事例研究(ミクロ経済政策･問題分析 III)

- 規制産業と料金･価格制度 -

(第7回 – 手法(3) 応用データ解析/基礎的手法)

2010年 6月 2日

戒能一成

0. 本講の目的

(手法面）

- **応用データ解析の手順や基本的な作業の流れ（Strategy）**を理解する
- 特にグラフ化や統計検定などの手法を用いた、データ解析手法の選択と検定･確認について理解する

（内容面）

- 計量経済学･統計学を実戦で応用する際の基礎的留意点を理解する (1)

1. 制度の効果を測るには

1-1. 政策分析の基本手順

– 料金･価格制度やその変更が及ぼす効果を推計するためには、以下の 2つの作業が必要

1) **制度変更による経済データへの影響経路と、因果関係･寄与度**の推定 (「モデル構築」)
→ 制度変更がどのような変化をもたらすか？

2) 制度の創設･変更と**同時に生じた経済データの「有意な変化」**の計測 (「モデル実証」)
→ 数量･価格や費用は本当に変化したか？
(→ 変化していれば余剰分析が応用可能)

1. 制度の効果を測るには

1-2. 政策分析の条件(1)

－ 制度（変更）の効果推計に際し充足すべき条件

1) **他の条件一定** “Ceteris Paribus”

→ 制度変更以外の外的要因変化の影響が、可能な限り十分除去されていること

2) **政策影響の独立性** “Unconfoundness”

→ 制度(変更)の影響が、制度の実施/非実施と独立と見なせること (影響の均質性)

3) **対照群･時間**の存在 “Overlap”

→ 制度(変更)が非実施の群･時間があること

1. 制度の効果を測るには

1-3. 政策分析の条件(2)

– 制度(変更)の効果推計に際し充足すべき条件
→ **分析手法･手順の選択や精度を規定**

時間	→	0	1	･･･	t (制度変更)	･･･	n (2010)
対象 ↓	X1	y10	y11	･･･	y1t (変更)	･･･	y1n (変更)
	X2	y20	y21	･･･	y2t (変更)	･･･	y2n (変更)
	X3	y30	y31	･･･	y3t (--)	･･･	y3n(--)
	X4	y40	y41	･･･	y4t (--)	･･･	y4n(--)

対照時系列比較？
→ 外的要因除去が必要

対照群横断比較？
→ 独立性が必要
(影響の均質性)

異質性が存在

外的要因(毎年度変化)の影響が存在

1. 制度の効果を測るには

1-4. 制度影響モデルの仮構築(1)

- 問題とする財サービスの費用、価格･料金、数量などについて、**制度が及ぼす影響経路･内容を、経済理論に基づく簡単な影響モデルで記述**
  → 費用、料金･価格、数量の変化
- 当該変化において、**外的要因**が存在する場合、(後で取除くことを目的に)外的要因の影響経路と内容を加味したモデルを構築
  → 需要変化(率)、一般物価･金利、他の制度

1. 制度の効果を測るには

1-5. 制度影響モデルの仮構築(2)

– 制度影響モデル(例: 投資影響による費用変化)

- C(t) = Cfix(t,H) + cval(t) * Q(t) + ε (t)

→ Y(t) = α 1 (or α 0) + β * X (t) + ε (t)

- Cfix(H) = △Cfixpo(H(1or0)) + Cfixtr
- cval(t) = cfuel(t) + cwaste(t)

C(t): t期実質総費用, Q(t): t期供給量, ε (t): 誤差項

Cfix (t,H) : t期固定費

△Cfixpo (H(1or0)) 政策実施(H(1))以降の実質減価償却費 + 同利払費変化 (政策影響部分)

Cfixtr 過去10年平均実質固定費 (不変)

cval (t) : t期可変費原単位

cfuel(t),cwaste(t) 実質単位燃料費･ゴミ処理費 (外部要因)

1. 制度の効果を測るには

1-6. 制度影響モデルの実測･修正

- 1-4. で構築した制度影響モデルを、実際の統計データを用いて実測する
- 実際の統計処理はパッケージ･ソフトで実施する（STATA, EViews, ･･･ ）
  → 重要なのは、**必要とされる前提条件に応じた適切な手法の選択と、検定結果などの解釈**
- 明らかに理論と矛盾する結果が出た場合には、1-4. に戻って制度影響モデルを再考する（ex. 正の価格弾力性, 負の所得効果･･･）

