

統計学I

早稲田大学政治経済学術院

西郷 浩

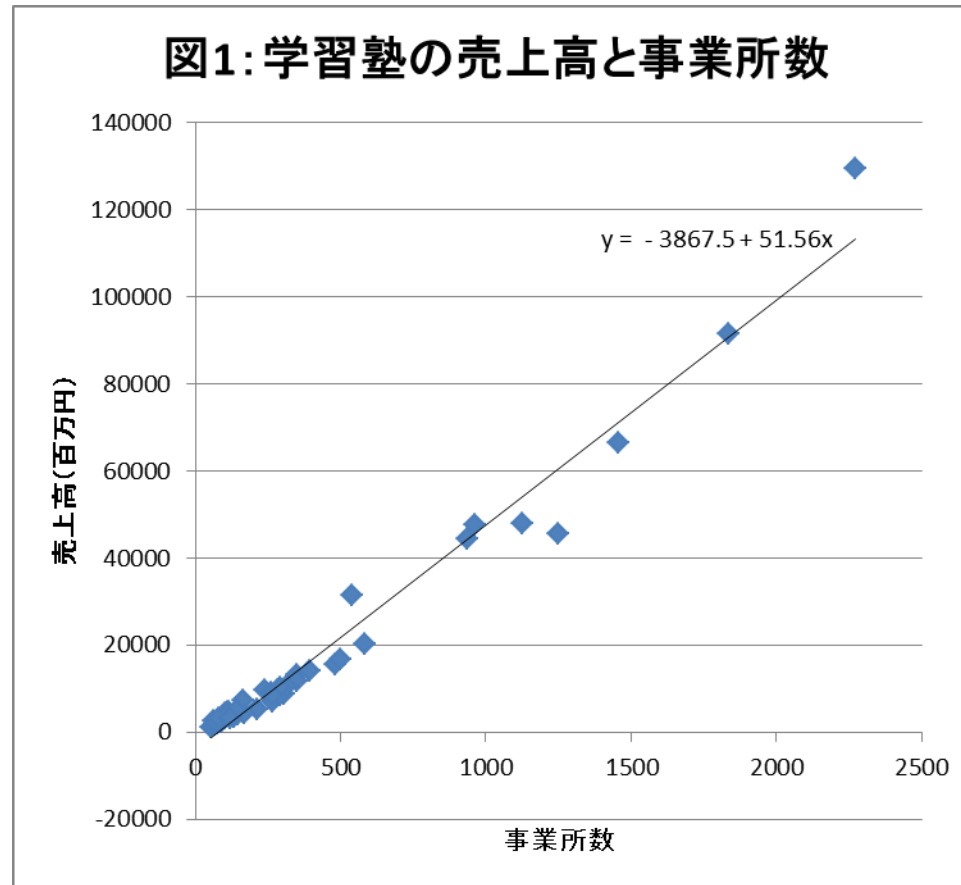
本日の目標

- 回帰分析の発展
 - 変数変換の利用
 - 対数変換
 - 複数(2つ)の説明変数
 - 平面の当てはめ
- PC実習

曲線的な関係(1)

- 都道府県別学習塾数と売上高
 - $y =$ (都道府県別学習塾売上高 100万円)
(平成28年)
 - $x =$ (都道府県別学習塾事業所数)
(平成28年)
 - 秋田県・鳥取県の売上高: x と書いてある。
 - 県内に事業所が少ない。→ 秘匿処置
 - これらの県のおおよその売上高を推定する。

曲線的な関係(2)



資料: 総務省・経済産業省「平成28年 経済センサス-活動調査」

曲線的な関係(3)

- 散布図からの所見
 - 正の相関がある。
 - x が小さいところに観察点が集中している。
 - x が大きくなるにつれて y 軸方向のばらつきも拡大する傾向がある。
 - 最小二乗法による回帰式
 - $y = -3867.5 + 51.6 x, R^2 = 0.97$
 - 秋田県 $x = 52 \rightarrow \hat{y} = -1186$ (負の売上高?)
 - 鳥取県 $x = 81 \rightarrow \hat{y} = 309$ (x が同じ宮崎県の3149より小さい)

