第2章

単回帰モデル（1要因のモデル）

$y = \beta_0 + \beta_1 x + u$

Ch.2　単回帰モデル

2-1　単回帰モデルの定義

◆（線形）単回帰モデル

The simple linear regression model

- 「変数 x で変数 y を説明」

切片（Intercept）　傾き（Slope parameter）

$$y = \beta_0 + \beta_1 x + u$$

被説明変数
・Dependent variable,
・explained variable,
・response variable,…

説明変数
・Independent variable,
・explanatory variable,
・regressor,...

誤差項
・Error term,
・disturbance,
・unobservables,...

2-1　単回帰モデルの定義

◆単回帰モデルの解釈

- 「xの変化にyがどう変化するかを調べる」
  - $\frac{\Delta y}{\Delta x} = \beta_1$ → 独立変数が1単位増加した場合、従属変数はどれだけ変化するか？

  ただし前提として
  - $\frac{\Delta u}{\Delta x} = 0$ → 独立変数が1単位増加したとき、他の変数すべてが等しいor変化しないと想定
    - 線形単回帰モデルは実際の適用可能性は非常に低いが、その議論は教育的観点から有用

2-1　単回帰モデルの定義

◆例2.1 大豆収穫量と肥料

$$yield = \beta_0 + \beta_1 fertilizer + u$$

← 降雨量、土地の質、害虫etc.

他のすべての要因を固定して、
収量に及ぼす肥料の影響を計測

◆例2.2 賃金方程式

$$wage = \beta_0 + \beta_1 educ + u$$

← 職業経験、在職期間、
仕事倫理、知性etc.

他のすべての要因を固定し、もう1年余分
に教育を受けたときの時給の変化を測定

2-1　単回帰モデルの定義

◆因果性の解釈はどんな時に可能か？

- 「条件付き平均独立」という仮定
  - $\mathrm{E}(u|x) = \mathrm{E}(u) = 0$ ← 「uの平均はxの値に影響を受けない」
    「説明変数xには、他の（観測できない）要因uの情報は含まれていない」
    「xとuは独立」
- 例：賃金方程式

  $wage = \beta_0 + \beta_1 educ + u$ ← 知性etc.
  - より多くの教育を受ける個人は平均的により知性的であるので、「条件付き平均独立」の仮定が成立することはほとんどない

2-1 単回帰モデルの定義

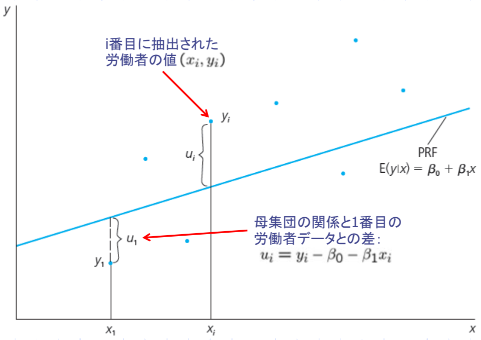
◆Population regression function (PRF)

- 条件付き平均独立仮定での単回帰モデル

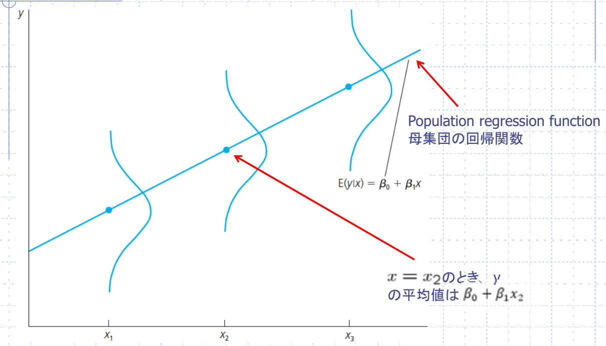
$$E(y|x) = E(\beta_0 + \beta_1 x + u|x)$$
$$= \beta_0 + \beta_1 x + E(u|x)$$
$$= \beta_0 + \beta_1 x$$

