# 回帰分析
# 単回帰

麻生良文

# 単回帰モデル
# simple regression model

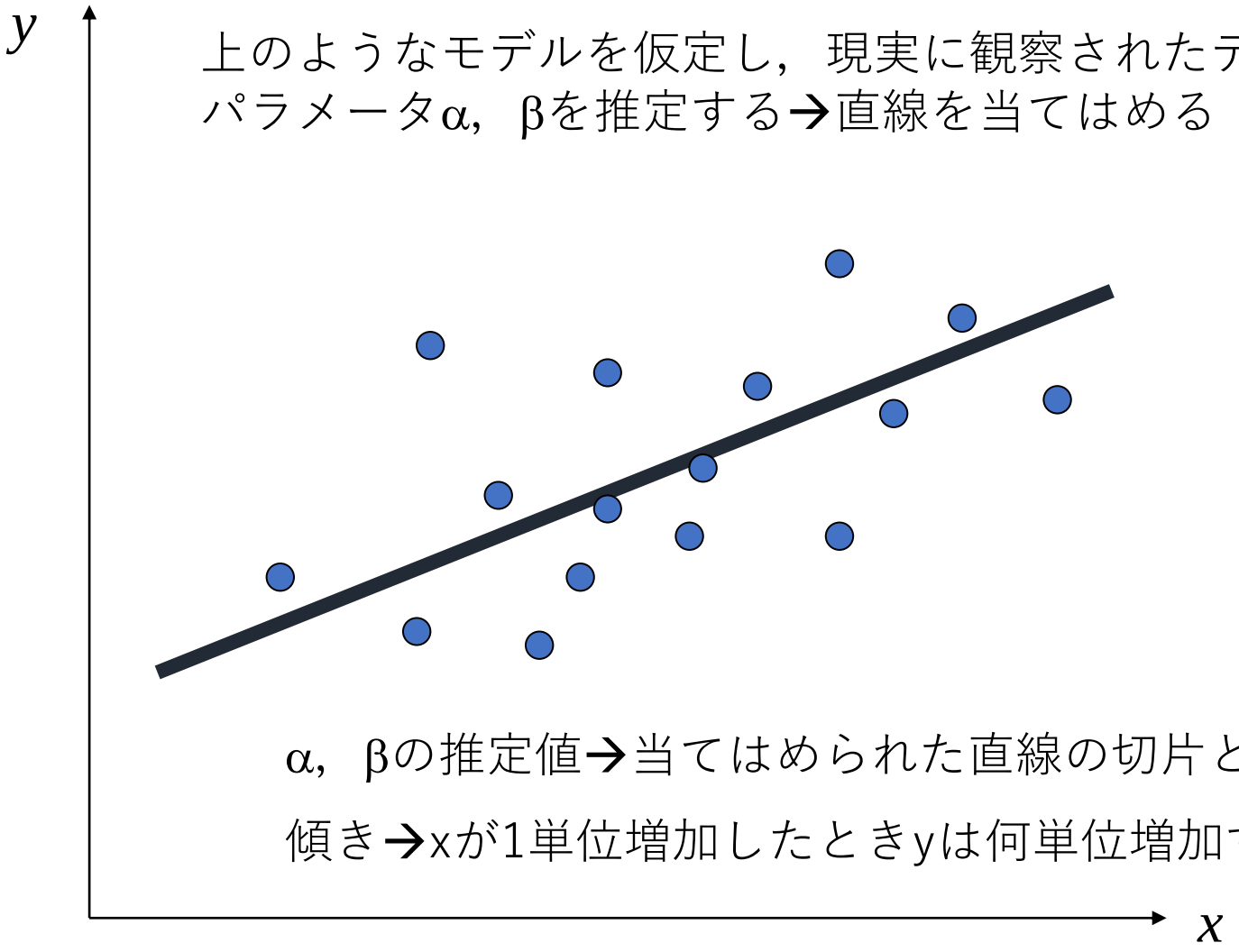
ここは単回帰モデルについての結果だけをまとめたものです。
詳細はreg.pdfを参照すること。

