

統計学I

早稲田大学政治経済学術院

西郷 浩

本日の目標

- 直線(および平面)の当てはめ
 - 回帰直線
 - 最小二乗法
 - 当てはまりの評価
 - 決定係数
 - 散布図・残差プロットの活用
 - PC実習

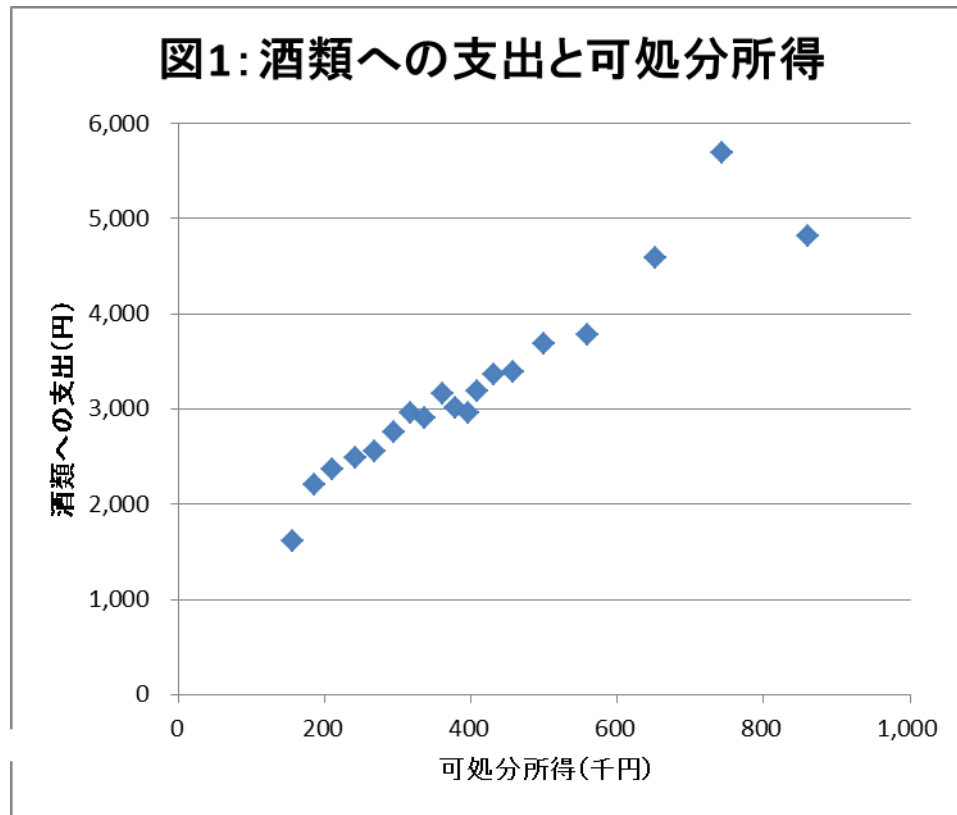
回帰直線(1)

- 2次元データ $(x_i, y_i) \ i = 1, 2, \dots, N$.
 - y を x の式であらわす。
 - 「 x の値がわかれば、 y の(おおよその)値がわかる」(変数間の関係の要約)

– 例

- $x_i = (i \text{ 番目の個体の可処分所得})$
- $y_i = (i \text{ 番目の個体の酒類への支出})$

回帰直線(2)



資料: 総務省統計局「平成26年全国消費実態調査」

回帰直線(3)

- どのような式が望ましいか。
 - 当てはまりの良いもの。
 - すべての観測点 (x_i, y_i) ($i = 1, 2, \dots, N$)からのズレ具合が小さいもの。
 - 要約(一種の単純化)であるから、数理的な側面が扱いやすい簡単なもの。
- ⇒ 直線 $y = a + bx$ が候補。

回帰直線(4)

- 目分量で直線を描く⇒客観的でない。
 - 同じデータに対していくつもの「妥当な」直線が描けることになってしまう。
- 客観的な基準の必要性
 - 「全観察点 (x_i, y_i) ($i = 1, 2, \dots, N$) と直線とのズレ」を数量的に定義し、ズレが最小になるように直線の位置を選ぶ(すなわち、 a, b の値を定める)。