  - これは被説明変数yの期待値が説明変数xの線形関数として表現できることを明示

2-1 単回帰モデルの定義



2-1 単回帰モデルの定義



2-2 最小二乗推定値の導出

◆最小二乗推定値の導出

- 回帰モデルの推定にはデータが不可欠
- n 個の観測値をもつランダムサンプル

(x_1, y_1) ← 第1観測値

(x_2, y_2) ← 第2観測値

(x_3, y_3) ← 第3観測値

⋮

(x_n, y_n) ← 第n観測値

$\{(x_i, y_i) : i = 1, \ldots, n\}$

i 番目の観測値である説明変数の値

i 番目の観測値である被説明変数の値

2-2 最小二乗推定値の導出

◆「可能な限り」データに沿う推定方法とは？

- 回帰残差Regression residuals

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- 回帰残差の平方和を最小化

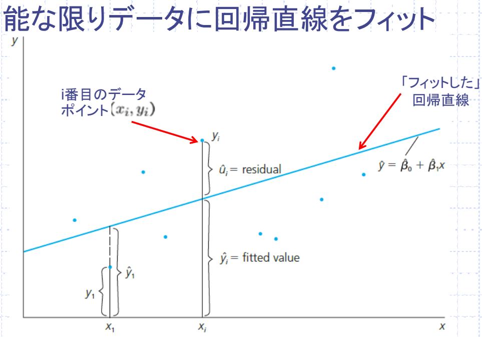
$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1$$

- 最小二乗法（OLS）による推定値

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

2-2 最小二乗推定値の導出

◆可能な限りデータに回帰直線をフィット



2-2 最小二乗推定値の導出

◆例2.3 CEOの給与と株主資本利益率(ROE)

$$salary = \beta_0 + \beta_1 roe + u$$

CEOの給与(1000ドル)　　当該企業の平均ROE

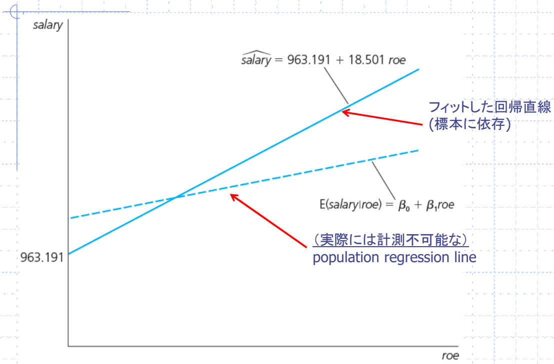
- フィットした回帰式

$$\widehat{salary} = 963.191 + 18.501\ roe$$

切片　傾き

- 因果性解釈
  - ROEが1%上昇すれば、CEOの給与は$18,501変化することを予測

2-2 最小二乗推定値の導出



2-2 最小二乗推定値の導出

◆例2.4 賃金と教育

$$wage = \beta_0 + \beta_1 educ + u$$

時給(ドル)　　教育年数

- フィットした回帰式

$$\widehat{wage} = -0.90 + 0.54\ educ$$

切片　傾き

- 因果性解釈
  - この標本では、1年の追加教育が0.54ドルの時給増加と関連

2-2 最小二乗推定値の導出

◆例2.5 投票結果と選挙運動費用

$$voteA = \beta_0 + \beta_1 shareA + u$$

候補者Aの得票率　　候補者Aの選挙運動費用比

- フィットした回帰

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

- 因果性解釈
  - 候補者Aの支出比率が1%増加すると、投票総数の46.4%を得る。

2-3 OLSの特性

◆OLSの特性

- 予測値と残差(Fitted values & residuals)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \hat{u}_i = y_i - \hat{y}_i$$

予測値　　実際値と予測値の差(=残差)

- OLS回帰の代数的性質

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} x_i \hat{u}_i = 0 \qquad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

