入門 統計学　第12章

回帰分析
― 多変量解析① ―

『入門　統計学　第2版　ー検定から多変量解析・実験計画法・ベイズ統計学までー』（オーム社）

※注：本書を購入された方へのサービスですので，教科書指定（参考図書は不可）していない授業での使用はお控えください。

12.1 多変量解析とは？ (multivariate analysis)

- 複数の変数※を同時に扱う分析手法の総称

※変量と変数は同じと考えてOK

- 変数間の因果関係の解明や予測，変数の削減，標本（個体）の分類などが目的

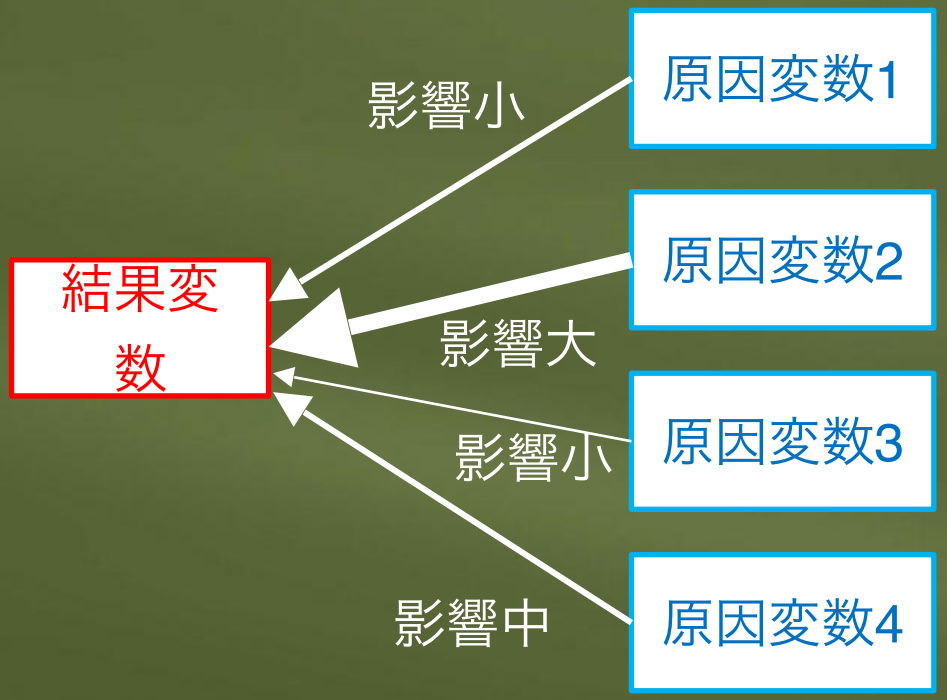
ひとまとめにして分析

多変量解析の目的と手法

	目　的（扱う章）	手　法
多変量解析	①因果関係の解明と予測（本章）	重回帰分析
	②分類（第13章）	ロジスティック回帰分析，クラスター分析
	③変数の削減（第14章）	主成分分析，因子分析

目的①

因果関係の解明と予測



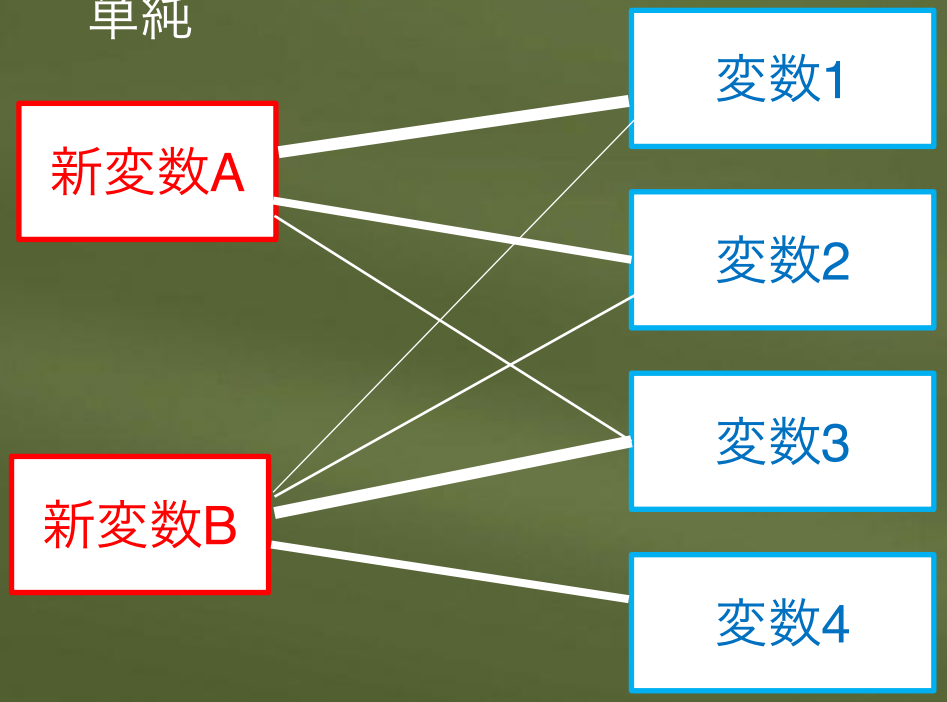
因果関係をモデル（数式）で捉え，予測に利用する

目的②

変数（次元）の削減

変数が沢山あって複雑

単純



総合指標を作成したり，構造を単純化する

目的③

分　類

個体や変数をグループ分けする

（手法によってはどちらに入るのか予測もできる）

12.2 回帰分析

(regression analysis)

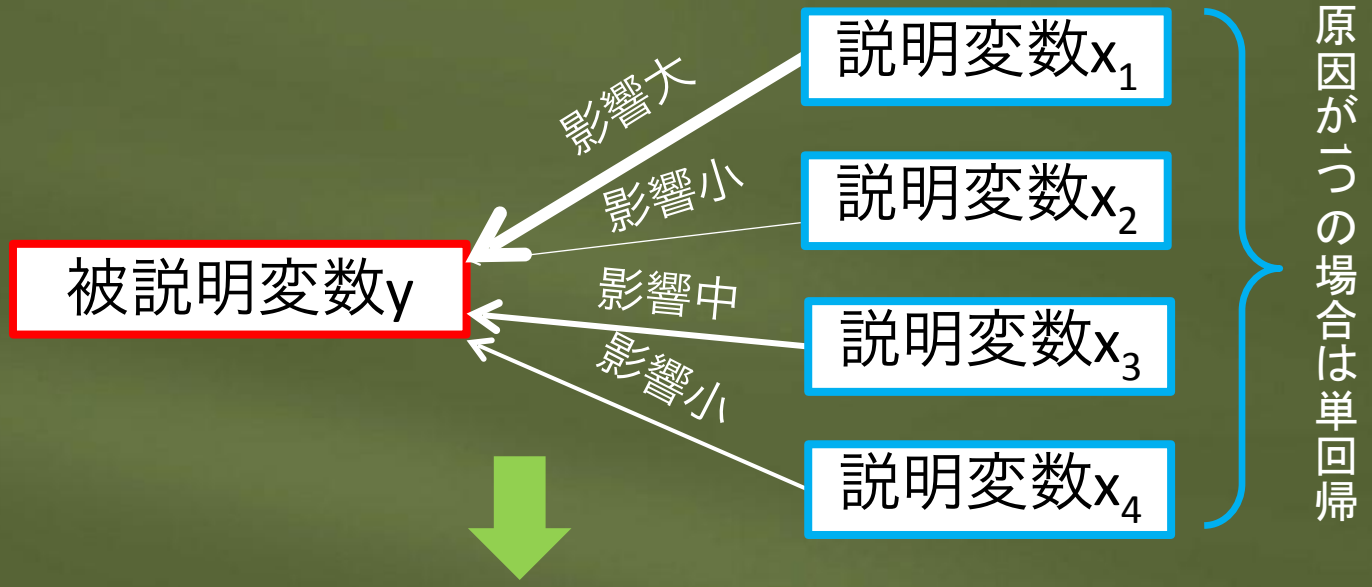
- 因果関係を解明する最も基本的手法
- 複数の原因変数から，1つの結果変数を説明するモデル（数式）を特定

