経済統計分析 10
回帰分析

今日のおはなし.

- 回帰分析 regression analysis
  - 2変数の関係を調べる手段のひとつ
  - 単回帰
  - 重回帰
  - 使用上の注意

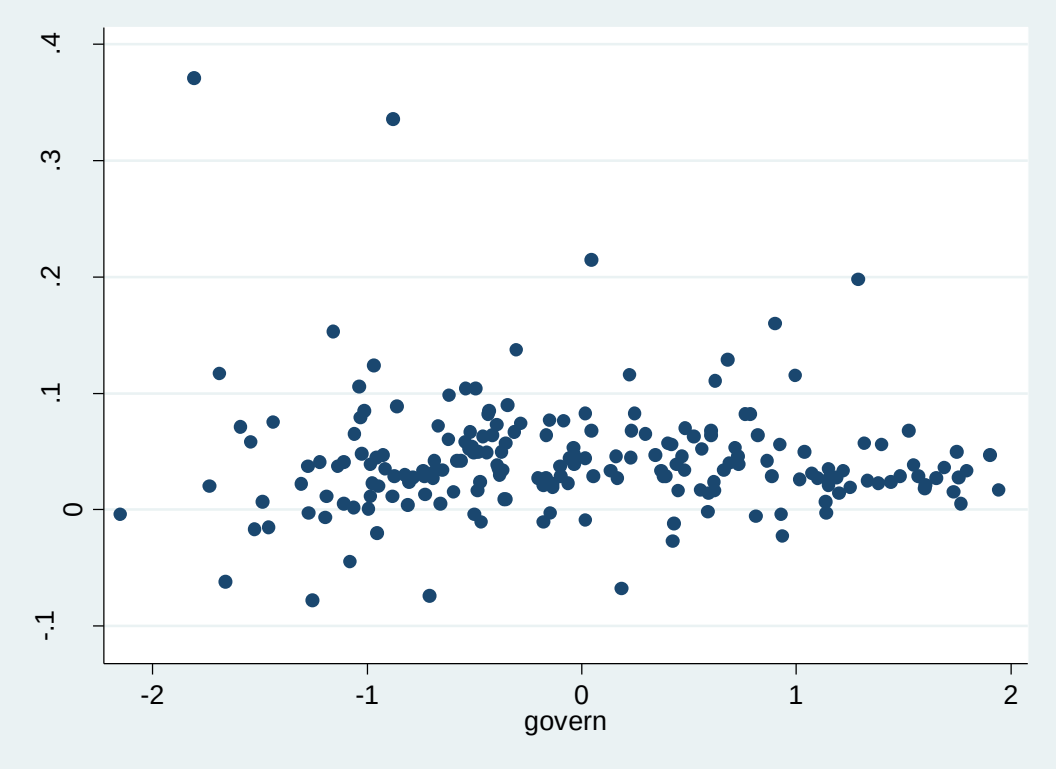
- 今日のタネ
  - 吉田耕作. 2006. 直感的統計学. 日経BP.
  - 中村隆英ほか. 1984. 統計入門. 東大出版会.
  - Stock, James H. and Mark W. Watson. 2006. Introduction to Econometrics. 2nd Revised International Ed, Prentice Hall.

なにができるようになりたいか

- ある変数が他の変数に与える効果の大きさの数量化
  - 確率論的な言葉遣いでは「同時分布の特性値の値を知りたい」
  - これまでの方法：散布図，共分散（相関係数），適合度検定，独立性検定
- 問題の設定
  - 母集団すべてを観測できず，標本のみ
  - 「平均的な関係」を推測する
  - 標本誤差の存在を認める
  - さしあたって2変数の関係

