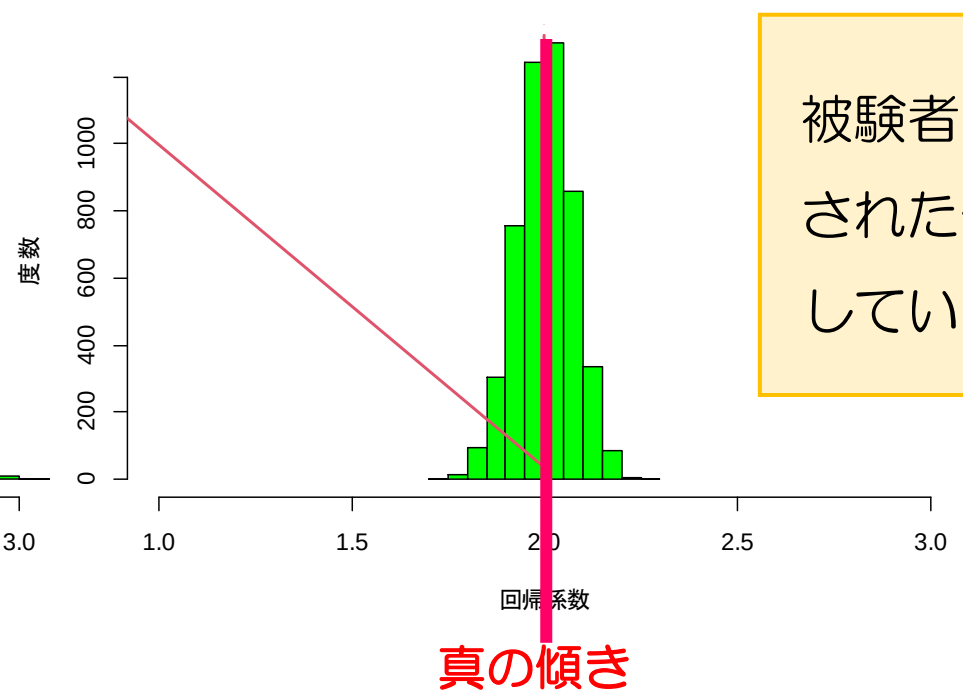
回帰式と被験者数(標本サイズ)の関係

真の傾きを2としたときに5000回シミュレーションを実施したときの結果

被験者数が20人の場合



被験者数が200人の場合



被験者数が増えるほど，計算された傾きが真の傾きに集中していることがわかる.

今回計算した傾きの信頼度を表すために用いられるのが信頼区間である.

信頼区間が0を含まなければ，有意性検定で有意になる

多重共線性

いま，奈良県における2016年と2017年の国別の外国人来訪者数から2018年の外国人来訪者数を予測できないかを検討している．

説明変数：2016年と2017年の国別外国人来訪者数，　　応答変数：2018年の国別外国人来訪者数

Model.1：2016年と2017年の両方を加えた場合の回帰式

(2018年の来訪者数) = -995.870 **- 1.082 × (2016年の来訪者数)** ＋ 1.954 × (2017年の来訪者数)

2016年の回帰係数が負値なので，2016年の来訪者数が増えるほど，2018年の来訪者が減る．

Model.2：2016年のみの回帰式

(2018年の来訪者数) = -18643.979 **+ 2.401 × (2016年の来訪者数)**

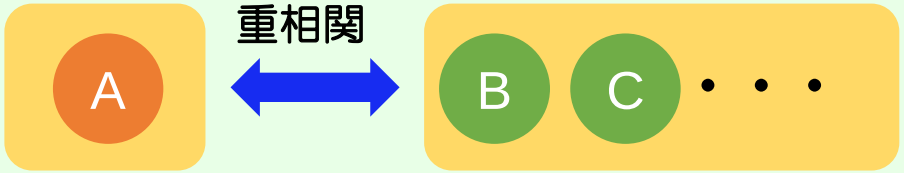
2016年の回帰係数が正値なので，2016年の来訪者数が増えるほど，2018年の来訪者が増える．

2016年＋2017年の回帰式における2016年の解釈と2016年のみの回帰式で解釈が逆になる．これは，2016年と2017年の来訪者の相関係数が0.994であり，非常に高いことに由来する．このように，同じような数値をとる説明変数を重回帰分析に用いた場合，相互に干渉してしまい，誤まった解釈をもたらすことがある．このことを**多重共線性**という．

多重共線性の診断：VIF (Variance Inflection Factor)

説明変数間に強い相関関係がある場合，変数間が干渉することで，実際の影響とは異なる(偏)回帰係数を推定したり，あてはまりが悪くなることを多重共線性という．それを評価する指標がVIFである．

VIFは，説明変数に用いられた「任意の変数A」と「変数A以外の変数」との重相関係数に基づいて計算できる．

重相関係数r ＝ 

を用いて $VIF = \frac{1}{1 - r^2}$ で計算できる

先ほどの事例におけるVIFの結果

奈良県の外国人来訪者のデータ

2016年	VIF = 84.443
2017年	VIF = 84.443

多重共線性があるため，VIFが10を大きく上回っている．

e-statによる平均貯蓄額のデータ

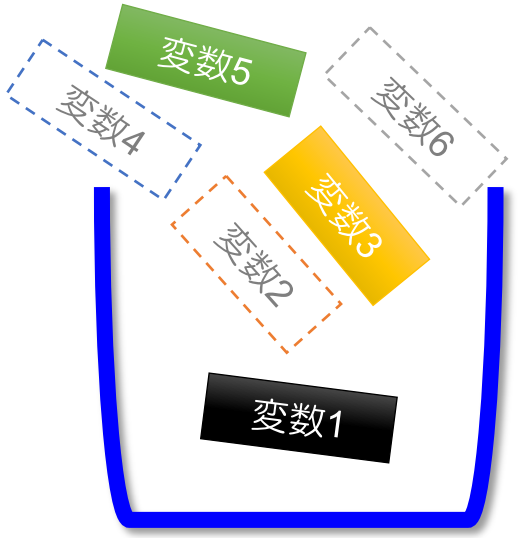
45歳以上就業者割合	VIF = 1.229
65歳以上割合	VIF = 4.272
300人以上事業所割合	VIF = 4.213
平均所得額	VIF = 2.068
未婚者割合	VIF = 4.309

VIFが10を上回る説明変数は存在しなかった．

重回帰分析における注意点：変数が多ければよいってものではない

回帰モデル(回帰式)

重回帰分析では，たくさんの説明変数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？

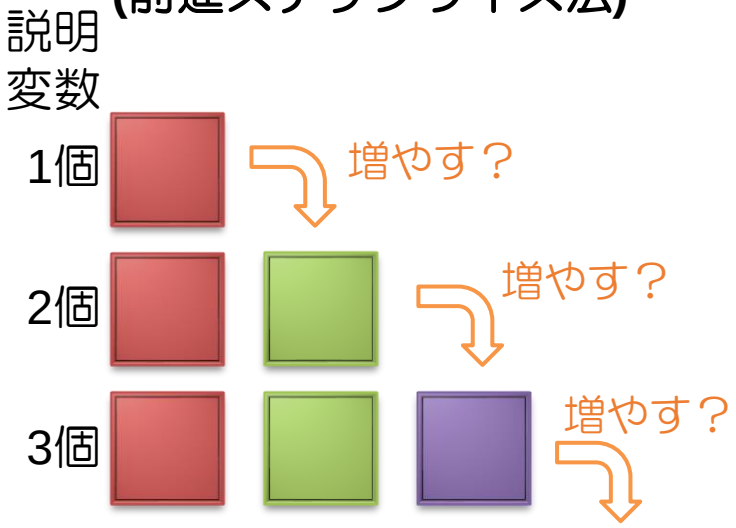


　必ずしもそうではない．説明変数が多いほど不必要な変数は単なるノイズでしかない．
　また，多重共線性の問題などがある．そのため，必要なものだけで回帰式をつくることが推奨されることが多い．そのための方法が変数選択である．

変数選択では，以下を決定したうえで実施しなければならない：
- どのようなアルゴリズムで変数を選択するのか (変数選択の方法)
- どのような基準で回帰モデル(回帰式)を評価するのか (評価基準)

