# 統計学01

早稲田大学政治経済学部
西郷 浩
第20回

# 本日の目的

- 重回帰分析
  - 重回帰式
  - 最小二乗法
  - 回帰係数に関する推定と検定

# 重回帰モデル

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

A matrix notation

$$y = X\beta + \varepsilon,$$

where

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{21} & \cdots & X_{p1} \\ 1 & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{2n} & \cdots & X_{pn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

# 最小二乘法

$$\min \sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \Leftrightarrow \min \sum_{i=1}^{n}\left(Y_i-b_1-b_2X_{i2}-\cdots-b_pX_{ip}\right)^2$$

正規方程式

$$\begin{cases} nb_1+\left(\sum X_{i2}\right)b_2+\left(\sum X_{i3}\right)b_3+\cdots+\left(\sum X_{ip}\right)b_p=\left(\sum Y_i\right) \\ \left(\sum X_{i2}\right)b_1+\left(\sum X_{i2}^2\right)b_2+\left(\sum X_{i2}X_{i3}\right)b_3+\cdots+\left(\sum X_{i2}X_{ip}\right)b_p=\left(\sum X_{i2}Y_i\right) \\ \cdots \\ \left(\sum X_{ip}\right)b_1+\left(\sum X_{i2}X_{ip}\right)b_2+\left(\sum X_{i3}X_{ip}\right)b_3+\cdots+\left(\sum X_{ip}^2\right)b_p=\left(\sum X_{ip}Y_i\right) \end{cases}$$

A matrix notation

$$X'Xb=X'y \Rightarrow b=\left(X'X\right)^{-1}X'y \quad \left(\text{denoted by } \hat{\beta}\right)$$

# 最小二乗推定量の性質

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$= (X'X)^{-1}X'(X\beta + \varepsilon)$$
$$= \beta + (X'X)^{-1}X'\varepsilon$$
$$E(\hat{\beta}) = \beta, V(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

$\hat{\beta}$が正規分布にしたがう。

とくに、

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$$

where $c_{jj} = (X'X)^{-1}_{jj}$

# 誤差の分散の推定

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_p X_{pi}$$

$$\text{Note: } \varepsilon_i = Y_i - \beta_1 - \beta_2 X_{2i} - \cdots - \beta_p X_{pi}$$

$$s^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n - p}$$

$$\text{Note: } E\left(s^2\right) = \sigma^2$$

# 回帰係数に関する推定と検定

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 c_{jj}}} \sim t(n-p)$$

区間推定（信頼係数0.95）

$$\left[\hat{\beta}_j - t_{0.025}(n-p)\sqrt{s^2 c_{jj}},\ \hat{\beta}_j + t_{0.025}(n-p)\sqrt{s^2 c_{jj}}\right]$$

検定（有意水準0.05）

$$\begin{cases} H_0 : \beta_j = a \\ H_1 : \beta_j \neq a \end{cases}$$

$$t = \frac{\beta_j - a}{\sqrt{s^2 c_{jj}}}$$

Reject $H_0$ if $t_{obs} \in W = \{t : t < -t_{0.025}(n-p) \text{ or } t > t_{0.025}(n-p)\}$

# 重回帰分析の実際(その1)

- データ
  - 大気中のオゾンの量とその影響要因
    - Ozone
    - Solar Radiation
    - Wind
    - Temperature

    (統計ソフトウェア R に内蔵されているデータセットのひとつ)