残差の合計は0　　説明変数と残差の共分散は0　　yとxの標本平均は回帰直線上にある

2-3 OLSの特性

TABLE 2.1 Fitted Values and Residuals for the First 15 CEOs

obsno	roe	salary	salaryhat	uhat
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	−438.7678
15	56.3	2011	2004.808	6.191895

x_i　y_i

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{u}_i = y_i - \hat{y}_i$$

12番目のCEOの給与は、その会社のROE情報からの予測値より$526,023 低い

2-3　OLSの特性

◆当てはまりの良さ(Goodness-of-Fit)

- 「説明変数xは被説明変数yをどの程度説明できるのか？」
- 変動の尺度

$$SST \equiv \sum_{i=1}^{n}(y_i-\bar{y})^2 \qquad SSE \equiv \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \qquad SSR \equiv \sum_{i=1}^{n}\hat{u}_i^{\,2}$$

Total Sum of Squares
被説明変数の全変動

Explained Sum of Squares
回帰で説明される変動

Residual Sum of Squares
回帰で説明されない変動

2-3　OLSの特性

◆全変動の分解

$$SST = SSE + SSR$$

全変動　説明可能部分　説明不能部分

- 「当てはまりの良さ」指標 (R-squared)

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- 決定係数R^2は、回帰によって説明される全変動の割合を測定

2-3　OLSの特性

◆例2.8 CEOの給与とROE

$$\widehat{salary} = 963.191 + 18.501\ roe$$
$$n = 209, \quad R^2 = 0.0132$$

この回帰式はCEO給与の全変動の1.3%を説明

◆例2.9 得票率と選挙活動費用

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$
$$n = 173, \quad R^2 = 0.856$$

この回帰式は選挙結果の全変動の85.6% を説明

- 注意：高いR^2は、必ずしも回帰式が因果性解釈を可能とすることを意味するものではない

2-4　計測単位と関数形

◆非線形関数：半対数形

- 例2.10 対数賃金の教育年数への回帰

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

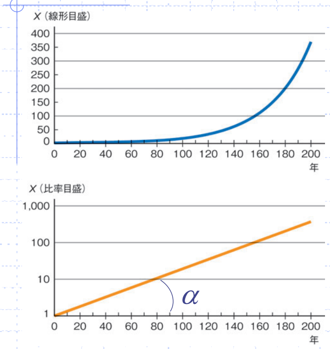
賃金の自然対数

- 回帰係数の解釈を変更

$$\beta_1 = \frac{\Delta\log(wage)}{\Delta educ} = \frac{1}{wage}\cdot\frac{\Delta wage}{\Delta educ} = \frac{\frac{\Delta wage}{wage}}{\Delta educ}$$

賃金の変化率　教育年数の1年増加

補論：対数表示について



- 上図は通常表示、下図は対数表示
- 両図ともに年平均3%成長率を表示
- 下図で成長率は傾きαで示される
- 成長率に注目するときには、下図の対数表示が便利

2-4　計測単位と関数形

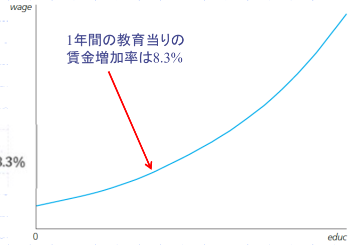
◆フィットした回帰

$$\widehat{\log}(wage) = 0.584 + 0.083\ educ$$

教育の1年追加は賃金を8.3%増加
(教育収益率＝8.3%)

例：

$$\frac{\frac{\Delta wage}{wage}}{\Delta educ} = \frac{\frac{+0.83\$}{10\$}}{+1\ year} = 0.083 = +8.3\%$$



2-4　計測単位と関数形

◆非線形関数：対数形

- 例2.11 CEO給与と企業の売上

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

CEO給与の自然対数　　当該企業売上の自然対数

- 回帰係数解釈の変更：

$$\beta_1 = \frac{\Delta \log(salary)}{\Delta \log(sales)} = \frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}}$$