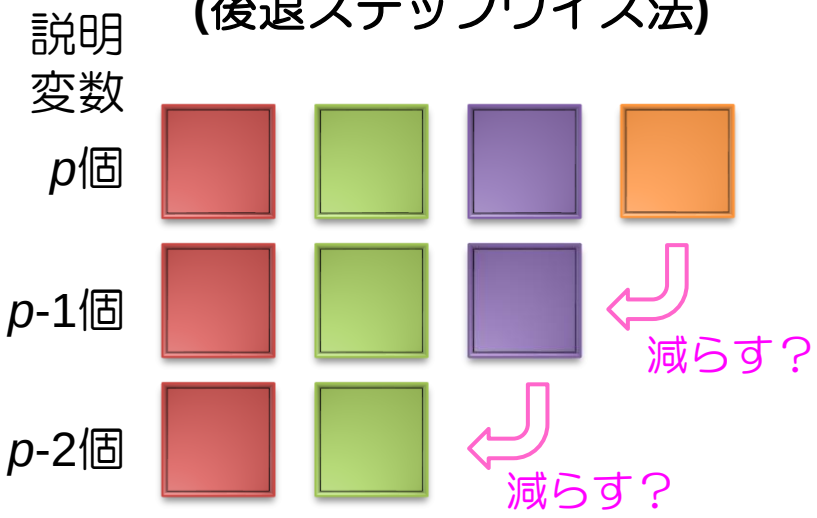
変数選択のアルゴリズム

変数増加法 (前進ステップワイズ法)



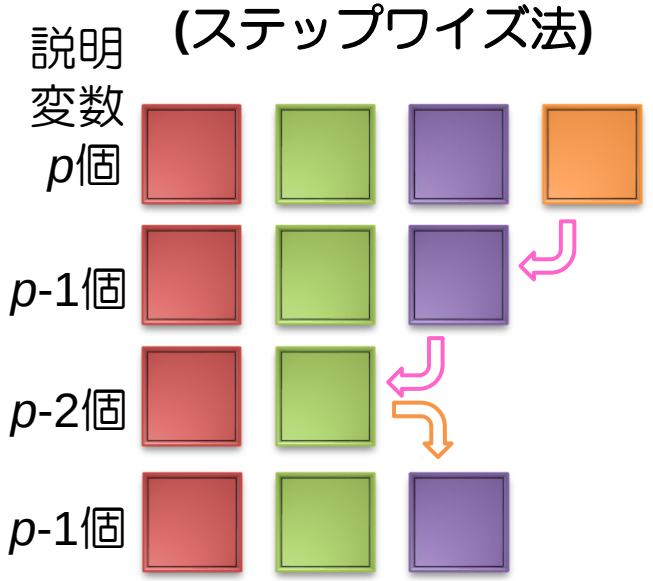
説明変数が1個の場合からスタートして，変数を追加したほうが良ければ増やし，そうでなければ変数の追加をしない．

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．

変数選択の基準

- 検定を用いる方法 (最近は推奨されない)
- 情報量規準を用いる方法 (最近はこちらが一般的)
  - — 赤池の情報量規準 (AIC; Akaike's Information Criteria)
  - — Bayes流情報量規準 (BIC; Bayesian Information Criteria)

情報量規準は，AICとBICのどちらをつかってもかまわない．ただし，AICのほうがBICよりも多くの変数を選択する(BICのほうがドラスティックに削除する)．

変数選択の留意点

(1) 変数増加法の落とし穴

標本サイズが小さい場合に，変数増加法を用いて変数選択を行う場合，結果の解釈が困難なモデルを選択することがしばしばある．また，本当は必要な説明変数が取り込まれる前に変数選択が終了する場合がある．

(2) 重回帰分析に用いることができる説明変数の数

重回帰分析解析に用いることができる説明変数の数は，標本サイズ(個体数)の1/10程度とされている(統計学的な意味はないが，慣例的に言われている)．

(3) 多数の説明変数がある場合の留意点

多数の調査項目(説明変数) が存在する場合には，全ての説明変数を用いて変数選択を行うのではなく，事前スクリーニングを行うことが推奨される．事前スクリーニングでは，説明変数毎に単変量解析(1個の説明変数による回帰モデルを推定する)を実施し，その(偏)回帰係数に対する検定(回帰係数が0であるか否かを評価する検定)のp値や回帰パラメータを用いる．このとき，有意水準αは0.10あるいは0.20であっても許容される．なお，慣例的に標本サイズの1/10が重回帰分析で用いることができる説明変数の目安として考えられている．

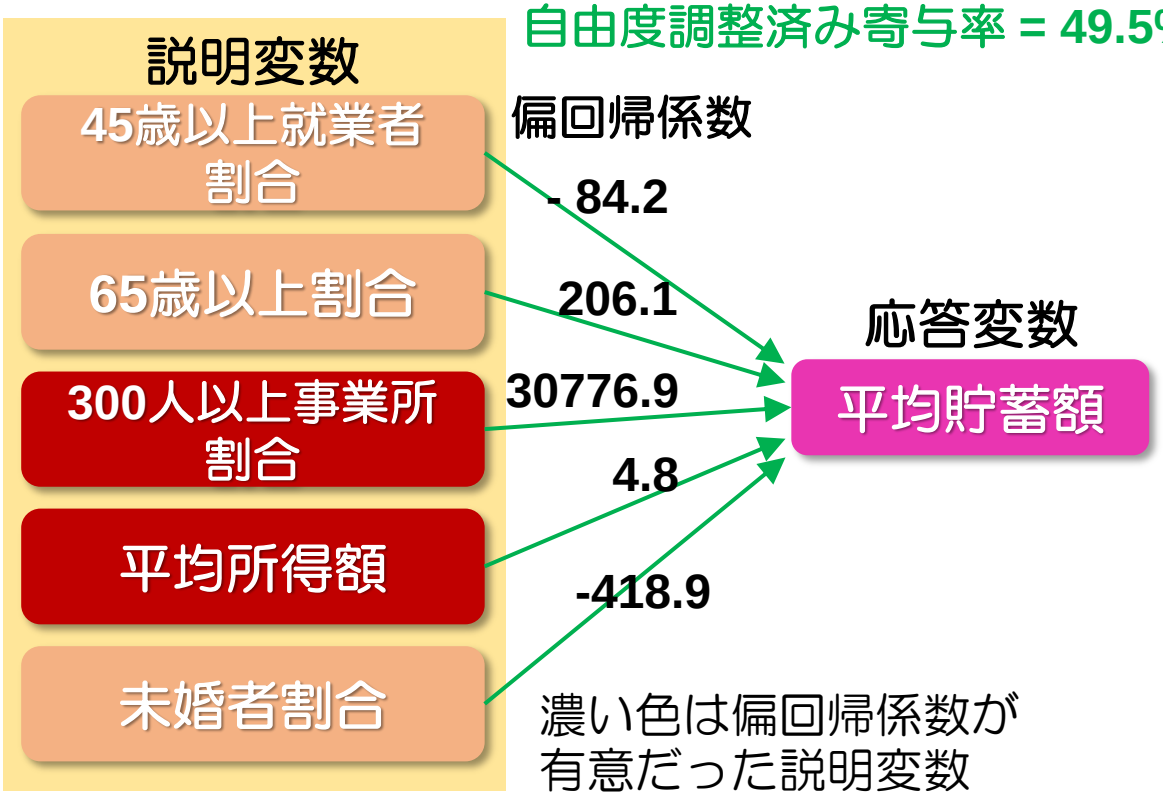
(4) 欠測が多い説明変数(調査項目)には注意が必要である

重回帰分析では，説明変数のなかで1個でも欠測があれば，その個体を削除しなければならない．そのため，欠測が多い説明変数をモデルに含めると，多くの個体を削除することになる．また，観測方法が煩雑な場合には，欠測が多くなる傾向にある．そのため，このような説明変数は，予め変数選択の候補から除外することが望ましい．