$$y = \alpha + \beta x + u$$

- $y$
  - 従属変数 (dependent variable)
  - 被説明変数(explained variable)
- $x$
  - 独立変数　(independent variable)
  - 説明変数 (explanatory variable)
- $u$
  - 誤差項　(error term)
  - 撹乱項　(disturbance term)
  - 他の要因，観察されない変数の影響，yの測定誤差

$$y = \alpha + \beta x + u$$

上のようなモデルを仮定し，現実に観察されたデータから，
パラメータα，βを推定する→直線を当てはめる




α，βの推定値→当てはめられた直線の切片と傾き

傾き→xが1単位増加したときyは何単位増加するか

# 重回帰モデル
# multiple regression model

説明変数が2個以上

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

係数 $\beta_j$ の意味

他の説明変数を一定に保っておいて，$x_i$ だけを1単位増加させたときに $y$ が何単位増えるか

他の要因をコントロールした $x_i$ 固有の影響

$$\beta_j = \frac{\partial y}{\partial x_j}$$

# 単回帰モデルにおける仮定

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

1. 線型モデル（パラメータに関し）
2. 誤差項の期待値は0
3. 誤差項は互いに独立
4. 誤差項の分散は一定（分散均一性）
5. 誤差項は正規分布に従う
    - BLUEの成立のためには5.の条件は不要

# 最小二乗法

- 残差平方和を最小にするようにパラメータを決定
  - $a,b$: 未知パラメータ $\alpha$, $\beta$ の推定値
  - $e_i$ : 残差

$$\min \quad S(a,b) = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$S(a,b)$の最小化のための条件

$$\frac{\partial S}{\partial a} = 0, \ \frac{\partial S}{\partial b} = 0$$

→正規方程式を解くと$a, b$が求められる

# 最小二乗推定量(OLS estimator)

係数の推定量

$$b = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

$$a = \bar{y} - b\bar{x}$$

誤差項の分散の推定量

$$s^2 = \frac{1}{n-2} RSS = \frac{1}{n-2} \sum_{i=1}^{n} {e_i}^2$$

誤差項の分散の推定量の平方根 $s$ は**回帰の標準誤差**（standard error of the regression; SER)とよばれる。

nはオブザベーション数，2は定数項を含んだ説明変数の個数

# 決定係数

TSS=ESS+RSS

全平方和＝回帰式で説明される部分の平方和 + 残差平方和

決定係数　$R^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- 回帰式の当てはまりの良さを表す指標
- 0から1の間の値
- 1に近いほど当てはまりが良いことを表す

# 回帰係数の確率分布

仮説　$H_0: \beta=\beta_0$　を考える

$H_0$ が正しければ

$$\frac{b-\beta_0}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$

しかし，誤差項の分散$\sigma^2$は未知のパラメータ。そこで，$\sigma^2$の最小二乗推定量$s^2$に置き換えた次の統計量は自由度$n-2$の$t$分布をする。

$$\frac{b-\beta_0}{\sqrt{s^2/S_{xx}}} = \frac{b-\beta_0}{s.e.(b)} \sim t(n-2)$$

ただし，$s^2 = RSS/(n-2)$ であった（RSSは残差平方和）。nはオブザベーション数，2は説明変数の個数（定数項とxの2つ）。

統計ソフトの回帰分析のoutputでは，$\beta_0=0$の場合のt値が出力されるのが普通

# Rでの回帰分析

- ここではwage1.csvを用いる（R入門で作成したファイル：wage1.xls, wage1.rawが元ファイルで変数名のヘッダー行を加えたもの）
- wage1.csvがimportされて，データ・フレームwage1ができているとして次のコマンドを実行する

