

統計学入門 第8回

早稲田大学政治経済学部
西郷 浩

1

本日の目標

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

2

2つ説明変数 (1)

- 2つの説明要因
 - $y_i = a + b x_i + c z_i$
 - y : 住宅地平均価格
(H22年7月1日、100円/m²)
 - x : 人口密度
(H22年10月1日、人/km²)
 - z : 1人当たり県民所得
(H20年度、1000円/人)

3

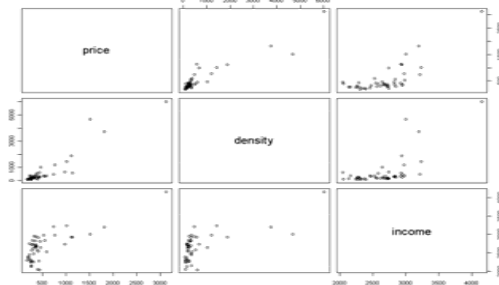
2つ説明変数 (2)

- 説明変数を2つ設ける理由
 - 住宅地価格への異なる影響
 - 混雑度: 人口密度
 - 所得水準: 1人当たり県民所得
 - 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る。

4

2つ説明変数 (3)

図1: 散布図行列 (住宅地平均価格、人口密度、1人当たり県民所得)



資料: 総務省統計研修所編『第61回日本統計年鑑 平成23年』

5

最小二乗法(1)

- 回帰係数 a, b, c をどう決めるか?
 $\Leftrightarrow$ 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。
 $\Leftrightarrow$
 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。
 $\Leftrightarrow$
 最小二乗法の考え方が使える。

6

最小二乗法(2)

最小二乗法

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

この最小化問題の解 $\Leftrightarrow$ 下の正規方程式の解

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

7

最小二乗法(3)

- 平方和の分解も成り立つ。

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

$$SS_0 \qquad SS_R \qquad SS_E$$

ただし、 $\hat{y}_i = a + b x_i + c z_i$
- したがって、 $R^2 = SS_R / SS_0$ も計算でき、意味も以前と同じである。

8

回帰平面の推定(1)

■ 平面の当てはめの結果

$$\hat{y} = -561 + 0.34x + 0.33z \quad R^2 = 0.90$$

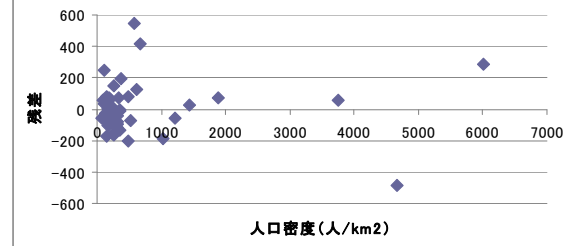
■ 係数の符号は常識に合う結果である。

- X (人口密度)の係数 > 0
 - 人口密度が高い⇒住宅地価格は高い。
- Z (1人当たり県民所得)の係数 > 0
 - 所得が多い⇒住宅地価格は高い。

9

回帰平面の推定(2)

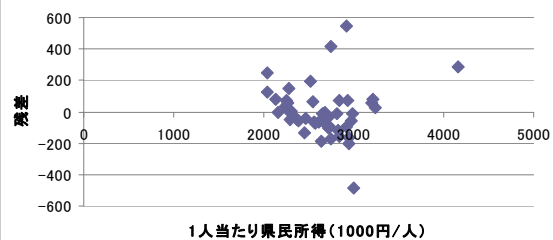
図2: 残差プロット(人口密度)



10

回帰平面の推定(3)

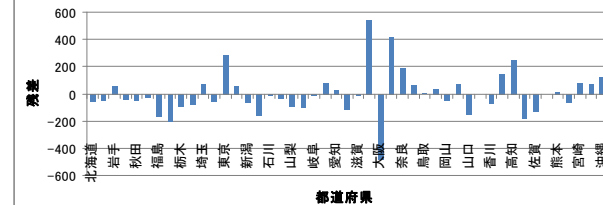
図3: 残差プロット(所得)