最小二乗法(1)

- 記号の整理

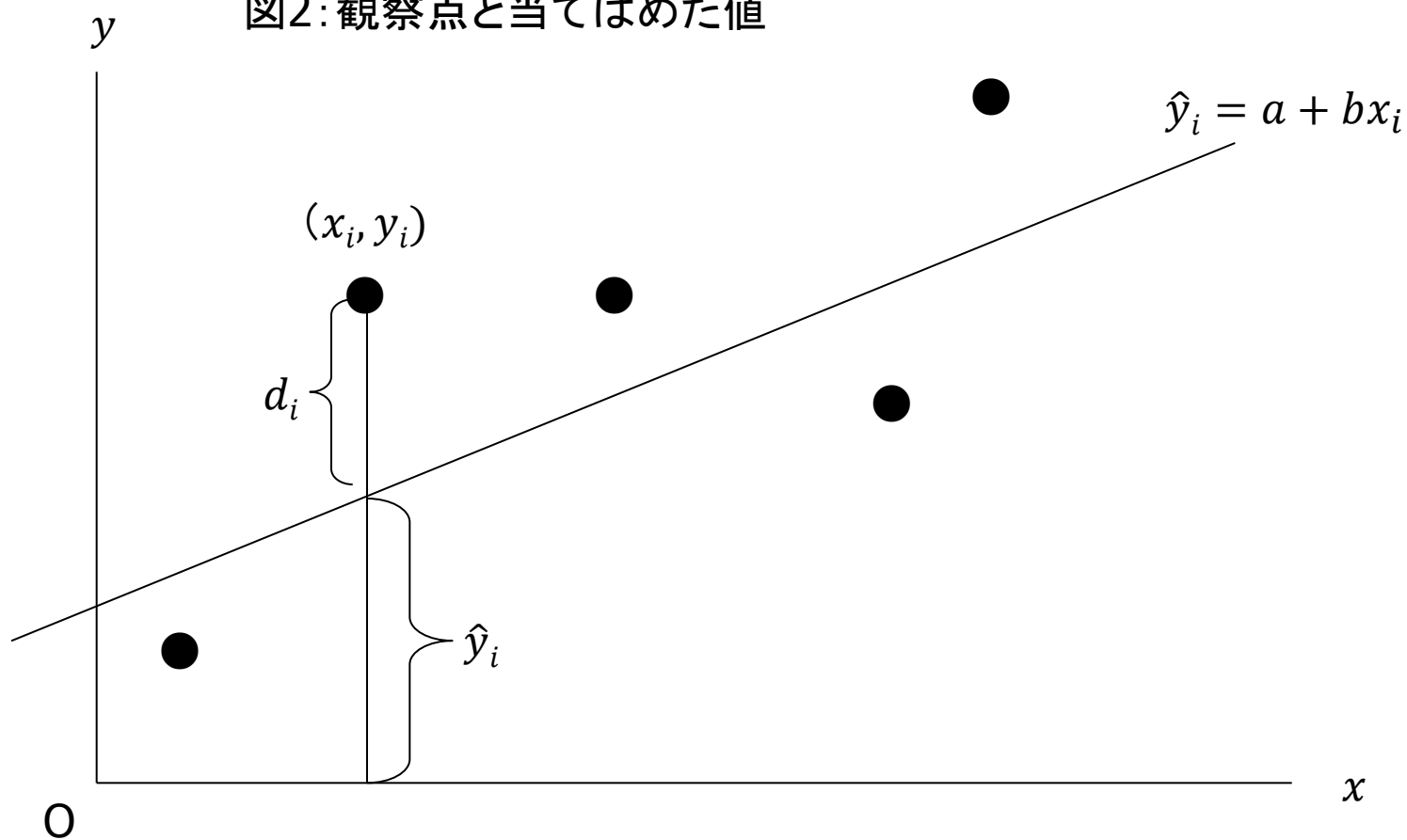
y_i = (観察された y の値)

$\hat{y}_i = a + bx_i$ = (x によって予想される y の値)

$d_i = y_i - \hat{y}_i = y_i - (a + bx_i)$
= (観察された y と、 x で予想される y の値との差)
(d_i を残差とよぶ)

最小二乗法(2)