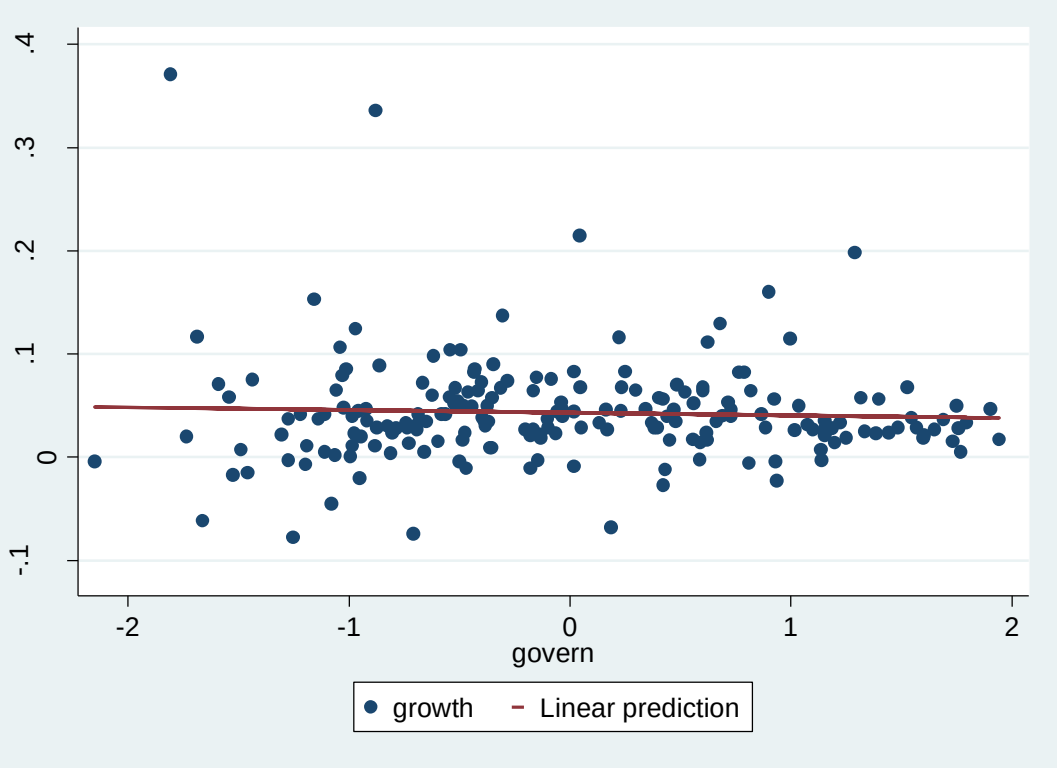
  - 「ある変数の値が1だけ増えたとき，影響を受ける他の変数の値は平均的にはどれくらい増えるか（減るか?）」を，統計的に推測しよう

まずは，散布図



- 例：統治状況と経済成長率（199カ国）
  - 統治状況が経済成長率に効果を与えると想定
  - 横軸が統治状況（原因となるもの），縦軸が経済成長率（結果となるもの）
  - 標本相関係数は-0.0478.

2つの変数が直線的に関係していたら?



- 例:統治状況と経済成長率(199カ国)
  - 統治状況と経済成長率の関係が直線的(線形)であったとして,それに誤差が乗っていると仮定してみたら,統治状況の改善が経済成長率に与える効果の大きさが分かるのでは?
  - 散布図の「真ん中」に直線を描いてみた.
  - 傾き-0.00262,切片0.043

回帰分析 regression analysis

- 回帰分析とは
  - ある変数（被説明変数 dependent variable）が，他の変数（説明変数 independent / explanatory variables）と誤差項（error）の関数であると仮定し，その関数のパラメタを推定する分析
  - 例：経済成長率を被説明変数とし，統治状況を説明変数とする1次関数を仮定した回帰分析
- 単回帰
  - 説明変数が1個だけ（定数を入れると2個）の回帰分析
  - 2個の変数は線形関係（1次関数）で表現される
  - 説明変数をx，被説明変数をy，誤差項をuとすると，