先ほどの事例における変数選択の結果

全説明変数を用いた場合

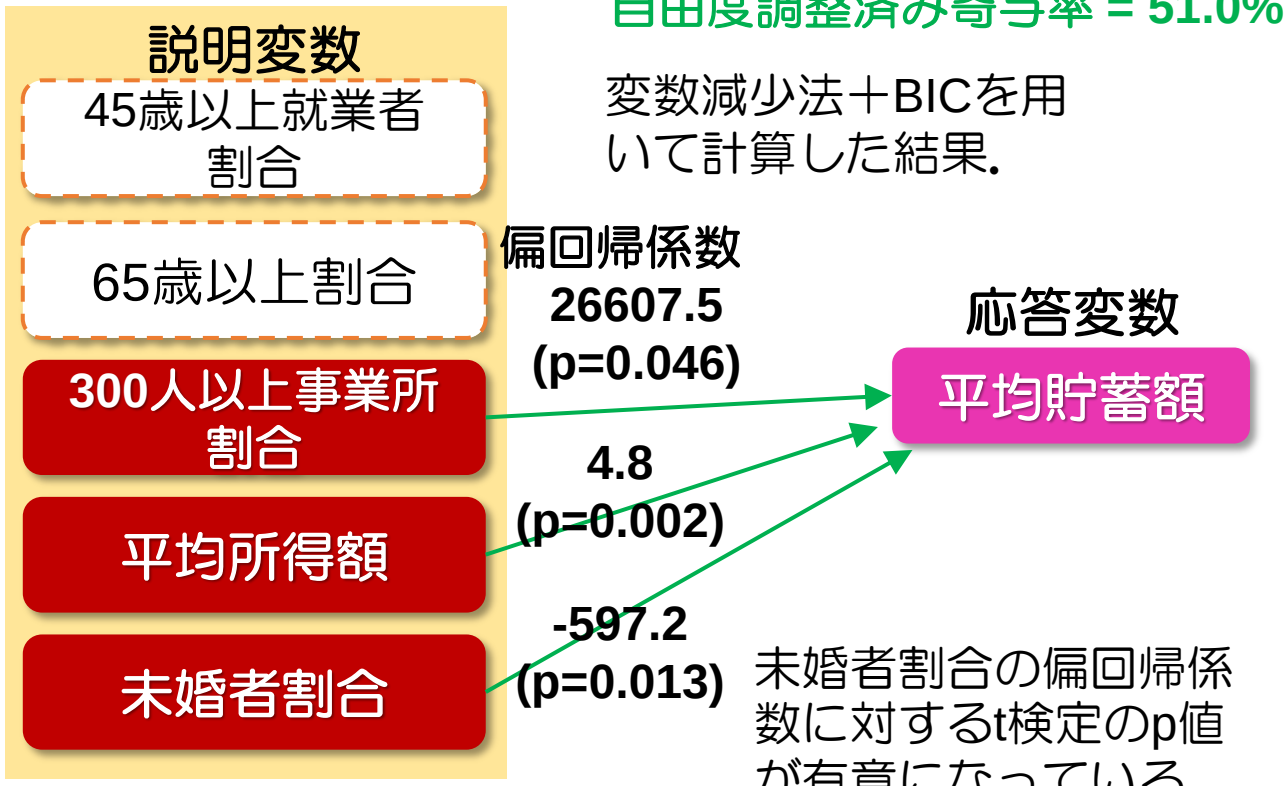
自由度調整済み寄与率 = 49.5%



変数選択後

自由度調整済み寄与率 = 51.0%

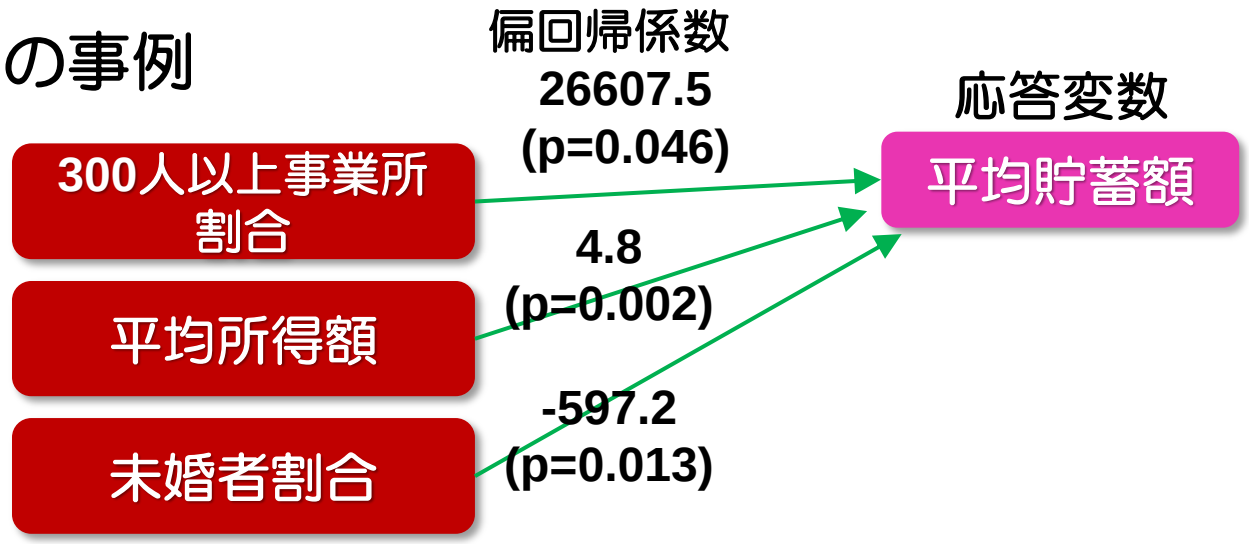
変数減少法＋BICを用いて計算した結果.



変数選択は，多重共線性に対する対処だけでなく，意味のない(ノイズとなっている)説明変数を除去することで重回帰式(重回帰モデル)の安定化を図るという意味でも重要である.

説明変数の影響の強さを見る：標準偏回帰係数

先ほどの事例



偏回帰係数は，説明変数の尺度に依存するため，説明変数(300人以上事業所割合，平均所得額，未婚者割合)のうち，どちらのほうが平均貯蓄額に影響するのかを判断することができない．説明変数の影響を相対的に評価する指標が**標準偏回帰係数(標準化係数)**である．

標準偏回帰係数の求め方

Step.1：各変数(すべての説明変数，応答変数)を標準化する．

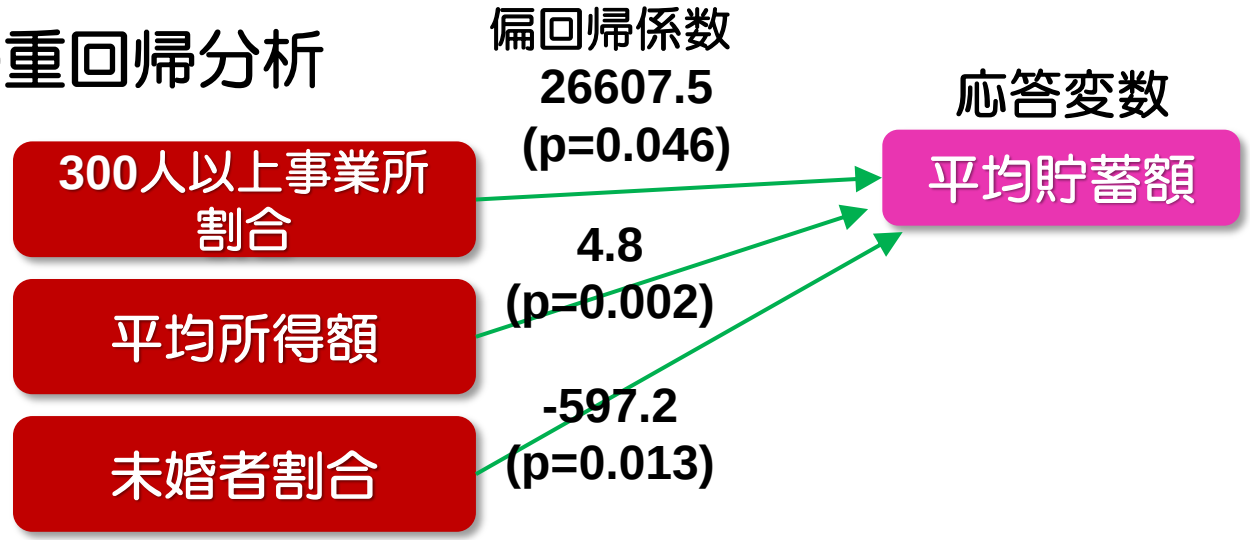
$$\text{標準化} = \frac{\text{(観測値)} - \text{(平均値)}}{\text{(標準偏差)}}$$