図2: 観測点と当てはめた値



最小二乗法(3)

- 残差 d_i は、第 i 観察点における、観察値と予想値との乖離を示す。
 - x と y との関係を「うまく」要約したいのなら、 $d_i = 0$ となることが望ましい。
 - 特定の(2つまでの) i について、 $d_i = 0$ とするのは可能である。しかし、全部 0 にすることは無理である。
 - 「観察点全体と直線とのズレ」を残差によって数量的に定義し、それを最小にする。

最小二乗法(4)

$$\text{最小二乗法： } \min_{a,b} \sum_{i=1}^N d_i^2 \Leftrightarrow \min_{a,b} \sum_{i=1}^N (y_i - a - bx_i)^2$$

この最小化問題の解 $\Leftrightarrow$ 下の方程式（正規方程式）の解

$$\begin{cases} N a + \left(\sum_{i=1}^N x_i \right) b = \left(\sum_{i=1}^N y_i \right) \\ \left(\sum_{i=1}^N x_i \right) a + \left(\sum_{i=1}^N x_i^2 \right) b = \left(\sum_{i=1}^N x_i y_i \right) \end{cases}$$

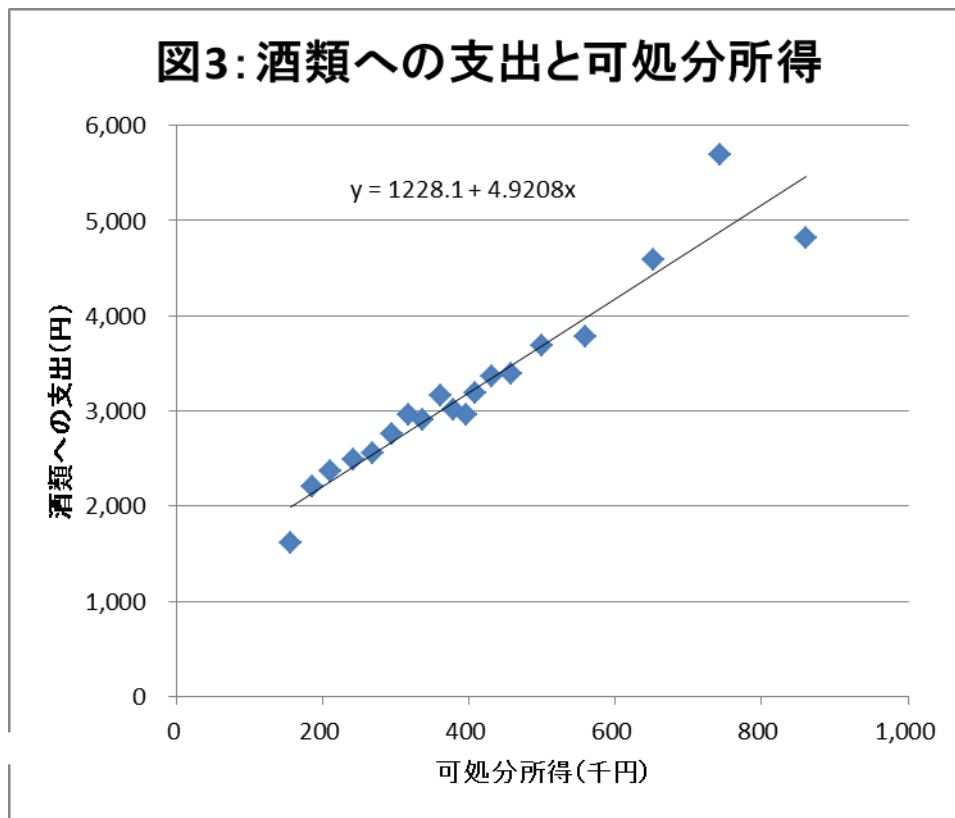
この連立方程式の解は下のようになる。

$$\begin{cases} b = s_{xy} / s_x^2 = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^N (x_i - \bar{x})^2 \\ a = \bar{y} - b\bar{x} \end{cases}$$

最小二乗法(5)

- 可処分所得(x)と酒類への支出 (y)とについて、最小二乗法によって切片 a と傾き b を求めると、
 - 傾き $b = 4.9$
 - 切片 $a = 1228$