  →原因が1つの場合は単回帰，複数は重回帰と呼ぶ

- 推定方法は最小２乗法（OLS)が基本
- 因果関係を把握すれば，予測も可能

重回帰分析の概念

結果（変数）　影響力（定数）　　原因（変数）



数式（モデル）で表す←予測にも使える

因果関係のある場合の変数の呼び方

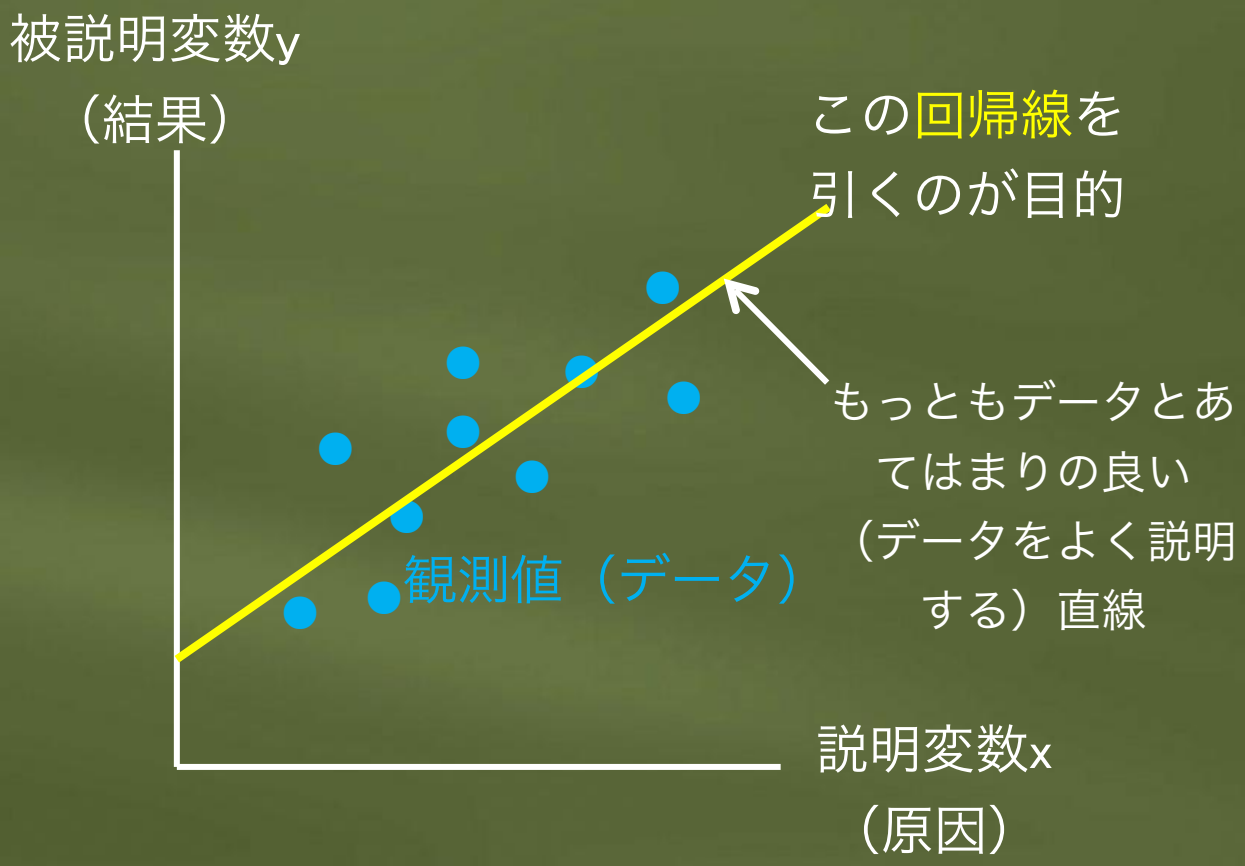
結　果	原　因
従属変数　（dependent variable）	独立変数　（independent variable）
目的変数　（objective variable）	説明変数　（explanatory variable）
被説明変数（explained variable）	予測変数　（predictor variable）
応答変数　（response variable）	
結果変数　（outcome variable）	
基準変数　（criterion variable）	