# 重回帰分析の実際(その1)

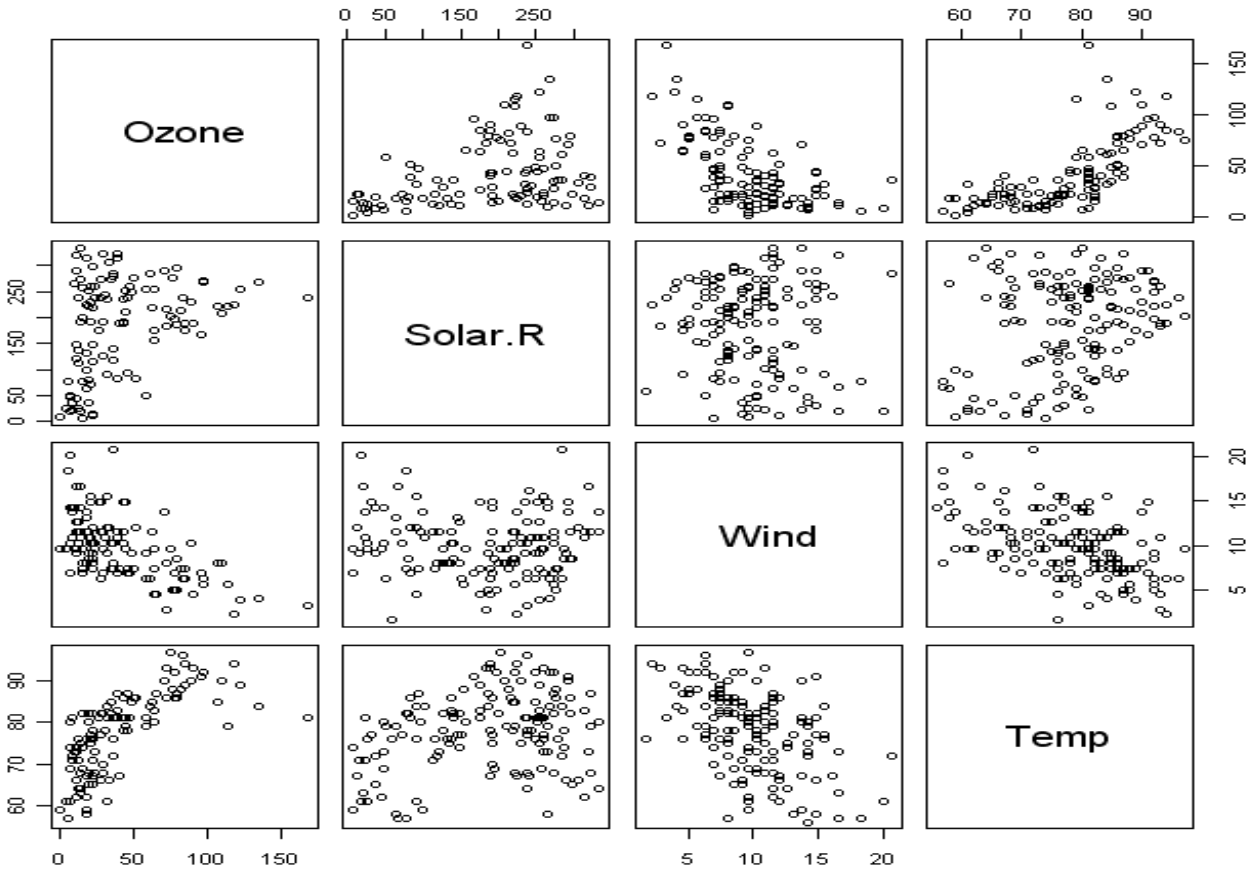


図1:散布図行列

# 重回帰分析の実際（その1）

```
Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
             Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)  -64.34208    23.05472  -2.791   0.00623 **
Solar.R        0.05982     0.02319   2.580   0.01124 *
Wind          -3.33359     0.65441  -5.094   1.52e-06 ***
Temp           1.65209     0.25353   6.516   2.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.18 on 107 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared: 0.6059,     Adjusted R-squared: 0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

# 重回帰分析の実際（その1）

- 残差の検討
  - $E(\varepsilon_i) = 0$
  - $V(\varepsilon_i) = \sigma 2$