Step.2：標準化された変数(すべての説明変数，応答変数)を用いて重回帰分析を行う．

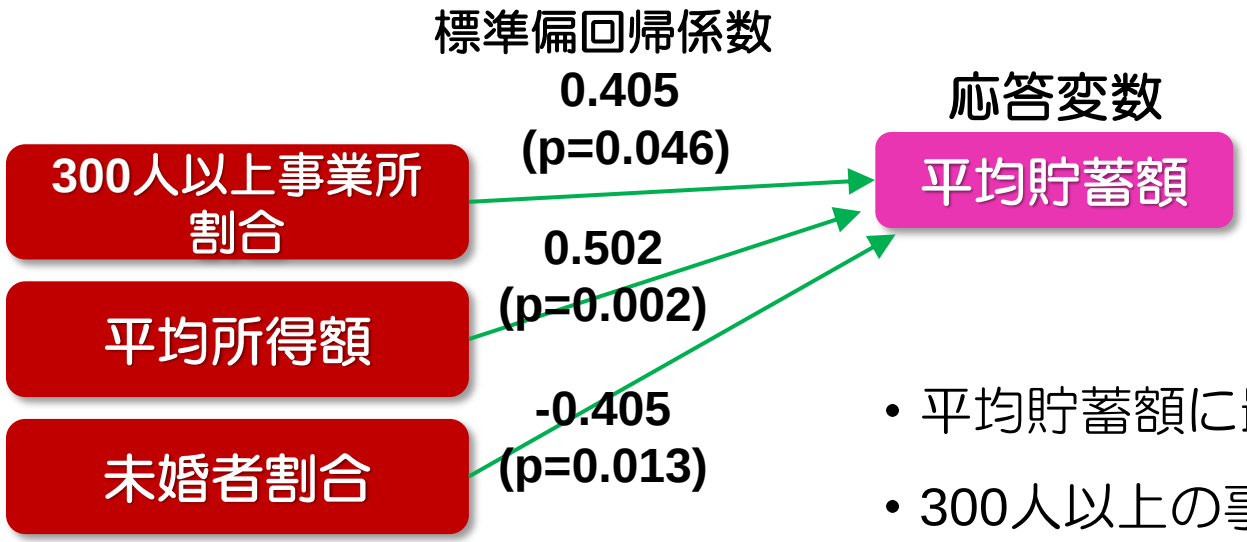
Step.3：標準化された変数による重回帰分析の偏回帰係数が標準偏回帰係数である．

先ほどの事例での標準偏回帰係数の結果

通常の重回帰分析



重回帰分析における標準偏回帰係数



- 平均貯蓄額に最も影響を及ぼすのが平均所得額である.
- 300人以上の事業者の割合，未婚者割合の影響は同程度である.

重回帰分析(その2)：Rによる重回帰分析の実践

Rにおける重回帰分析

ここでは，Cドライブの**Fukuoka_Seminor**というフォルダにある**multi_regress.csv**というCSVファイルを読み込む．これは，**e-stat**を用いて作成した**46**都道府県(東京都を除く)の**1**世帯当たりの平均貯蓄額，平均貯蓄額，**300**人以上事業所割合，未婚者の割合である．

Input

```
> dat <- read.csv("C:/Fukuoka_Seminor/multi_regress.csv",fileEncoding = "cp932")
> head(dat)
```

Output

```
   Pref Savings Income Office Work   Single Senior
1 北海道   11918   2538   0.16 32.9 25.94550   28.1
2 青森県    8624   2358   0.10 29.7 24.81216   29.0
3 岩手県   12689   2671   0.12 34.5 24.46407   29.6
4 宮城県   12154   2863   0.18 32.3 27.45849   24.6
5 秋田県   10419   2409   0.10 31.1 22.07448   32.6
6 山形県   12640   2539   0.11 28.3 22.94568   29.9
```

ここで，
- Pref　：道府県
- Savings：1世帯当たりの平均貯蓄額
- Income　：平均所得額
- Office　：300人以上事業所割合(%)
- Work　：45歳以上の就業者割合(%)
- Single：未婚者の割合(%)
- Senior：65歳以上の割合(%)

関数rownames()を用いて，道府県名(Pref)を行名にする.

Input

```
> rownames(dat) <- dat$Pref
> head(dat)
```

Output

```
       Pref Savings Income Office Work   Single Senior
北海道 北海道   11918   2538   0.16 32.9 25.94550   28.1
青森県 青森県    8624   2358   0.10 29.7 24.81216   29.0
(以下省略)
```

Input

```
> dat <- dat[,-1]
> head(dat)
```

Output

```
      Savings Income Office Work   Single Senior
北海道   11918   2538   0.16 32.9 25.94550   28.1
青森県    8624   2358   0.10 29.7 24.81216   29.0
岩手県   12689   2671   0.12 34.5 24.46407   29.6
宮城県   12154   2863   0.18 32.3 27.45849   24.6
秋田県   10419   2409   0.10 31.1 22.07448   32.6
山形県   12640   2539   0.11 28.3 22.94568   29.9
```

```
      Savings Income Office Work   Single Senior
北海道   11918   2538   0.16 32.9 25.94550   28.1
・・・
```

1世帯当たりの平均貯蓄額(Savings)を応答変数，その他の変数を説明変数とした重回帰分析を行う．

重回帰分析(および単回帰分析)を実施する関数は，`lm()`である．書式は次の通りである．

`lm(`*formula*`,` `data=`*dataframe*`)`

- 引数*dataframe*は，重回帰分析に用いるデータフレーム名を表している．

- 引数*formula*の記載方法は次の通りである．

応答変数 ~ 説明変数1 + 説明変数2 + ・・・

なお，データフレーム名において，応答変数以外の変数をすべて説明変数に用いる場合には，「.」で短縮できる．

[今回のFormulaの記載方法1]

```
Savings ~ Income + Office + Work +Single + Senior
```

[今回のFormulaの記載方法2]

```
Savings ~.
```

Input

```
> mdl <- lm(Savings~., data=dat)
> summary(mdl)
```

関数summary()は，重回帰分析の結果を表示するための関数である．

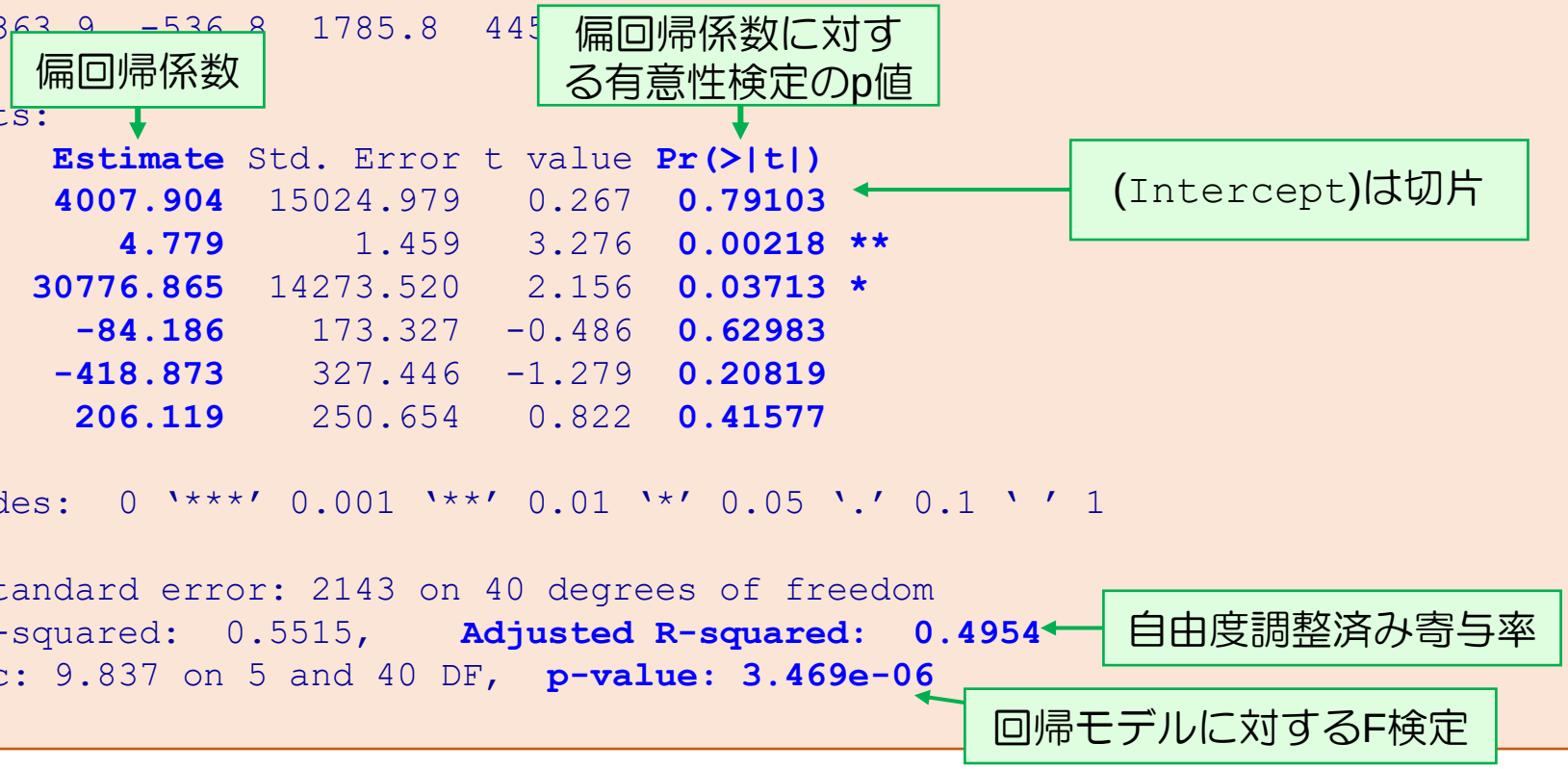
Note：見て欲しい部分を青色で表示しています．

```
Call:
lm(formula = Savings ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3354.7 -1363.9  -536.8  1785.8  445

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4007.904  15024.979   0.267  0.79103
Income          4.779      1.459   3.276  0.00218 **
Office      30776.865  14273.520   2.156  0.03713 *
Work          -84.186    173.327  -0.486  0.62983
Single       -418.873    327.446  -1.279  0.20819
Senior        206.119    250.654   0.822  0.41577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2143 on 40 degrees of freedom
Multiple R-squared:  0.5515,    Adjusted R-squared:  0.4954
F-statistic: 9.837 on 5 and 40 DF,  p-value: 3.469e-06
```