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

であり，β_0とβ_1の値を推定する．
- 重回帰
  - 説明変数が2個以上ある回帰分析

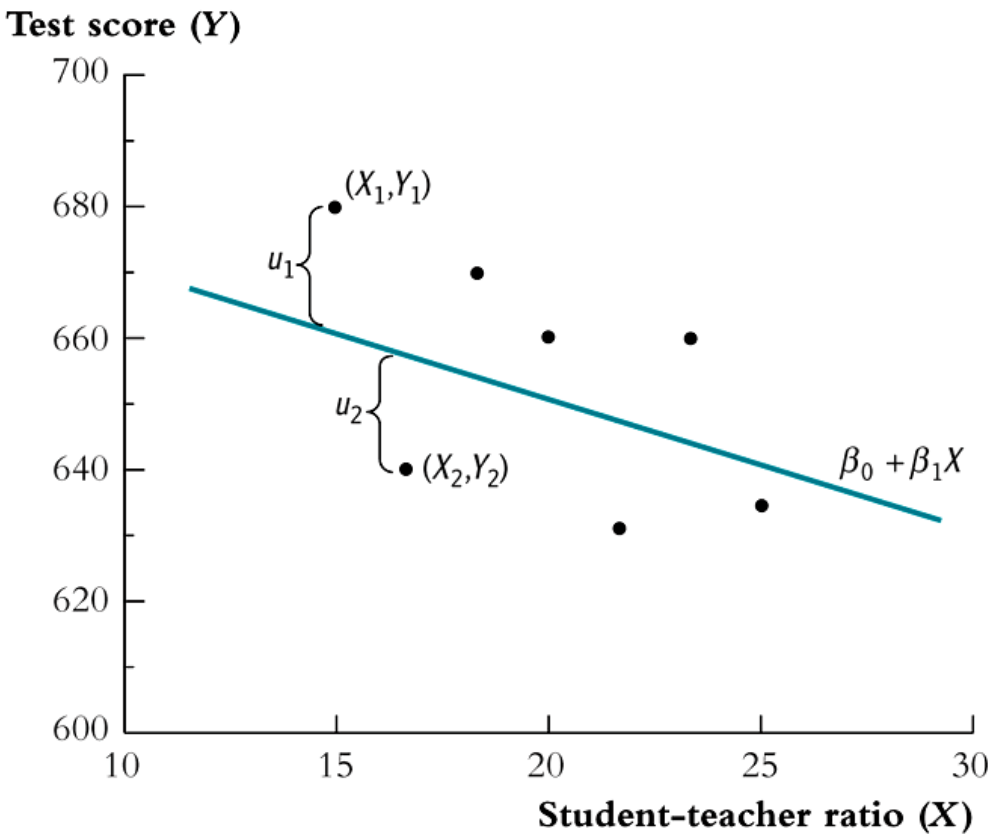
線形回帰モデル linear regression model

- $y_i = \beta_0 + \beta_1 x_i + u_i$
  - x_i：説明変数，独立変数，共変数，y_i：被説明変数，u_i：誤差項
  - $\beta_0 + \beta_1 x_i$：回帰線．x_iが分かったときのy_iの平均的な値
  - β_0：切片（intercept），β_1：傾き（slope）．合わせて係数（parameter）とも
- 誤差項 error term
  - 「その他の要因」を代表する確率変数．平均的な値（$\beta_0 + \beta_1 x_i$）と実現値（y_i）の差を説明するもので，x_i以外のすべての要因を含む
  - 誤差項は観測できない
- 傾き
  - x_iの値が1だけ増えたときのy_iの平均的な増加分（期待値の変分）
  - おもに注目される
  - 「因果関係」を推定したいが，実際には「相関」を計測

線形回帰モデル

FIGURE 4.1 Scatter Plot of Test Score vs. Student-Teacher Ratio (Hypothetical Data)

The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the i^{th} point to the population regression line is $Y_i - (\beta_0 + \beta_1 X_i)$, which is the population error term u_i for the i^{th} observation.



- Stock and Watson, 2003.

線形回帰モデル：例

- $y_i = \beta_0 + \beta_1 x_i + u_i$
  - x_i：説明変数は統治状況（05年）
  - y_i：被説明変数は経済成長率（per capita, 05→06年）
  - u_i：誤差項はその他の要因. 技術・教育・地政・発展段階などなどなど
  - $\beta_0 + \beta_1 x_i$：統治状況が分かったときの経済成長率の平均的な値
  - β_1：傾きは統治状況が1増えたときの経済成長率の変化の大きさ
- 注意点
  - 統治状況と経済成長率が1次関数の関係にあるのは「前提」
  - この前提が正しいかどうかは分からない（all models are wrong!）
  - 変数を「変形」したものなど含めれば, 1次近似として有効
- 発想
  - もし, データが1次関数の関係から発生しているものと考えれば, そのときのパラメタはどれほどであろうか?