  - $Cov(\varepsilon_i, \varepsilon_j) = 0$
  - $\varepsilon_i$ が正規分布に従う。

# 重回帰分析の実際(その1)

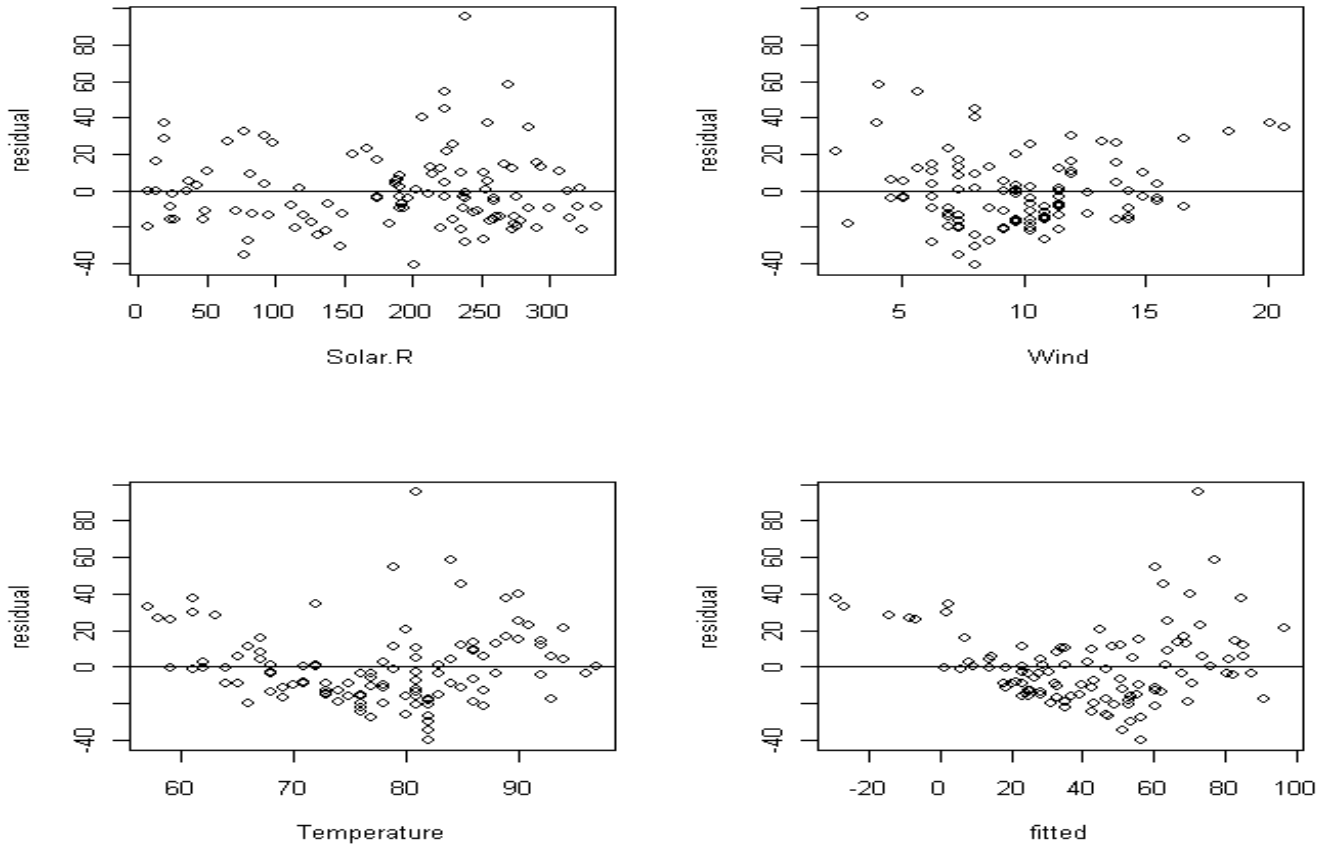


図2:残差プロット

# 重回帰分析の実際(その1)

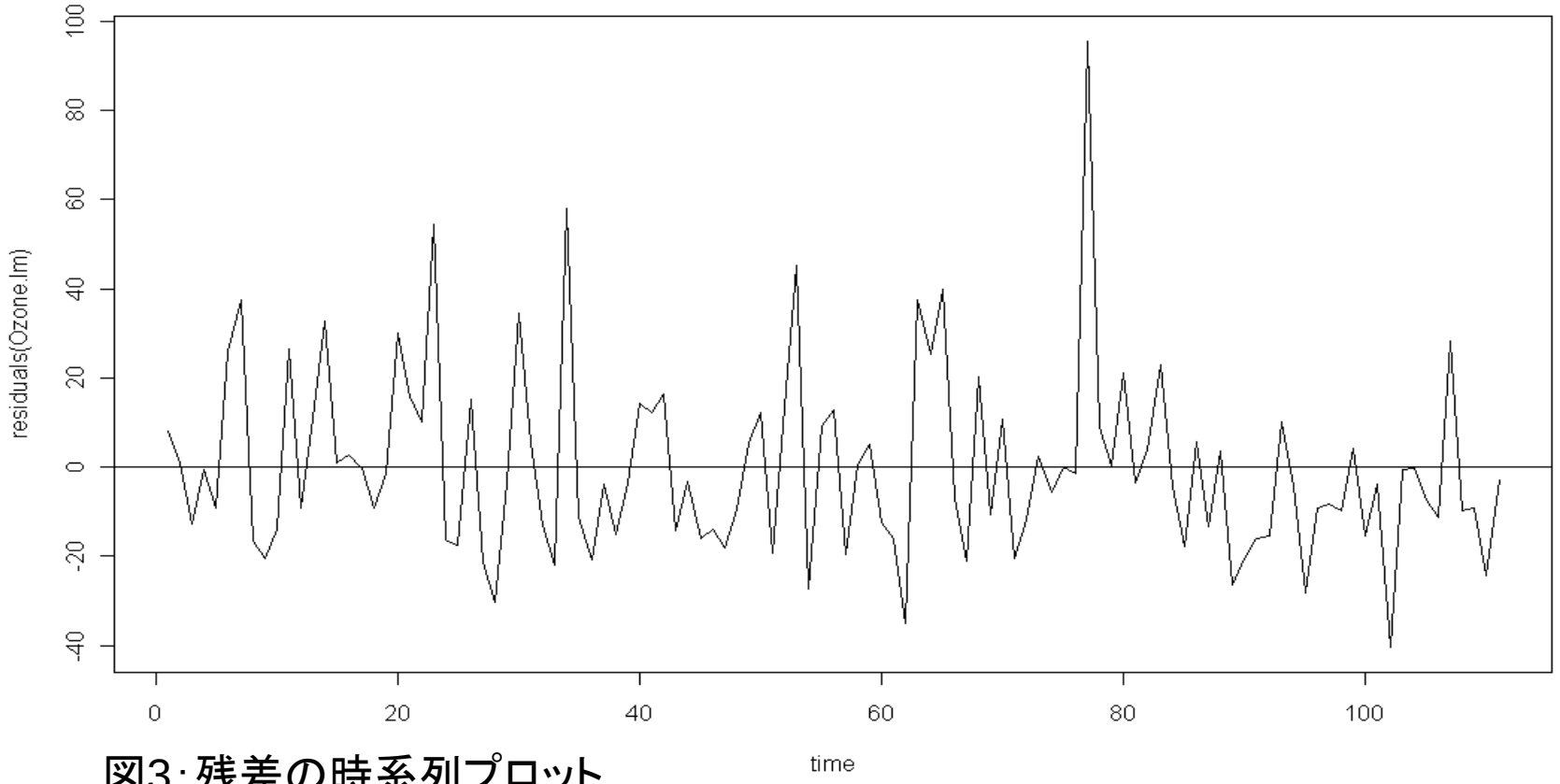


図3:残差の時系列プロット

# 重回帰分析の実際(その1)

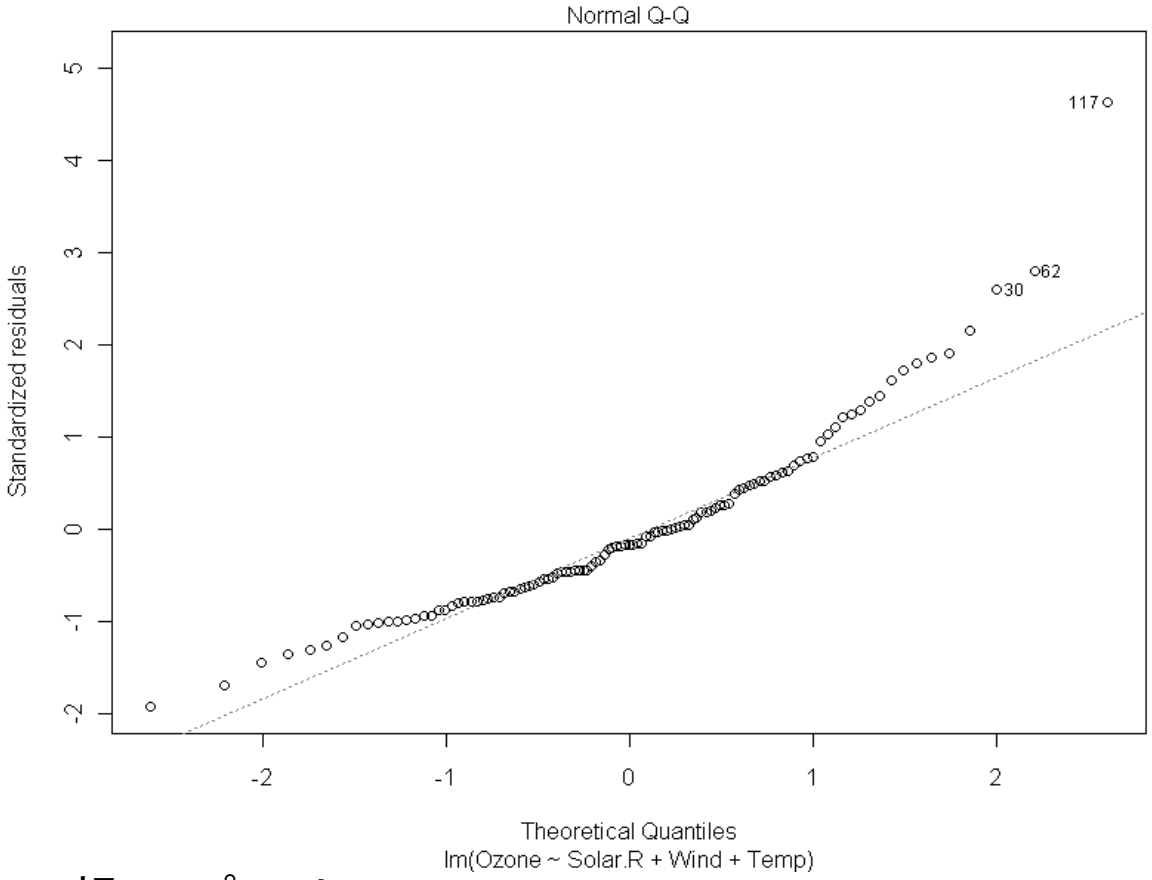


図4:正規QQプロット

# 重回帰分析の実際(2)

- ビールの需要関数(「統計学入門」で使用)
  - ビール購入数量
  - ビール価格指数/消費者物価指数
  - 実質可処分所得
  - データ
    - 総務省統計局「家計調査」
    - 総務省統計局「消費者物価指数」

# 重回帰分析の実際(2)

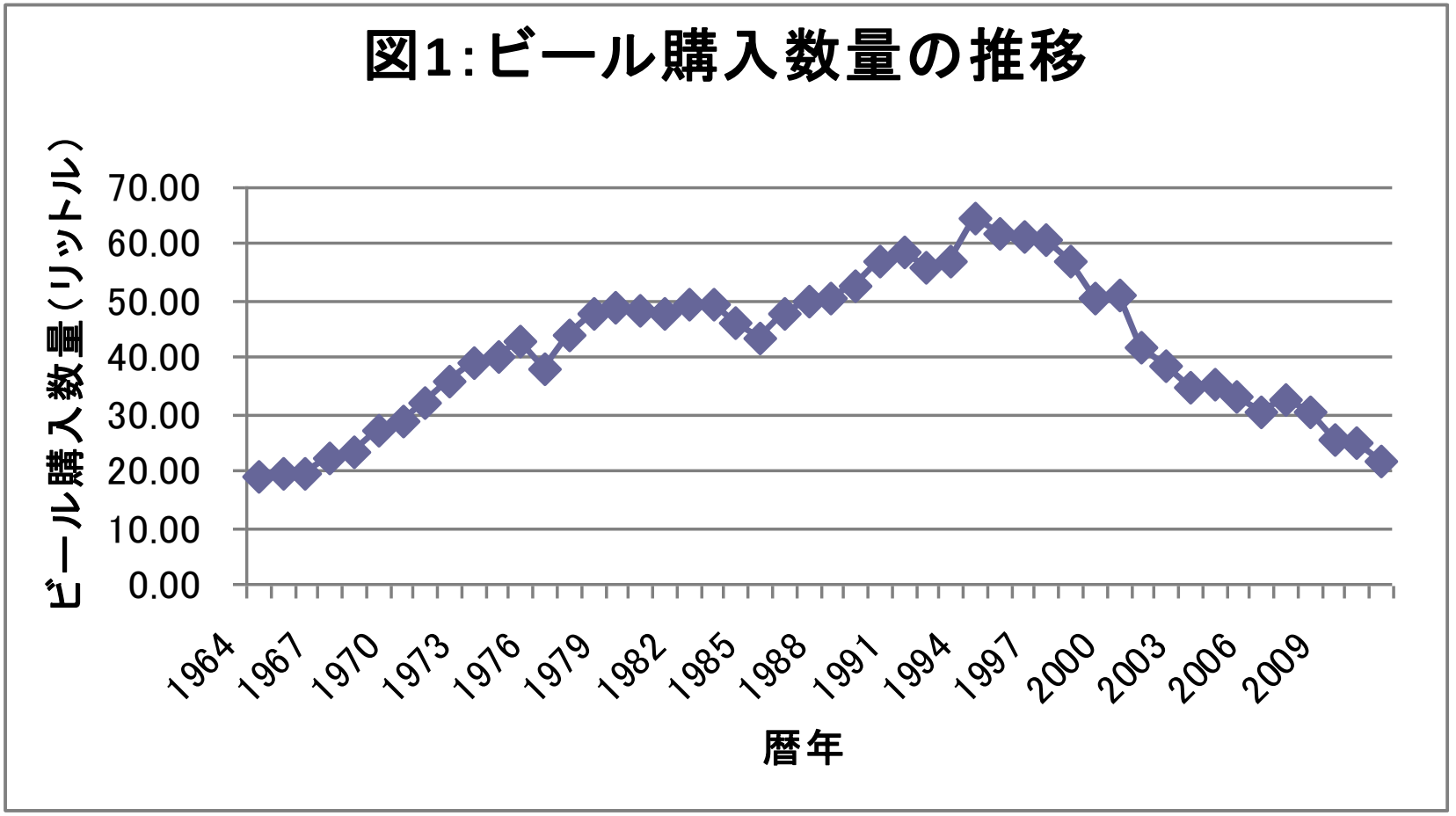
図1:ビール購入数量の推移
ビール購入数量(リットル)
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
暦年

# 重回帰分析の実際(2)

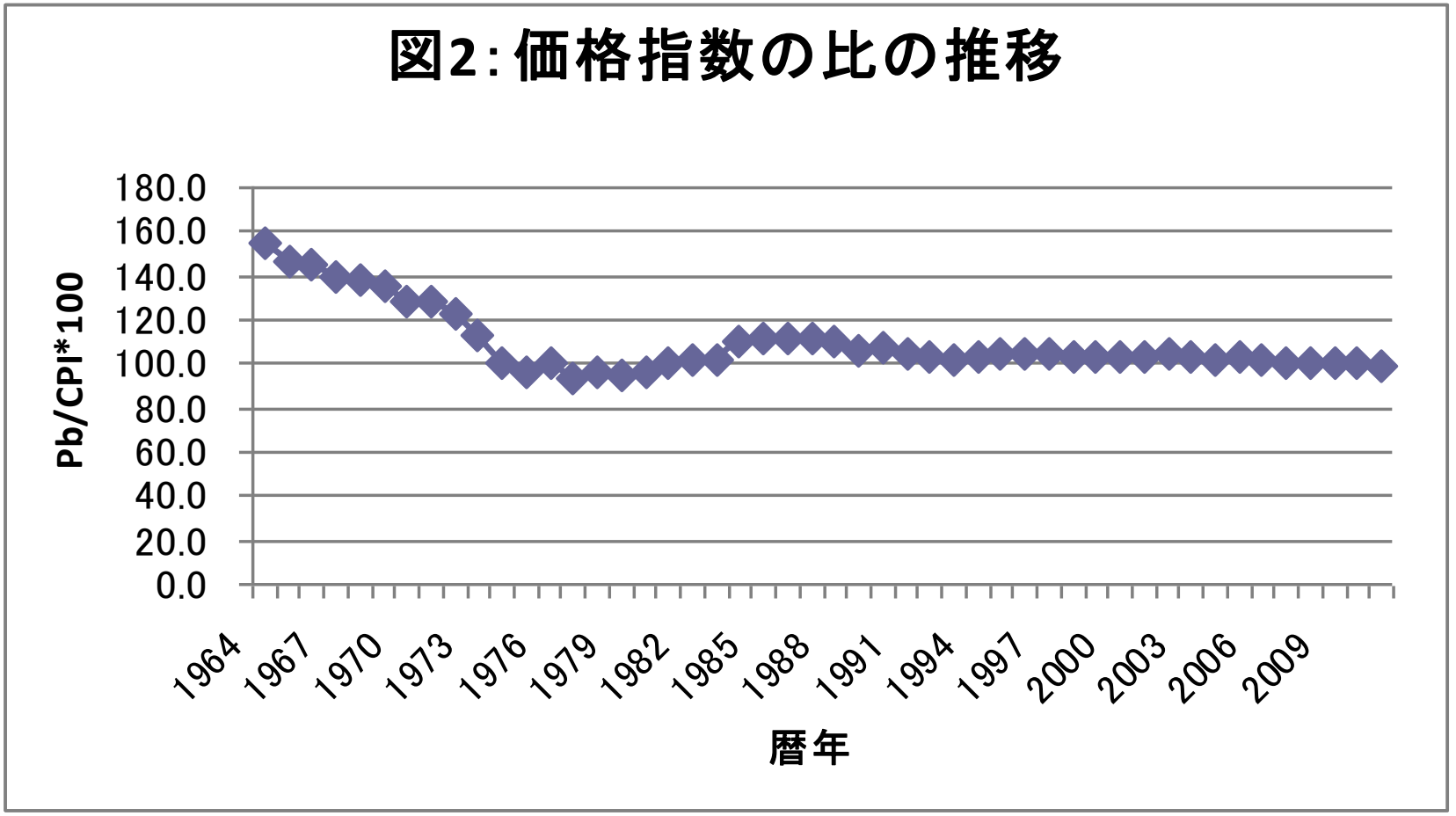
図2：価格指数の比の推移
Pb/CPI*100
180.0
160.0
140.0
120.0
100.0
80.0
60.0
40.0
20.0
0.0