本章ではこの名称を使用

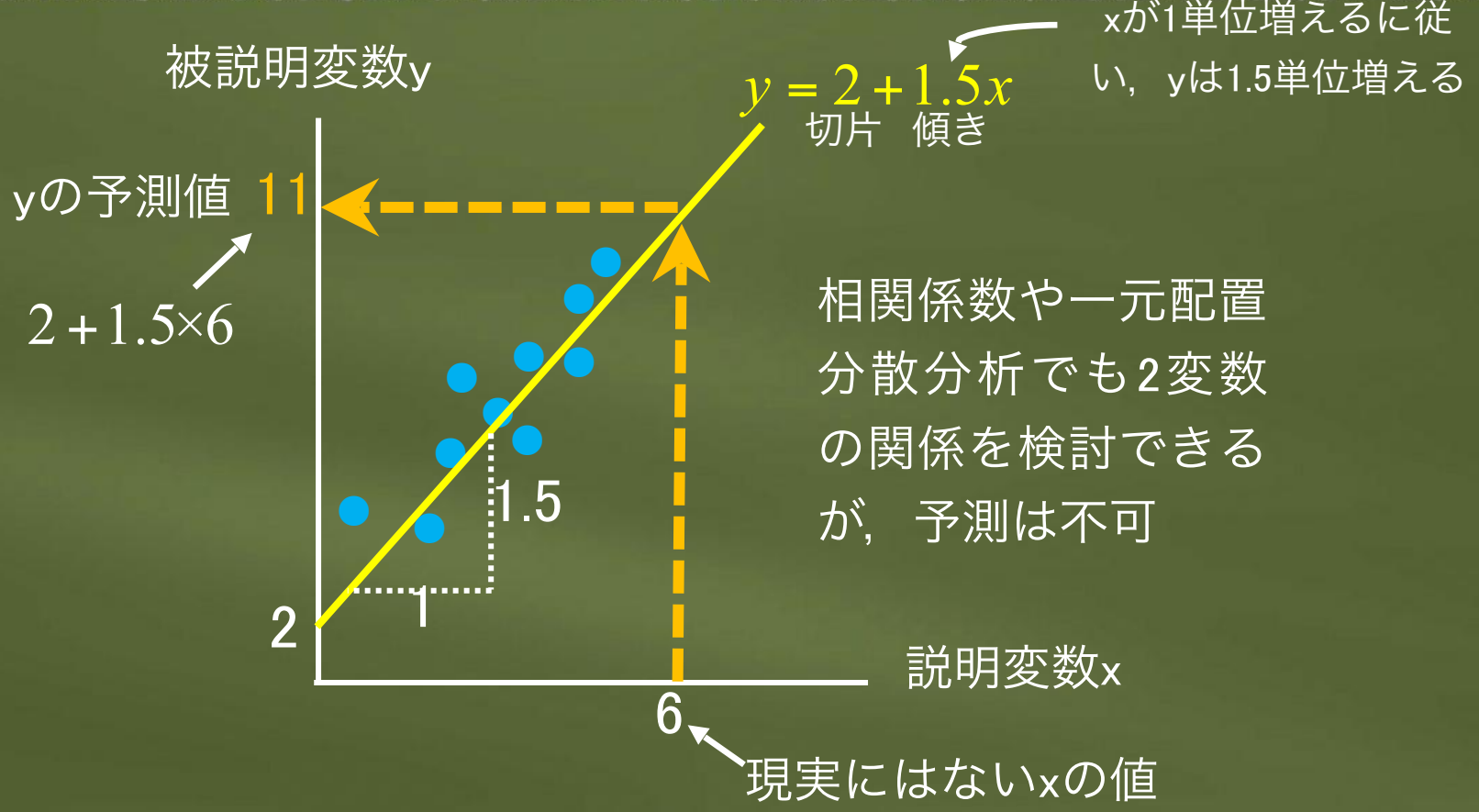
どれでも良いが手法によって大体決まっている

単回帰モデル（回帰線）

～説明変数が1つ～



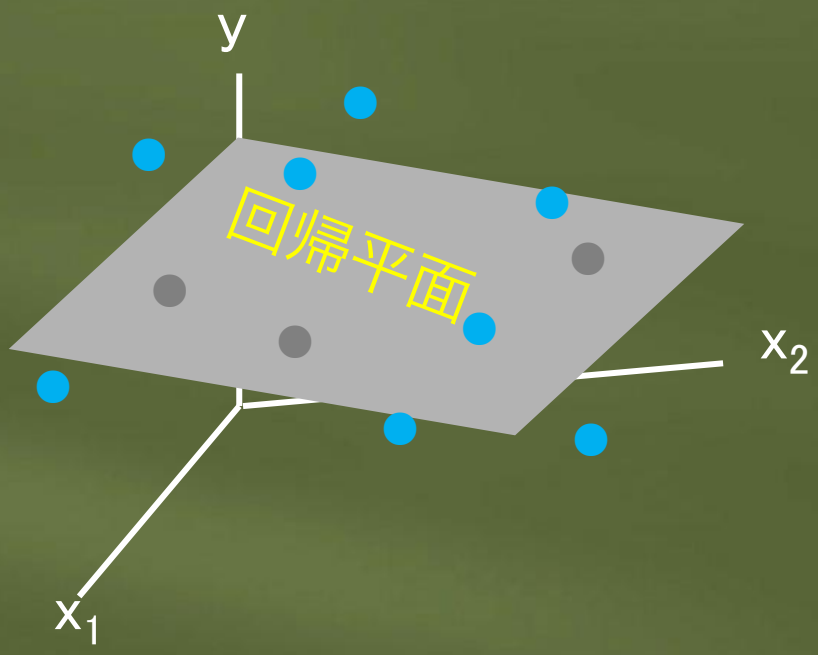
回帰線が特定されれば予測も可能



相関係数や一元配置分散分析でも2変数の関係を検討できるが，予測は不可

重回帰モデル（回帰平面）

～説明変数が複数～



説明変数xが3つ以上は回帰超平面となり図示不可

回帰線や回帰平面の式の書き方

傾き　切片

中学校で習った直線式 $y = ax + b$

①xが複数出てくるので，傾きと説明変数は切片の後ろ

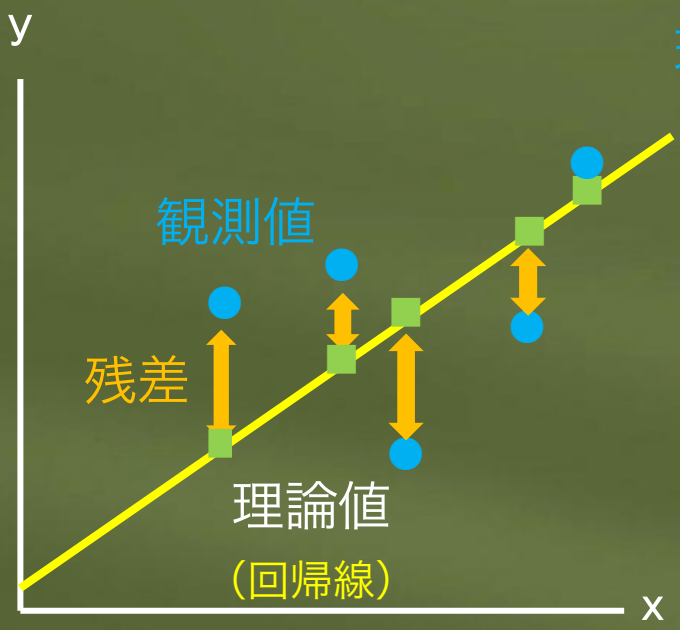
②変数をアルファベット，定数をギリシャ文字

回帰線の式　$y = \alpha + \beta x$

回帰平面の式　$y = \alpha + \beta_1 x_1 + \beta_2 x_2$

定数項　回帰係数　2は2番目の変数という意味

モデル式には誤差項を含める



ズレ

現実 ⟷ 理論　　誤差項

$$y_i = \alpha + \beta x_i + u_i$$

i番目の変数の値

単回帰モデル
（誤差項を含めることで観測値と直線との関係を全て表現）

注：誤差uとは母集団の概念（観測できない）で,
その実現値（実際に観測される値）を残差$\hat{u}$と呼ぶ

回帰モデルと推定式

単回帰モデル $y_i = \alpha + \beta x_i + u_i$

現実の値

真の値（母数；パラメータ）

誤差項

観測値（データ）から推定

←直線式なので誤差項は付けない

推定回帰線 $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

（^は推定値の意味）

これを特定することが回帰分析の目的

$\hat{y}_i$：被説明変数の理論値

$\hat{\alpha}$：定数項の推定値

$\hat{\beta}$：回帰係数の推定値

では，どのようにしてパラメータ（αやβ）を推定するのでしょうか？

12.3　パラメータの推定

- パラメータの推定法にはいくつかある

  →最尤法（次章で紹介），モーメント法など

- もっとも基本的最小２乗法（OLS）を解説

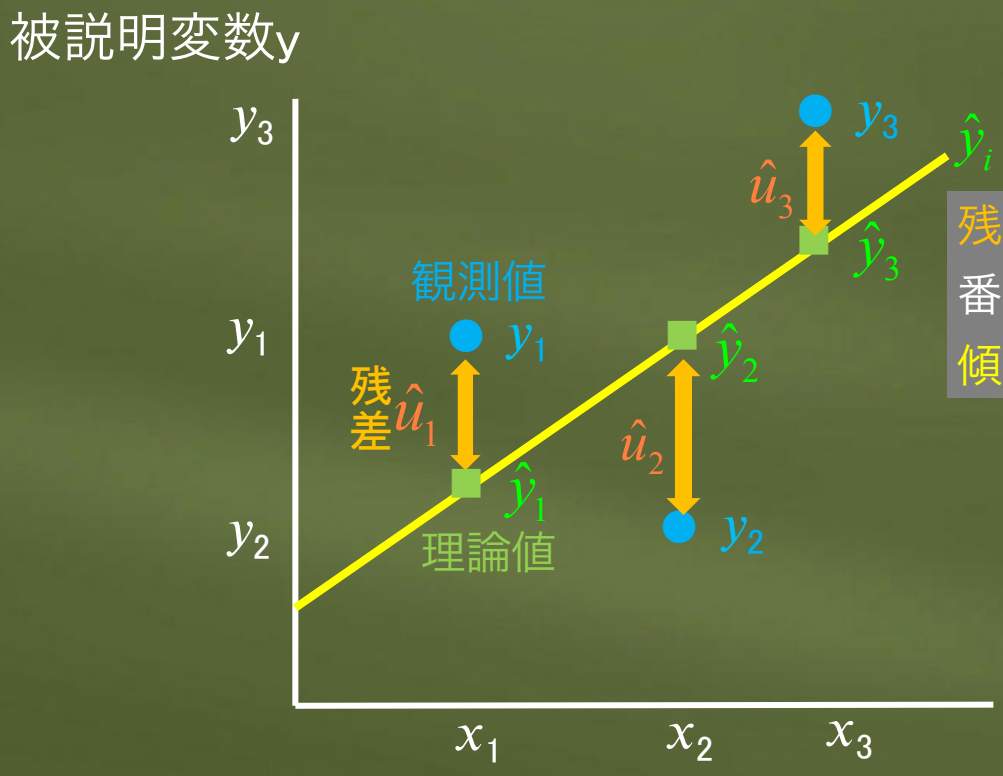
  理由①：標準的仮定（後掲）が満たされたとき

  適切な推定量が得られる

  理由②：連立方程式から求めることができる

単回帰でOLSを解説…

残差が一番小さくなるパラメータを見つける



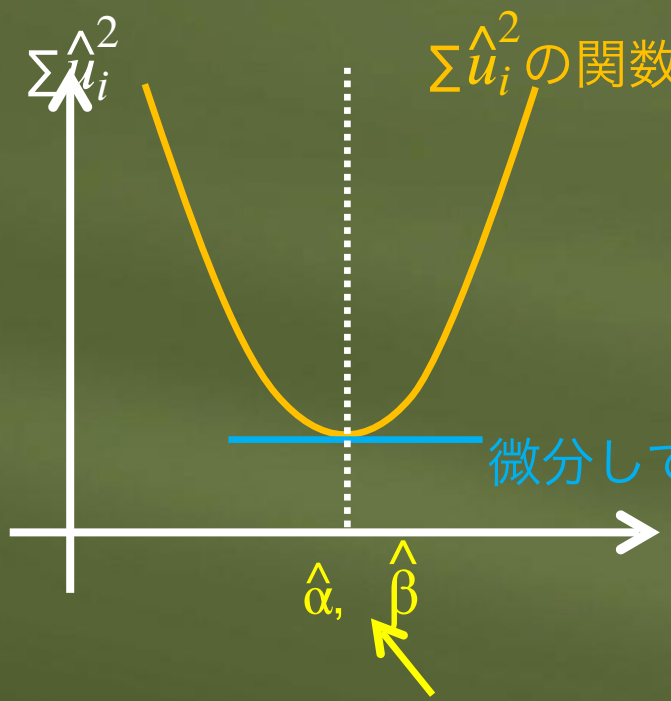
$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

残差（$\hat{u}_i = y_i - \hat{y}_i$）が番小さくなる切片 α や傾き β を見つける

残差には正負が混在するので平方（2乗）和 $\sum \hat{u}_i^2$ で考える

最小2乗法

（残差の2乗の和を最小とする）



導関数が接線の傾き
であることを利用

α（定数項），β（回帰係数）
↑これから値を探すので変数として扱う

正規方程式①

①残　差　$\hat{u}_i = y_i - y_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$

観測値　理論値　　推定回帰線

②残差平方和　$\sum \hat{u}_i^2 = \sum (y_i - y_i)^2 = \sum \left[y_i - (\hat{\alpha} + \hat{\beta} x_i) \right]^2$

2次関数の微分　$(y^2)' = 2y'y$

③$\hat{\alpha}$で偏微分　$\dfrac{\partial \sum \hat{u}_i^2}{\partial \hat{\alpha}} = \sum \left[2(y_i - \hat{\alpha} - \hat{\beta} x_i)'(y_i - \alpha - \beta x_i) \right] = -2 \sum (y_i - \alpha - \beta x_i)$

-1

注：偏微分とは，多変数関数を1つの変数のみで微分し，他の変数は定数として扱う。

$\hat{\beta}$で偏微分　$\dfrac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}} = \sum \left[2(y_i - \hat{\alpha} - \hat{\beta} x_i)(y_i - \alpha - \beta x_i)' \right] = -2 \sum x_i (y_i - \alpha - \beta x_i)$

$-x_i$

正規方程式②

④2本の導関数＝0とおく

$\Sigma\hat{\alpha}=n\alpha$

$$\begin{cases} \sum(y_i-\hat{\alpha}-\hat{\beta}x_i)=\sum y_i-n\alpha-\sum\beta x_i=0 \\ x_i\sum(y_i-\hat{\alpha}-\hat{\beta}x_i)=\sum x_iy_i-\alpha\sum x_i-\beta\sum x_i^2=0 \end{cases}$$

⑤両式を整理すると…**正規方程式**が得られる

$$\begin{cases} n\hat{\alpha}+\hat{\beta}\sum x_i=\sum y_i \\ \hat{\alpha}\sum x_i+\hat{\beta}\sum x_i^2=\sum x_iy_i \end{cases}$$