線形回帰モデルにおける統計的推測

- $y_i = \beta_0 + \beta_1 x_i + u_i$
  - β_0とβ_1の真の値がわかっていれば、x_iとu_iの実現値に応じてy_iの値を計算できる
  - 手許にあるデータは(x_i, y_i)の（無作為抽出）標本だけであり、ここからβ_0とβ_1を推測する
  - もう1 つの確率変数u_iは実現値もわかっていない
  - (x_i, y_i, u_i)が線形の関係にあるかどうかも（ほんとうは）定かではないが，ここでは仮定
  - β_0とβ_1の真の値を標本から統計的に推測するから，仮説検定や信頼区間の形成という手続きが可能
- では，β_0とβ_1の真の値をどのように推測するのか？
  - 切片と傾きの一致推定量を計算するにはどのようにすればよいのか?

最小2乗法 OLS(Ordinary Least Squared)

- 最も有名な推定量の1つ
  - いくつかの条件の下で, 切片と傾きは一致推定量になる
- 発想
  - 誤差が平均的にはゼロであれば, 散布図の「真ん中」に回帰線があるはず
  - 回帰線からの「乖離」がなるべく小さくなるように, 直線を引けばよい
  - 「乖離」の合計を小さくすればよいが, そのまま足すと計算がめんどう
  - 「乖離」の2乗の和を最小にするようが計算が簡単
- 式で書くと.
  - 推定量をb_0, b_1として, 次を最小化するものを選ぶ

$$\min \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 x_i \right) \right]^2$$

  - 最小化問題になるので, b_0, b_1で偏微分してゼロとおけばよい
  - 正規方程式：式が2つ, 未知数が2つ

最小2乗法

- 正規方程式を解くと（計算は電子計算機に任せる），

$$b_1 = \frac{\sum_{i=1}^{n}\left[x_i - \bar{x}\right]\left[y_i - \bar{y}\right]}{\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2}, b_0 = \bar{y} - b_1\bar{x}$$

- 標本共分散，標本分散を用いると，$b_1 = \frac{s_{xy}}{s_x^2}$
  - 例：統治状況の分散は.864707，共分散は-.00227　→割ってみると-0.00262
- 別の解釈
  - 推定式の両辺とx_iの共分散を計算してみよう

最小2乗法の基礎用語

- OLS回帰線
  - OLSによって得られた係数推定値で描かれる回帰線
- 当てはめ値 fitted value
  - 所与のx_iに対するy_iのOLS回帰線上の値. 期待値のようなもの.
- 残差 residual
  - 各観測値と, 対応する当てはめ値との差.
  - 誤差の推定量として用いられることも.
- 係数の標準誤差 standard error
  - 標本平均が確率変数であったのと同様にOLS 推定量も確率変数.
  - 同じ母集団であってもサンプルが異なればOLS 推定値は異なる.
  - それゆえ, OLS推定量も標本分布をもち, 標準偏差がある.

なんでOLS推定量なの?

- 望ましい性質
  - ある仮定のもとでは, OLS推定量は真の係数の一致推定量
  - さらにある仮定のもとでは, 有効推定量
  - BLUE: Best Linear Unbiased Estimator
- じっさい, よく使われているし.
  - 実証分析を進めるうえでの共通言語のひとつ
  - パッケージソフトも多い. MS-Excel にも組み込み関数がある
  - 収束計算が不要で, 「手計算」が比較的容易だったという事情も.
- 拡張
  - 説明変数が2個以上
  - 関数形が線形に限らない

当てはまりのよさ:回帰の標準誤差

- 残差 residuals の性質
  - 残差の和はゼロ
    - 観測値は予測値と残差の和だから
  - 残差と説明変数は無相関
  - 残差と予測値は無相関
- 回帰の標準誤差 standard error of regression
  - 誤差項の標準偏差の推定値
  - (残差平方和 / n-2)の平方根

$$\mathrm{SER} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}}$$

当てはまりのよさ:決定係数

- 定義
  - 決定係数 R^2 :説明変数の変動が全変動に占める比率
  - 全変動 = 説明変数の変動 + 残差の変動

$$R^2 = \frac{\hat{Y}\text{の標本分散}}{Y\text{の標本分散}} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2}$$