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
暦年

# 重回帰分析の実際(2)

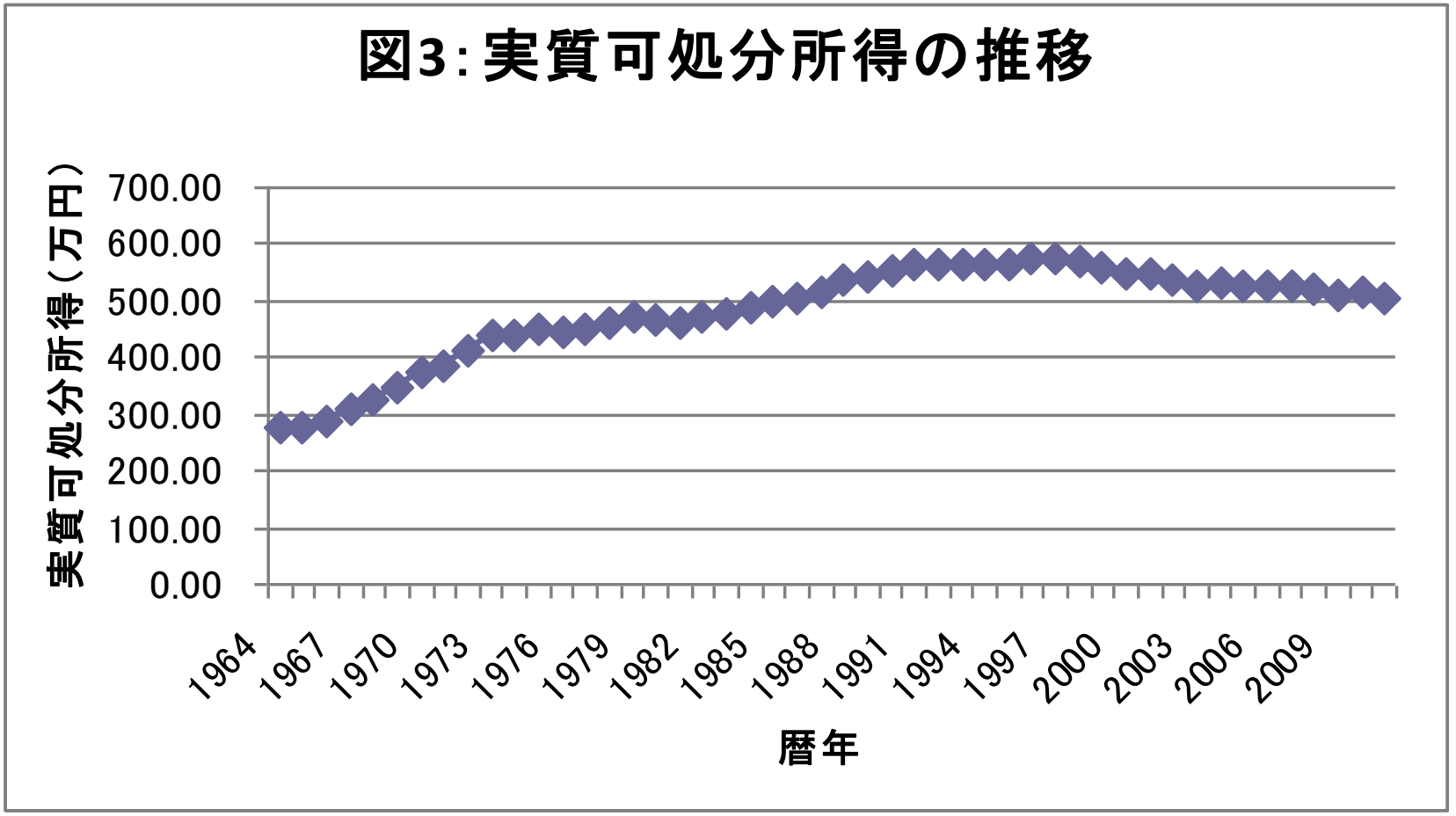
図3:実質可処分所得の推移
実質可処分所得(万円)
700.00
600.00
500.00
400.00
300.00
200.00
100.00
0.00
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
暦年

# 重回帰分析の実際(2)

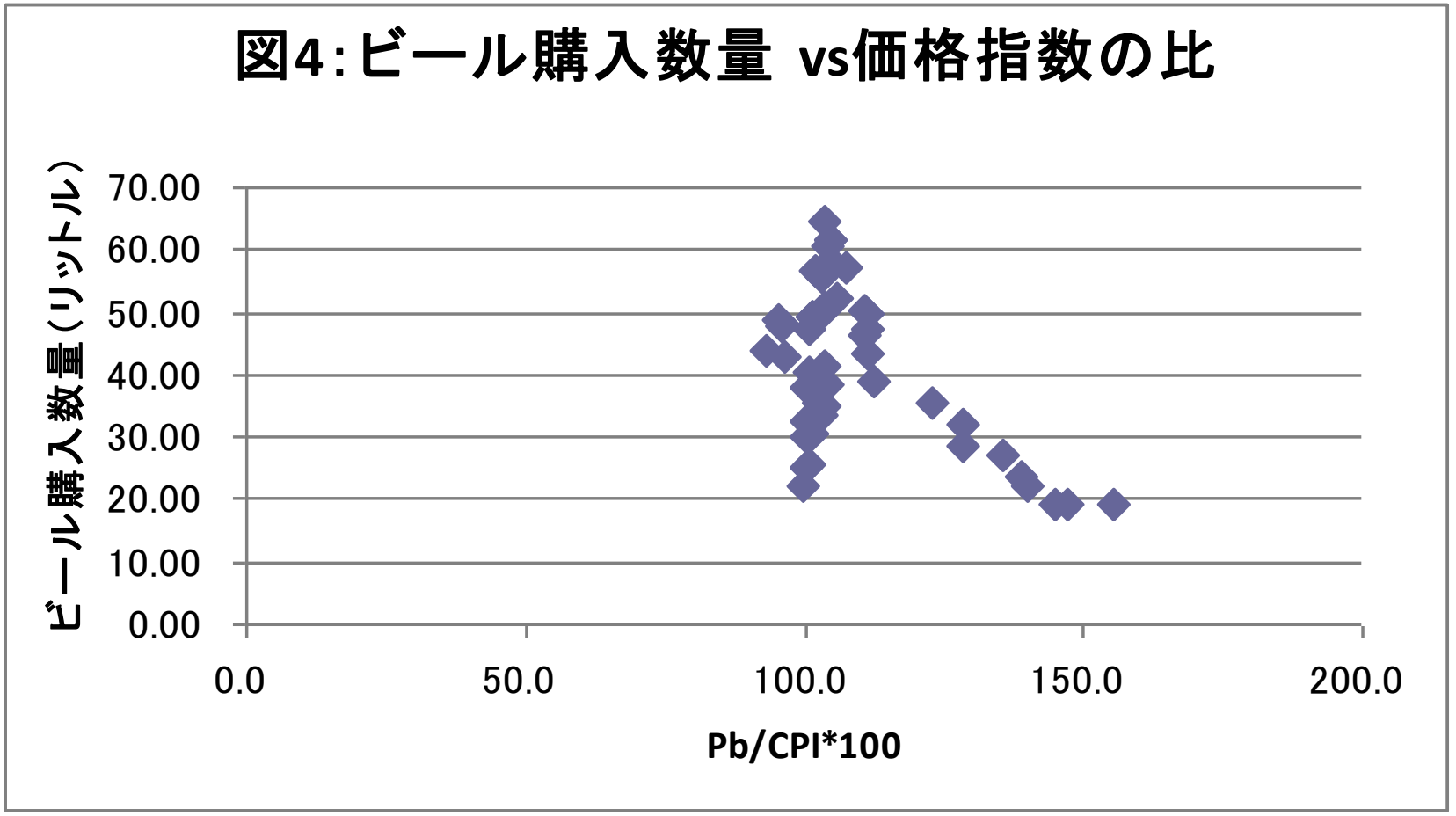
図4:ビール購入数量 vs価格指数の比
ビール購入数量(リットル)
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.0
50.0
100.0
150.0
200.0
Pb/CPI*100

# 重回帰分析の実際(2)

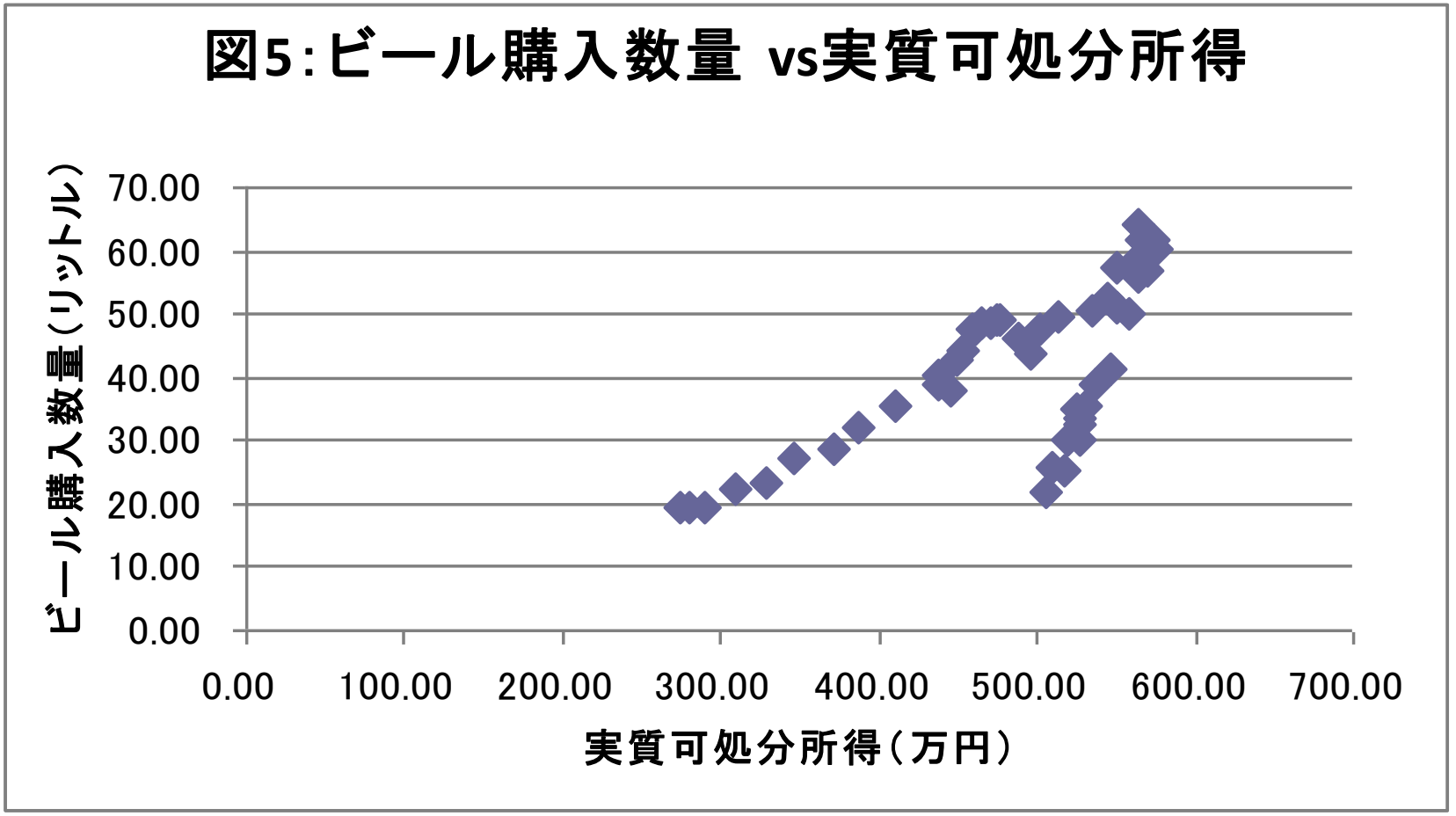
図5:ビール購入数量 vs実質可処分所得
ビール購入数量(リットル)
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
100.00
200.00
300.00
400.00
500.00
600.00
700.00
実質可処分所得(万円)

# 重回帰分析の実際(2)

計測結果

$$Q = -8.95 + 0.02\left(\frac{p_b}{CPI} \times 100\right) + 0.11y \quad R^2 = 0.51$$