2. 応用データ解析の基礎(1): 線形回帰モデル

2-1. 線形回帰モデルとは

- 最も簡単な線形回帰モデルは、被説明変数(例: 費用)を説明変数(前期固定資産、燃料費・・・)で**最小二乗法により回帰分析**したモデル

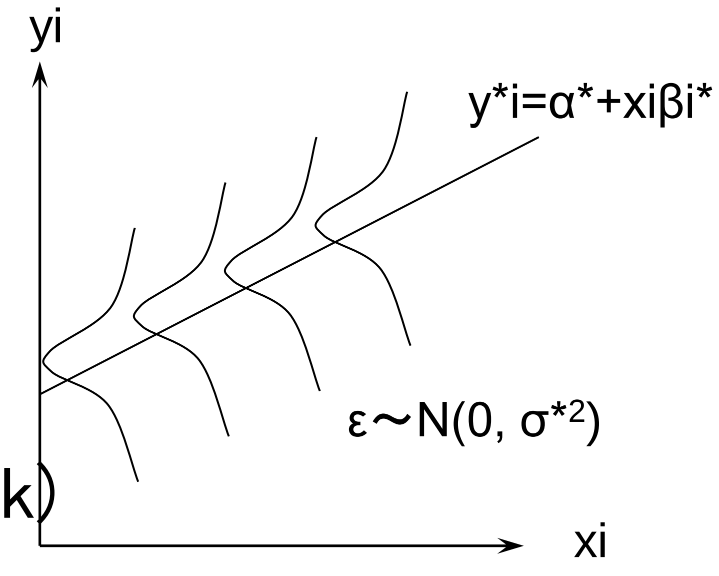
$$y = \alpha + x'\beta + \varepsilon$$

$$\rightarrow \quad y^* = \alpha^* + x'\beta^*$$

$$\alpha^* = y - x'\beta^*$$

$$\beta^* = (x'x)^{-1}x'y$$

$$\sigma^{*2} = (y-y^*)'(y-y^*)/(n-k)$$



- 最も簡単で扱いやすい手法だが・・・

2. 応用データ解析の基礎(1): 線形回帰モデル

2-2. 線形回帰モデルと前提条件(1)

– 線形回帰モデルが適用できる前提条件は 4つ

#1: **線形性** Linearity

– 適切な変換で $y = \alpha + x'\beta + \varepsilon$ 型になること

→ 適用困難例と対処

– yが**離散値**(0, 1), **切断値**($y_i \mid y_i > 0$)

→ ダミー変数・切断変数モデル回帰

→ 平均措置効果(ATE; matching 他)

– y がCES型($= (K^{\delta} + L^{\delta})^{\gamma}$)等**連続非線形**

→ 非線形回帰 (数値解析法)

2. 応用データ解析の基礎(1): 線形回帰モデル

2−3. 線形回帰モデルと前提条件(2)

#2: **説明変数の外生性** Strict Exogeniety

– 説明変数 X が誤差項 ε と独立であること

⇔ $E(\varepsilon_i \mid X) = 0$ (i = 1 to n)

→ 適用困難例と対処

– 説明変数 X が**誤差項ε と相関**あり

(XとYが需給均衡・同時決定の場合など)

→ 操作変数法 Instrumental Variable

Xとは相関があるが εとは相関がない変数 Z を探して併用回帰

2. 応用データ解析の基礎(1): 線形回帰モデル

2-4. 線形回帰モデルと前提条件(3)

#3: **説明変数の非多重共線性** No Multicolinarity

– 説明変数 xi が他の xj (i≠j)の組合わせで

表現できないこと ⇔ rank $X_{kxn}' X_{nxk} = k$

→ 適用困難例と対処

– 説明変数 X の間での**相関高**

→ 主成分回帰

→ 一部変数除去 (= モデルの見直し)

(ex. ダミー変数は全ての分類に設定できない

∵ 少なくとも分類の 1つは他の補集合)

2. 応用データ解析の基礎(1): 線形回帰モデル

2-5. 線形回帰モデルと前提条件(4)

#4: **誤差項の均一分散性** Homoskedasticity

- 誤差項 ε の分散は全て σ^2 で共分散なし

  $\Leftrightarrow E(\varepsilon'\varepsilon \mid X) = \sigma^2 I$
- 通常さらに 誤差項ε は正規分布 $N(0, \sigma^2 I)$ と仮定する

→ 適用困難例と対処

- **分散が不均一**

  → 不均一分散回帰 Heterosked. robust
- **系列相関**あり **[重要]**

  → 時系列分析法 Time Series Analysis

2. 応用データ解析の基礎(1): 線形回帰モデル

2-6. 線形回帰モデルと実用上の問題

- 現実の料金･価格制度の分析という視点からは、**線形回帰モデルの前提条件が成立しない場合多**

#1 **線形性:** **成立しない場合有**

(→ “凸/凹型” Convex/Concave, 離散型など)

#2 説明変数の外生性: (回避可能)

#3 説明変数の非多重共線性: (回避可能)

#4 **誤差項の均一分散性:** **ほぼ確実に成立せず**

(→ 殆どの場合「時系列相関」あり, 粘着性など)

→ 分析手法として**時系列分析･パネルデータ分析**が有効 (後述)

3. 応用データ解析の基礎(2): 線形回帰と検定

3−1. 決定係数・自由度修正済決定係数

- **決定係数 R^2** ；最も一般的な精度指標
  - 推計式 $y^* = \alpha^* + x'\beta^*$ が、実際の y の変動のどの程度を説明しているかを表す係数
    → $0 \leqq R^2 \leqq 1$, $R^2 = 1 - (y-y^*)^2/(y'(I-x(x'x)^{-1}x)y)$
  - 但し、説明変数 X をたくさん使うと R^2 は実際の精度と無関係に大きくなるので、**自由度修正済決定係数 R^2 (Adjusted R^2)** が用いられる
    → $\text{Adj. } R^2 = 1 - (n-1)/(n-k)(1 - R^2)$
    n: 試料数　k:説明変数数　$\text{Adj.}R^2 \leqq 1$