曲線的な関係(4)

- 直線を当てはめるに無理がある。
 - 曲線的な傾向がある。
 - そのまま最小二乗法を適用するのは危険である。

変数変換の利用(1)

- もともとの関係が、直線ではない(曲線である)可能性がある。
 - 直線による近似に無理がある。
- 曲線的な関係をどのようにあつかうか。
 - 変数変換によって直線化する。
 - 常用対数変換: $y' = \log_{10} y$
 - 逆数変換: $y' = 1/y$
 - ベキ乗変換: $y' = y^p$

変数変換の利用(2)

- なかでも、対数変換がよく使われる。
 - 理由: 解釈がしやすい。
 - 説明:
 - x が1%増加すると、 y が $b\%$ 変化する。
 - 乗法モデル $y = a x^b$
 - » b は弾力性とよばれる。
 - 通常のモデルは加法のモデルと呼ばれる。
 $y = a + bx$ (x が1単位増加すると y が b 単位変化)

変数変換の利用(3)

- 対数の性質をもちいると、

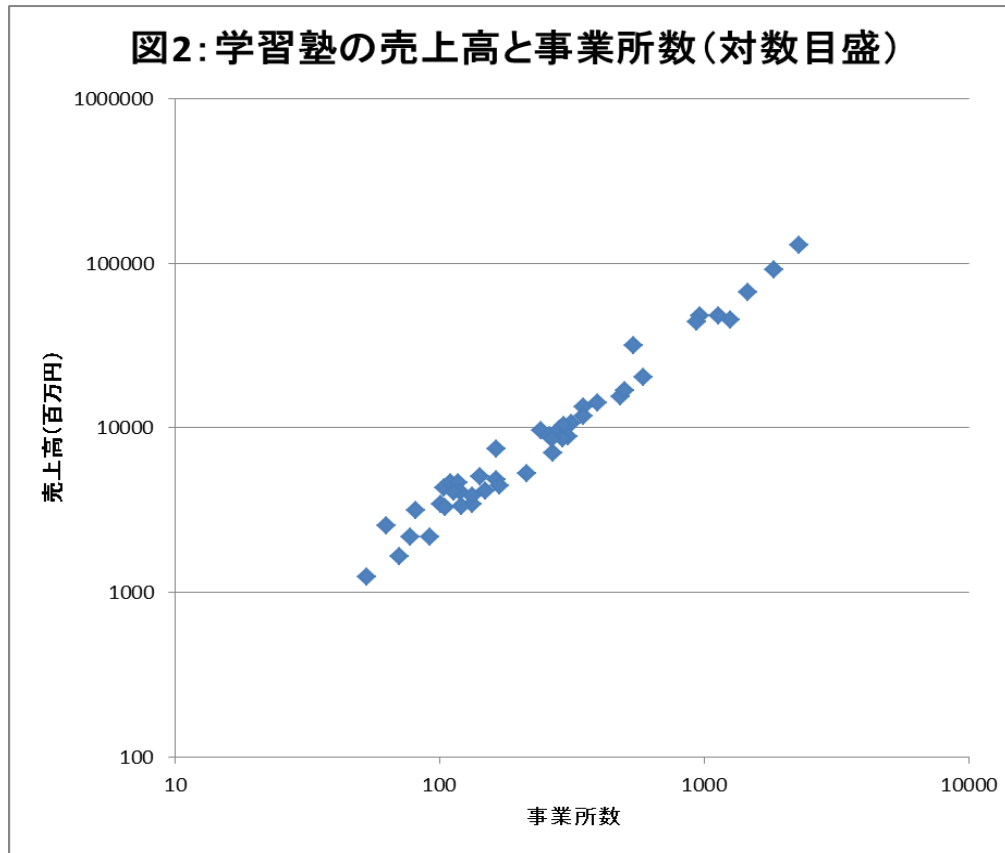
$$y = ax^b$$

$$\Leftrightarrow \log_{10} y = \log_{10} a + b \log_{10} x$$

$$y' = \log_{10} y, x' = \log_{10} x, a' = \log_{10} a \text{ とすれば、} y' = a' + bx'$$