---

attach(wage1) # wage1内の変数に直接アクセスできるようにする

wage1.lm <- lm(wage ~ educ) #回帰分析を行い，その結果をwage1.lmに保存する

summary(wage1.lm)　#wage1.lmに保存された要約統計量を出力する

---

# 以下は説明なのでタイプする必要はない

- wage1のデータフレームでの分析終了→ detach(wage1)を忘れない：　他のデータフレーム内の変数と混同する可能性

# Rでの回帰分析(2)

前のページのコードの解説

- lm( y ~ x1 + x2 + x3 )

　被説明変数をy, 説明変数をx1, x2, x3 とする回帰分析（複数の説明変数がある場合，説明変数を ＋ でつなぐ）

- wage1.lm <- lm(wage ~ educ)

　回帰分析の結果をwage1.lm というオブジェクトに代入せよという命令。オブジェクト名はwage1.lm でなくてもかまわない。また，.lm という拡張子に意味があるわけではなく，lm( )の結果だということを忘れないようにつけただけの名前。回帰分析の結果をこのように名前を付けて保存しておくと，後で繰り返し利用できるので便利。

- summary(wage1.lm)　回帰分析の結果の要約を出力する

# Rでの回帰分析(3)

summary(wage1.lm)で次のような結果が表示されます

```
> summary(wage1.lm)

Call:
lm(formula = wage ~ educ)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3396 -2.1501 -0.9674  1.1921 16.6085

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.90485    0.68497  -1.321    0.187
educ         0.54136    0.05325  10.167   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.378 on 524 degrees of freedom
Multiple R-squared:  0.1648,    Adjusted R-squared:  0.1632
F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

係数の推定値(estimate)，標準誤差(std. error)，t値(t- value), p値が出力される
educ の係数の推定値0.54136，標準誤差 0.05325，t値は10.167 など　t値，p値は重回帰の際に解説

回帰の標準誤差：残差の標準偏差の推定値

$R^2$ 決定係数

# Rでの回帰分析(3)

回帰分析の結果はsummary(object)で取り出せたが，他の情報も取り出せる

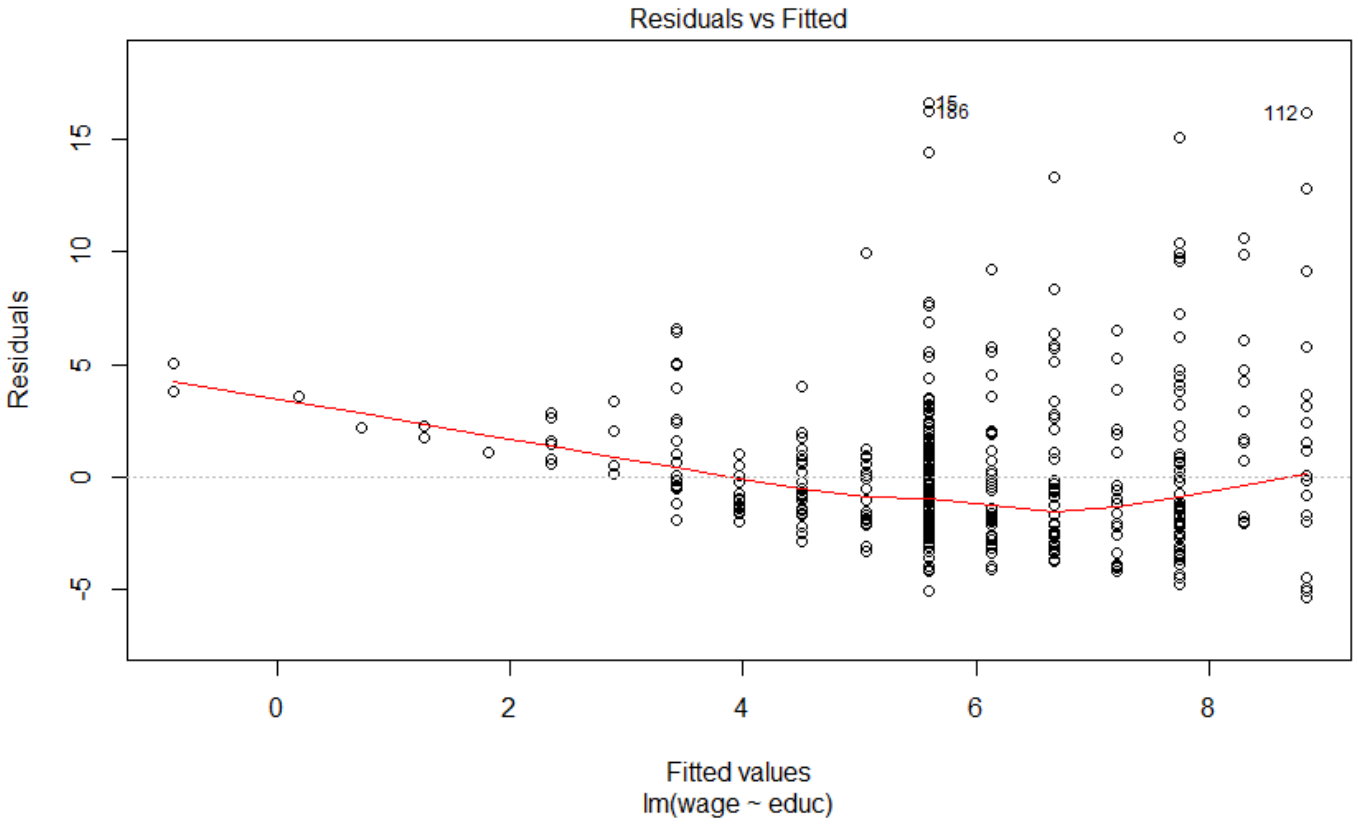
summary(object) 回帰分析の結果の要約
coef(object) 係数の推定値
resid(object) 残差
fitted(object) 回帰モデルの推定値
deviance(object) 残差平方和
plot(object) 残差のチェックのためのグラフ
confint(object) 係数の信頼区間