給与の変化率

売上1%増加

- 対数変化は常に変化率の変化

2-4　計測単位と関数形

◆CEO給与と企業売上：推定結果

$$\widehat{\log}(salary) = 4.822 + 0.257 \log(sales)$$

- 1%の売上増 → 0.257%のCEO給与増
  - 例：

$$\frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}} = \frac{\frac{+2{,}570\$}{1{,}000{,}000\$}}{\frac{+10{,}000{,}000\$}{1{,}000{,}000{,}000\$}} = \frac{+0.257\%\ salary}{+1\%\ sales} = 0.257$$

- log-log形は弾性値一定
  - （⇔半対数形は半弾性値）

2-5　OLS推定量の期待値と分散

◆OLS推定量estimatorsの期待値と分散

- 推定回帰係数はランダム変数
  - 係数はランダム（無作為）標本からの算出

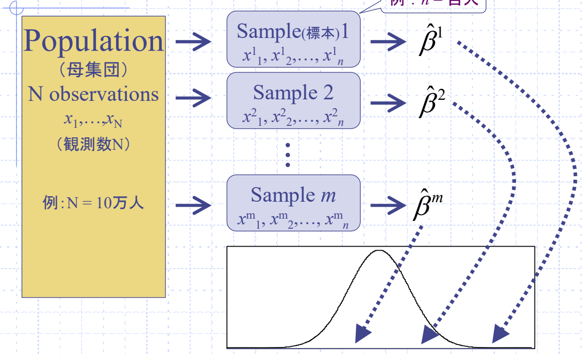
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

データは標本に依存→データはランダム

- ポイント：繰り返し標本のもとで各推定量が
  - 「平均的にどのような値になるのか」
  - 「散らばり具合はどれくらい大きいのか」

$$E(\hat{\beta}_0) = ?,\ E(\hat{\beta}_1) = ? \quad Var(\hat{\beta}_0) = ?,\ Var(\hat{\beta}_1) = ?$$

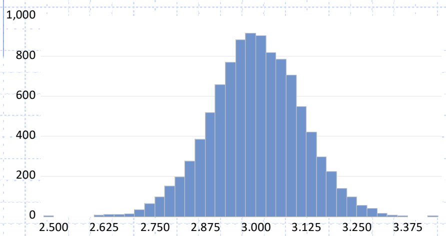
OLS推定値の分布



OLS推定値の分布

◆Monte Calro実験

- $y = 2 + 3x + u$
- $N = 100{,}000,\ n = 100,\ m = 10{,}000$



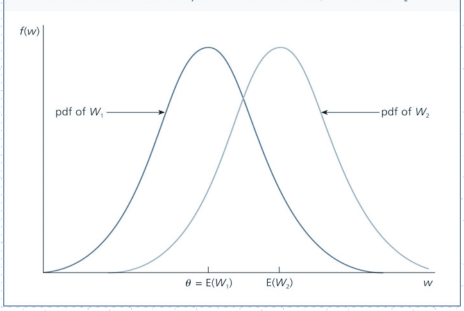
Series: BETASER
Sample 1 10000
Observations 10000

Mean	3.001257
Median	3.001224
Maximum	3.429046
Minimum	2.497484
Std. Dev.	0.109880
Skewness	-0.052916
Kurtosis	3.031889
Jarque-Bera	5.090462
Probability	0.078455

不偏性

FIGURE C.1

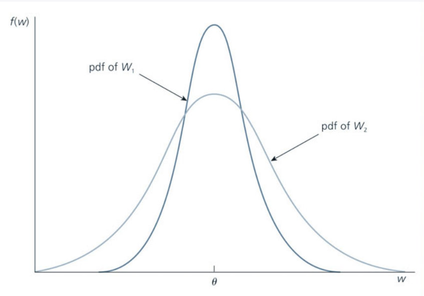
An unbiased estimator, W_1, and an estimator with positive bias, W_2.