»「対数変換してほぼ直線関係で近似できれば、変化率の間の関係が安定的である」ことを意味する。

変数変換の利用(4)



資料: 総務省・経済産業省「平成28年 経済センサス-活動調査」

変数変換の利用(5)

- 対数変換したデータに最小二乗法を適用
 - 推定結果
 - $\log_{10} \hat{y} = 1.21 + 1.14 \log_{10} x$ $R^2 = 0.97$
 - 推定結果を以下のように書き換える。
 - $\hat{y} = 10^{1.21} x^{1.14}$
 - 秋田県と鳥取県の売上高の推定値
 - 秋田県 $\hat{y} = 10^{1.21} \times 52^{1.14} \approx 1449$
 - 鳥取県 $\hat{y} = 10^{1.21} \times 81^{1.14} \approx 2402$

変数変換の利用(6)

- 弾力性 $=1.14 > 1$
 - y の変化率(増加率) $> x$ の変化率(増加率)
→ 散布図が尻上がり形状
 - 「 y が x に対して弾力的である」という。
- 弾力性 $=(y \text{ の変化率})/(x \text{ の変化率})$
 - 弾力性 > 1 : 弾力的(尻上がり)
 - 弾力性 $= 1$: 比例関係
 - $0 < \text{弾力性} < 1$: 非弾力的(頭打ち)

変数変換の利用(7)

– 加法モデルと乗法モデル

- どちらを用いるかは経験的に(実際に当てはめて)判断する場合が多い。

• 対数以外の変数変換も利用される。

- 対数変換は係数の解釈(弾力性と解される)がしやすいので多用される。

2つ説明変数 (1)

- 都道府県別の住宅地の価格
 - 決定要因
 - 宅地を求める人の数 → 人口密度
 - 所得水準 → 一人当たり県民所得
 - 両者は異なる側面を捉えている。
 - 2つの変数が異なる影響をもつので、同時に説明要因に取り入れたい。

2つ説明変数 (2)

図3: 住宅地平均価格と人口密度

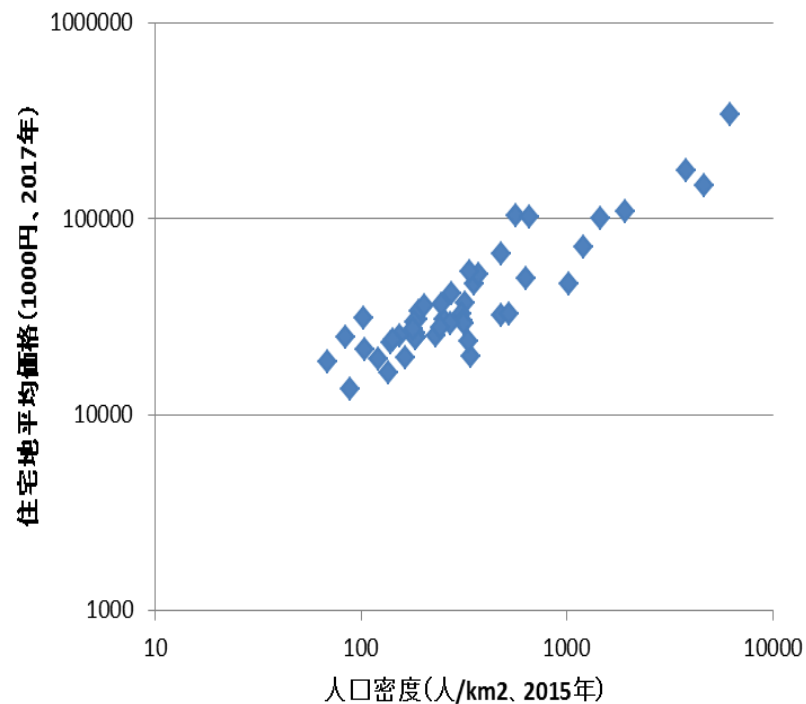
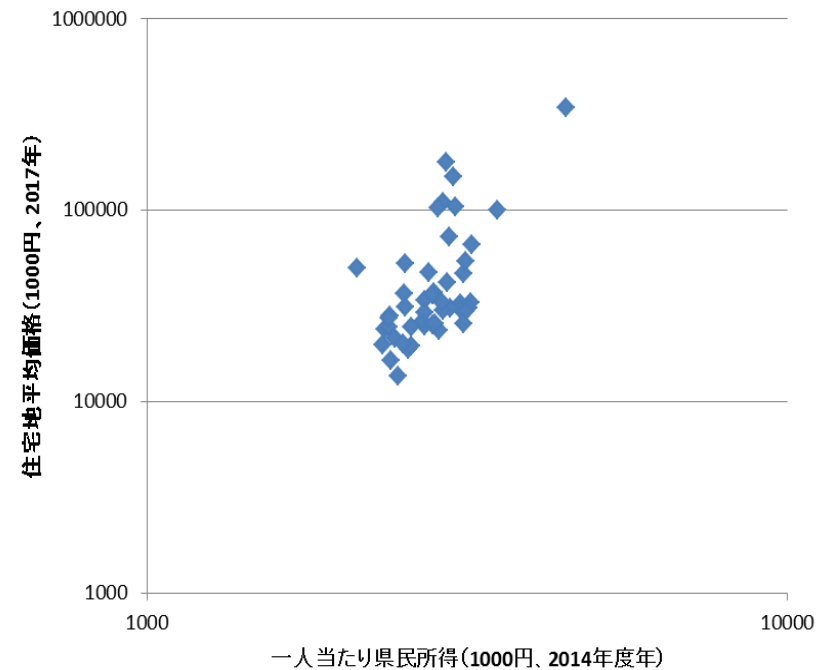