最小二乗法(6)

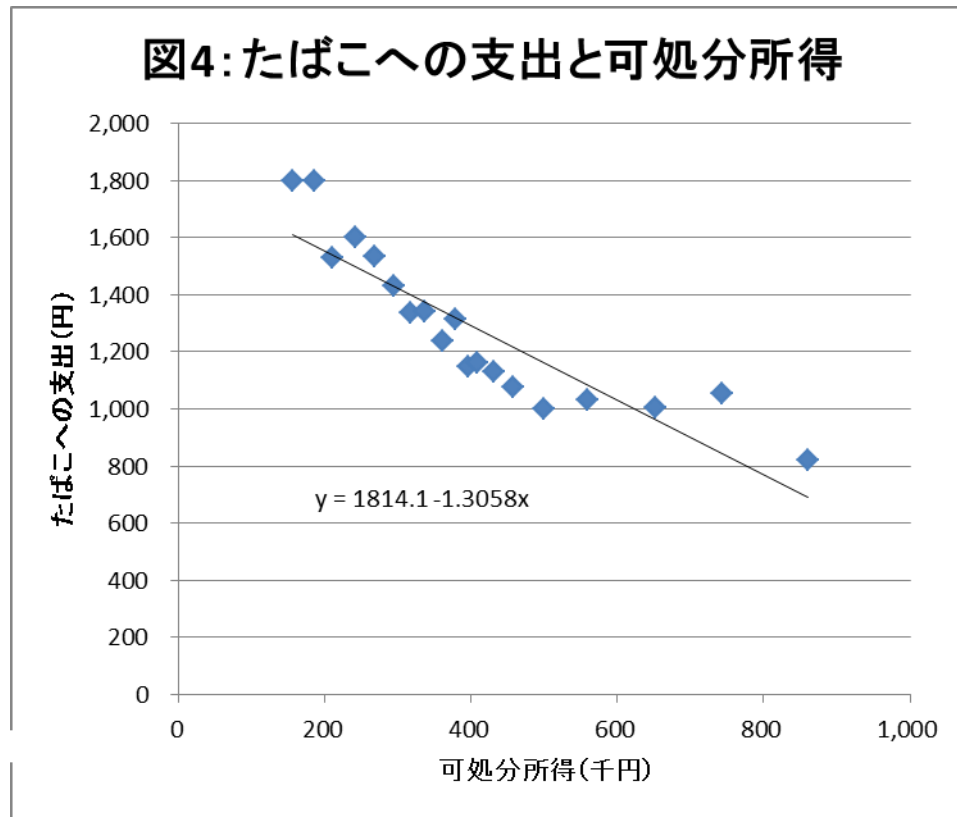


資料: 総務省統計局「平成26年全国消費実態調査」

当てはまりの評価(1)

- 対象によって当てはまりに差がある。
 - A: x =(可処分所得), y =(酒類)
 - B: x =(可処分所得), y =(たばこ)見た目の当てはまりはどちらが良い？
- このような相違の評価方法は？
 - 決定係数
 - 散布図・残差プロットの活用

当てはまりの評価(2)



資料:総務省統計局「平成26年全国消費実態調査」

決定係数(1)