この連立方程式を解けばパラメータが求まる

注：説明変数が複数の重回帰では，連立させる方程式の本数が増えるだけ

12.4　モデルの評価

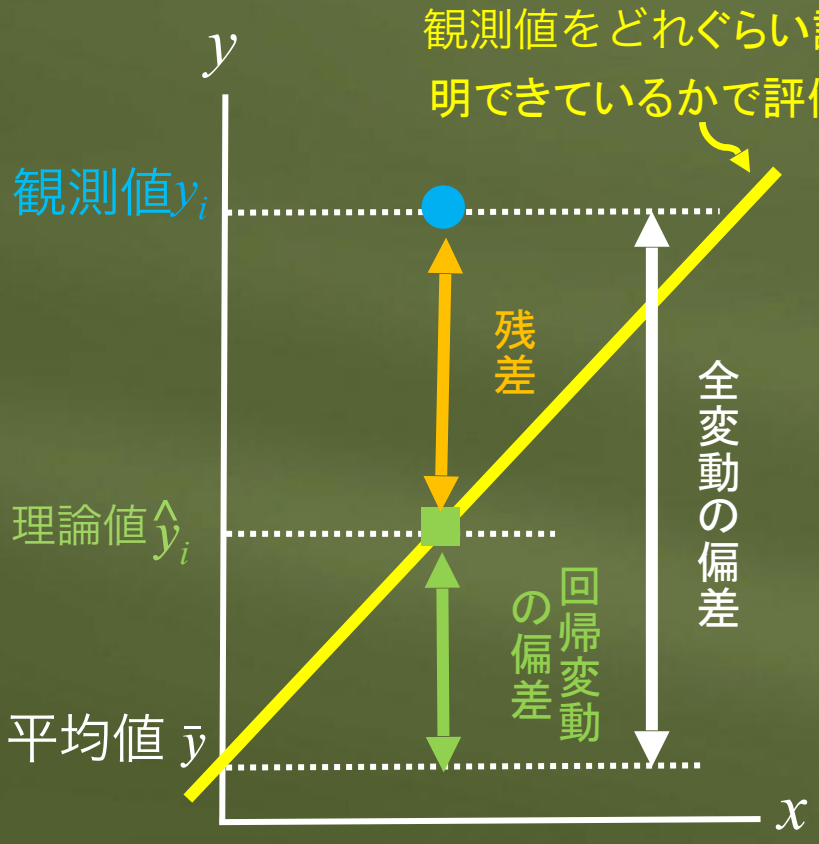
- どれぐらい良いモデルが推定できたのかを，2つの方法で評価

①決定係数：モデルが観測値をどれくらい説明できているかを（0～1の係数で）評価

②回帰係数の検定：回帰係数が統計的に有意かどうかを判定（定数項はあまり気にしない）

モデルの評価①

決定係数



回帰線で説明できる部分

$$\text{決定係数 } R^2 = \frac{\text{回帰変動}}{\text{全変動}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

観測値の動き

決定係数（$0 \leqq R^2 \leqq 1$）の解釈：

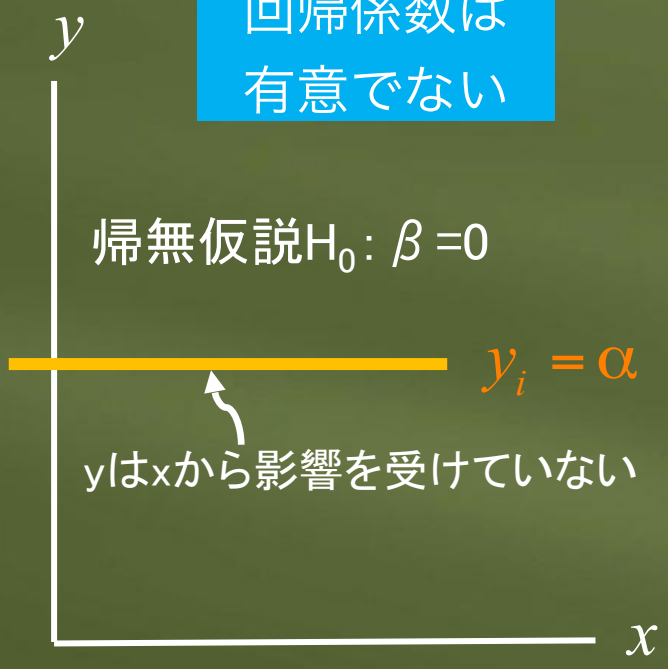
1に近いと「はてはまりが良い」，「適合度が高い」，「説明力が高い」

注：いくつ以上が良いという基準はない

モデルの評価②

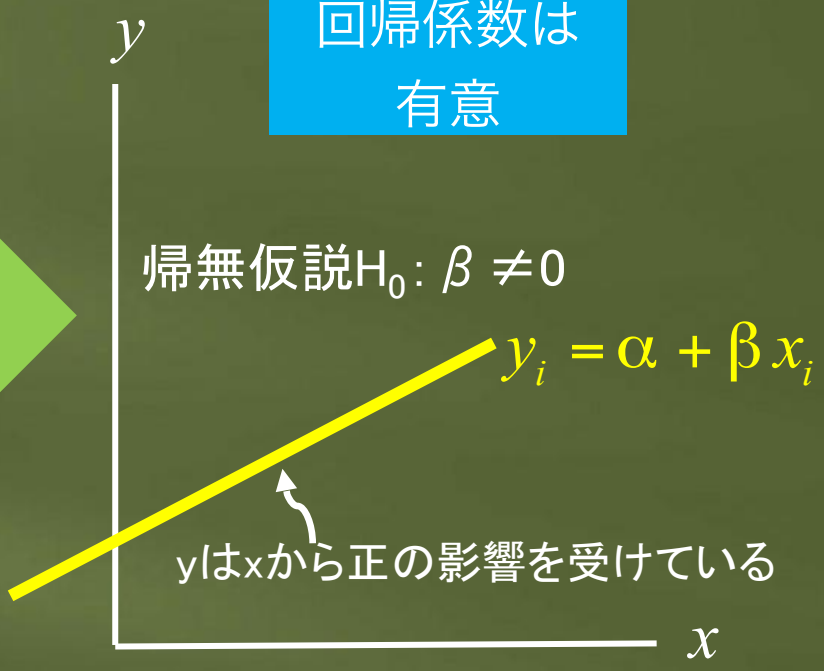
回帰係数の検定

回帰係数は
有意でない



検定

回帰係数は
有意



回帰係数の推定値は正規分布

$$y_i = \alpha + \beta x_i + u_i$$

$$u_i \sim N(0, \sigma^2)$$

誤差項は正規分布に従う
と仮定（次スライドで図示）

推定

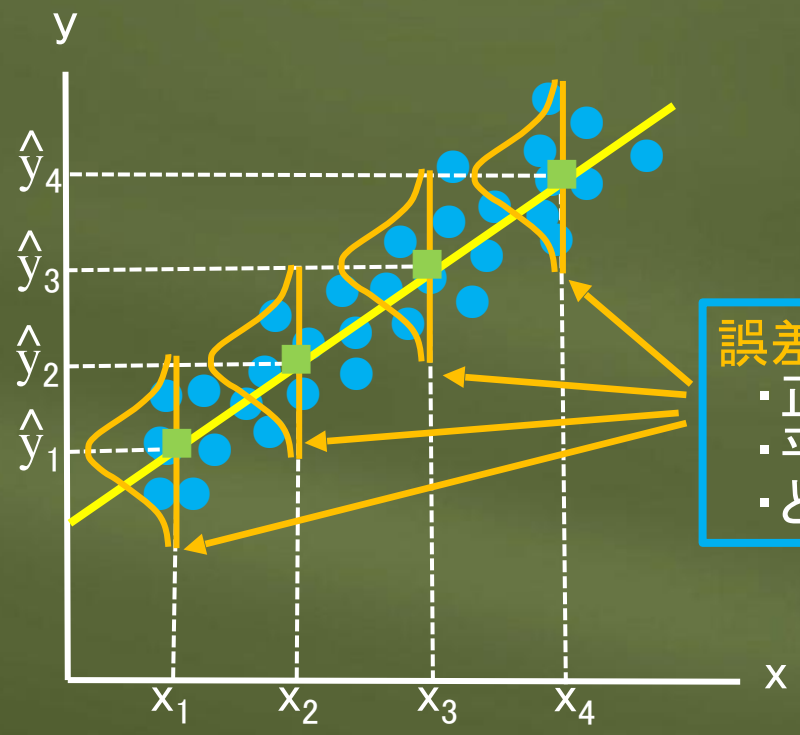
$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

推定値は誤差を含んでいるので，
パラメータの推定値も正規分布に従う

$$\hat{\alpha} \sim N(\alpha, \sigma_{\hat{\alpha}}^2) \qquad \hat{\beta} \sim N(\beta, \sigma_{\hat{\beta}}^2)$$

真のバラツキである母標準誤差は不明なので，不偏標準誤差を用いる→t分布を考える

誤差項の性質



誤差項（●と■のズレ）の性質
- 正規分布に従う
- 平均はゼロ
- どこでも分散は均一

回帰係数の検定統計量

帰無仮説は $\beta = 0$

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\sqrt{\dfrac{\sum \hat{u}_i^2 / (n-k-1)}{\sum (x_i - \bar{x})^2}}}$$