図4: 住宅地平均価格と一人当たり県民所得



資料: 総務省(2018)『第68回日本統計年鑑』表20-11、表2-3、表3-10

2つ説明変数 (3)

- 2つの説明要因をもつ回帰式

$$-\log y_i = a + b \log x_i + c \log z_i$$

- y : 住宅地平均価格
(2017年7月1日、円/m²)
- x : 人口密度
(2015年10月1日、人/km²)
- z : 一人当たり県民所得
(2014年度、1000円/人)

—どのように回帰係数 a, b, c を求めるか。

最小二乗法(1)

- 回帰係数 a, b, c をどう決めるか？
⇔ 回帰平面の位置をどう決めるか？
 - 当てはまりがもっともよくなるように。
⇔
説明変数で説明できない部分(残差)が全体としてもっとも小さくなるように。
⇔
最小二乗法の考え方が使える。

最小二乗法(2)

最小二乗法

以下の説明では、 $\log y_i$ などをあらためて y_i などと表記している。

$$\min \sum_{i=1}^N d_i^2 \Leftrightarrow \min \sum_{i=1}^N (y_i - a - b x_i - c z_i)^2$$

この最小化問題の解 $\Leftrightarrow$ 下の正規方程式の解

$$\begin{cases} N a + \left(\sum_i x_i\right) b + \left(\sum_i z_i\right) c = \left(\sum_i y_i\right) \\ \left(\sum_i x_i\right) a + \left(\sum_i x_i^2\right) b + \left(\sum_i x_i z_i\right) c = \left(\sum_i x_i y_i\right) \\ \left(\sum_i z_i\right) a + \left(\sum_i z_i x_i\right) b + \left(\sum_i z_i^2\right) c = \left(\sum_i z_i y_i\right) \end{cases}$$

最小二乗法(3)

– 平方和の分解も成り立つ。

$$\sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^N d_i^2$$

$SS_T \qquad \qquad \qquad SS_R \qquad \qquad \qquad SS_E$

$$\text{ただし、} \hat{y}_i = a + bx_i + cz_i$$

– したがって、 $R^2 = SS_R / SS_T$ も計算でき、意味も以前と同じである。

回帰平面の推定(1)