---

コマンドラインで，coefficients(wage1.lm)またはcoef(wage1.lm)とタイプすると推計された係数が出力される

coef(wage1.lm)[1] coef(wage1.lm)[2] で係数ベクトルの1番めの要素と2番めの要素が出力される

# 残差分析　R

plot(wage1.lm)で出力される図の１部



# 残差分析の必要性

- 回帰分析の前提
  - 誤差項の系列相関は無い
  - 誤差項の分散は一定（分散均一性）
  - 説明変数と誤差項は独立
- これらの前提が成り立つ場合に，係数の信頼区間や検定などの統計的推測の正しさが保証されます
- 逆に言えば，これらの前提が成立しない場合，統計的推測は正しくありません
- 残差の分析によって，回帰分析の前提が満たされないことが分かった場合，回帰式の定式化のやり直しが必要になります
- これらの問題は重回帰分析の際に扱います（今の段階では気にしなくていいです）。

# 非線形効果のとらえ方

線形回帰：パラメータに関し線形モデル

→説明変数や被説明変数を次のように変形することで非線形効果を捉えられる（どんな形になるかグラフを描いてみて確かめておくこと）

なお，計量経済学的分析では，xまたはyの対数変換がよくつかわれる（この理由は次のページ）

$$y = a + b \ln(x)$$
$$\ln(y) = a + b\, x$$
$$\ln(y) = a + b \ln(x)$$
$$y = a + b_1 * x + b_2 * x^2$$
$$y = a + b / x$$
$$y = a + b_1 / x + b_2 * x$$

# 対数

- 定義　$y = \ln x \Leftrightarrow x = \exp(y)$

- 重要な公式

$$\ln xy = \ln x + \ln y$$
$$\ln x^{\alpha} = \alpha \ln x$$

- 対数値の変化=もとの変数の比率での変化

$$\ln x(1+h) - \ln x = \ln(1+h) \approx h$$

  - 導出は対数関数の1次近似式から
  - h=0.01, 0.02 として，電卓，excel ,あるいはRで確かめてみよ
  - Rではコンソール画面で log(1.01)とタイプすると1.01の対数値を返してくれる

- ln は自然対数：　$e$=2.7182..を底（てい）とする対数
- $\exp(y) = e^{y}$
- レジュメでは，lnと書いたり，logと書いたりしますが，全て自然対数だと思ってください。

# 対数(2)

- 対数の性質から次の式が成り立つ

$$y = a + b\ln x \Rightarrow b = \frac{\Delta y}{\Delta \ln x} = \frac{\Delta y}{\Delta x/x} \qquad (1)$$

$$\ln y = a + bx \Rightarrow b = \frac{\Delta \ln y}{\Delta x} = \frac{\Delta y/y}{\Delta x} \qquad (2)$$

$$\ln y = a + b\ln x \Rightarrow b = \frac{\Delta \ln y}{\Delta \ln x} = \frac{\Delta y/y}{\Delta x/x} \qquad (3)$$

- 係数bの意味
  (1) $x$の1%*の変化がyの何単位の変化をもたらすか