効率性

FIGURE C.2

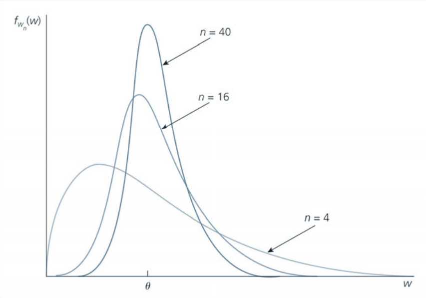
The sampling distributions of two unbiased estimators of θ.



一致性

FIGURE C.3

The sampling distributions of a consistent estimator for three sample sizes.



2-5 OLS推定量の期待値と分散

◆線形回帰モデルの標準仮定

- 仮定 SLR.1 (Linear in parameters)

$$y = \beta_0 + \beta_1 x + u$$ ← 母集団ではyとxの関係は線形

- 仮定 SLR.2 (Random sampling)

$$\{(x_i, y_i) : i = 1, \ldots, n\}$$ ← データは母集団から得られたランダムな（無作為）標本

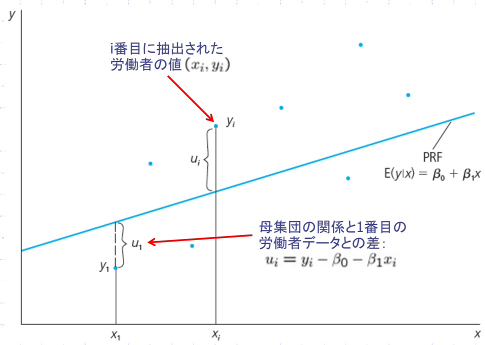
$$y_i = \beta_0 + \beta_1 x_i + u_i$$ ← よって「各データポイントは母集団の式に沿う」と仮定

2-5 OLS推定量の期待値と分散

◆無作為抽出について：賃金と教育

- 例：母集団ーA国の全労働者
  - 母集団での賃金と教育の線形関係の存在を想定
  - 母集団から労働者1名を無作為抽出
    - 事前にどの労働者抽出されるか不明
      →抽出された労働者の賃金と教育は無作為random
  - （労働者を母集団に戻し）同様にn回抽出
  - 抽出された労働者の賃金と教育のデータを用いて賃金と教育の線形関係を推定

2-5 OLS推定量の期待値と分散



2-5 OLS推定量の期待値と分散

◆線形回帰モデルの仮定（続き）

- 仮定 SLR.3 (説明変数の標本変動)

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$$

  - 説明変数の値すべてが同じではない
    - そうでなければ、「説明変数値の違いがどのように被説明変数値の違いにつながるか」の分析は不可能
- 仮定 SLR.4 (Zero conditional mean)

$$E(u_i | x_i) = 0$$

  - 説明変数の値は、観察されない要因uの平均に関する情報を含まない

2-5 OLS推定量の期待値と分散

◆定理2.1（OLS推定量の不偏性）

$$SLR.1-SLR.4 \quad \Rightarrow \quad E(\hat{\beta}_0)=\beta_0,\ E(\hat{\beta}_1)=\beta_1$$

- 不偏性の解釈
  - 推定された係数$\hat{\beta}_i$は無作為抽出の結果である標本に依存し、<u>真の値β_iより大きくも小さくもなる</u>。
    - 標本によっては$\hat{\beta}_i$はβ_iとかなり異なる場合もある。
  - しかし平均的には、母集団でのyとxの関係を表す値β_iと等しくなる。
    - 「平均的」とは、「無作為標本の抽出→推定」を何度も繰り返した場合を意味する。