- 平面の当てはめの結果

- $\log \hat{y}_i = 0.22 + 0.53 \log x_i + 0.88 \log z_i \quad R^2 = 0.84$

- 係数の符号は常識に合う結果である。

- x (人口密度)の係数 > 0

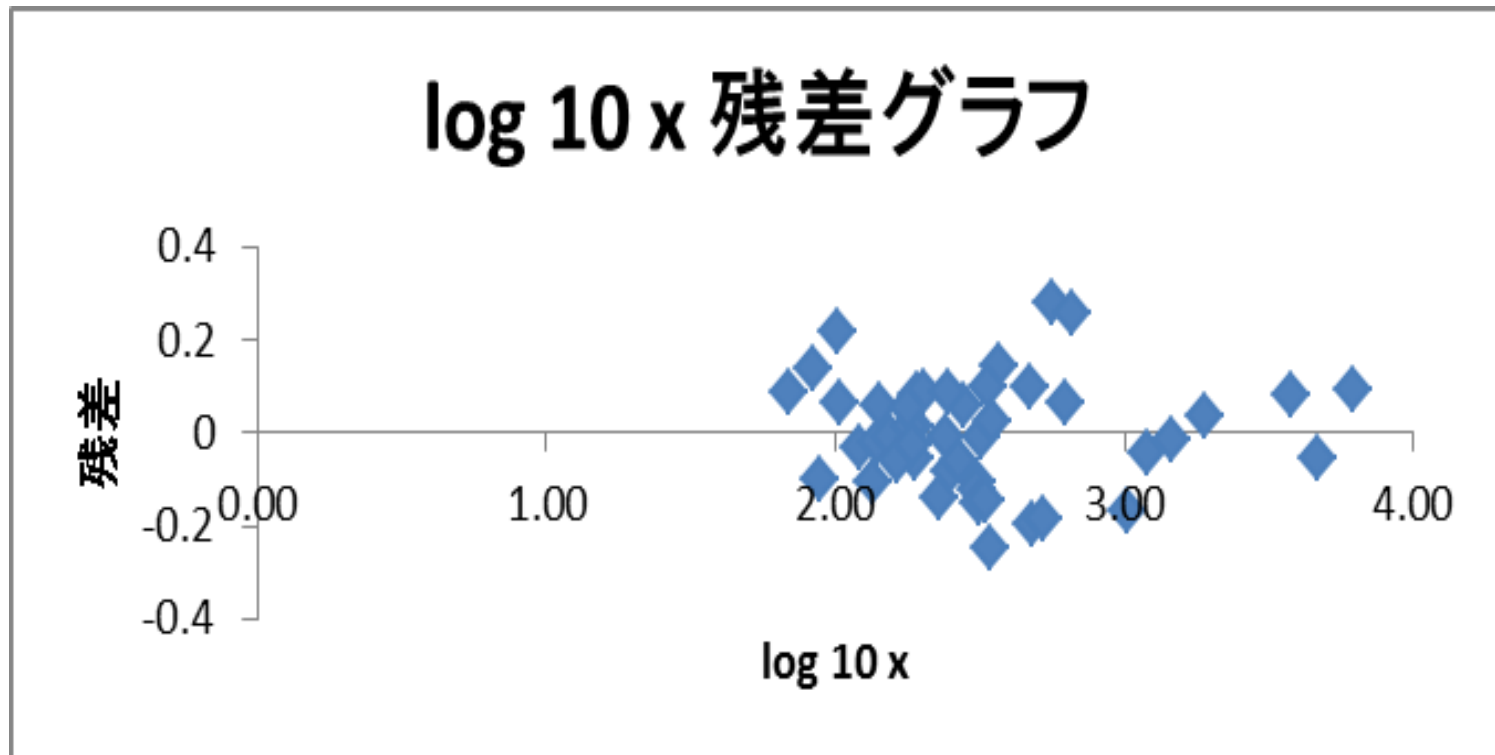