偏回帰係数の95%信頼区間を計算する

偏回帰係数の95%信頼区間は，関数`confint()`を用いて計算できる．

`confint(`*model*`)`

- 引数*model*は，関数`lm()`で推定されたモデルを表している．

Input

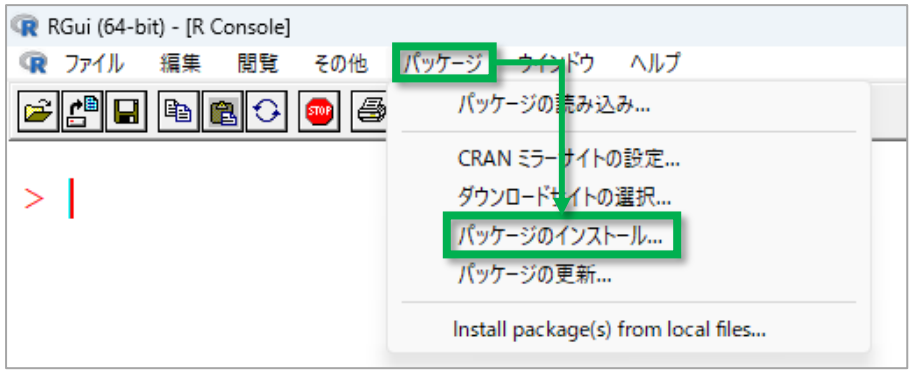
```
> confint(mdl)
```

Output

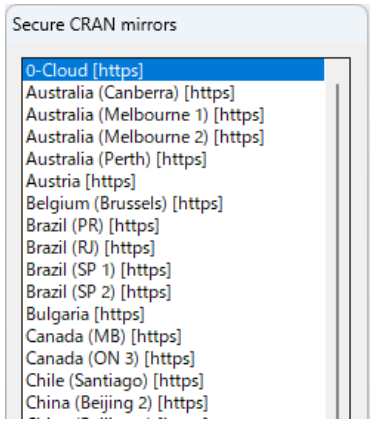
```
                   2.5 %       97.5 %
(Intercept) -26358.71169 34374.519157
Income           1.83052     7.728461
Office        1929.00599 59624.724941
Work          -434.49346   266.121059
Single       -1080.66701   242.920756
Senior        -300.47079   712.708579
```

多重共線性を確認するためのIVFを計算する.

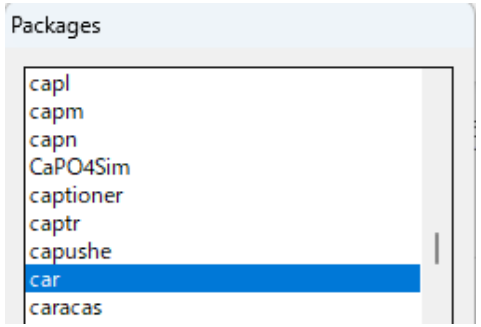
IVFを計算するための関数は，Rのデフォルトのパッケージには存在せず，パッケージcarをインストールしなければならない.



「パッケージ」→「パッケージのインストール」を選択



ダウンロードサイトを選択(どこでもいい)



carを選択してOKボタンを押す

一度インストールすれば，そのパソコンで同じ作業を行う必要はない.

なお，パッケージ名がわかる場合には，`install.packages("car")`で行えば，ダウンロードサイトを選択するだけで良い.

パッケージを利用するには，次のように入力して読み込まなければならない.

Input

```
> library(car)
```

パッケージのより見込みは，利用する場合には，その都度行わなければならない(ただし，Rの起動中は再読込は不要).

VIFの計算は，関数`vif()`を用いて計算できる．

`vif(`*model*`)`

- 引数*model*は，関数`lm()`で推定されたモデルを表している．

Input

```
> vif(mdl)
```

Output

```
  Income   Office     Work   Single   Senior
2.068310 4.212760 1.229495 4.390321 4.271970
```

VIFが10を超える変数がないことから，多重共先生は存在しないことが示唆される．

AICを用いた変数減少法による変数選択を行う

変数選択には，パッケージMASSを用いる(MASSパッケージはデフォルトで入っているので，インストールの必要はない).

Input

```
> library(MASS)
> sel <- stepAIC(mdl)
```

関数stepAIC()の書式は以下の通りである.

confint (*model, direction, k*)

- *model*：関数lm()で推定されたモデル
- *direction*：ステップワイズ法のアルゴリズム

  direction = "forward" ：変数増加法

  direction = "backward" ：変数減少法 (default)

  direction = "both" ：変数増減法

- *k*：情報量規準

  k = 2 ：赤池の情報量規準 (AIC)

  k = log(n) ：Bayes流情報量規準 (BIC)

関数stepAIC()の出力(上記)は，ステップワイズ法のプロセスを表しているが，意味はないので無視しても問題ない.

Input

```
> summary(sel)
```

```
Call:
lm(formula = Savings ~ Income + Office + Single, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3186.1 -1460.8  -572.6  1721.2  4601.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12253.899   6608.094   1.854  0.07072 .
Income          4.806      1.427   3.369  0.00163 **
Office      26607.592  12959.221   2.053  0.04632 *
Single       -597.204    228.918  -2.609  0.01253 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2111 on 42 degrees of freedom
Multiple R-squared:  0.5429,    Adjusted R-squared:  0.5103
F-statistic: 16.63 on 3 and 42 DF,  p-value: 2.863e-07
```

Work，Seniorがモデルから除外され，すべての説明変数における偏回帰係数に対する検定が有意になっていることに注目(有意にならない場合もある)

全データでの自由度調整済み寄与率が0.4954なので，わずかに寄与率が上昇している.

重回帰分析のための便利マクロを使ってみる

いままでに説明した工程を一度に処理するためのマクロを用意したので，それを使ってみることにする．

`Multiple.regress` (*y, cov, dir, IC, Scale*)

- *y*：応答変数　　• *cov*：説明変数

- *dir*：ステップワイズ法のアルゴリズム

  `direction = "forward"` ：変数増加法　　`direction = "backward"` ：変数減少法 (default)

  `direction = "both"` ：変数増減法

- *IC*：情報量規準

  `IC = "AIC"` ：赤池の情報量規準 (AIC) (default)　　`IC = "BIC"` ：Bayes流情報量規準 (BIC)

- *Scale*：標準化するか否か

  `Scale = "Y"` ：標準化する (default)　　`Scale = "N"` ：標準化しない

Note：マクロ`Multiple.regress()`では，パッケージ`MASS`, `aod`，`car`が必要である．`aod`，`car`は，デフォルトではインストールされていないので，インストールする必要がある．

標準偏回帰係数での結果 (AICによる変数減少法)

	dat[,1]	dat[,-1]				
	Savings	Income	Office	Work	Single	Senior
北海道	11918	2538	0.16	32.9	25.94550	28.1
青森県	8624	2358	0.10	29.7	24.81216	29.0
岩手県	12689	2671	0.12	34.5	24.46407	29.6
・・・	・・・・	・・・・・・・・・・・・・・・・・・・・				

Input

```
> library(MASS)
> library(aod)
> library(car)
> source("C:/Fukuoka_Seminor/Multiple.Regression.R")
> Y <- dat[,1]
> X <- dat[,-1]
> result <- Multiple.regress(Y,X)
```

標準化しない場合(通常の重回帰分析)では,

```
result <- Multiple.regress(Y,X,scale="N")
```

と入力する.

```
> names(result)
```

```
[1] "model"   "Fit"     "vif.all" "vif.sel"
```

Output (偏回帰係数の表示)

```
> result$model
```

```
                                Coef.All p.value.all                          Coef.SW p.value.sel
Income     0.499 [   0.191,    0.807]       0.001    0.502 [   0.201,    0.802]       0.001
Office     0.469 [   0.029,    0.908]       0.031    0.405 [   0.007,    0.803]        0.04
Work      -0.057 [  -0.294,    0.180]       0.627                                         -
Single    -0.284 [  -0.732,    0.165]       0.201   -0.405 [  -0.718,   -0.092]       0.009
Senior     0.180 [  -0.262,    0.622]       0.411                                         -
```

result$modelでは，変数選択前の偏回帰係数および95%信頼区間(Coef.All)およびp値(p.value.All)，ならびに，変数選択後の偏回帰係数および95%信頼区間(Coef.SW)およびp値(p.value.sel)が入っている.