- 平方和の分解

$d_i = y_i - \hat{y}_i$ であるから、

$$y_i = \hat{y}_i + d_i$$

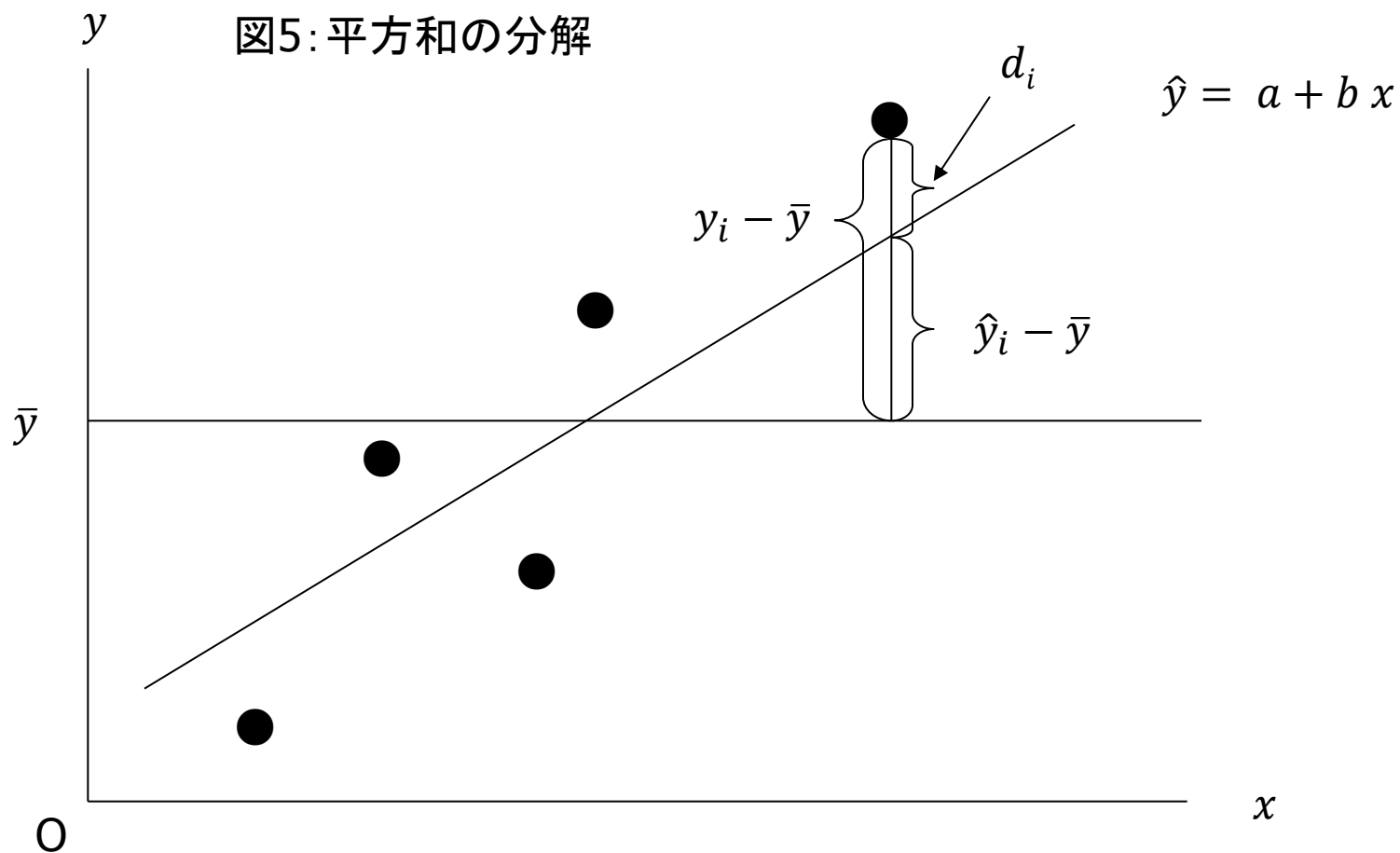
両辺から $\bar{y}$ を引いて、

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + d_i$$

残差 d_i の性質を利用すると、

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

決定係数(2)



決定係数(3)

$$SS_T = \sum_{i=1}^N (y_i - \bar{y})^2 = (\text{観察される}y\text{の変動})$$

$$SS_R = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 = (\text{回帰式で説明できる}y\text{の変動})$$

$$SS_E = \sum_{i=1}^N d_i^2 = (\text{回帰式で説明できない}y\text{の変動})$$

とすれば、

$$SS_T = SS_R + SS_E$$

決定係数(4)

平方和（二乗和）なので、

$$SS_T \geq 0, SS_R \geq 0, SS_E \geq 0$$

したがって、

$$\begin{aligned} R^2 &= \frac{SS_R}{SS_T} = \frac{\text{（回帰式で説明できる}y\text{の変動）}}{\text{（観察される}y\text{の変動）}} \\ &= 1 - \frac{SS_E}{SS_T} \end{aligned}$$