- 性質
  - 0から1のあいだの値を取る
  - データが回帰線上に並んでいるとき, 決定係数は1
  - データが説明変数によって全く説明されないとき, 決定係数は0
  - R^2が大きいほど, Y_iの予測がうまくできている

OLS推定量の仮説検定

- 手続き
  1. 仮説を立てる.
  2. 有意水準を決める.
  3. 検定統計量（test statistics）を計算する.
  4. p値を求めて，棄却/受容を判定する.
- 検定する帰無仮説
  - H_0:「傾きの値が～だ」
  - 最もしばしば用いられるのは「傾きの値がゼロだ」
    - 「説明変数は被説明変数に影響を与えていない」
  - 平均値の検定と同じなので，t-検定を用いる
    - 「傾きがゼロだ」に対応するt-値，p値は自動的に出力されることが多い

OLS推定量の仮説検定

- 切片, 傾きの推定量の標準誤差
  - 推定量の標準偏差の推定量

$$\mathrm{SE}(b_1) = \sqrt{\frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^{n}(x_i-\bar{x})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}}$$

  - これは「分散不均一に頑健な標準誤差」と呼ばれるもの
  - MS-Excelの組込み関数の計算方法は異なる
    - 「分散均一 homoskedasticity を仮定した標準誤差」と呼ばれる
    - より強い仮定を必要とするので,「分散不均一に頑健な標準誤差 robust standard error」を使うほうが好ましいが...
  - データは母集団から抽出された標本なので, 標本が異なれば推定される傾きや切片の値も異なる

係数についてのt検定

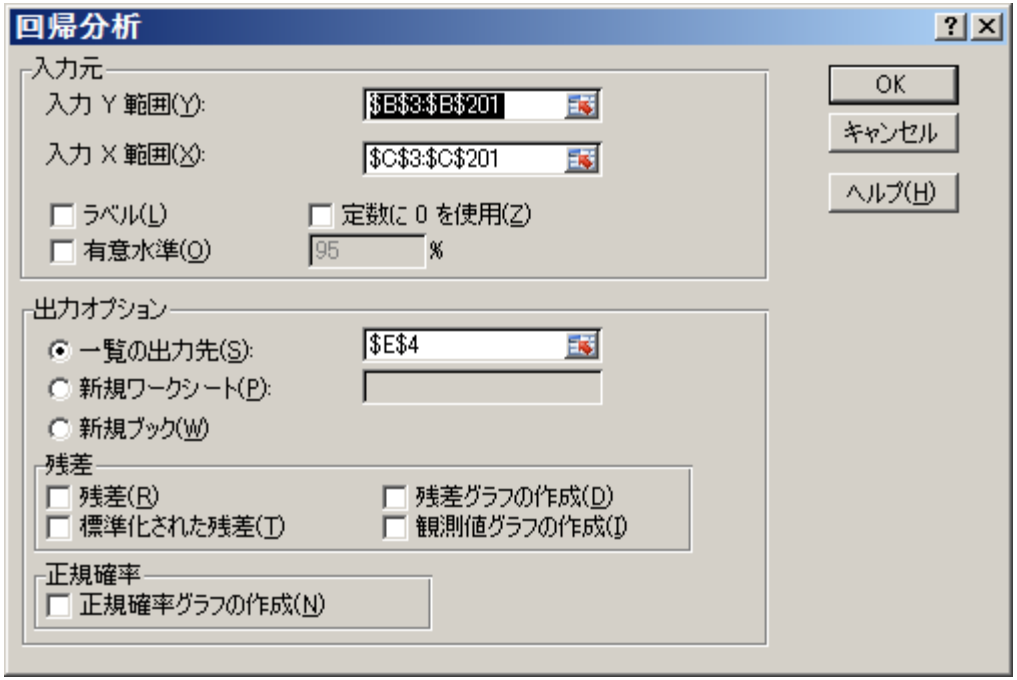
- 検定統計量：t値
  - 帰無仮説が正しいとき，サンプルサイズが十分に大きく，各観測値がi.i.d.であれば，標準正規分布に従う

$$t = \frac{\text{推定値} - \text{仮説の値}}{\text{推定量の標準誤差}} = \frac{b_1 - \beta_{1,0}}{\mathrm{SE}(b_1)} \xrightarrow{d} N(0,1)$$