- 符号条件：満たされていない。
- $R^2$ の値：低くはない。

# 重回帰分析の実際(2)

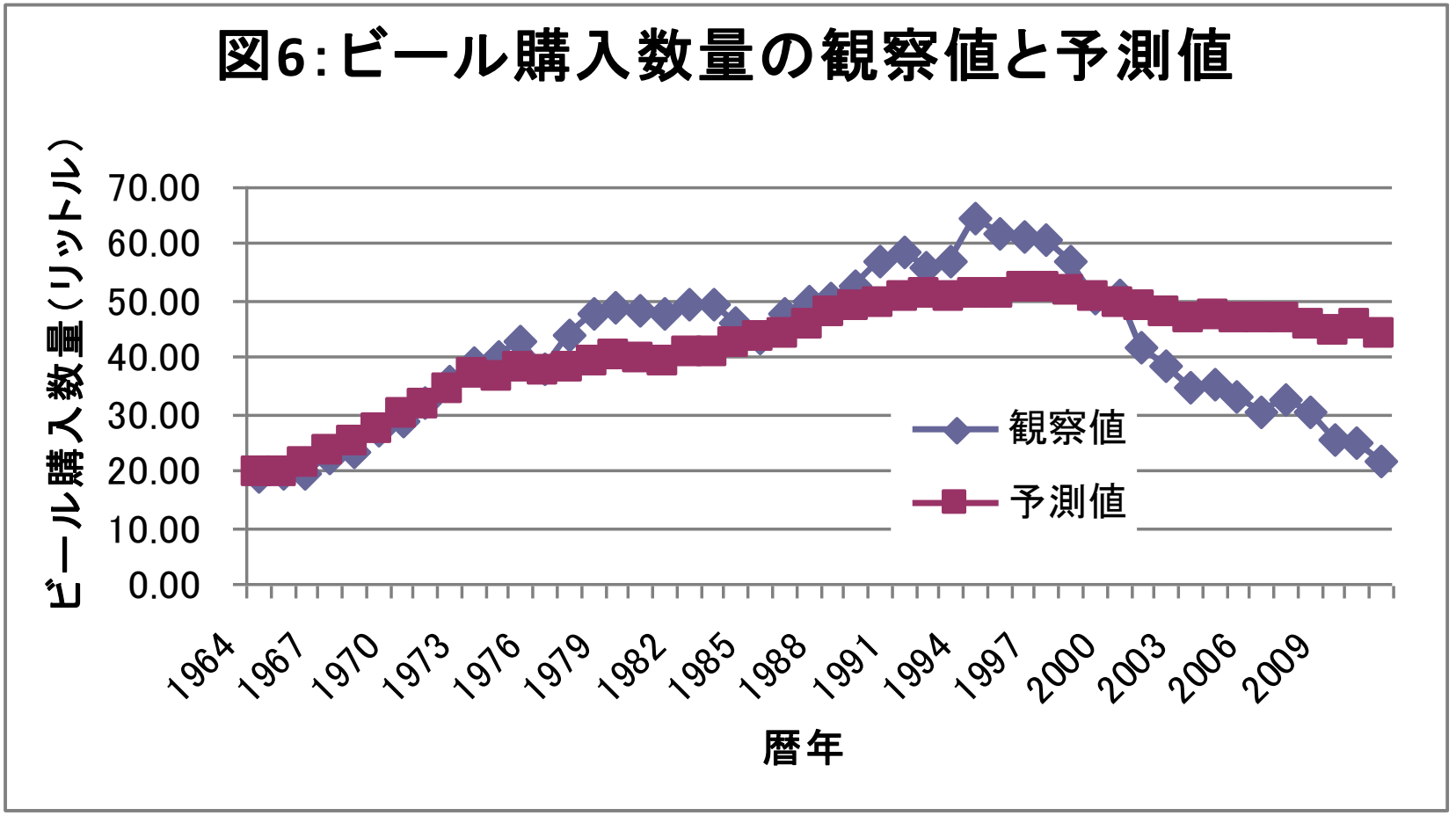
図6:ビール購入数量の観察値と予測値
ビール購入数量(リットル)
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
暦年
観察値
予測値

# 重回帰分析の実際(2)

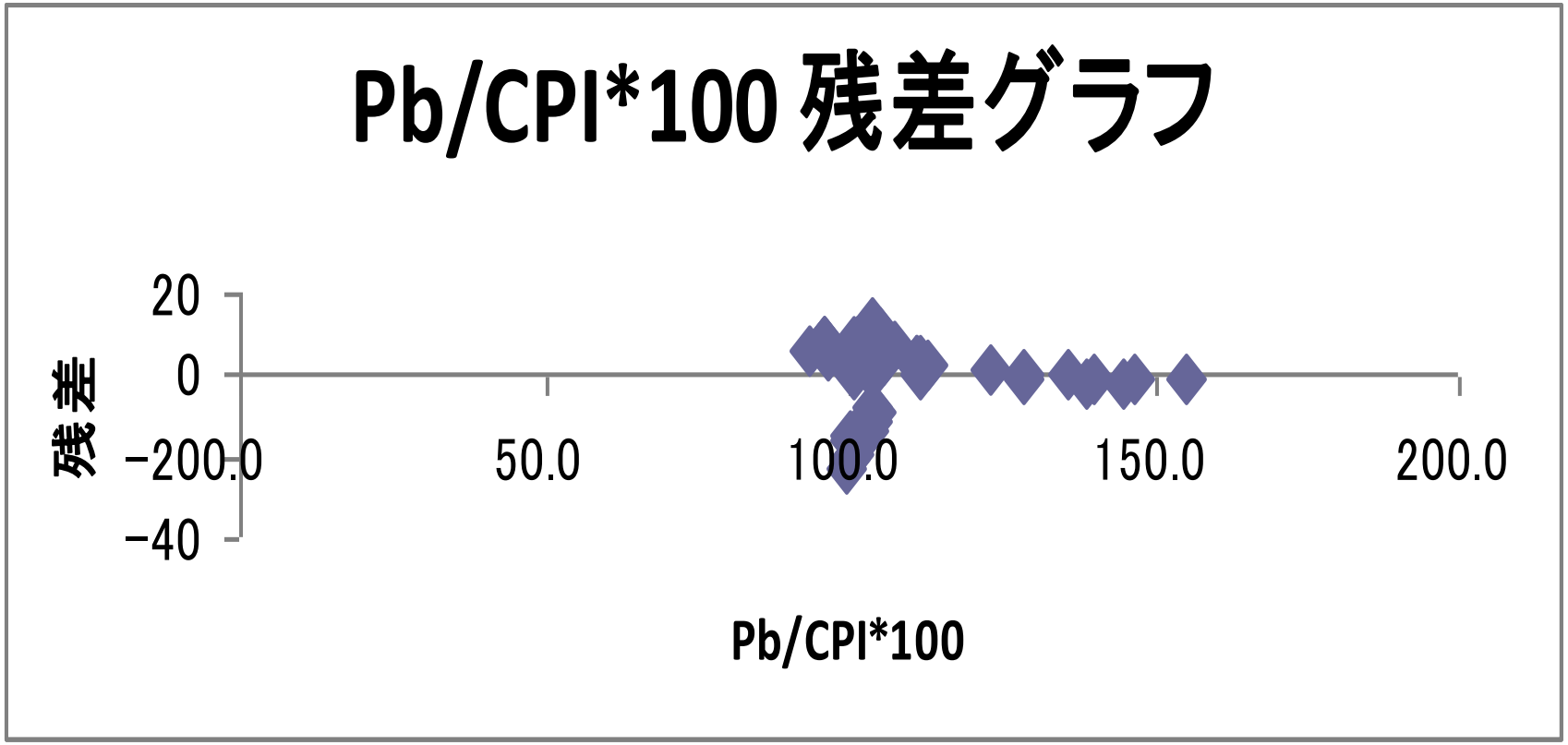


図7：残差プロット1

# 重回帰分析の実際(2)

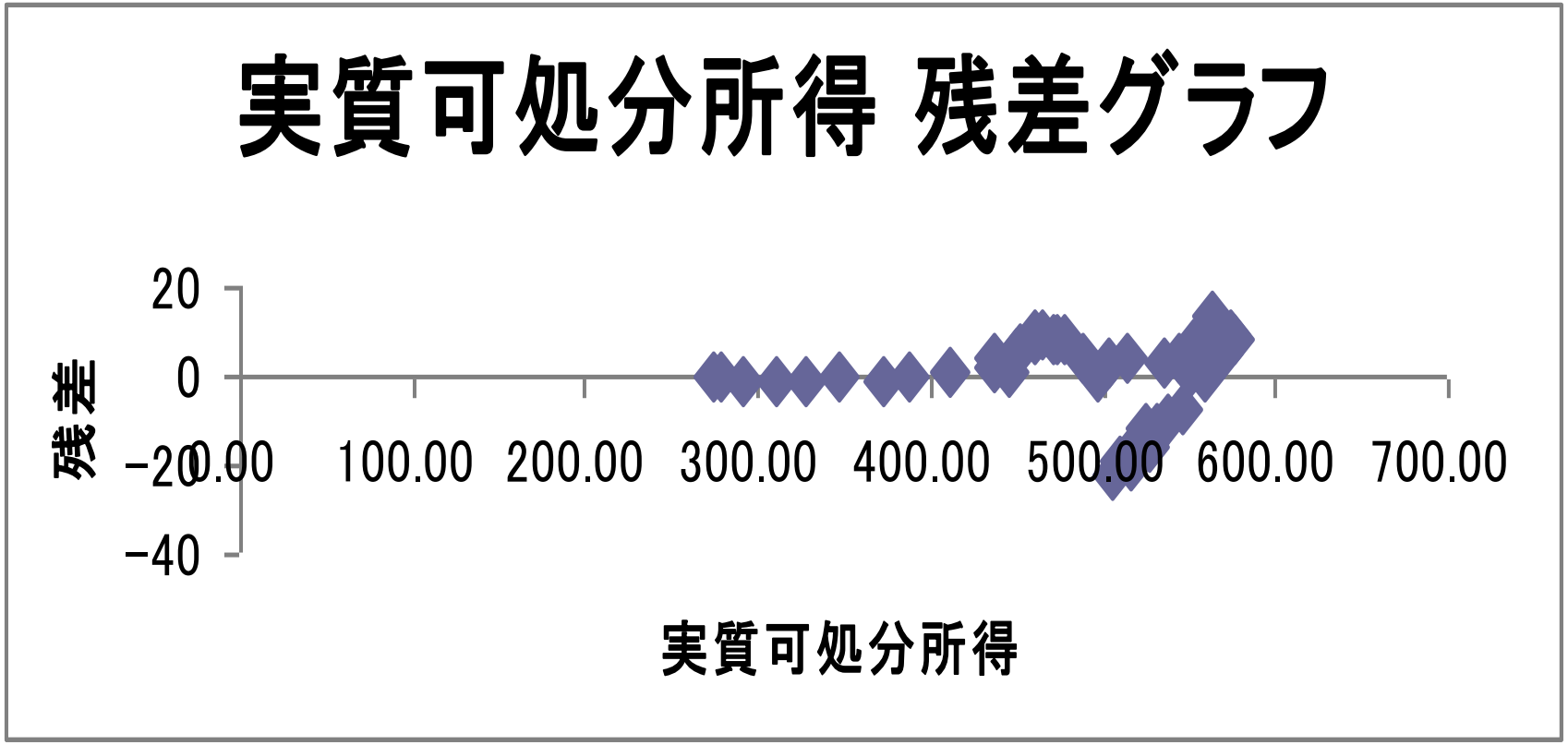


図8：残差プロット2

# 重回帰分析の実際(2)

1994年までのデータで推定した回帰式

$$Q = 0.48 - 0.12\left(\frac{P_b}{CPI} \times 100\right) + 0.12y \quad R^2 = 0.96$$

- 1994年までは世帯で消費するビールの数量は比較的単純な仕組み（相対価格と実質所得）でうまく説明できる。
- 回帰式の当てはまりがよくなくなったのは、1995年以降のビール市場の急変が原因と考えられる。

# 重回帰分析の実際(2)

- ダミー変数
  - D=0 （1994年まで）
  - D=1 （1995年以降）

$$Q = a + b\left(\frac{P_b}{CPI}\times 100\right) + cy$$

$$+ d\,D + e\,D\left(\frac{P_b}{CPI}\times 100\right) + f\,D\,y$$