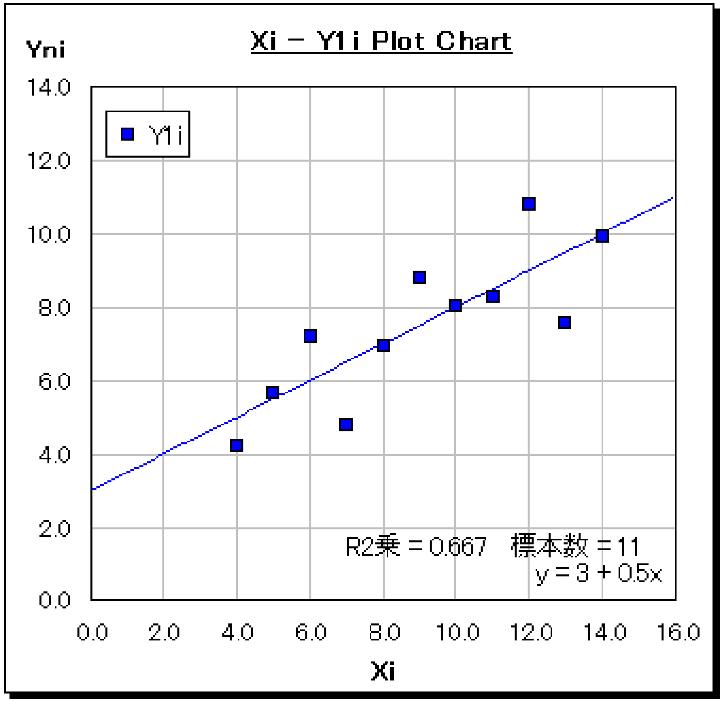
3. 応用データ解析の基礎(2): 線形回帰と検定

3-2. グラフ化(=可視化)による考察の重要性(1)

- 記述統計量(=X,Yの平均･分散等)と決定係数のみに頼ると危険、必ず**グラフ化(=可視化)**すべき

- Anscombe（‘73） Yni = 3.0 + 0.5Xi **Adj.R^2 =0.666**

i	Xi	Y1i	Y2i	Y3i
1	10.0	8.04	9.14	7.46
2	8.0	6.95	8.14	6.77
3	13.0	7.58	8.74	12.74
4	9.0	8.81	8.77	7.11
5	11.0	8.33	9.26	7.81
6	14.0	9.96	8.10	8.84
7	6.0	7.24	6.13	6.08
8	4.0	4.26	3.10	5.39
9	12.0	10.84	9.13	8.15
10	7.0	4.82	7.26	6.42
11	5.0	5.68	4.74	5.73
平 均	**9.00**	**7.50**	**7.50**	**7.50**
分 散	**3.32**	**2.03**	**2.03**	**2.03**



3. 応用データ解析の基礎(2): 線形回帰と検定

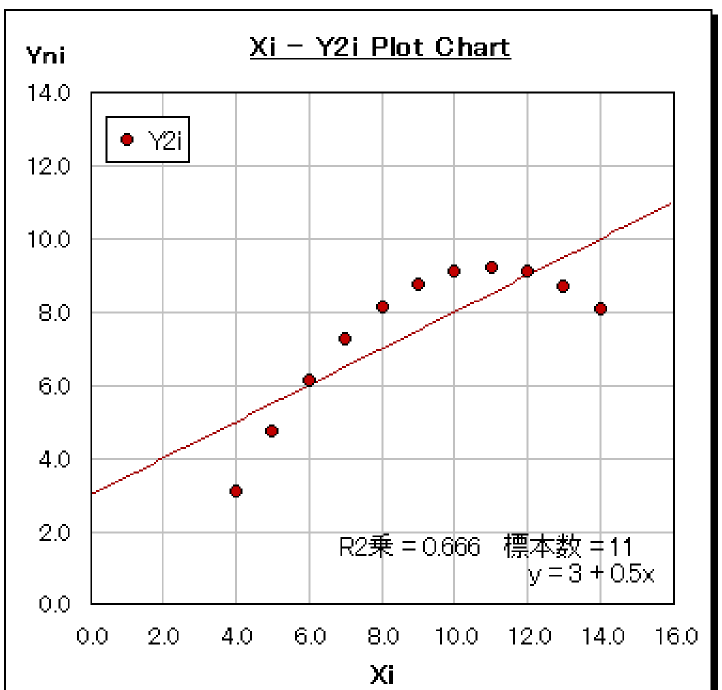
3-3. グラフ化(=可視化)による考察の重要性(2)