Output (適合結果の表示)

```
> result$Fit
```

```
       Adj.Rsquared      p.value
All       0.4954192 3.468543e-06
Select    0.5102552 2.862790e-07
```

result$Fitでは，変数選択前後での自由度調整済み寄与率(Adj.Rsquared)およびF検定のp値(p.value) が入っている.

Output (変数選択前のVIF)

```
> result$vif.all
```

```
  Income   Office     Work   Single   Senior
2.068310 4.212760 1.229495 4.390321 4.271970
```

Output (変数選択後のVIF)

```
> result$vif.sel
```

```
  Income   Office   Single
2.037500 3.577859 2.210727
```

発展的な回帰分析：ロジスティック回帰分析

応答変数の種類によって回帰分析の名前は変わる

名前	応答変数の形式	例	係数の解釈
重回帰分析	量的	人口の変化量	回帰係数 (標準回帰係数)
ロジスティック回帰分析	2値	高齢者の認知症の有無	オッズ比
ー 名義ロジスティック	名義	交通事故の種類	オッズ比
ー 比例オッズモデル	順序	介護度	オッズ比
Poisson回帰分析	計数	ある日の交通死亡事故の件数	率比
Cox比例ハザード・モデル	生存時間	がん患者の生存期間	ハザード比

■ (偏)回帰係数は説明変数の尺度に依存する．そのため，重回帰分析では，すべての変数を標準化したもとで計算する回帰モデルの係数は標準(偏)回帰係数(標準化係数)と呼ばれる．標準(偏)回帰係数の絶対値の大きさを利用することで，従属応答に対する各説明変数の影響を評価できる．

■ ロジスティック回帰，Poisson回帰，Cox比例ハザード・モデルでは，指数関数$A=\exp(\beta)$を計算することで，それぞれ，オッズ比，率比，ハザード比を計算できる．それぞれの解釈は下記のとおり：

- **オッズ比**：変数Xが1上がるとA倍$y=1$になる(例：A倍治療が成功する)．
- **率比**：変数Xが1上がるとA倍計数が上がる(例：A倍ポリープが検出される)．
- **ハザード比**：変数Xが1上がるとA倍イベントリスクが高まる(例：A倍死亡リスクがあがる)

オッズ比とは

	満足	非満足	合
取水地A	291 (O_{11})	125 (O_{12})	p_A=0.700
取水地B	270 (O_{21})	146 (O_{22})	p_B=0.649

要因

結果

取水地A：井戸水
取水地B：河川水

オッズとは，ある結果が生じる比率とある結果が生じない比率との比である．

取水地A(要因あり)のオッズ：$\mathrm{odds_A} = \frac{p_A}{1-p_A}$，　取水地B(要因なし)のオッズ：$\mathrm{odds_B} = \frac{p_B}{1-p_B}$

$$\frac{0.700}{1-0.700} = 2.333$$

$$\frac{0.649}{1-0.649} = 1.849$$

オッズ比$\mathrm{odds_A}/\mathrm{odds_B}$は，結果(アウトカム)に対してどれぐらい要因が寄与しているかを表す．

➡ 要因があるとない場合に比べて〇〇倍の結果が生じるか．

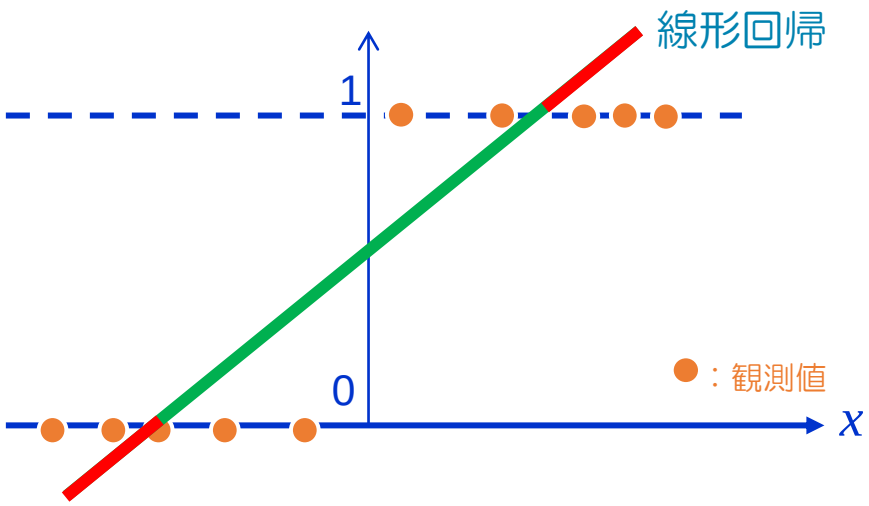
オッズ比の公式

$$OR = \frac{O_{11} \times O_{22}}{O_{12} \times O_{21}}$$

事例の場合には，$OR = \frac{O_{11} \times O_{22}}{O_{12} \times O_{21}} = \frac{291 \times 146}{125 \times 270} = 1.259$
なので，

取水地Aの水道水は取水地Bに比べて1.259倍満足されているといえる．

ロジスティック回帰分析の動機



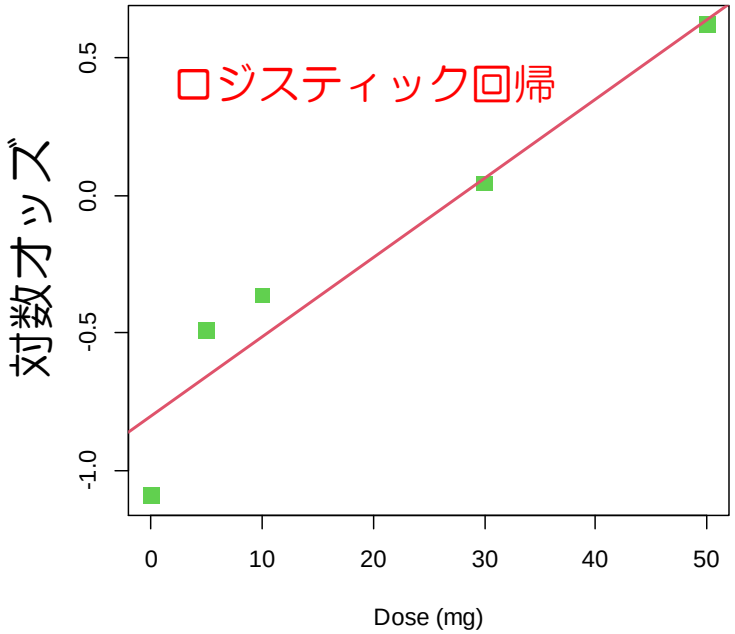
2値で与えられた応答変数に線形回帰分析を行うと．．．

➡応答変数の確率が1を超えたり，0を下回ってしまう(赤色の部分)

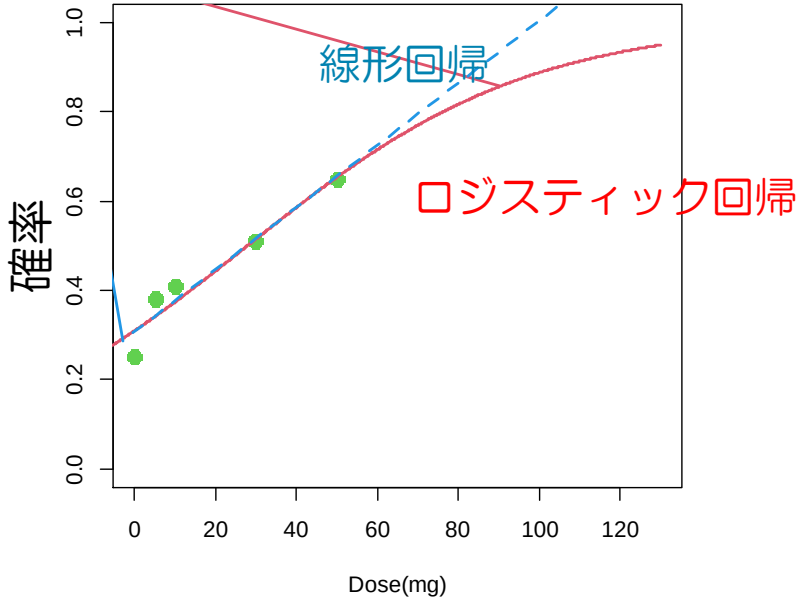
応答変数に対する対数オッズに対する回帰モデルを考える

$$\log\frac{\Pr(y=1)}{1-\Pr(y=1)}=\log\frac{p}{1-p}=\beta_0+\beta_1 x$$

薬剤量(mg)	有効	無効	有効率
0	79	235	0.252
5	81	132	0.380
10	169	243	0.410
30	318	305	0.510
50	379	204	0.650