# 重回帰分析の実際(2)

1994年まで

$$Q = a + b\left(\frac{P_b}{CPI} \times 100\right) + cy$$

1995年以降

$$Q = (a + d) + (b + e)\left(\frac{P_b}{CPI} \times 100\right) + (c + f)y$$

# 重回帰分析の実際(2)

- ダミー変数の係数
  - 1994年から1995年にかけての回帰係数の変化をあらわす。

$$Q = 0.4 - 0.11\left(\frac{P_b}{CPI}\times 100\right) + 0.12y$$

$$-317.1D + 0.75D\left(\frac{P_b}{CPI}\times 100\right) + 0.42Dy$$

$$R^2 = 0.97$$

# 重回帰分析の実際(2)

1994年まで

$$Q = 0.4 - 0.11\left(\frac{P_b}{CPI} \times 100\right) + 0.12y$$

1995年以降

$$Q = -316.6 + 0.63\left(\frac{P_b}{CPI} \times 100\right) + 0.54y$$

# 重回帰分析の実際(2)

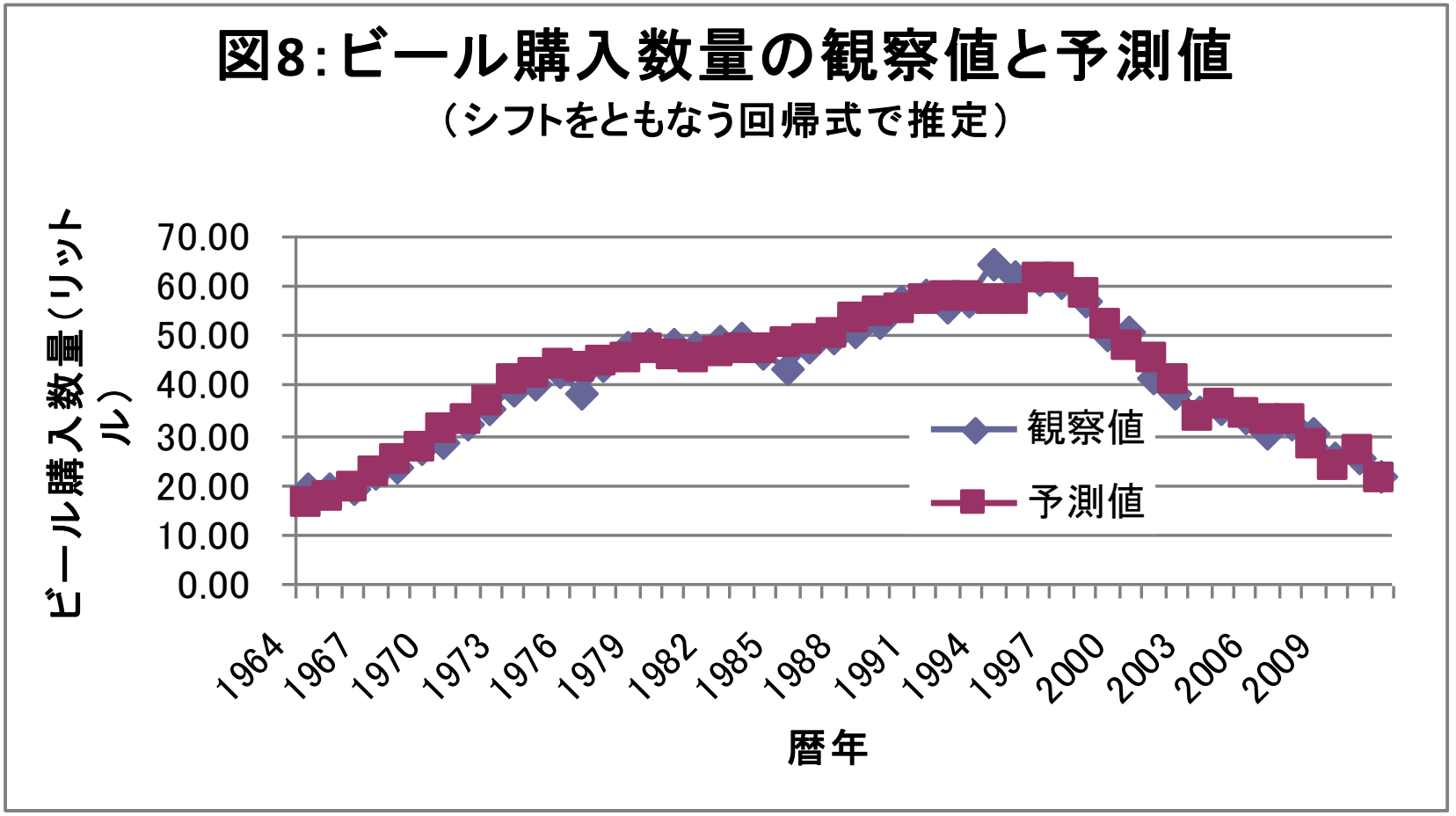
図8:ビール購入数量の観察値と予測値
（シフトをともなう回帰式で推定）
ビール購入数量（リットル）
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