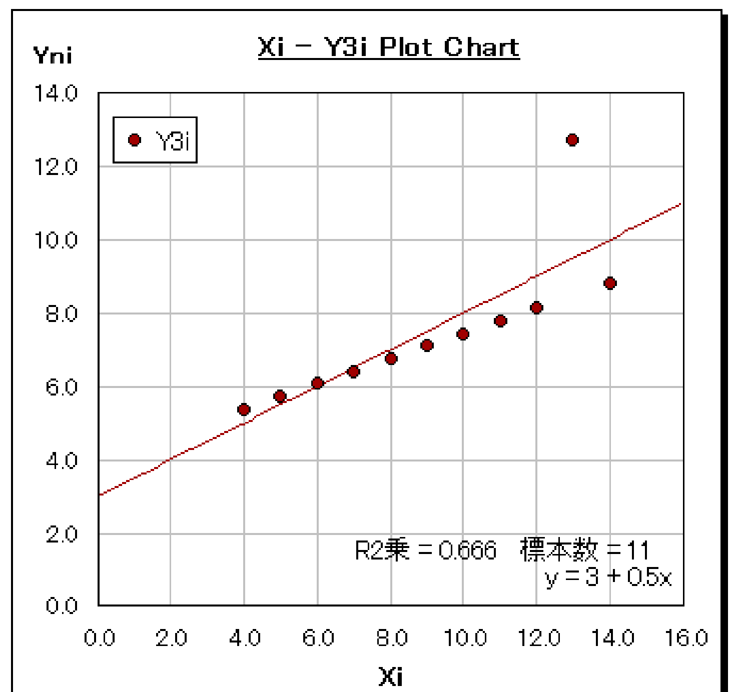
- Y2：前提 #1（線形）に問題有（要変数変換）

  $Y2i = -6.00 + 2.78\ Xi - 0.13\ Xi^2 + \varepsilon\ i$　　Adj.R^2 = 0.999

- Y3：前提 #1, #4(均一分散）に問題有（特異値）

  $Y3i = +4.01 + 0.35\ Xi + 4.24\ DM\#10 + \varepsilon\ i$　Adj.R^2 = 0.999



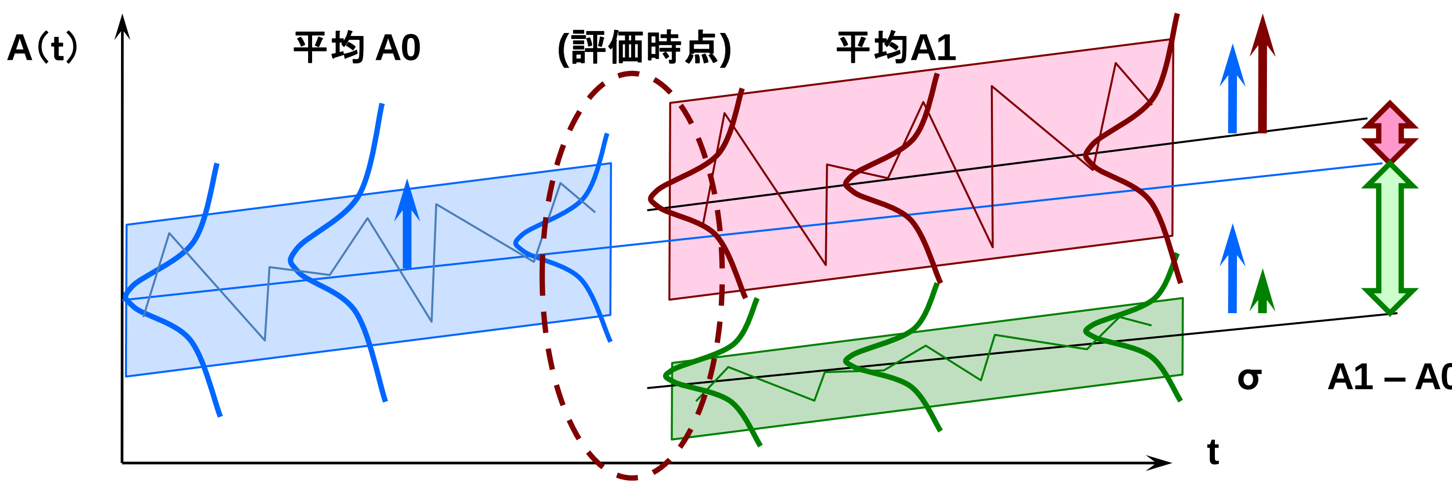


3. 応用データ解析の基礎(2): 線形回帰と検定

3-4. 統計検定の基礎(1)

- ある 2つの値の間に差があるかを判定するには条件を揃えた上で**当該試料の「ばらつき」と比べ「差」が十分大きい(= 「A1≠A0」)** かを判定する

- 仮に試料の「ばらつき(標準偏差などの指標)」と比べ「A1－A0」が小さければ差があるとは言えず



3. 応用データ解析の基礎(2): 線形回帰と検定

3-5. 統計検定の基礎(2)

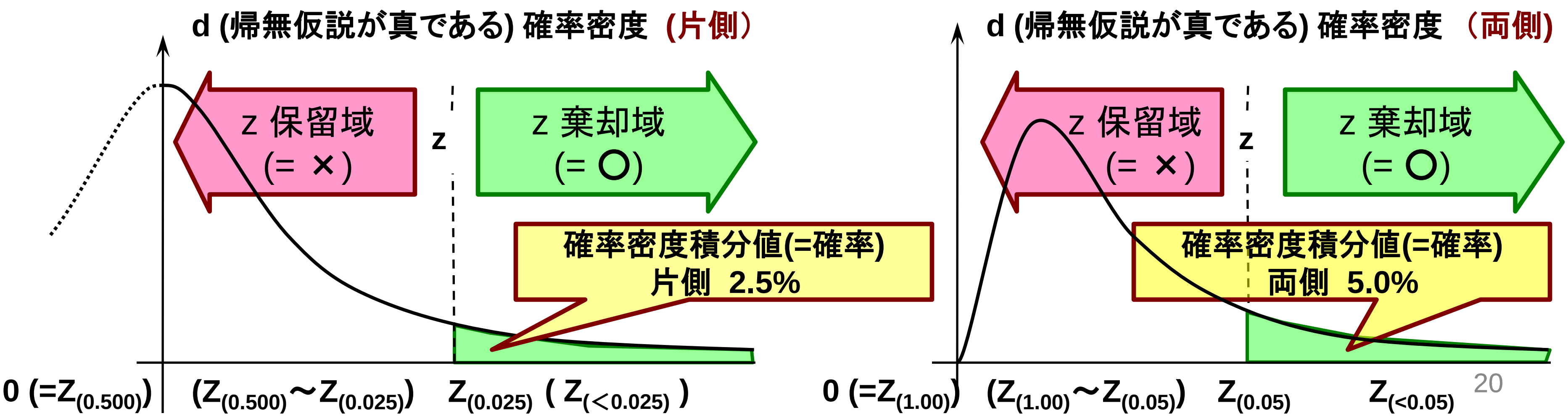
- 統計検定の多くは、検定したい内容を否定する仮説(帰無仮説: H_0)を敢えて設けた上で、当該**帰無仮説が統計的に見て「真」である確率が十分に小さいといえるか否か**を判定
  → 帰無仮説が「真」の確率が十分小
  ⇒ 内容を否定する仮説が「棄却」 ⇒ ○
- つまり 「**背理法**」
- 通常「**95%有意**」(= 確率 5%以下, “*”)が、稀に「99%有意」(同 1%以下,”**“)が用いられる