$\hat{\beta}$の不偏標準誤差

自由度
（kは説明変数の数）

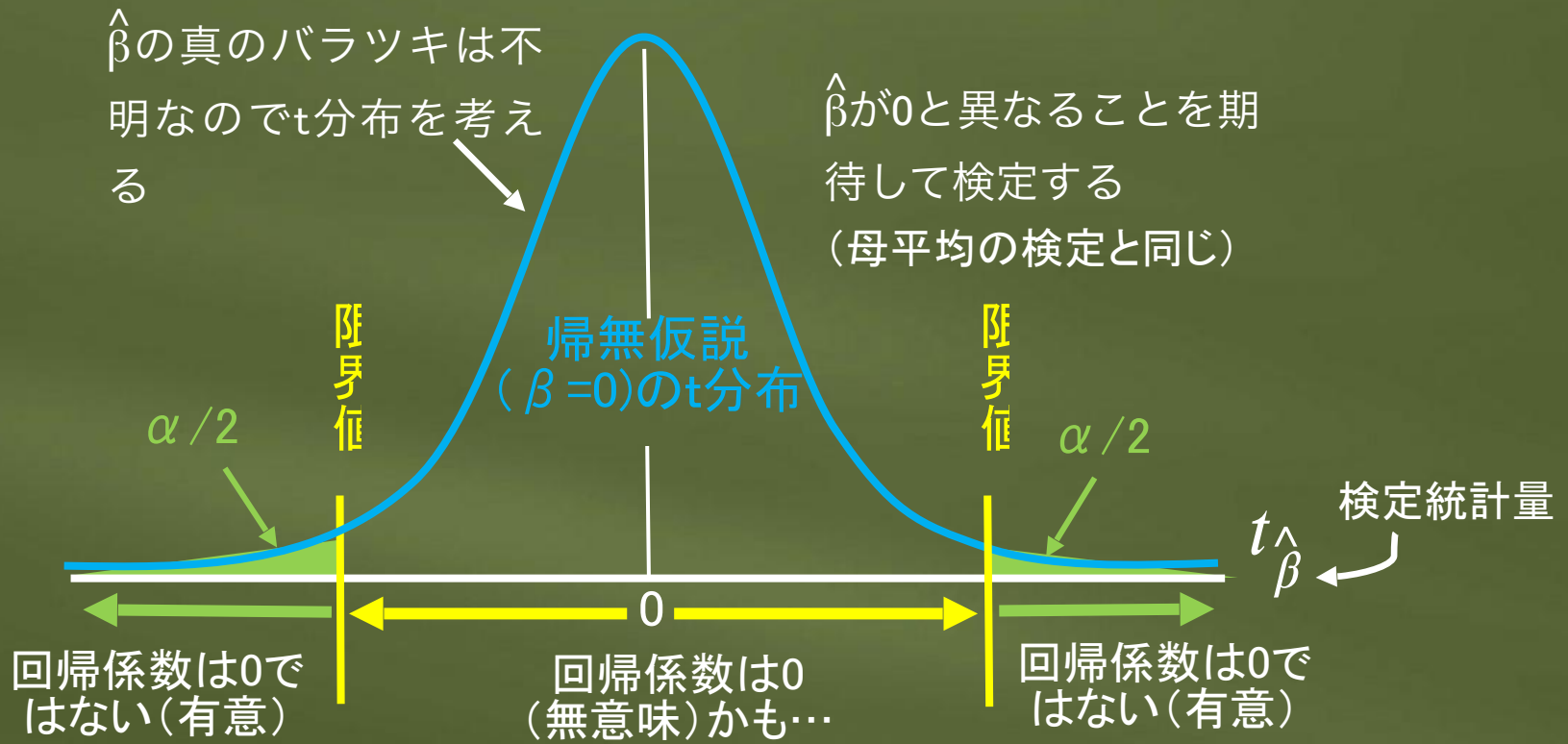
xの変動に対する残差の不偏分散

$\hat{\beta}$にはxだけでなくyも関係するため複雑

※注：定数項 $\hat{\alpha}$ も検定できるが，あまり気にしなくて良い

というわけで…

t検定を用いる



例題（単回帰分析）を解いてみよう

例題のデータ

x_i	3	1	4	2
y_i	3	-2	5	-1

正規方程式

$$\begin{cases} n\hat{\alpha} + \hat{\beta}\sum x_i = \sum y_i \\ \hat{\alpha}\sum x_i + \hat{\beta}\sum x_i^2 = \sum x_i y_i \end{cases}$$

①正規方程式を解くのに必要な値をExcelなどを使って計算（n=4）

x_i	y_i		$x_i y_i$
3	3	9	9
1	-2	1	-2
4	5	16	20
2	-1	4	-2
総和 Σ **10**	**5**	**20**	**25**

正規方程式に代入

$$\begin{cases} 4\hat{\alpha} + 10\hat{\beta} = 5 \\ 10\hat{\alpha} + 30\hat{\beta} = 25 \end{cases}$$

⇩ パラメータが推定できる

$$\hat{\alpha} = -5.0, \quad \hat{\beta} = 2.5$$

$$\hat{y} = -5.0 + 2.5x$$

例題（続き）　モデルの評価

x_i	y_i			$(-)^2$	$(y-)^2$
3	3	2.5	1.25	1.56	3.06
1	-2	-2.5	1.25	14.06	10.56
4	5	5.0	1.25	14.06	14.06
2	-1	0.0	1.25	1.56	5.06
			総和Σ	31.25	32.75

②決定係数

$$R^2 = \frac{\text{回帰変動}}{\text{全変動}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{31.25}{32.75} = 0.95$$

※評価：推定モデルの決定係数は0.95となり，観測値に対して高い説明力を持っている（データと良く適合している）といえる。

x_i	y_i		$=(y-)^2$	$(x-)^2$
3	3	2.5	0.25	0.25
1	-2	-2.5	0.25	2.25
4	5	5.0	0.00	2.25
2	-1	0.0	1.00	0.25
		総和Σ	1.50	5.00

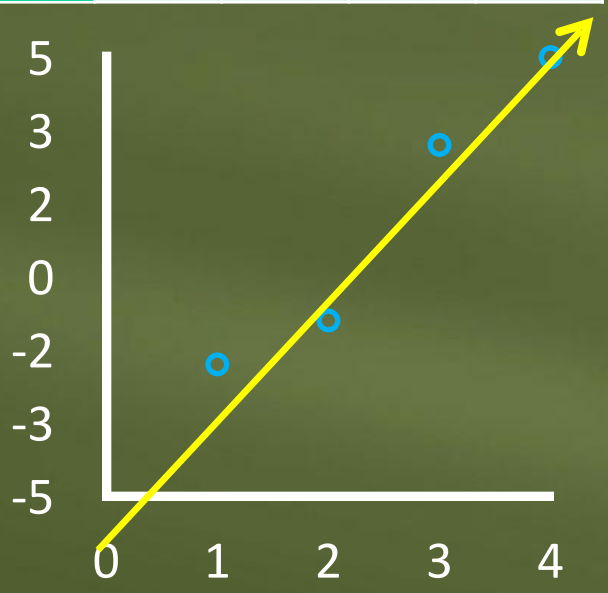
③回帰係数のt検定

$$t_{\hat{\beta}} = \frac{\hat{\beta}}{\sqrt{\frac{\sum \hat{u}_i^2/(n-k-1)}{\sum(x_i - \bar{x})^2}}} = \frac{2.50}{\sqrt{\frac{1.50/(4-1-1)}{5.00}}} = \frac{2.50}{0.387} = 6.46$$

※評価：両側5%の限界値（$\nu=2$）である4.303よりも大きいため，回帰係数は統計的に有意にゼロから離れているといえる。

例題　まとめ（結果の書き方）

x_i	3	1	4	2
y_i	3	-2	5	-1



t検定の結果を＊で示すことがある（**：1%，*：5%水準で有意）

$$\hat{y} = -5.0^{*} + 2.5^{*} x_i$$

$$(-4.71) \quad (6.46)$$

t値か標準誤差を（）で示す

$$n = 4, \quad R^2 = 0.95$$

決定係数も示しておく

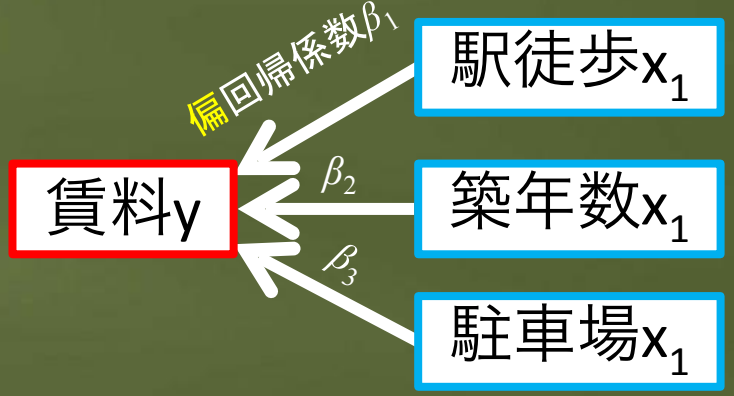
12.5　重回帰分析

- パラメータの求め方は正規方程式の本数が増えるだけで単回帰と全く同じ
- 計算は大変なのでソフトウェアを使用（Excel分析ツールとRコマンダーを紹介）
- 重回帰で注意しなければならない点のみを整理して解説（重回帰特有のモデル評価と係数の解釈法）

12.5　重回帰分析の事例

ある駅周辺のアパート（1K）の賃料と条件

物件番号	y: 賃料（万円/月）	x_1: 駅徒歩（分）	x_2: 築年数（年）	x_3: 駐車場（有=1）
1	4	3	4	0
2	2	4	10	0
3	7	3	4	1
4	6	5	6	1
5	8	1	4	1
6	3	3	9	0