- 人口密度が高い $\Rightarrow$ 住宅地価格は高い。

- z (一人当たり県民所得)の係数 > 0

- 所得水準が高い $\Rightarrow$ 住宅地価格は高い。

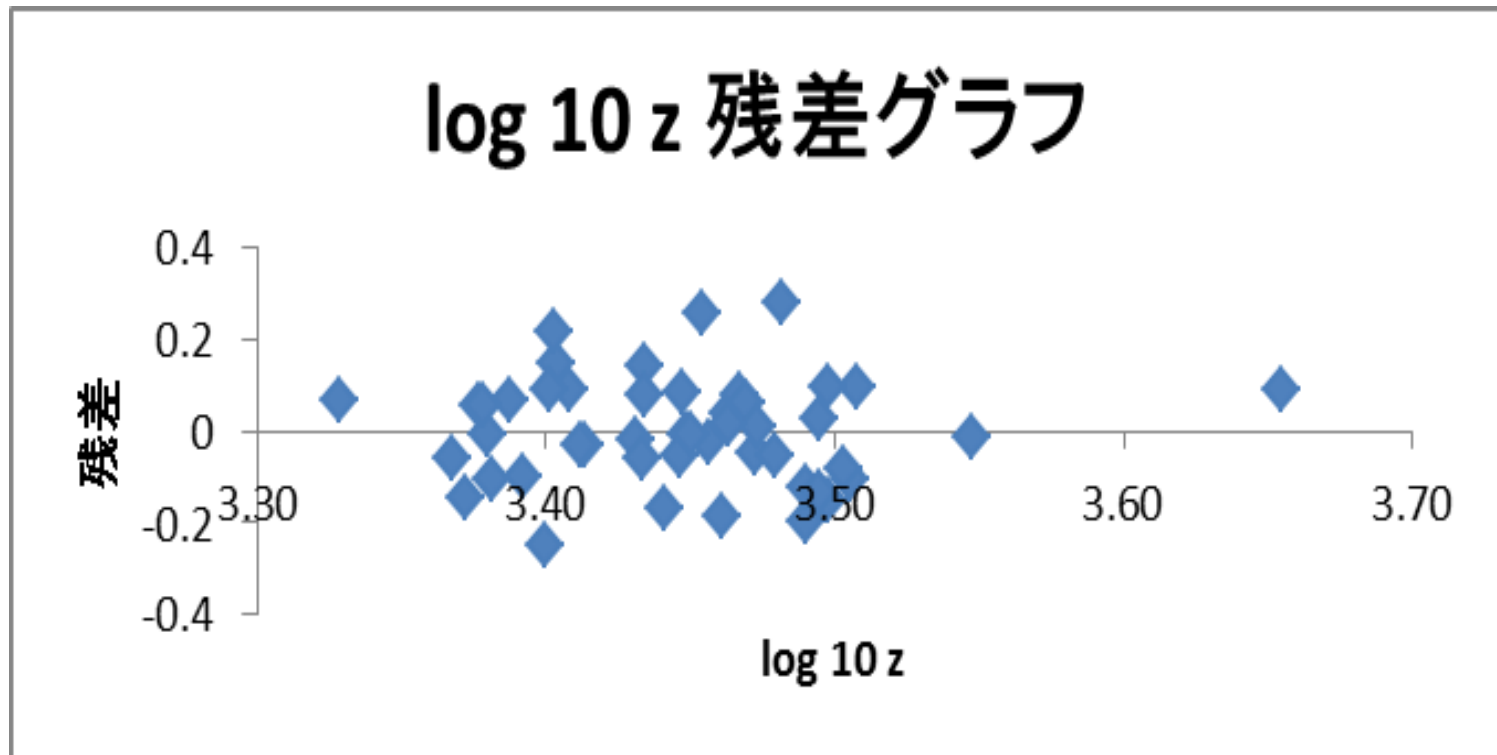
回帰平面の推定(2)