3. 応用データ解析の基礎(2): 線形回帰と検定

3−6. 統計検定の基礎(3)

– 95%・片側検定の場合、確率(= 確率密度積分値)が2.5%となる点 $Z_{(0.025)}$ に対し**帰無仮説に対応する検定統計値 Z (= 試料の「ばらつき」に対する検定対象値の比) の大小**を判定

– $Z < Z_{(0.025)}$ なら帰無仮説が「真」の確率大 ⇒ ×



3. 応用データ解析の基礎(2): 線形回帰と検定

3−7. 回帰係数の有意性の検定（⇒β ≠0? ）

– （Student） **t−検定 ; β ≠0? [重要]**

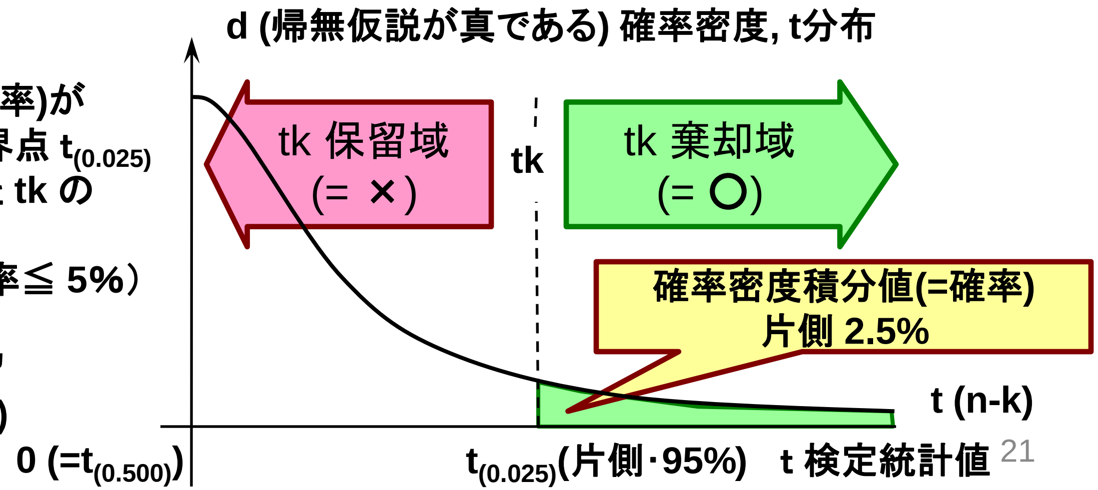
$$tk = \beta^*_k \;/\; (\sigma^{*2} \cdot (x' x)^{-1}_{kk})^{0.5} \quad (\text{t値})$$

回帰係数k　回帰係数k に対応する試料のばらつき具合

tk 〜 t(n−k) 自由度 n−k の t分布, 片側

– 結果を **p値 (tk に対応する確率)** で表すこと多し

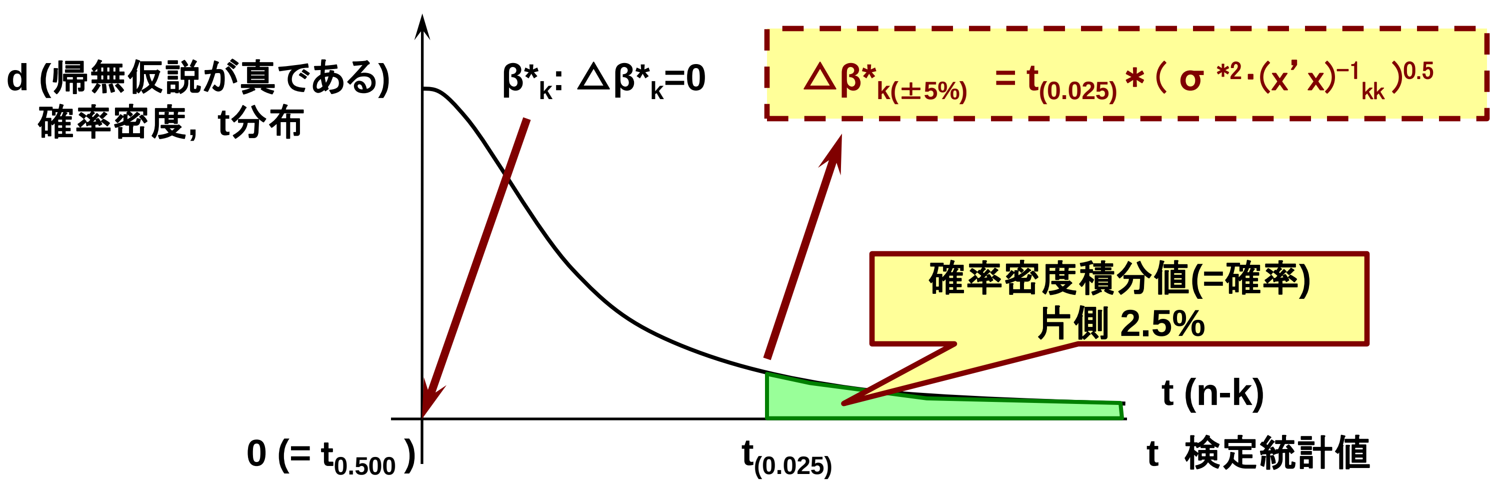
- 確率密度の総和(不定積分)は 1
- 確率密度の +∞ からの積分値(=確率)が 2.5%(95%･片側の場合)となる臨界点 $t_{(0.025)}$ に対し、仮説(帰無仮説)に対応した tk の大小を判定
- tk ≧ $t_{(0.025)}$ （= 帰無仮説「真」の確率≦ 5%） の場合帰無仮説を棄却 (= ○)
- tk ＜ $t_{(0.025)}$ の場合帰無仮説を保留 (= 帰無仮説「真」の確率＞ 5%, ×)