2-5 OLS推定量の期待値と分散

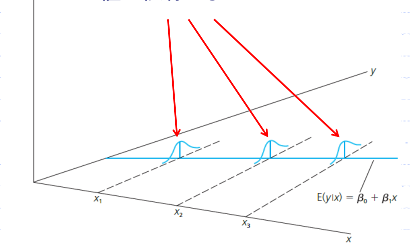
◆OLS推定量の分散

- 標本によって、推定値は真の母集団の値に近くも遠くにもなる。
  - 平均的に、推定値は真の母集団の値からどれほど離れているか？（標本変動性）
  - 標本変動性は、推定量の分散によって測定
    $Var(\hat{\beta}_0),\ Var(\hat{\beta}_1)$
- 仮定 SLR.5 (均一分散Homoskedasticity)
  $Var(u_i|x_i)=\sigma^2$
  - 説明変数の値は、観察されない要因uの変動性についての情報を含まない

2-5 OLS推定量の期待値と分散

◆均一分散の図解

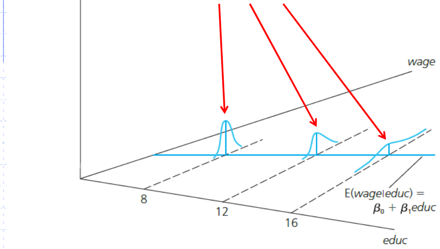
観察されない要因uの変動性（＝分散）は、説明変数の値に依存しない



2-5 OLS推定量の期待値と分散

◆不均一分散の例：賃金と教育

- 賃金に影響を及ぼす（観察されない教育以外の）要因の分散は、教育の水準とともに増加する



2-5 OLS推定量の期待値と分散

◆定理 2.2 (OLS推定量の分散)

- 仮定SLR.1 - SLR.5のもとで

$$Var(\hat{\beta}_1)=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sigma^2}{SST_x}$$

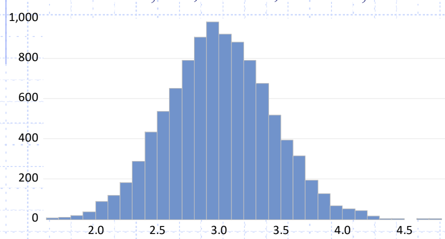
$$Var(\hat{\beta}_0)=\frac{\sigma^2 n^{-1}\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sigma^2 n^{-1}\sum_{i=1}^{n}x_i^2}{SST_x}$$

- 推論：
  - 推定された回帰係数の標本変動性（=分散）は、
    - 観測されない要因uの変動性σ^2が小さいほど小さい
    - 説明変数の変動SSTxが大きいほど小さい
      - 標本サイズnが大きいほど小さい

OLS推定値の分布

◆Monte Calro実験

- $y=2+3x+u$
- $N=100{,}000,\ n=10,\ m=10{,}000$



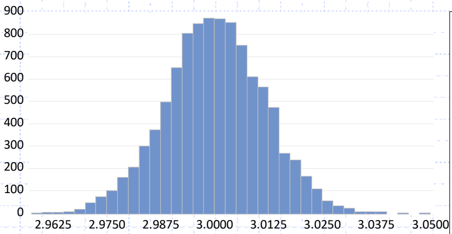
Series: BETASER
Sample 1 10000
Observations 10000

Mean	3.003766
Median	2.997587
Maximum	4.758092
Minimum	1.637265
Std. Dev.	0.419043
Skewness	0.058148
Kurtosis	2.952597
Jarque-Bera	6.571541
Probability	0.037412

OLS推定値の分布

◆Monte Calro実験

- $y = 2 + 3x + u$
- $N = 100{,}000,\ n = 10{,}000,\ m = 10{,}000$



Series: BETASER	
Sample 1 10000	
Observations 10000	
Mean	3.000157
Median	3.000119
Maximum	3.047796
Minimum	2.959836
Std. Dev.	0.011373
Skewness	-0.001241
Kurtosis	3.048835
Jarque-Bera	0.996245
Probability	0.607671