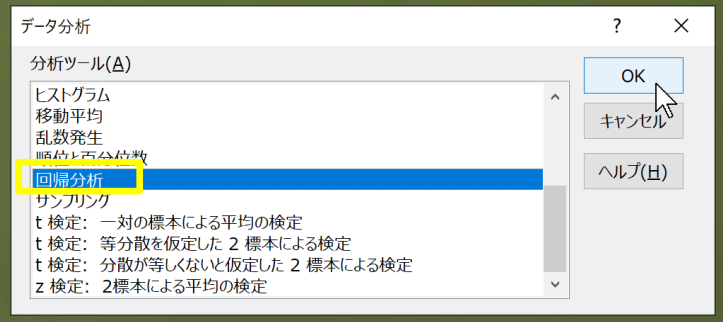


パラメータの求め方は正規方程式の本数が増えるだけで単回帰と全く同じ

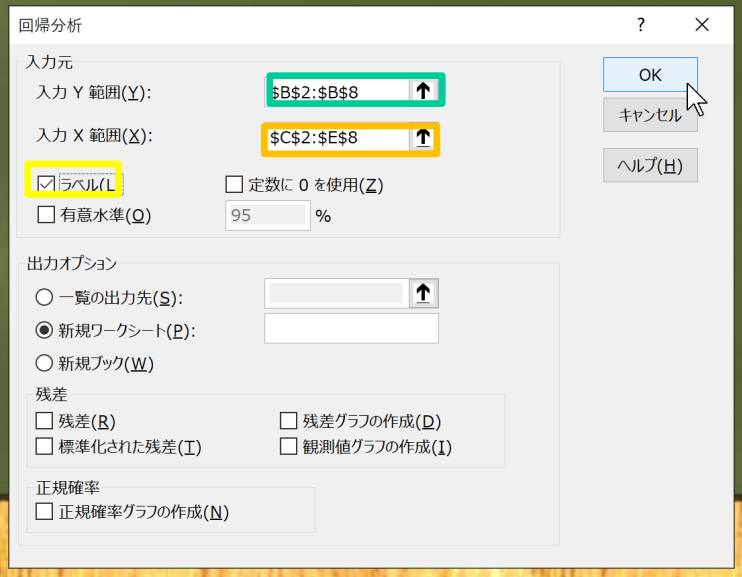
Excel分析ツールによる回帰分析

物件番号	y: 賃料（万円/月）	x_1: 駅徒歩（分）	x_2: 築年数（年）	x_3: 駐車場（有=1）
1	4	3	4	0
2	2	4	10	0
3	7	3	4	1
4	6	5	6	1
5	8	1	4	1
6	3	3	9	0

説明変数は隣同士

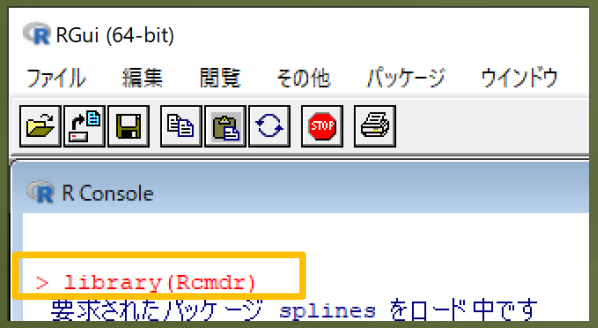


変数を指定

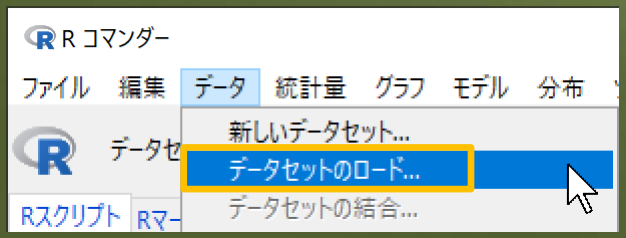


Rコマンダーによる回帰分析

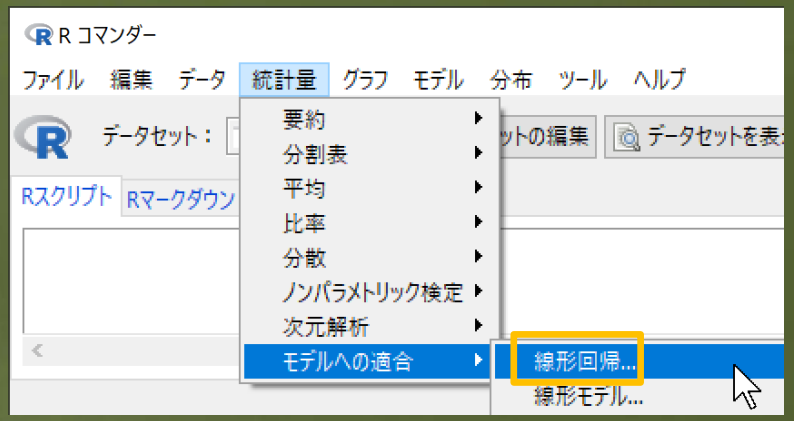
Rコマンダーの起動



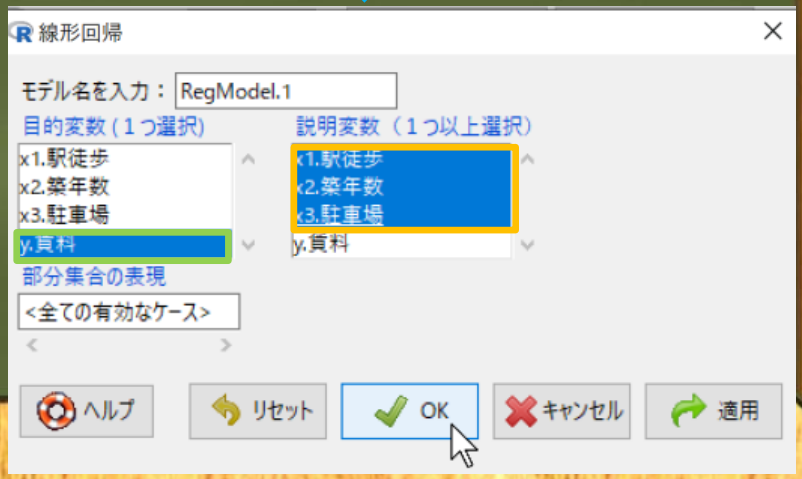
既存のデータセットのロード
（あるいは新しいデータセットを自分で入力）



線形回帰を指定



変数を指定



出力内容（分析ツール）の解説1/2

概要

回帰統計	
重相関 R	0.998
重決定 R2	0.996
補正 R2	0.991
標準誤差	0.230
観測数	6

←①自由度修正済み決定係数

②分散分析（回帰モデル全体の検定）

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	3	27.894	9.298	176.182	0.006
残差	2	0.106	0.053		
合計	5	28.000			

p値

①自由度修正済み決定係数

決定係数は説明変数が増えると自動的に大きくなる

回帰変動

(=全変動-残差平方和)

残差平方和

どんどん引いて行く

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i})^2}{\sum(y_i - \bar{y})^2}$$

全変動

説明変数の数kが増えた分だけ割り引

自由度修正済み

決定係数

$(-\infty \leqq \bar{R}^2 \leqq 1)$

注意

$$\bar{R}^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2 / (n-k-1)}{\sum(y_i - \bar{y})^2 / (n-1)} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

説明変数が増えた分だけ小さくする

②分散分析

推定した重回帰モデル全体が，統計的に見て意味があるかどうかを検定

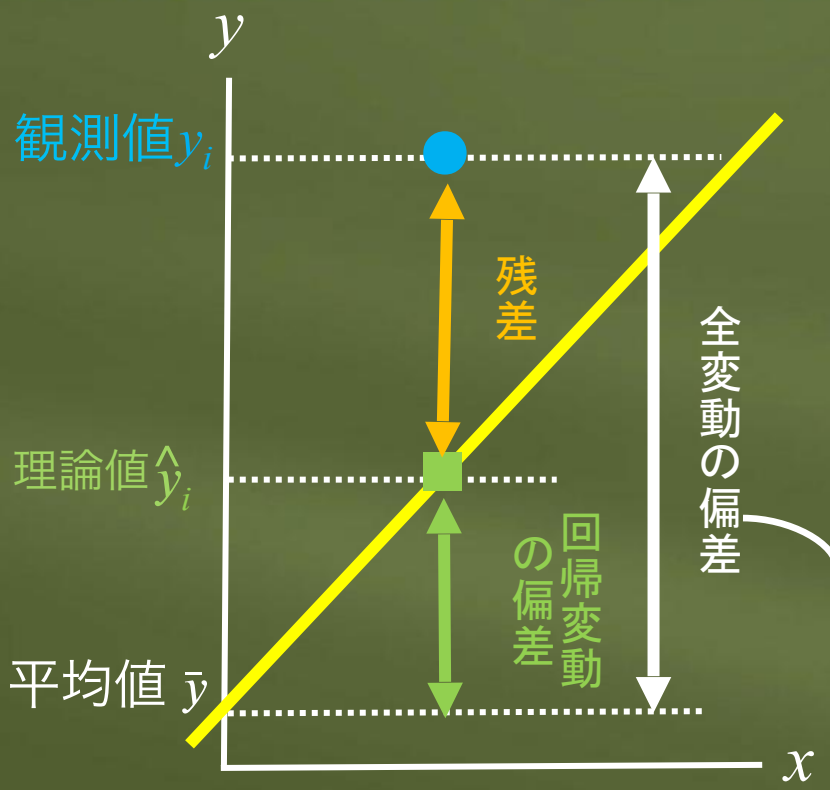
帰無仮説H_0：全ての係数が0　（$\beta_1=\beta_2=\beta_3=0$）全係数をまとめて検定