  - 「傾きがゼロだ」を検定するときには，推定値をその標準誤差で割った値の絶対値が1.96より大きければ，有意水準5%で棄却できる
- 注意
  - 各観測値がi.i.d.に正規分布に従うとき，t統計量は「t分布」にexactに従う
  - 観測値が正規分布に従うとは限らないし，サンプルサイズが大きければt分布は標準正規分布で近似されるので，ここでは標準正規分布を用いている．
  - 「傾きがゼロだ」という帰無仮説を棄却できるとき，係数が「統計的に有意にゼロと異なる (statistically significantly different from zero)」と言い，略して「統計的に有意だ statistically significant」とも言われる
  - 統計的有意性は，政策的な重要さとは直接関係ない

MS-Excel de 回帰分析

- MS-Excel 2007でやってみた
  - データ→データ分析→回帰分析
  - 欠損値が混じっているとエラーが返ってくるなんて!
    - 系列の並べ替えを使って欠損値を除去してから
    - こういうウィンドウが開くはず



MS-Excel de 回帰分析

- 出力はこうなります（桁だけ揃えた）
  - 被説明変数：05→06年の経済成長率
  - 説明変数：05年の統治状況，定数項

概要

回帰統計	
重相関 R	0.048
重決定 R2	0.002
補正 R2	-0.003
標準誤差	0.051
観測数	199

▶ 係数推定値，標準誤差のほか，「係数がゼロ」という帰無仮説に対するt統計量，p値が出力される
▶ ここでは，「傾きがゼロ」という仮説は棄却できず，「傾きはゼロと統計的に有意には異ならない」

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	1	0.001	0.001	0.447	0.505
残差	197	0.515	0.003		
合計	198	0.516			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	0.0431	0.0036	11.8831	0.0000	0.0359	0.0502	0.0359	0.0502
X 値 1	-0.0026	0.0039	-0.6682	0.5048	-0.0103	0.0051	-0.0103	0.0051

重回帰

- 説明変数を2個以上に増やす
  - 定数項を説明変数と解釈すれば3個以上
  - 線形関係の仮定はそのまま
- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
  - たとえば説明変数が2個のケース
  - β_1：他の条件を一定として，x_{1i}が1増えたときのy_iの変化分
  - β_2：他の条件を一定として，x_{2i}が1増えたときのy_iの変化分
- 最小2乗推定
  - 残差平方和を最小にする，という方針は同じ

$$\min \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \right) \right]^2$$

  - 単回帰と同じく，一定の仮定のもとで一致・不偏推定量
  - OLS推定量はここでは明示的には表現しない（行列表現）

多重共線性 multi-colinearity

- 完全な多重共線性
  - ある説明変数が，他の説明変数（と定数）の1次関数で表現されること
  - 例：x_{1i}とx_{2i}がつねに同じ値を取る
  - 例：x_{1i}を100倍するとx_{2i}になる（パーセント表記）
  - 例：x_{1i}を1から引くとx_{2i}になる
  - ダミー変数（0か1の値を取る）を使うときにありがち
    - 例：x_{1i}が男性ダミー，x_{2i}が女性ダミー
- 完全な多重共線性が発生していると推定できない
  - 論理的に無理：「他の条件を一定として」を考えられないから
- 不完全な多重共線
  - 説明変数の間の相関係数が極めて高い（0.99など）
  - 理論的には問題はないものの，推定値が不安定になりがち

自由度修正済み決定係数 adjusted-R^2

- 決定係数

$$R^2 = \frac{\hat{Y}\text{の標本分散}}{Y\text{の標本分散}} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2}$$

  - 説明変数の数を増やすと，ほぼ自動的に決定係数が上昇
  - サンプルサイズが大きくないとき，「当てはまり」の指標としては不適切
- 自由度修正済み決定係数
  - 「説明変数が多い」という要因を修正したもの

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2}$$

  - 決定係数より小さな値を取る
  - 負の値を取るときもある：説明変数の数が多いとき
  - 説明変数の数が増えても，自動的に増加するわけではない
  - 決定係数が高くなっても，説明変数の追加が適切だとは限らない