暦年
観察値
予測値

# 重回帰分析の実際(2)

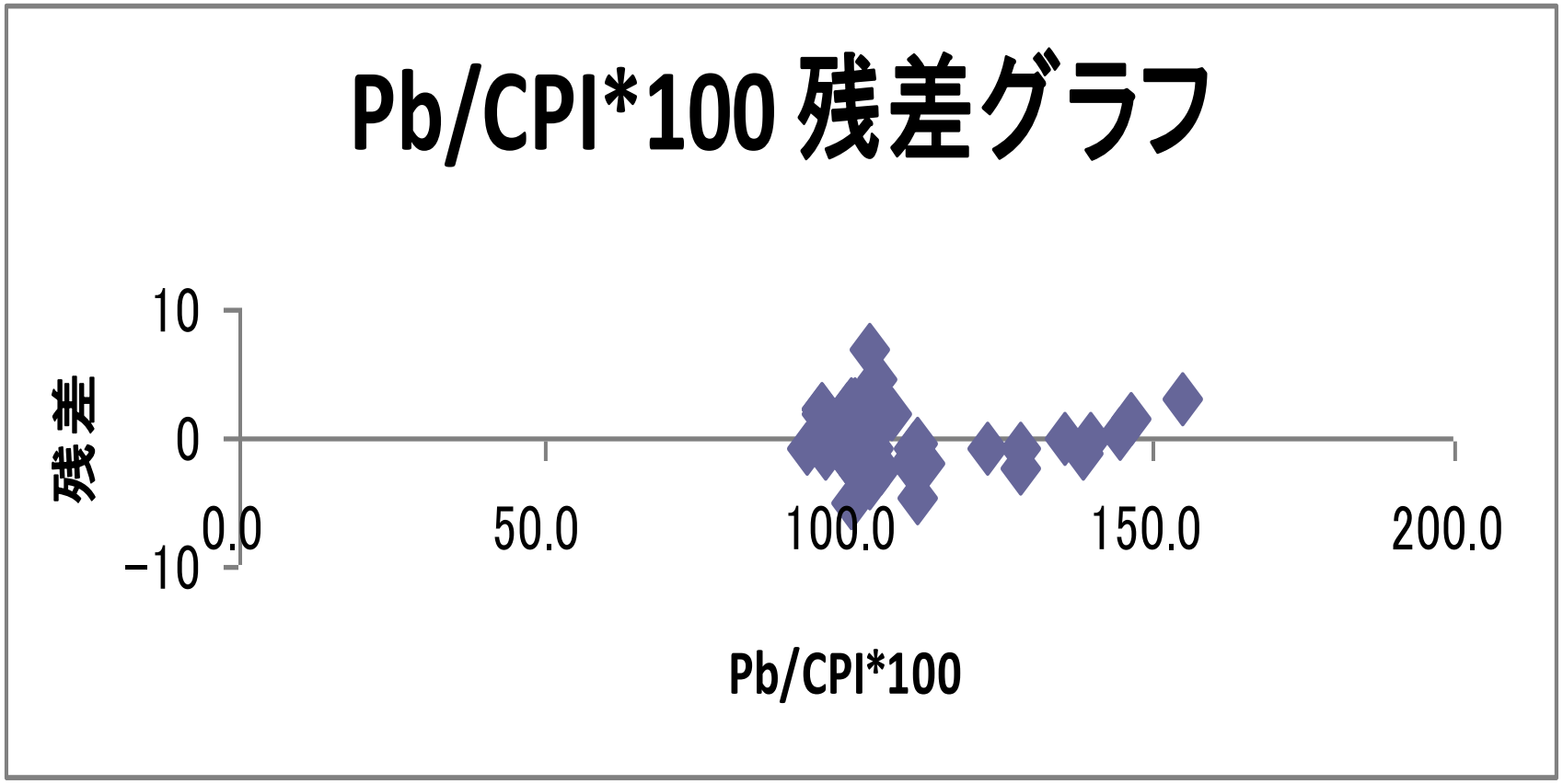


図9：残差プロット3

# 重回帰分析の実際(2)

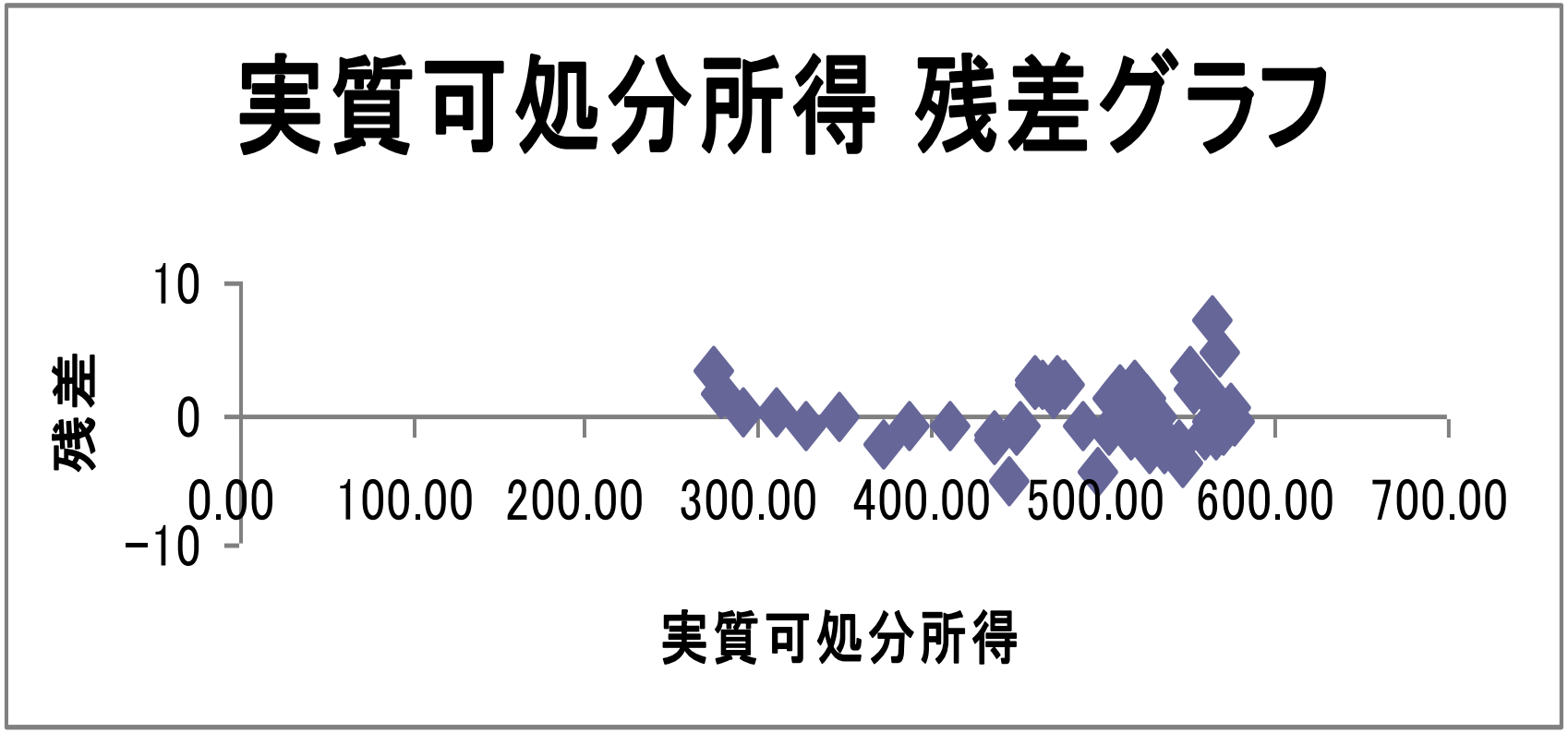


図10：残差プロット4

# 重回帰分析の実際(2)

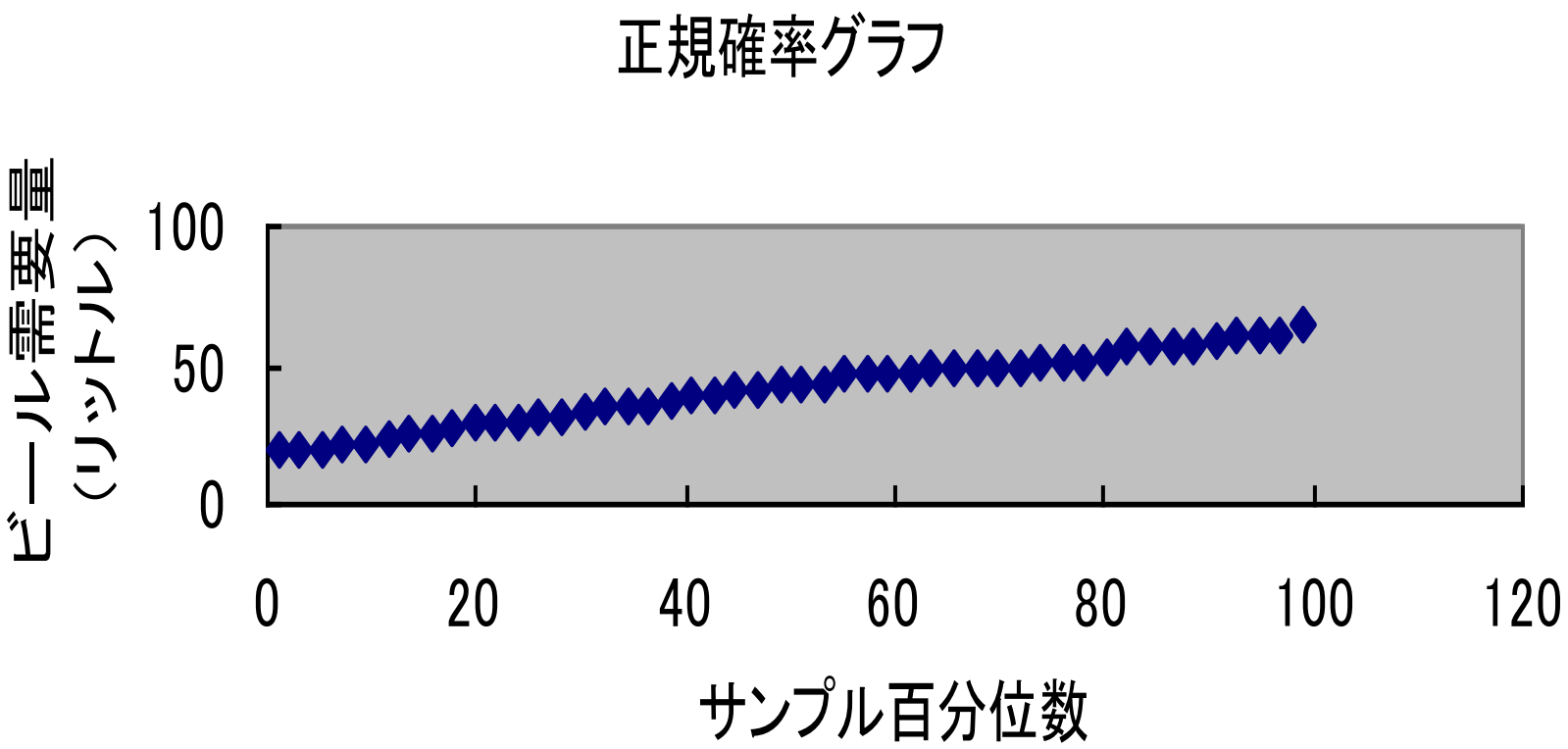


図11：正規確率紙

# 重回帰分析の実際(2)

表1：Excel の出力

概要

| 回帰統計 | |
|---|---|
| 重相関 R | 0.983341 |
| 重決定 R2 | 0.96696 |
| 補正 R2 | 0.963027 |
| 標準誤差 | 2.473136 |
| 観測数 | 48 |

分散分析表

| | 自由度 | 変動 | 分散 | 測された分散 | 有意 F |
|---|---|---|---|---|---|
| 回帰 | 5 | 7518.195 | 1503.639 | 245.8372 | 5.97E-30 |
| 残差 | 42 | 256.8888 | 6.116401 | | |
| 合計 | 47 | 7775.084 | | | |

| | 係数 | 標準誤差 | t | P-値 | 下限 95% | 上限 95% | 下限 95.0% | 上限 95.0% |
|---|---|---|---|---|---|---|---|---|