対立仮説H_1：H_0でない（いずれかの係数が0でない）

分散分析の検定統計量 $F=\dfrac{\text{回帰変動}/\text{自由度}}{\text{残差平方和}/\text{自由度}}=\dfrac{\sum(\hat{y}_i-\bar{y})^2/k}{\sum(y_i-\hat{y}_i)^2/(n-k-1)}$

もう一度，図で確認しておきましょう…

分散分析と決定係数の違い（復習）



分散分析
（自由度は省略）$F = \frac{\text{回}}{\text{残差平方}}$

決定係数
（自由度は省略）$R^2 = \frac{\text{回}\quad\text{動}}{\text{全}\quad\text{動}}$

全変動　回帰変動　残差平方和

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y})^2$$

決定係数　分散分析

出力内容（分析ツール）の解説2/2

	係数	標準誤差	t	P-値	下限95%	上限95%
切片	6.138	0.403	15.230	0.004	4.404	7.873
x1:駅徒歩（分）	-0.407	0.087	-4.671	0.043	-0.782	-0.032
x2:築年数（年）	-0.232	0.053	-4.374	0.048	-0.461	-0.004
x3：駐車場（有=1）	3.167	0.239	13.232	0.006	2.137	4.197

②偏回帰係数

①回帰係数のt検定

係数の区間推定

①回帰係数のｔ検定と変数選択

- 事例では全て有意となったが，もし有意とならなかった変数があったらどうする？

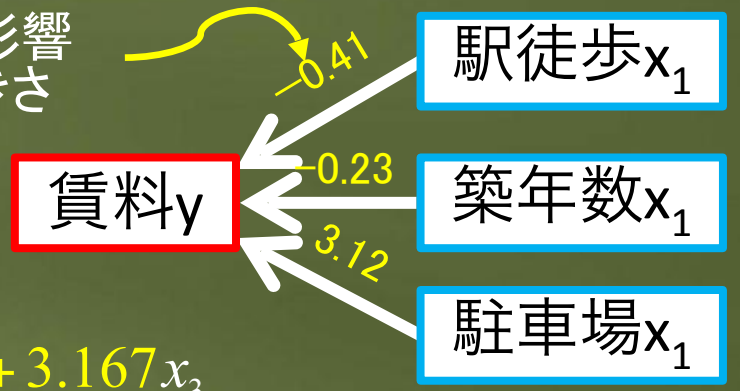
  ①モデルから取り除く　←一般的

  ②モデルに残す場合もある←入っているべき重要な変数（他の変数への影響を抑えるためのコントロール変数）として残しておく

※悩んだら…必要な変数が入っていないよりは不要な変数が入っていた方が害は少ない

②偏回帰係数

偏回帰係数：ほかの変数の影響
を取り除いた後の影響の大きさ



$$\hat{y} = 6.138 - 0.407x_1 - 0.232x_2 + 3.167x_3$$

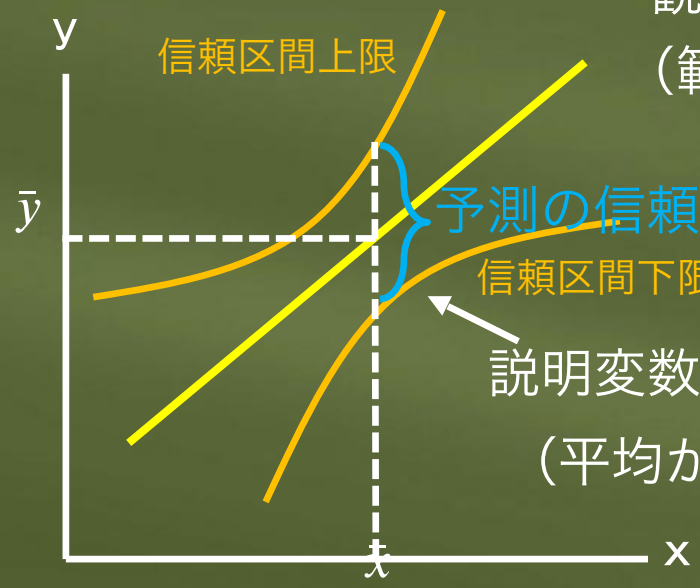
x_1やx_2が一定のとき，x_3の変化によって生じるyの変化
（駅距離と築年数が同じ場合，駐車場があるアパートは無いアパートよりも3.167円高い）

予　測

駅徒歩3分で，築年数5年で，駐車場がないアパートの賃料は？$6.138-0.407\times 3-0.232\times 5+3.167\times 0=3.757$（万円）

観測データの範囲にない外挿予測

（範囲内の予測を内挿予測と呼ぶ）



説明変数の平均$\bar{x}$前後の予測精度が高い

（平均から離れた外挿予測は精度が低い）

標準回帰係数

駅徒歩の負の影響は築年数よりも大きい？

$$\hat{y} = 6.138 - 0.407 \text{駅} \quad \text{徒} - 0.2\text{歩} \quad + 3.167$$

それぞれ単位や分散が異なるため偏回帰係数は相互比較できない

（変数内で）標準化して回帰させる

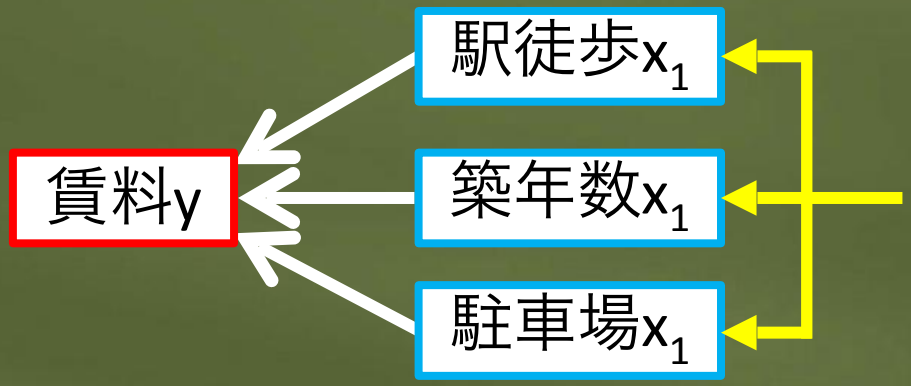
物件番号	y: 賃料	x_1: 駅徒歩	x_2: 築年数	x_3: 駐車場
1	-0.46	-0.14	-0.87	-1.00
2	-1.39	0.69	1.55	-1.00
3	0.93	-0.14	-0.87	1.00
4	0.46	1.51	-0.07	1.00
5	1.39	-1.79	-0.87	1.00
6	-0.93	-0.14	1.14	-1.00
平均	0.00	0.00	0.00	0.00
標準偏差	1.00	1.00	1.00	1.00

$\beta_1^*=-0.229,\ \beta_2^*=-0.266,\ \beta_3^*=0.733$

築年数の影響の方が大きかった
（もっとも大きいのは駐車場の有無）

12.6　モデル推定における問題

①多重共線性（マルチコ）

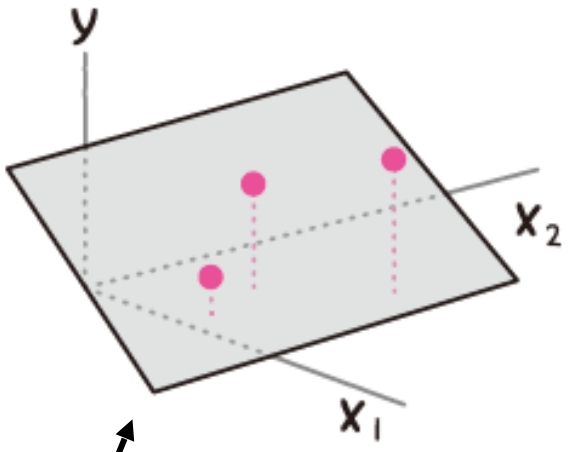


本来，互いに独立（相関係数0）していなければならない

独立していない（多重共線性がある）と…
- 決定係数が高いのにt値が低くなる
- 標本サイズで推定値が大きく変化する
- 係数の符号が理論と逆になる

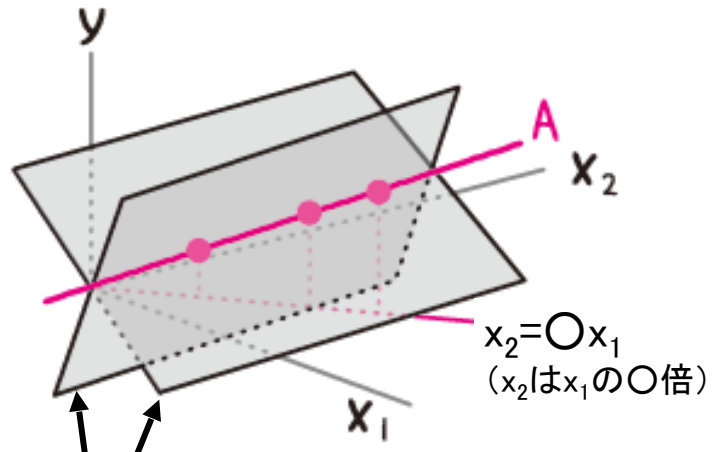
多重共線性の発生原因

x_1とx_2が独立している場合



うまくバラつくので回帰平面が1つに定まる

完全な多重共線性がある場合



いくつもの回帰平面が直線Aを通るので1つに定まらない

多重共線性の見つけ方

変数ごとの多重共線性の深刻さを表す指標

($\hat{\beta}_i$の)分散拡大要因 $VIF_i = \dfrac{1}{1-R_i^2}$

(ダミー変数には使えない)