MS-Excel で重回帰

- 被説明変数：05→06年の経済成長率
- 説明変数：05年の1人当たりGDP，05年の統治状況，定数項

概要

回帰統計	
重相関 R	0.092
重決定 R2	0.008
補正 R2	-0.002
標準誤差	0.051
観測数	199

- 自由度修正済み決定係数（補正R2）が出力される
- ここでも，「傾きがゼロ」という仮説は棄却できず，「傾きはゼロと統計的に有意には異ならない」

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	2	0.004	0.002	0.829	0.438
残差	196	0.511	0.003		
合計	198	0.516			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	0.0467	0.0049	9.4730	0.0000	0.0370	0.0565	0.0370	0.0565
X 値 1	0.0000	0.0000	-1.1007	0.2724	0.0000	0.0000	0.0000	0.0000
X 値 2	0.0017	0.0055	0.3104	0.7566	-0.0092	0.0126	-0.0092	0.0126

MS-Excel で重回帰

- 被説明変数：05→06年の経済成長率
- 説明変数：05年の1人当たりGDP（x_{1i}），05年の統治状況（x_{2i}），定数項
- イラクとアゼルバイジャンを除外（異常値っぽい）

回帰統計	
重相関 R	0.161
重決定 R2	0.026
補正 R2	0.016
標準誤差	0.040
観測数	197

- 「傾きがゼロ」という仮説は棄却される
  - OLS推定値は異常値にひっぱられやすい
- すでに経済成長している国の成長率は低い
  - 「収束仮説」に整合的
- 統治状況のよい国の成長率は高い

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	2	0.008	0.004	2.566	0.079
残差	194	0.312	0.002		
合計	196	0.320			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	0.0455	0.0039	11.7392	0.0000	0.0379	0.0532	0.0379	0.0532
X 値 1	0.0000	0.0000	-2.1277	0.0346	0.0000	0.0000	0.0000	0.0000
X 値 2	0.0091	0.0044	2.0596	0.0408	0.0004	0.0178	0.0004	0.0178

非線形関数への拡張

- 非線形関数
  - 1次関数以外の関数
  - 2乗項, 3乗項の入る多項式, 対数, 逆数がよく用いられる
  - ありとあらゆるパターンに対応可能なわけではない
- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
  - 説明変数を「変形したもの」をみなせばよい
  - 例: $x_{2i} = x_{1i}^2$
  - 例: $x_{2i} = \log(x_{1i})$
  - このとき, 傾きの解釈が変化
    - 「他の条件を一定として, x_{2i}が1増えたときのy_iの変化分」には変わりない
    - 例:「他の条件を一定として, $\log(x_{1i})$ が1増えたときのy_iの変化分」
    - 例:「他の条件を一定として, x_{1i}^2 が1増えたときのy_iの変化分」??

OLS推定量が一致性を持つ条件

- 4条件
  - 説明変数で条件付けられた誤差項の期待値がゼロ
  - 観測値はi.i.d.
  - 説明変数と誤差項は母分散を持ち，4次モーメントが有限
  - 完全な多重共線性がない