3. 応用データ解析の基礎(2): 線形回帰と検定

3-8. 回帰係数の信頼区間推定

- 95%水準での t検定の考え方を拡張して、逆に **回帰係数β ^{*k} が信頼できる確率95%の範囲**(= β $^*_k$ との差が 0 と言える確率が片側2.5%以上の範囲、「**信頼区間**」) を推計できる
- **$\beta^*_{k(\pm 5\%)} = \beta^*_k \pm t_{(0.025)} * (\sigma^{*2} \cdot (x'x)^{-1}{}_{kk})^{0.5}$**



3. 応用データ解析の基礎(2): 線形回帰と検定

3−9. 平均値の差の検定(⇒∀β =0の際, α 1≠α 0?)

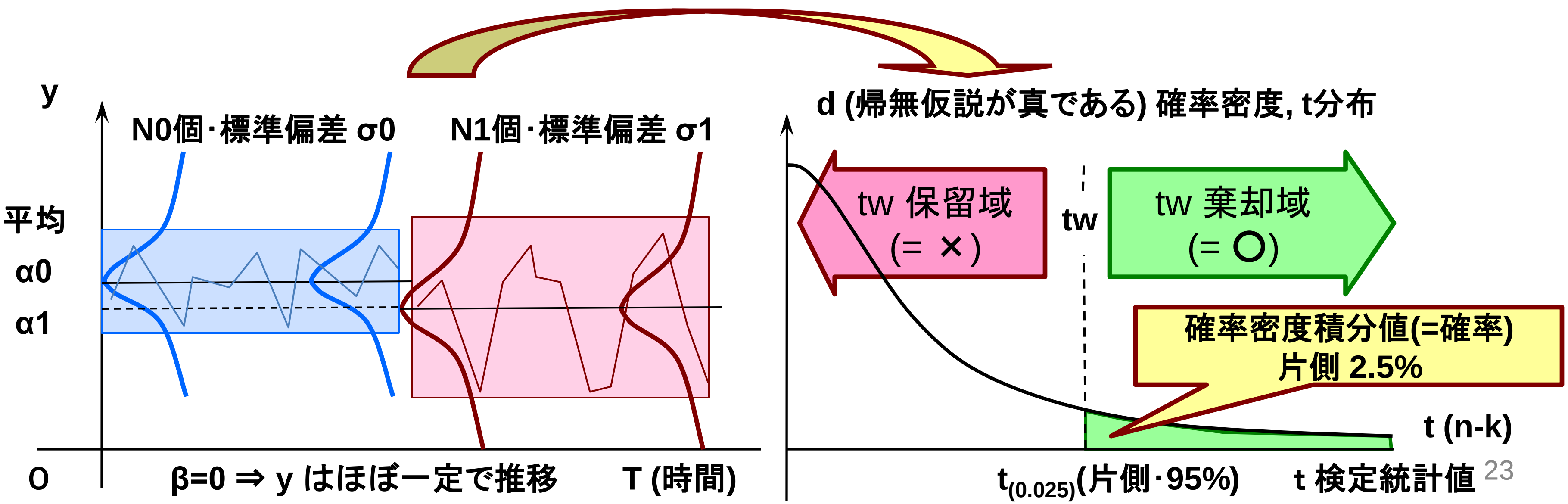
- **Welch−t検定; α 1≠α 0 ?**

$$tw = (\alpha_1 - \alpha_0) \,/\, (\sigma^{*}{}_1^{2}/N_1 + \sigma^{*}{}_0^{2}/N_0)^{0.5}$$

平均値の差 / 状態1·0 の「ばらつき」の合成値

tw ～ t(v) 自由度v の t分布, 片側

$$v = (\sigma_1/N_1+\sigma_0/N_o)^2 \,/\, (\sigma_1^{2}/(N_1^{2}\cdot(N_1-1)) + \sigma_0^{2}/(N_0^{2}\cdot(N_0-1)))^{0.5}$$



3. 応用データ解析の基礎(2): 線形回帰と検定

3-10. 平均値の差の検定の応用（簡易定常化法）

– 分析対象 y が複数の説明変数 X から影響を受けている場合でも、**$\beta_i \gg \beta_{others}$ ならば、（X_i の y への影響が他の X より卓越する場合）**
y/X_1 はほぼ一定となり、Welch t−検定が使える

$$y = \alpha + X_i * \beta_i + X_j * \beta_j + \varepsilon$$

$$y/X_i = \beta_i + \boxed{X_j/X_i * \beta_j + \alpha/X_i + \varepsilon/X_i}$$

$$\rightarrow \ll \beta_i$$

$$y/X_i = \beta_i + \varepsilon' \quad (= X_j/X_i * \beta_j + \alpha/X_i + \varepsilon/X_i)$$