図5: 人口密度(対数変換)の残差プロット

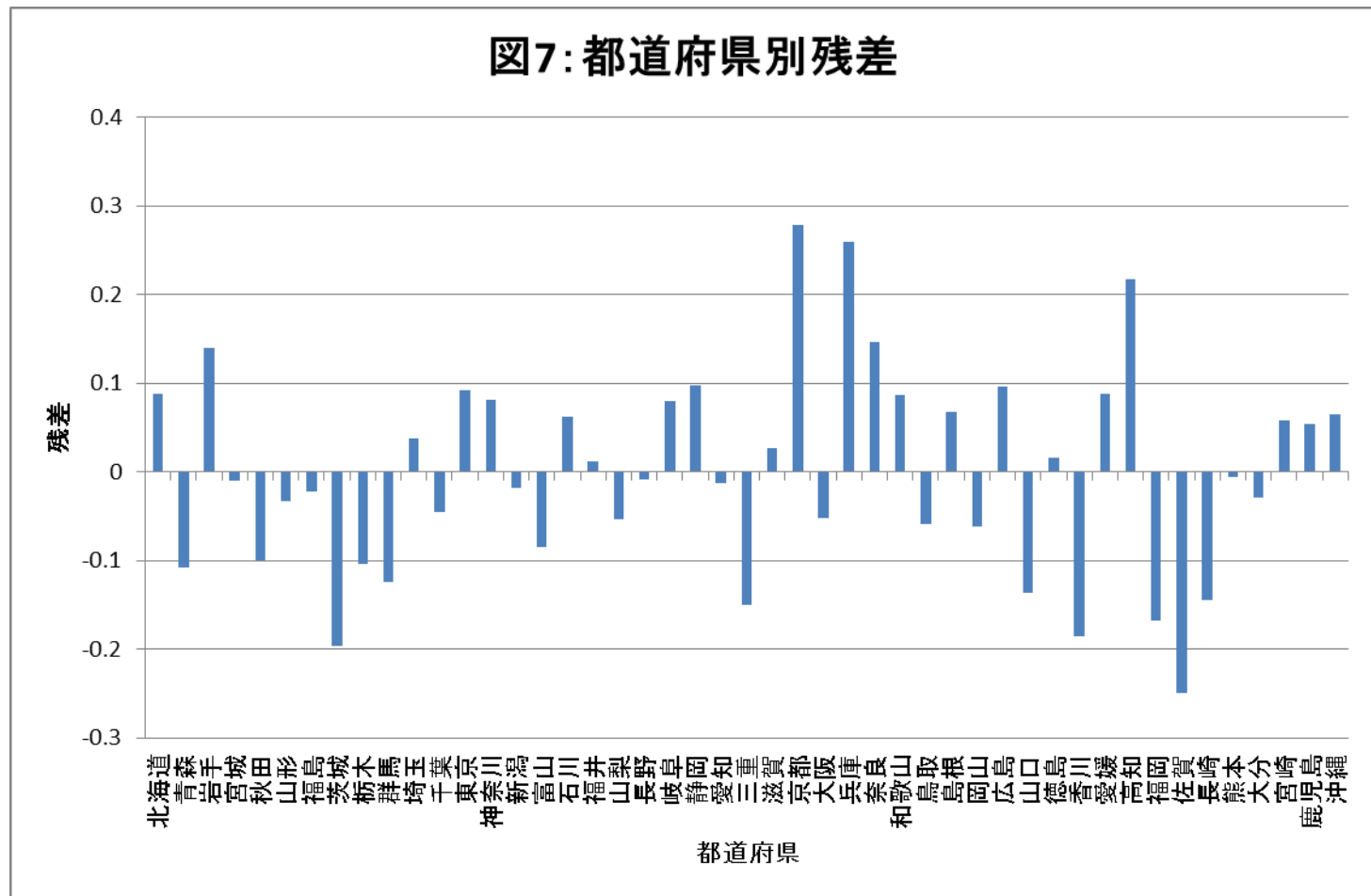


回帰平面の推定(3)

図6: 一人当たり県民所得(対数変換)の残差プロット



回帰平面の推定(4)



PC実習

- 変数変換
 - 散布図における対数変換
 - 変数変換を利用した回帰分析
- 分析ツールを利用した回帰式の推定
 - 散布図の描画(省略)
 - 回帰係数の推定
 - 残差プロットの描画(省略)