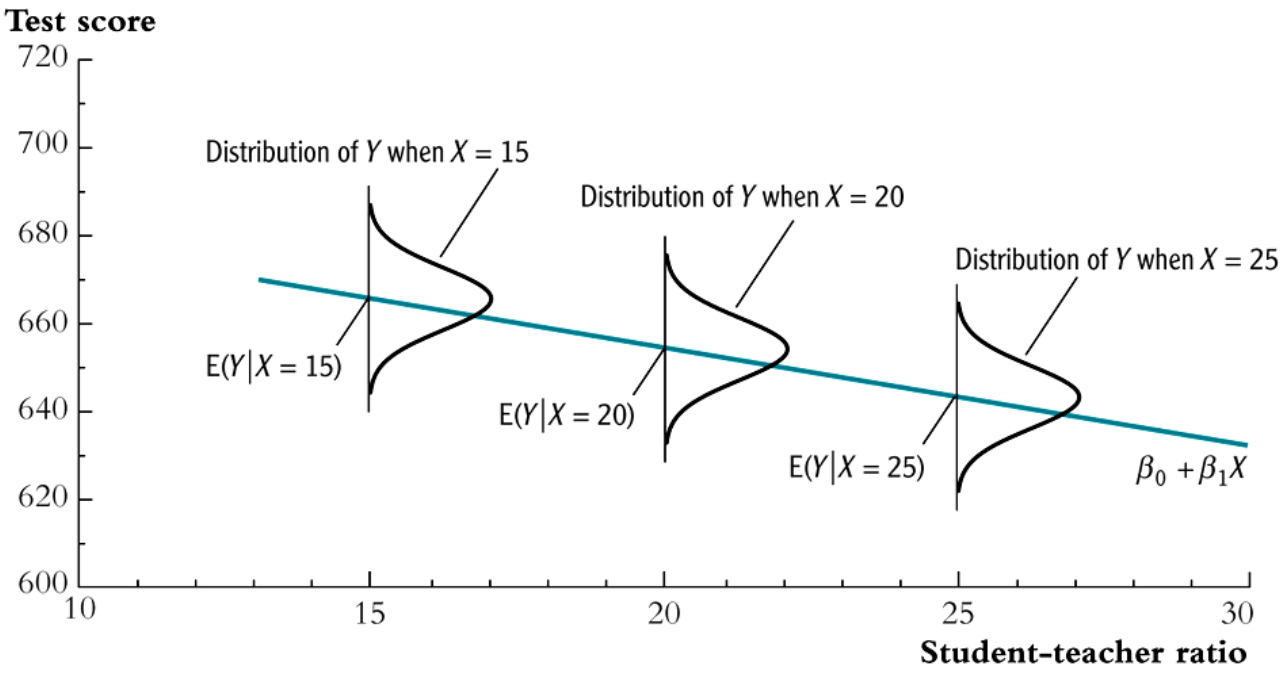
- Remarks
  - すべての条件が厳密に満たされることは，ほとんどない．
  - 「観測値はi.i.d.」：時系列的，地域的な相関は避けられない？
  - 「母分散と4次モーメント」：統計学上の技術的な仮定として，満たされているものとする
  - 「多重共線性」：計算の途中でエラーが返ってくる

説明変数の外生性 exogeneity

- 説明変数で条件付けられた誤差項の期待値がゼロ
  - =「誤差項と説明変数が相関を持たない」
- 回帰分析の発想
  - 被説明変数と説明変数が1次関数の関係にあり，ここに誤差が乗ったものがデータとして観測されていると考える
  - 誤差は「noise」として足されているだけだから，データの「真ん中」を通るように直線を引けば，本来の1次関数を復元できる
  - 「データの真ん中に直線がある」=「誤差項の条件付き期待値がゼロ」
- 逆に言うと．
  - 誤差項の条件付き期待値がゼロでないところがあれば，データの真ん中に（直接には観測できない）直線が通っているとは限らない
  - データの真ん中に直線を引いても，本来の1次関数を復元できるわけではない　→推定値に偏り（bias）をもたらす

誤差項の条件付き分布

FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line



The figure shows the conditional probability of test scores for districts with class sizes of 15, 20, and 25 students. The mean of the conditional distribution of test scores, given the student-teacher ratio, $E(Y|X)$, is the population regression line $\beta_0 + \beta_1 X$. At a given value of X, Y is distributed around the regression line and the error, $u = Y - (\beta_0 + \beta_1 X)$, has a conditional mean of zero for all values of X.

誤差項とは?

- 誤差項が表しているもの
  - 説明変数に含まれてはいないが, 被説明変数に影響を与える要因全て
  - 実験データに見られる「純粋なランダムさ」だけではない
  - もし観測できるものなら, 説明変数に追加するのが解決方法のひとつ
  - 誤差項がどのような要因を代表しているのか?

- 例:統治状況と経済成長率
  - 経済成長率に影響するのは統治状況だけか?
  - 他の要因もいろいろ:人的資本(教育), 衛生, 言語, 貯蓄率などなど
  - 教育水準と統治状況は相関がありそう
    - 統治状況が高い値を示す国では, 教育水準が高い　→誤差項が大きな値
  - OLSで「真ん中に」直線を通すと, 上に偏っているかも
    - 教育水準の効果を反映しているだけで, 統治状況の効果ではないかも