11

回帰平面の推定(4)

図4: 残差プロット(都道府県)



12

回帰平面の推定(5)

- 残差プロット
 - 大阪・京都・兵庫の残差が大きい。
 - 何が原因か。
- 散布図を良く見ると...
 - 曲線的な関係が見て取れる。

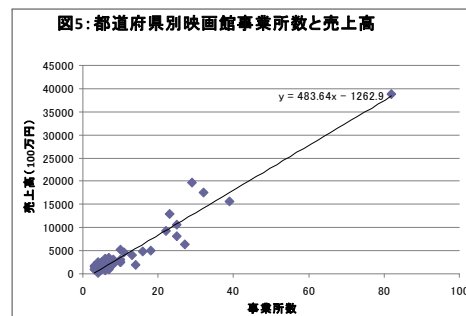
13

曲線的な関係(1)

- 都道府県別映画館数と売上高
 - x = (都道府県別映画館事業所数) (平成22年)
 - y = (都道府県別映画館売上高 100万円) (平成22年)
 - 島根県・徳島県の売上高: x と書いてある。
 - 県内に事業所が2つしかない。→ 秘匿処置
 - これらの県のおおよその売上高を推定する。

14

曲線的な関係(2)



15

曲線的な関係(3)

- 散布図からの所見
 - 正の相関関係がある。
 - 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
 - $y = -1263 + 484x, R^2 = 0.93$
 - $x = 2 \rightarrow \hat{y} = -296$ (負の売上高?)

16

曲線的な関係(4)

- 直線を当てはめるには無理がある。
 - 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。

17

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: $y' = y^p$

18

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。
 - $y = ax^b$ (乗法モデル)
 - b は弾力性とよばれる。
 - 注: 通常のモデル($y = a + bx$, x が1単位増加すると y が b 単位変化する)を加法モデルとよぶことがある。

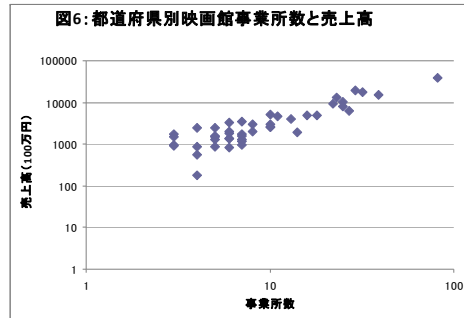
19

変数変換の利用(3)

- 対数の性質をもちいると、
 - $y = ax^b$
 - $\Leftrightarrow \log_{10} y = \log_{10} a + b \log_{10} x$
 - $y' = \log_{10} y$, $x' = \log_{10} x$, $a' = \log_{10} a$ とすれば、 $y' = a' + bx'$
 - 要は、「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」ということ。

20

変数変換の利用(4)



資料: 経済産業省「平成22年 特定サービス産業実態調査」

21

変数変換の利用(5)

$$\log_{10} y = 2.2 + 1.2 \log_{10} x$$

$$R^2 = 0.77$$

すなわち、

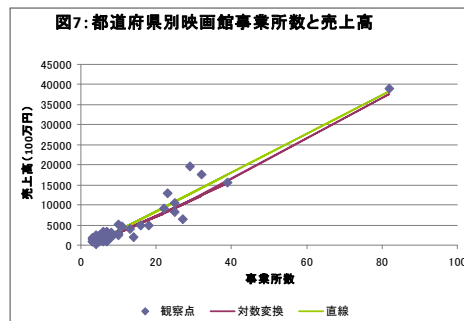
$$y = 10^{2.2} x^{1.2}$$

$x = 2$ のときの予測

$$y = 10^{2.2} \times 2^{1.2} = 443$$

22

変数変換の利用(6)



23

変数変換の利用(7)

- 弾力性 = 1.2 > 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 の変化率) / (x 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)

24



変数変換の利用(8)

- 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)がしやすいので多用される。

25



PC実習

- 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)
- 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

26