確率pに変換



ロジスティック回帰では，対数オッズに対して直線を当てはめる(最小2乗法ではない点に注意)．

確率pに変換した後に描かれる曲線のことを，ロジット曲線(ロジスティック曲線)という．

ロジスティック回帰分析における調整オッズ比 (1/2)

いま，取水地(井戸水，河川水）とともに，被験者の年齢（65歳以上／65歳未満)についても調査した．
そのときの，水道水に対する満足度について調査した．

年齢	取水地	満足／不満足		
		満足(1)	不満足(0)	合計
65歳以上(1)	井戸水(1)	77	7	84
	河川水(0)	65	18	83
	合計	142	25	167
65歳未満(0)	井戸水(1)	70	17	87
	河川水(0)	60	27	87
	合計	130	44	174

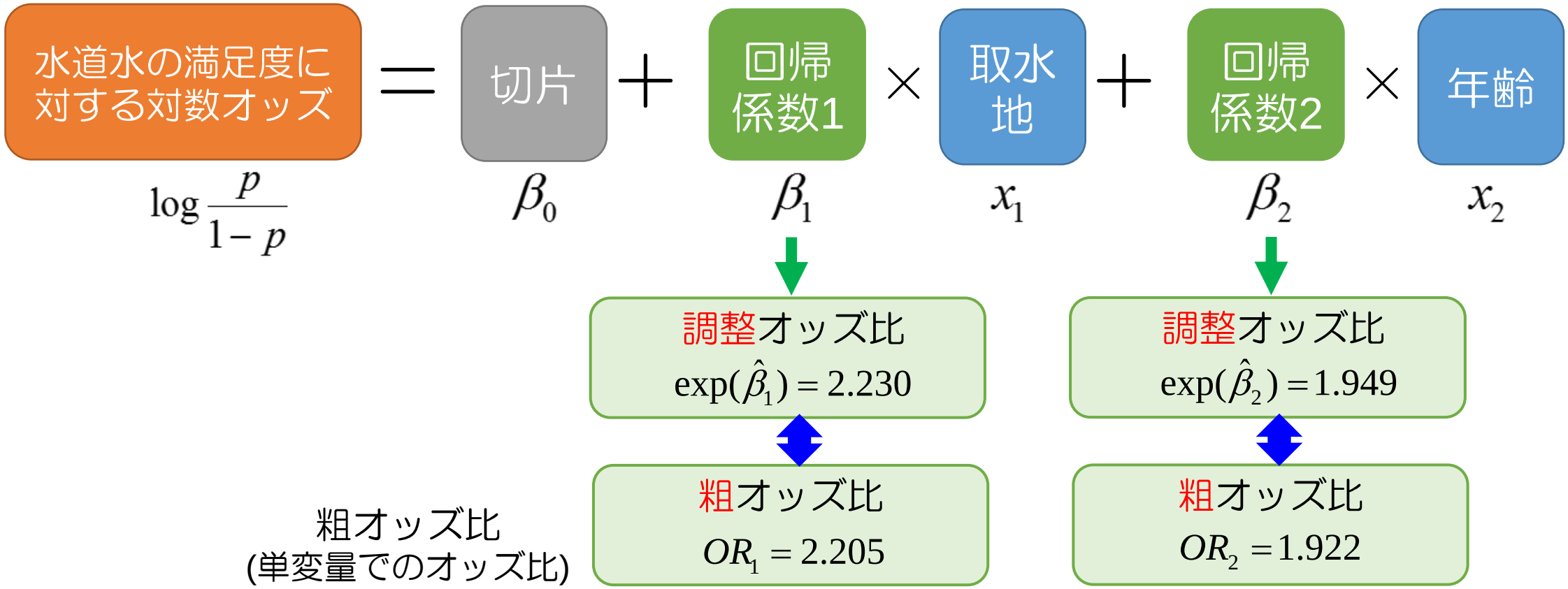
(多重)ロジスティック回帰分析のモデルは

水道水の満足度に対する対数オッズ $\log \frac{p}{1-p}$ ＝ 切片 β_0 ＋ 回帰係数1 β_1 × 取水地 x_1 ＋ 回帰係数2 β_2 × 年齢 x_2

井戸水：1
河川水：0

65歳以上：1
65歳未満：0

ロジスティック回帰分析における調整オッズ比 (2/2)



粗オッズ比は，他の要因(説明変数)を何も考慮しない．これに対して，調整オッズ比では，

- 取水地のオッズ比で年齢による違いを考慮 (2.230倍 水道水のほうが河川水よりも満足)
- 年齢のオッズ比では取水地による違いを考慮 (1.922倍 65歳以上のほうが65歳未満よりも満足)

のように調整したうえで，オッズ比を解釈できる．

3水準以上のカテゴリカル変数(順序・名義尺度)の取り扱い

ある説明変数が3水準(A，B，C)が存在する場合，ロジスティック回帰モデルでは，2個の変数で表現する

変数B　カテゴリがBならば1，それ以外は0　　変数C　カテゴリがCならば1，それ以外は0

ロジスティック回帰モデル：$\log\frac{p}{1-p} =$ その他の独立変数 ＋ $\beta_B \times$ 変数B ＋ $\beta_C \times$ 変数C

	変数B	変数C
水準Aの場合	0	0
水準Bの場合	1	0
水準Cの場合	0	1

例：交通事故の発生の有無に，天気が(晴，曇，雨)が影響しているかどうかを検討している場合，

事故有に対する対数オッズ ＝ $\beta_{曇}$ × 曇か否か ＋ $\beta_{雨}$ × 雨か否か

で表される．このとき，次のように解釈できる．

$\exp(\beta_{曇})$　:曇りの日は，晴れの日よりも何倍，交通事故が発生するか．

$\exp(\beta_{雨})$　:雨の日は，晴れの日よりも何倍，交通事故が発生するか．

におけるロジスティック回帰分析

ここでは，Cドライブの`Fukuoka_Seminor`というフォルダにある`logistic_regress.csv`というCSVファイルを読み込む．これは，e-statを用いて作成した46都道府県(東京都を除く)の1世帯当たりの平均貯蓄額(14,500万円以上，14,500万円未満)，平均所得額(多，中，少)，未婚者の割合(中央値以上，中央値未満)である．

Input

```
> dat <- read.csv("C:/Fukuoka_Seminor/logistic_regress.csv")
> head(dat)
```

今回は，`fileEncoding = "cp932"`を引数にしていないが，これは，ファイル`logistic_regress.csv`に日本語が入っていないためである．

Output

```
 Savings Income Single
1       0      B      1
2       0      A      0
3       0      B      0
4       0      B      1
5       0      A      0
6       0      B      0
```

`Savings` ：1世帯当たりの平均貯蓄額 (1：14,500万円以上，0：14,500万円未満)
`Income` ：平均所得額(A：少ない，B：中程度，C：多い)
`Single` ：未婚者の割合(1：中央値以上，0：中央値未満)

1世帯当たりの平均貯蓄額(Savings)を応答変数，その他の変数を説明変数としたロジスティック回帰分析を行う．

ロジスティック回帰分析を実施するには，パッケージMASS(デフォルトでインストールされている)のなかの関数glm()を用いる．関数glm()は，一般化線型モデル(GLM; Generalized linear model)の関数だが，ロジスティック回帰分析は，一般化線型モデルに包含される回帰分析の一つである．

```
glm(formula, data=dataframe, family=binomial())
```