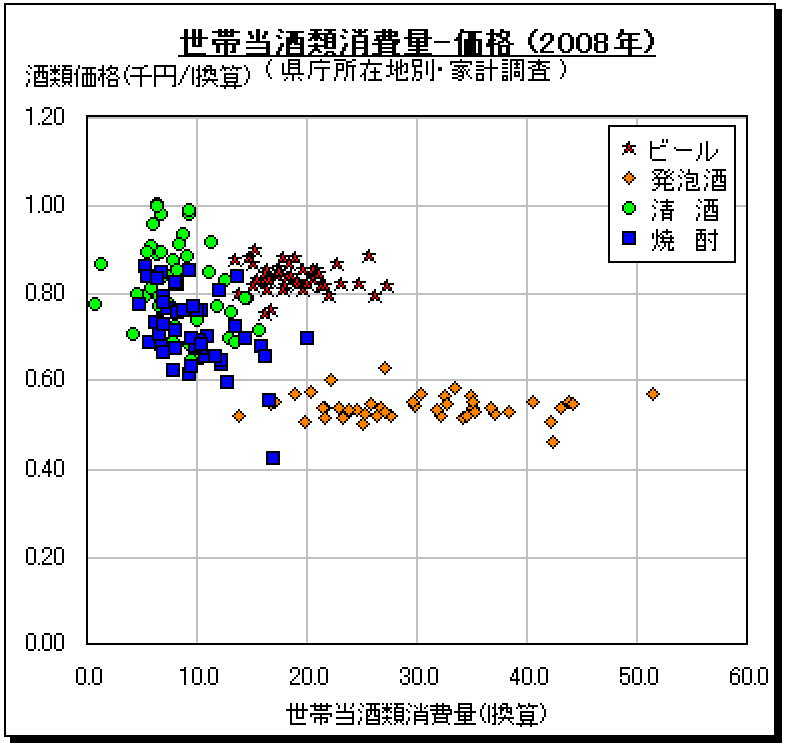
→ ほぼ一定なら Welch t−検定が適用可

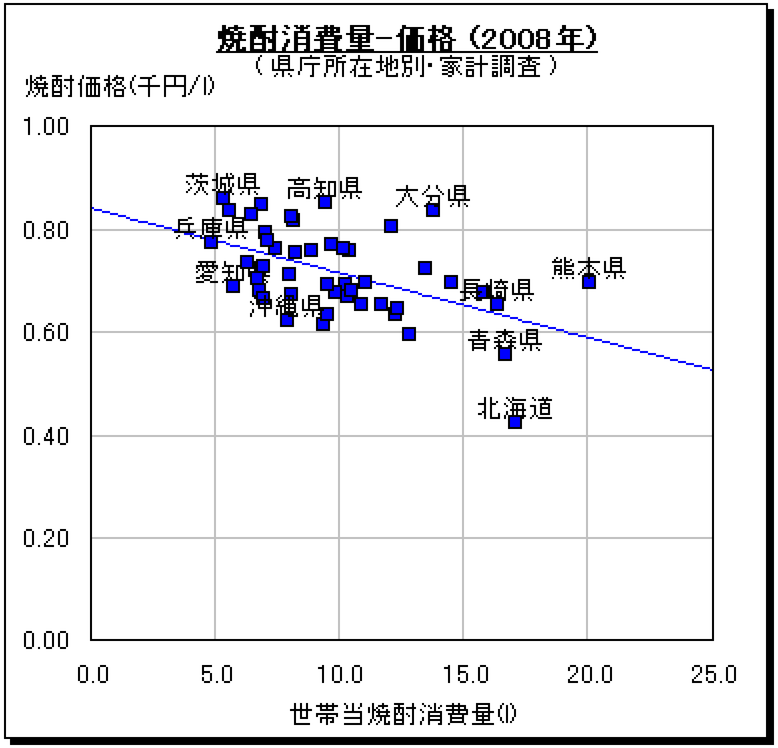
4. 応用データ解析の基礎(3): 実戦編

4-1. 回帰分析と結果の解釈(1) STATA

– 例: 酒類消費量(家計調・県庁所在地別・2008)

→ まず **P-Qグラフ(価格-数量)**を書いてみる





4. 応用データ解析の基礎(3): 実戦編

4-2. 回帰分析と結果の解釈(2) STATA

– 焼酎購入量(家計調・県庁所在地別・2008)

lsaq: 消費量(対数, l) lsap: 価格(対数, ¥/l)

lexp: 消費支出(対数) lpdp: 人口密度(対数)

lbeep,lsesp,hhpsp: ビール・清酒・発泡酒価格(対数)

↑ 適切な代替財は?

```
. reg lsaq lsap lbeep lsesp lhpsp lexp lpdp

      Source |       SS       df       MS              Number of obs =      47
-------------+------------------------------           F(  6,    40) =    7.15
       Model |  2.67358707     6  .445597846           Prob > F      =  0.0000
    Residual |  2.49448231    40  .062362058           R-squared     =  0.5173
-------------+------------------------------           Adj R-squared =  0.4449
       Total |  5.16806938    46  .112349334           Root MSE      =  .24972

------------------------------------------------------------------------------
        lsaq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lsap |  -1.427614   .3314886    -4.31   0.000    -2.097578   -.7576508
       lbeep |   3.301218   1.130336     2.92   0.006     1.016723    5.585713
       lsesp |   .3204504   .3194426     1.00   0.322    -.3251671     .966068
       lhpsp |  -.5590815   .7909589    -0.71   0.484    -2.157669    1.039506
        lexp |  -.6802628   .4823904    -1.41   0.166     -1.65521    .2946847
        lpdp |  -.1579981   .0537782    -2.94   0.005    -.2666878   -.0493084
       _cons |   2.657071   10.10848     0.26   0.794    -17.77292    23.08706
------------------------------------------------------------------------------
```