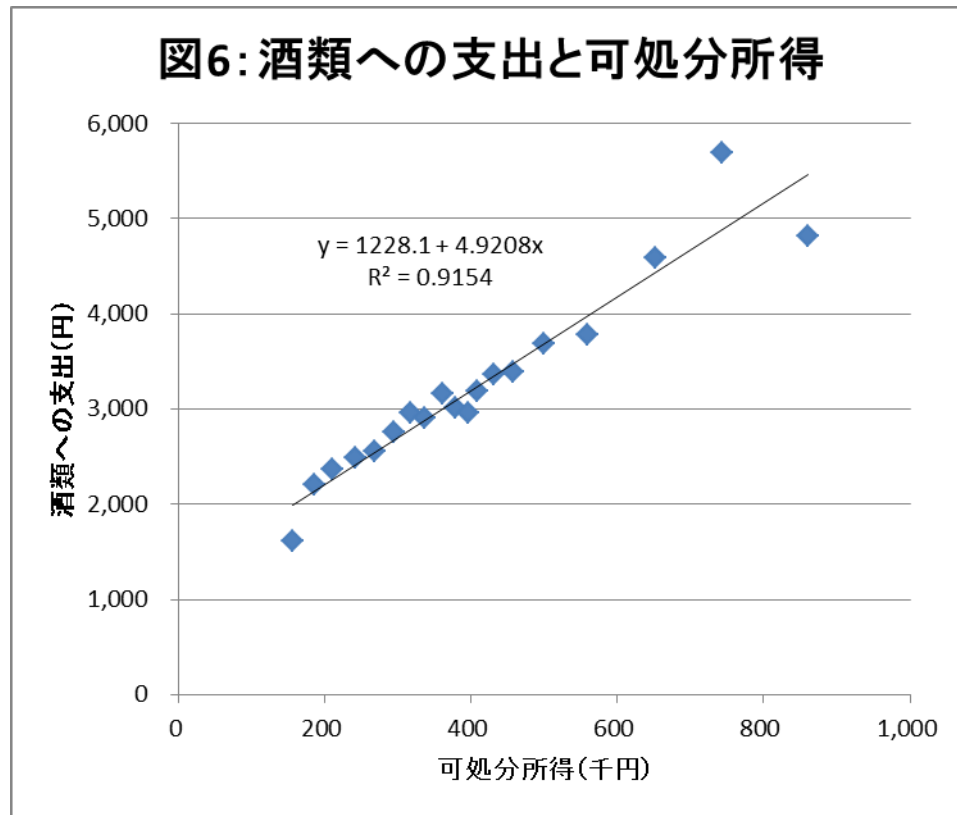
とすれば、 $0 \leq R^2 \leq 1$ であり、

これが1に近いほど当てはまりがよいことをあらわす。

決定係数(5)

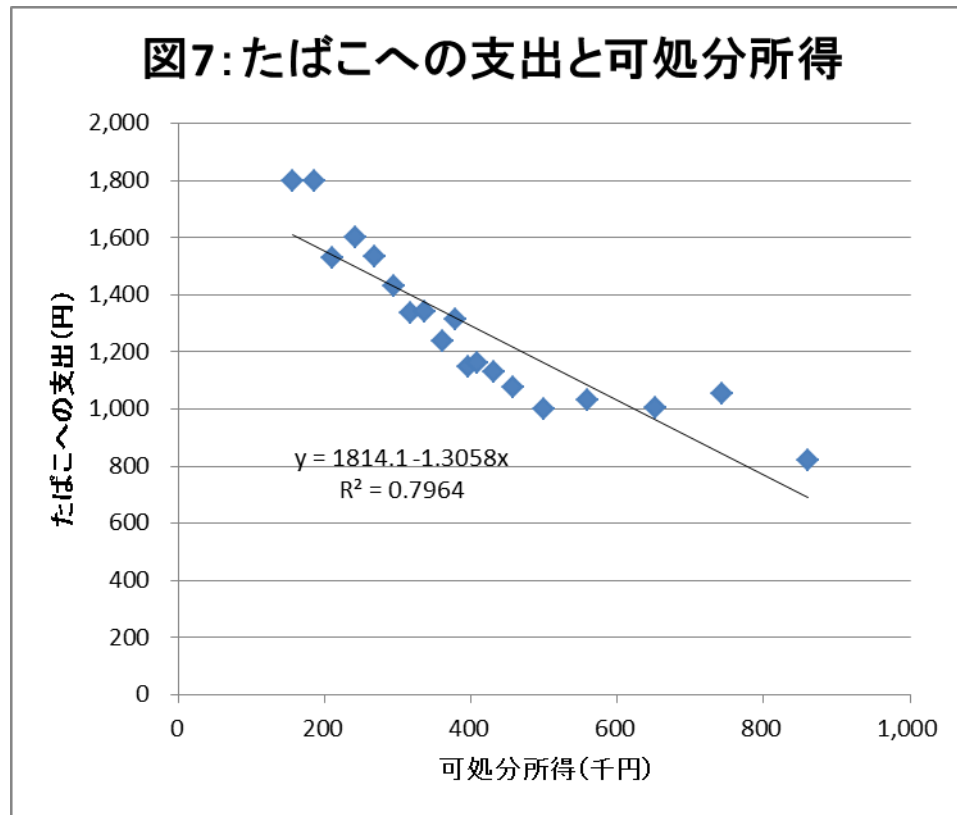
- 決定係数
 - A: 可処分所得と酒類 $R^2 = 0.92$
 - B: 可処分所得とたばこ $R^2 = 0.80$
- 決定係数と相関係数との関係
 - $R^2 = r_{xy}^2$
 - この関係式を使って決定係数を定義する教科書もある。
 - 直線関係の強さの尺度であることに注意する。