- 引数*dataframe*は，重回帰分析に用いるデータフレーム名を表している．
- 引数*formula*の記載方法は次の通りである．

応答変数 ~ 説明変数1 + 説明変数2 + ・・・

なお，データフレーム名において，応答変数以外の変数をすべて説明変数に用いる場合には，「.」で短縮できる．

[今回のFormulaの記載方法1]

```
Savings ~ Income + Single
```

[今回のFormulaの記載方法2]

```
Savings ~.
```

Note：ロジスティック回帰分析は，応答変数が2値の場合のみに用いることができるので，2値以上の場合にはエラーになる．なお，関心のある事象を1，そうでない事象を0とすれば，1が起きる確率として計算されるので，データをそのようにコーディングすることが推奨される．

```
> library(MASS)  パッケージMASSの読み込み
> mdl <- glm(Savings~., data=dat, family=binomial())  ロジスティック回帰の実行
> summary(mdl)  結果の表示
```

```
Call:
glm(formula = Savings ~ ., family = binomial(), data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9132  -0.6128   0.5913   1.0102   1.8946

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.57742    0.82263  -1.918  0.05517 .
IncomeB      1.98430    0.89043   2.228  0.02585 *
IncomeC      3.23287    1.10341   2.930  0.00339 **
Single      -0.07803    0.68495  -0.114  0.90930
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 63.421  on 45  degrees of freedom
Residual deviance: 51.381  on 42  degrees of freedom
AIC: 59.381

Number of Fisher Scoring iterations: 4
```

偏回帰係数

偏回帰係数に対する有意性検定のp値

(Intercept)は切片

IncomeB (IncomeがBならば1，そうでないならば0)

IncomeC(IncomeがCならば1，そうでないならば0)

偏回帰係数の95%信頼区間は，重回帰分析と同様に関数confint()を用いて計算できる．

```
> confint(mdl)
```

```
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) -3.522290 -0.1389853
IncomeB      0.384716  4.0175135
IncomeC      1.275327  5.7186652
Single      -1.457794  1.2667811
```

ロジスティック回帰分析では，偏回帰係数ではなく，オッズ比で解釈することが多い．オッズ比(厳密には調整オッズ比)は，次のように計算できる．

```
> Coef <- mdl$coefficients  偏回帰係数を変数Coefに代入
> CI <- confint(mdl)95%信頼区間を変数CIに代入
> Coef.CI <- cbind(Coef, CI) 変数Coefと変数CIを列方向に連結
> exp(Coef.CI) オッズ比を計算
```

```
                  Coef      2.5 %      97.5 %
(Intercept)  0.2065080 0.02953173   0.8702408
IncomeB      7.2739304 1.46919697  55.5627798
IncomeC     25.3522423 3.57987160 304.4982183
Single       0.9249324 0.23274904   3.5494088
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.57742    0.82263  -1.918  0.05517 .
IncomeB      1.98430    0.89043   2.228  0.02585 *
IncomeC      3.23287    1.10341   2.930  0.00339 **
Single      -0.07803    0.68495  -0.114  0.90930
```

偏回帰係数に対する有意性検定より，以下のように解釈できる

- 平均所得額(Income)のp値がいずれも0.05未満なので，平均所得額によって，平均貯蓄額が14,500万円以上の住民の割合に違いが認められる.
- 独身者(Single)のp値がいずれも0.05以上なので，独身者によって，平均貯蓄額が14,500万円以上の住民の割合に違いが認められない.

```
                 Coef      2.5 %      97.5 %
(Intercept)  0.2065080 0.02953173   0.8702408
IncomeB      7.2739304 1.46919697  55.5627798
IncomeC     25.3522423 3.57987160 304.4982183
Single       0.9249324 0.23274904   3.5494088
```

オッズ比は次のように解釈できる.

- 平均所得額が中程度(B)の県は，低い県(A)に比べて，平均貯蓄額が14,500万円以上の住民が約7.3倍多い.
- 平均所得額が高い(C)の県は，低い県(A)に比べて，平均貯蓄額が14,500万円以上の住民が約25.4倍多い.
- 独身者が中央値以上の県は中央値未満の県に比べて，平均貯蓄額が14,500万円以上の住民が約0.9倍多い.

変数選択は，重回帰分析と同様に関数stepAIC ( ) を用いて計算できる.

```
> Sel <- stepAIC(mdl)   変数選択を実行 (AICを用いた変数減少法)
> summary(sel)  変数選択結果を表示
```

```
Call:
glm(formula = Savings ~ Income, family = binomial(), data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8930  -0.6039   0.6039   1.0258   1.8930

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6094     0.7746  -2.078   0.0377 *
IncomeB       1.9772     0.8877   2.227   0.0259 *
IncomeC       3.2189     1.0954   2.938   0.0033 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 63.421  on 45  degrees of freedom
Residual deviance: 51.394  on 43  degrees of freedom
AIC: 57.394

Number of Fisher Scoring iterations: 4
```

Singleがモデルから除外され，すべての説明変数における偏回帰係数に対する検定が有意になっていることに注目 (有意にならない場合もある)

```
> Coef <- Sel$coefficients  偏回帰係数を変数Coefに代入
> CI <- confint(Sel)95%信頼区間を変数CIに代入
> Coef.CI <- cbind(Coef, CI) 変数Coefと変数CIを列方向に連結
> exp(Coef.CI) オッズ比を計算
```

変数選択後 (今回の結果)

	Coef	2.5 %	97.5 %
(Intercept)	0.200000	0.0307379	0.7587157
IncomeB	7.222222	1.4657381	54.8896029
IncomeC	25.000000	3.5804226	295.5141332

変数選択前 (前回の結果)

	Coef	2.5 %	97.5 %
(Intercept)	0.2065080	0.02953173	0.8702408
IncomeB	7.2739304	1.46919697	55.5627798
IncomeC	25.3522423	3.57987160	304.4982183
Single	0.9249324	0.23274904	3.5494088

わずかではあるが，変数選択後に調整オッズ比が変化している．変数選択後のほうの結果のほうが，信頼性がおけるので，こちらを採用した方が良い．

ロジスティック回帰分析のための便利マクロを使ってみる

いままでに説明した工程を一度に処理するためのマクロを用意したので，それを使ってみることにする．

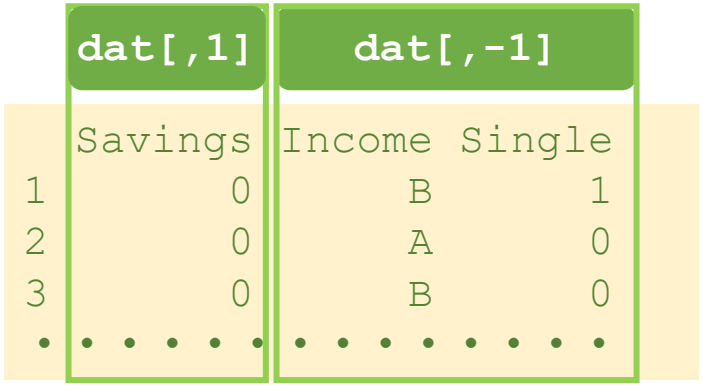
`Logistic.Mult(`*y, cov, dir, IC*`)`

- *y*：応答変数　　• *cov*：説明変数
- *dir*：ステップワイズ法のアルゴリズム
  `direction = "forward"` ：変数増加法　　`direction = "backward"` ：変数減少法 (default)
  `direction = "both"` ：変数増減法
- *IC*：情報量規準
  `IC = "AIC"` ：赤池の情報量規準 (AIC) (default)　　`IC = "BIC"` ：Bayes流情報量規準 (BIC)

Note：マクロ`Logistic.Mult()`では，パッケージ`MASS`, `aod`が必要である．　`aod`は，デフォルトではインストールされていないので，インストールする必要がある．

Input

```
> library(MASS)
> library(aod)
> source("C:/Fukuoka_Seminor/Logistic.Regression.R")
> Y <- dat[,1]
> X <- dat[,-1]
> result <- Logistic.Mult(Y,X)
> result
```



Output

```
                          OR.All p.value.all                          OR.SW p.value.sel
IncomeB    7.274 [ 1.469, 55.563]       0.012    7.222 [ 1.466, 54.890]       0.012
IncomeC 25.352 [ 3.580, 304.498]                25.000 [ 3.580, 295.514]
Single     0.925 [ 0.233,  3.549]       0.909
```

resultでは，変数選択前のオッズ比および95%信頼区間(OR.All)およびp値(p.value.all)，ならびに，変数選択後のオッズ比および95%信頼区間(OR.SW)およびp値(p.value.sel)である．

IncomeCのp.value.all，p.value.selには，p値が入っていない．これは，平均所得額(Income)が説明変数として意味があるか否かを評価しているためである．つまり，IncomeB，IncomeCの2つのダミー変数に意味があるか否かを評価している(glm()での偏回帰係数では，個々のダミー変数を評価している)．p値が有意水準0.05を下回っていることから，平均所得額(Income)に意味があると解釈できる．