βi (係数)

t値·p値

4. 応用データ解析の基礎(3): 実戦編

4-3. 回帰分析と結果の解釈(3) STATA

– 焼酎購入量(家計調·県庁所在地別·2008)

lsaq: 消費量(対数, l) lsap: 価格(対数, ¥/l)

lexp: 消費支出(対数) lpdp: 人口密度(対数)

lbeep: ビール価格(対数)

```
. reg lsaq lsap lbeep lexp lpdp

      Source |       SS       df       MS              Number of obs =      47
-------------+------------------------------           F(  4,    42) =   10.56
       Model |  2.59129468     4  .647823671           Prob > F      =  0.0000
    Residual |   2.5767747    42  .061351779           R-squared     =  0.5014
-------------+------------------------------           Adj R-squared =  0.4539
       Total |  5.16806938    46  .112349334           Root MSE      =  .24769

------------------------------------------------------------------------------
        lsaq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lsap |  -1.452444     .32794    -4.43   0.000    -2.114253   -.7906338
       lbeep |   3.232778   1.085017     2.98   0.005     1.043126     5.42243
        lexp |  -.7018367   .4758033    -1.48   0.148    -1.662047    .2583733
        lpdp |  -.1476392   .0525584    -2.81   0.008    -.2537063    -.041572
       _cons |   .0037443   9.692547     0.00   1.000    -19.55661     19.5641
------------------------------------------------------------------------------
```

二乗和· k, n-k ·平均二乗和

推計式説明分·残差分

F検定結果

R^2· Adj.R^2

残差平方和

t値·p値

βi (係数)

$\sqrt{\sigma^2(xx)^{-1}}$(標準誤差)

95%信頼区間上限·下限

4. 応用データ解析の基礎(3): 実戦編

4-4. 回帰分析と結果の解釈(4) STATA

– 焼酎購入量(家計調・県庁所在地別・2008)

理論と整合するか？(1)

$e_{qx,px} + e_{qx,py} + e_{qx,I} = 0$（需要関数の同次性条件）

Min(−2.11+1.04−1.66) Max(−0.79+5.42+0.26)

= −2.73 ～+4.89

```
. reg lsaq lsap lbeep lexp lpdp

      Source |       SS       df       MS              Number of obs =      47
-------------+------------------------------           F(  4,    42) =   10.56
       Model |  2.59129468     4  .647823671           Prob > F      =  0.0000
    Residual |   2.5767747    42  .061351779           R-squared     =  0.5014
-------------+------------------------------           Adj R-squared =  0.4539
       Total |  5.16806938    46  .112349334           Root MSE      =  .24769

------------------------------------------------------------------------------
        lsaq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lsap |  -1.452444     .32794    -4.43   0.000    -2.114253   -.7906338
       lbeep |   3.232778   1.085017     2.98   0.005     1.043126     5.42243
        lexp |  -.7018367   .4758033    -1.48   0.148    -1.662047    .2583733
        lpdp |  -.1476392   .0525584    -2.81   0.008    -.2537063    -.041572
       _cons |   .0037443   9.692547     0.00   1.000    -19.55661     19.5641
------------------------------------------------------------------------------
```

βi (係数)

t値・p値

95%信頼区間上限・下限

4. 応用データ解析の基礎(3): 実戦編

4-5. 回帰分析と結果の解釈(5) STATA

– 焼酎購入量(家計調・県庁所在地別・2008)

理論と整合するか？(2) 人口密度を外すと・・・

$e_{qx,px} + e_{qx,py} + e_{qx,I} = 0$ (需要関数の同次性条件)

Min(−2.42+1.05−1.85) Max(−1.07+5.75+0.21)

= −3.22 ～ +4.89

```
. reg lsaq lsap lbeep lexp

      Source |       SS       df       MS              Number of obs =      47
-------------+------------------------------           F(  3,    43) =    9.87
       Model |  2.10718286     3  .702394287           Prob > F      =  0.0000
    Residual |  3.06088652    43  .071183408           R-squared     =  0.4077
-------------+------------------------------           Adj R-squared =  0.3664
       Total |  5.16806938    46  .112349334           Root MSE      =   .2668

------------------------------------------------------------------------------
        lsaq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lsap |  -1.741788   .3353633    -5.19   0.000    -2.418113   -1.065464
       lbeep |   3.405312    1.16685     2.92   0.006     1.052134     5.75849
        lexp |  -.8154809   .5106551    -1.60   0.118    -1.845315    .2143532
       _cons |  -.4769214   10.43869    -0.05   0.964    -21.52855     20.5747
------------------------------------------------------------------------------
```

βi (係数)

t値・p値

95%信頼区間上限・下限

4. 応用データ解析の基礎(3): 実戦編

4-6. 回帰分析と結果の解釈(6) STATA

– **不均一分散**最小二乗法 (Heterosked. robust)

→ 回帰係数βｉは同じ、**標準誤差が異なる**

(←分散均一性検定が棄却：清酒の例)

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lsesq

         chi2(1)      =    17.26
         Prob > chi2  =   0.0000

. regress lsesq lsesp lexp lpdp, robust

Linear regression                                      Number of obs =     47
                                                       F(  3,    43) =   4.26
                                                       Prob > F      = 0.0101
                                                       R-squared     = 0.1756
                                                       Root MSE      = .49607
```

$\sqrt{(x'x)^{-1}x'\Omega x(x'x)^{-1}}$

lsesq	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lsesp	-.7272809	.4543773	-1.60	0.117	-1.64362	.1890582
lexp	2.65602	1.511482	1.76	0.086	-.3921736	5.704214
lpdp	-.0517519	.0654726	-0.79	0.434	-.1837901	.0802862
_cons	-30.95186	23.00742	-1.35	0.186	-77.35076	15.44703