決定係数(6)



資料: 総務省統計局「平成26年全国消費実態調査」

決定係数(7)



資料: 総務省統計局「平成26年全国消費実態調査」

散布図・残差プロットの活用(1)

- 散布図の活用

- 散布図に回帰直線を描き込む。

- 観察点全体への回帰直線の当てはまり具合は適切か？

- 直線による近似に無理がないか。

- » 無理があると判断できるようなら、変数変換などの対応策を講じる(次回)。

散布図・残差プロットの活用(2)

- 残差プロットの活用

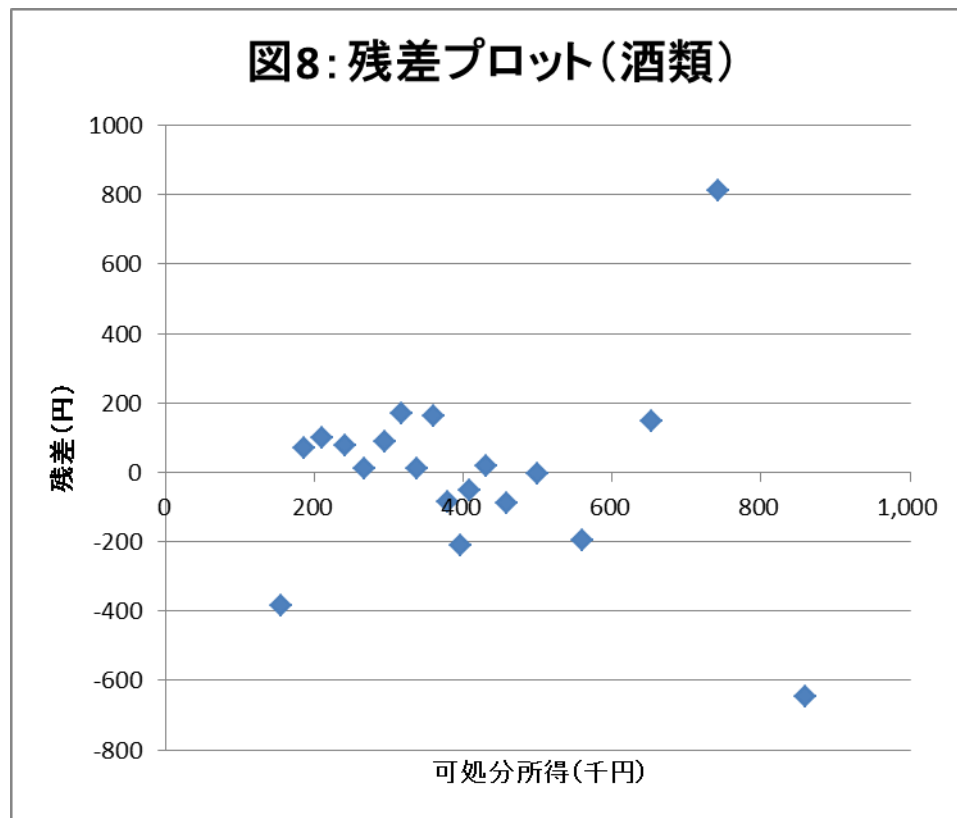
- 残差プロットの例:

- (x_i, d_i) を座標平面にあらわしたもの。
 - 説明要因 x_i の値を所与としたときに、 $d_i = y_i - \hat{y}_i$ をグラフにしたもの。
 - 何らかの規則性あり。→ 問題あり。
 - 読み取りにはコツがいる。

散布図・残差プロットの活用(3)

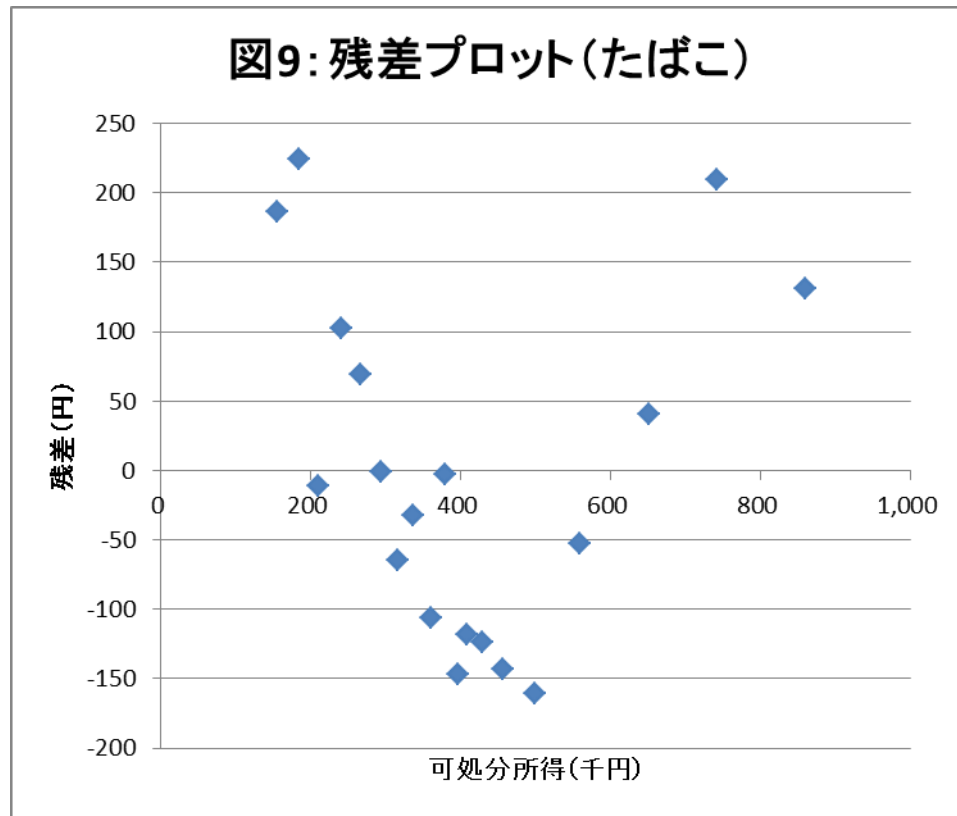
- A: 残差プロット(酒類への支出)
 - 横軸: 可処分所得
 - 縦軸: 回帰式(酒類)からえた残差
- B: 残差プロット(たばこへの支出)
 - 横軸: 可処分所得
 - 縦軸: 回帰式(たばこ)からえた残差

散布図・残差プロットの活用(4)



資料: 総務省統計局「平成26年全国消費実態調査」

散布図・残差プロットの活用(5)



資料: 総務省統計局「平成26年全国消費実態調査」

散布図・残差プロットの活用(6)

－散布図・残差プロットによる吟味

- 目視による判断なので主観的になりやすい。
- しかし、決定係数の大小のみで当てはまりを判断するのは危険なので、必ず散布図・残差プロットも確認する必要がある。

PC実習

- 回帰直線の計算
 - 関数の利用
 - 分析ツールの利用
- 残差プロットの描画