| 切片 | 0.437433 | 8.556411 | 0.051123 | 0.95947 | -16.8301 | 17.70497 | -16.8301 | 17.70497 |
| 実質可処分所得 | 0.121794 | 0.008806 | 13.83072 | 3.1E-17 | 0.104022 | 0.139565 | 0.104022 | 0.139565 |
| Pb/CPI*100 | -0.11441 | 0.044212 | -2.5877 | 0.013212 | -0.20363 | -0.02518 | -0.20363 | -0.02518 |
| D | -317.051 | 46.44033 | -6.82705 | 2.57E-08 | -410.771 | -223.33 | -410.771 | -223.33 |
| Dy | 0.42084 | 0.048521 | 8.673401 | 6.57E-11 | 0.322921 | 0.518759 | 0.322921 | 0.518759 |
| D pb/CPI | 0.746701 | 0.631089 | 1.183194 | 0.243387 | -0.52689 | 2.020291 | -0.52689 | 2.020291 |

# 重回帰分析：まとめ

- 重回帰分析
  - 単回帰分析と似ている。
    - t検定で係数が0と異なるか同かを検定する。
  - 誤差項の性質を残差プロットで確認する。
    - 説明変数のどの部分でも、残差の平均が0に近いか。
    - 説明変数のどの部分でも、残差のバラツキが一定か。
  - もっと詳しい内容は「計量経済学」で扱われる。