  (2) $x$の1単位の変化がyの何%*の変化をもたらすか
  (3) $x$の1%の変化がyの何%の変化をもたらすか
  ＊: $\Delta x/x$の1単位の変化は$x$の100%の変化であることに注意；(1), (2)は不正確な表現です

# 回帰分析の解釈

- 係数の意味
  - wage = a + b* educの場合
    - 教育年数(educ)が1年増加すると賃金(wage)は何単位増加するか
  - log(wage) = a + b*educ の場合
  - educが1単位増加したとき，賃金の対数値が何単位増加するか → 賃金が何%増加するか
- 教育年数の効果
  - educを連続変数とするより，高卒，大卒のようなカテゴリー分けの方が適切？
- 因果関係（代替的なモデルが考えられる）
  - 教育年数 →人的資本の蓄積の効果
  - 教育年数 → その人の能力の証
    - 高学歴者は学業に耐えられるだけの能力をもともと備えていた
    - スクリーニングの機能だけ（人的資本の蓄積ではない）

# みせかけの関係

- Wooldridge の chapter2 example 2.12 （meap93.raw）
- 生徒の成績と高校のlunch programの関係
  - lnchprg ：perc. of studs. in sch. lunch prog（昼食補助プログラムに参加している生徒の比率）
  - math10 ：perc studs passing MEAP math（数学の学力テスト）
  - ミシガン州の高校： 408校, 1992-1993年
  - 他の条件が一定なら，昼食への補助 → 生徒の成績にプラスの影響のはず
- 推計結果（マイナスの影響）

$$\text{math10} = 32.14 - 0.319\ \text{lnchprg} \qquad n=408,\ R^2=0.171$$

- そのまま解釈すると昼食補助プログラムへの参加率が高いほど学力は低い。しかし，これはみせかけの関係。貧困家庭の比率が高いと昼食補助プログラムへの参加率が高く，また貧困家庭の生徒は学力が低いという関係をとらえたものと考えられる→みせかけの相関関係
- 回帰分析は，説明変数から被説明変数の因果関係を常に表すわけではないことに注意

# 問題1

- データ：wage1.xls, wage1.raw　（教育年数と賃金の関係）

1. 次の単回帰を行え
   - 被説明変数：wage（賃金），説明変数：educ（教育年数）
   - educの係数を解釈せよ。
2. 次の単回帰を行え
   - 被説明変数：lwage（賃金の対数値），説明変数：educ（教育年数）
   - educの係数を解釈せよ

# 問題2

- データ：wage2.xls, wage2.raw
- 変数の説明はwage2.desをみること（wage: 月給，IQ: IQのスコア）

1. wageとIQの要約統計量を求めよ
   - summary( 変数名)を用いる
2. wage，lwage(wageの対数値），IQのヒストグラムを描け
   - hist(変数名）；図はWord等にexportする
3. IQ（横軸）とwage（縦軸）の散布図を描け。またIQを横軸，縦軸をlwageにした場合の散布図も描け。
4. 被説明変数をwage，説明変数をIQにした回帰分析を行え。
5. 被説明変数をlwage（wageの対数値），説明変数をIQにした回帰分析を行え。