2-5 OLS推定量の期待値と分散

◆誤差分散の推定

- uの分散はxに依存しない（=無条件分散と同値）

$$Var(u_i|x_i) = \sigma^2 = Var(u_i)$$

- 残差分散を計算することで誤差分散の推定可能
  - 誤差分散の不偏推定は、観測数nから回帰係数の数を減算することで得られる

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$ ○ 誤差分散の不偏推定

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$ × 誤差分散のバイアス推定

2-5 OLS推定量の期待値と分散

◆定理2.3（標準分散の不偏性）

$$SLR.1 - SLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

- 回帰係数の標準誤差（≠標準偏差）の計算
  - 未知のσ^2にσ^2の推定値を代入

$$se(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2 / SST_x}$$

$$se(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 n^{-1}\sum_{i=1}^{n}x_i^2 / SST_x}$$

- 回帰係数の推定された標準偏差は「標準誤差」と呼ばれ、回帰係数がどれほど正確に推定されたかの指標となる

2-6 原点からの回帰

◆切片なしの回帰式：

$$\tilde{y} = \tilde{\beta}_1 x \quad (2.63).$$

◆最小化問題をFOC（1階条件）で解くと、OLS推定量は

$$\tilde{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad (2.66) \quad \Leftrightarrow \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (2.19)$$

*切片の存在は常に$E(u) = 0$を満たすようにβ_0を調整

2.7 2値説明変数への回帰

◆xが0または1のどちらかに等しい

$$y = \beta_0 + \beta_1 x + u$$

$$E(y|x=0) = \beta_0 \quad E(y|x=1) = \beta_0 + \beta_1$$

- この回帰により、xの状態によってyの平均値が異なる

$$\beta_1 = E(y|x=1) - E(y|x=0)$$

  - xが2値である場合、OLSの統計的性質は変わらない
    - ⇔yが2値である場合、OLSの統計的性質は変わることに注意

2.7 2値説明変数への回帰

◆2-7a 反実仮想の結果、因果関係、政策分析

- 政策分析において、処置効果を次のように定義

$$\tau_i = y_i(1) - y_i(0)$$

  - あるiについて、我々は$y_i(1)$か$y_i(0)$のどちらかのみを観測し、両方を観測することはない。よって、実際にτ_iを観測することはないことに注意
- 平均処置効果ATEを次のように定義

$$\tau_{ate} = E[y_i(1)] - E[y_i(0)]$$

2.7 2値説明変数への回帰

- x_i を2値政策変数とする

  $y_i = (1 - x_i)y_i(0) + x_i y_i(1)$ ← $x_i = 0$のとき$y_i = y_i(0)$ $x_i = 1$のとき$y_i = y_i(1)$

  $\rightarrow y_i = y_i(0) + [y_i(1) - y_i(0)]x_i$

  - ここで $y_i(0) = \alpha_0 + u_i(0)$とし、定義 $\tau_i = y_i(1) - y_i(0)$より

  $y_i = \alpha_0 + \tau x_i + u_i(0)$

  - したがって y を x に回帰させると、(一定の)処置効果の推定値が得られる
  - 無作為割付である限り、OLSEは処置効果 τ の不偏推定量

49

2.7 2値説明変数への回帰

◆無作為割付Random assignment

- 対象は処置群と対照群に無作為に分けられ、両群間に処置以外の系統的な差異はない
- 実際、ランダム化比較試験(RCTs)は実施コストが高く、倫理的問題が生じる可能性も
- 現実的にはRCTsはしばしば実現不可能だが、もし無作為割付が可能であれば、この種の実験を検討することは有用。これは無作為割付に対する潜在的な障害(多変量回帰でコントロールできうるもの)の識別に役立つ

50

2.7 2値説明変数への回帰

◆例2.14 職業訓練プログラムの評価

- 実質所得は、職業訓練プログラムへの参加を示す2項変数に回帰

  $\widehat{re78} = 4.55 + 1.79train$
  $n = 445, R^2 = 0.018$

  ← $re78$は1978年の実質所得 (単位:千$)、$train$は同プログラム参加であれば1、そうでなければ0の2値変数

  - 研修に参加した人は、参加しなかった人に比べて収入が1,790ドル多い
  - これは、参加しなかった人の平均収入4,550ドルと比較して39.3%の増加

51