← xiを被説明変数として，xi以外の説明変数に回帰させた決定係数

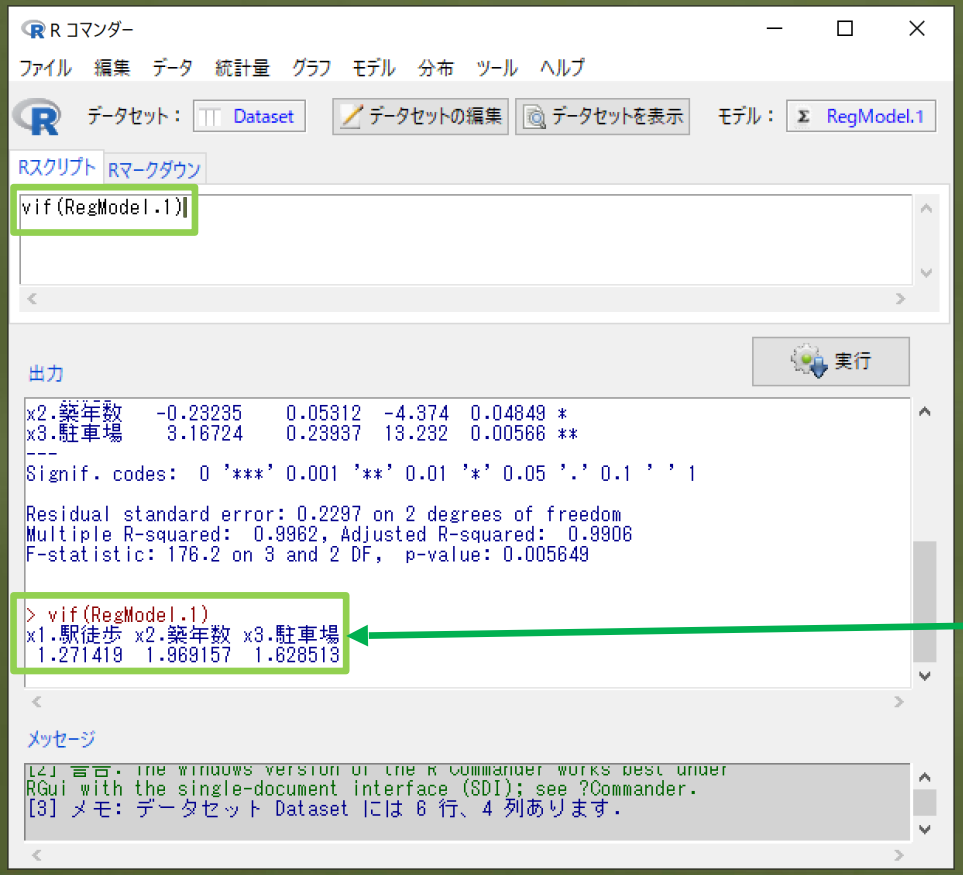
VIF≧10ならば，多重共線性が発生していると判断(決まりはない)

発生していた場合の対処：

- モデルから外す
- 標本サイズを大きくする
- 主成分分析(14章)で独立した変数を作成

Rコマンダー内でVIFを求める

（分析ツールではかなり面倒）



モデルへの適合→線形回帰を実行した後に，
Rスクリプト内で[vif(RegModel.●]と入力して実行

事例のVIFを求めたところ，いずれも小さい値であったため，多重共線性は発生していないと思われる（駐車場はダミー変数なので不明）

モデル推定における問題

②不均一分散

標準的仮定が成立するときにOLSを用いると，最も適切な推定量（最良線形不偏推定量：BLUE）を得られる

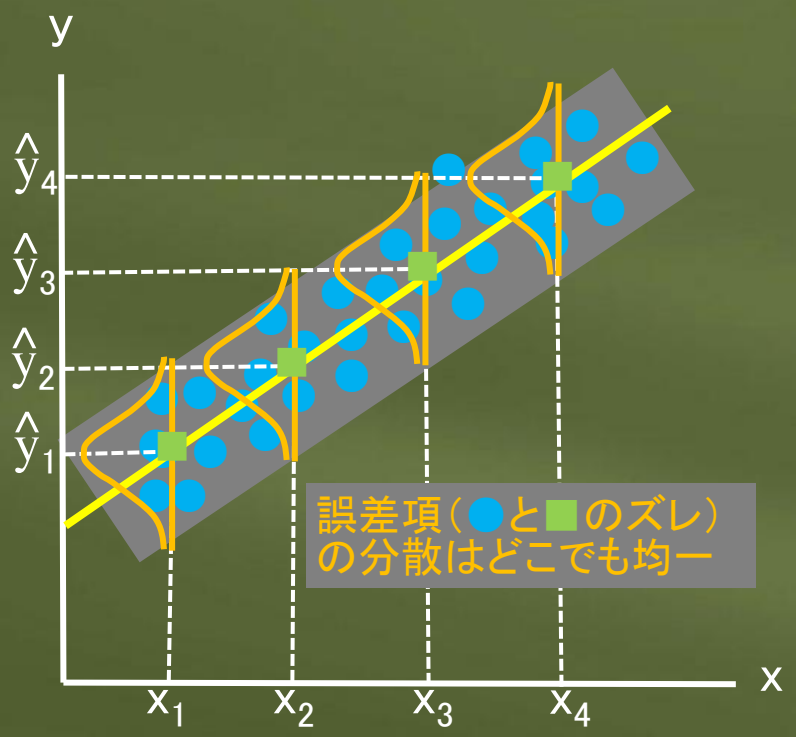
回帰モデルの標準的仮定
①説明変数は非確率変数
②誤差の平均はゼロ
③誤差の分散は均一
④誤差は互いに独立

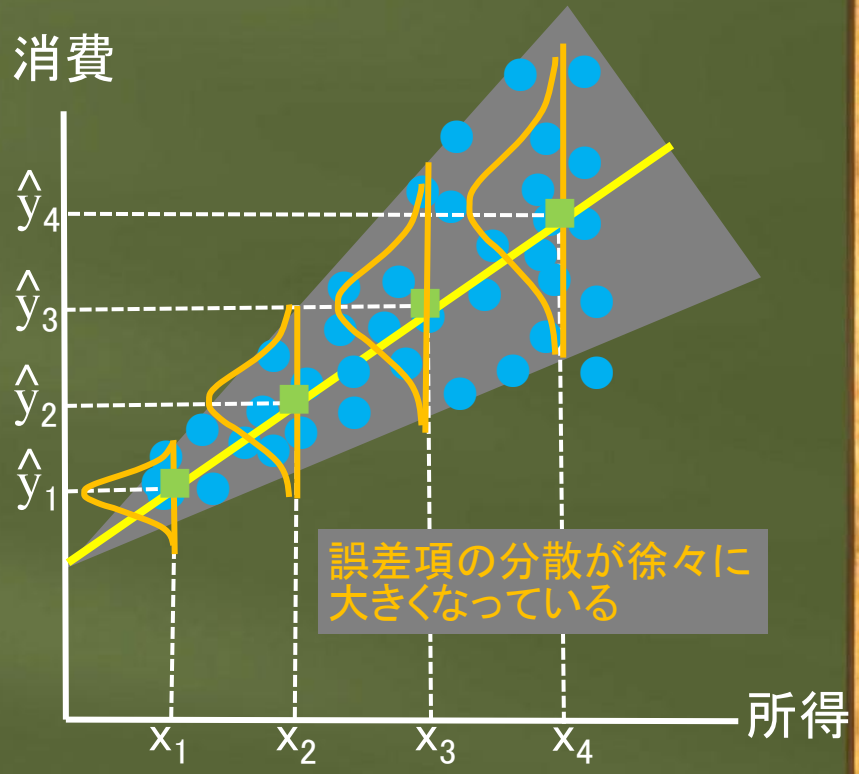
クロスセクション（横断面）データで成り立たないことが多い（不均一分散）

最良線形不偏推定量：

不偏性（真の値を反映）と効率性（分散が最小）と一致性（データを増やせば真の値に近づく）を兼ね備えている

分散不均一の場合の誤差項





不均一分散の症状と対処

- 推定量の分散は最小ではなくなるため，効率性を失う（BLUEでなくなる）

  →t検定の結果が信頼できない

- 対処法

  ・被説明変数yを比の形にしたり対数にする

  ・推定方法を変える（一般化最小2乗法，加重最小2乗法）

  ・頑健（ロバスト）な標準誤差を推定して検定

以上で第12章は終了です。