統計学
回帰分析、最小二乗法1

担当:　　長倉　　大輔
（ながくら　だいすけ）

回帰分析、最小二乗法1

- 回帰分析

回帰分析は 2 つ以上の変数間の定量的な関係を
分析するのに用いられる。

- 回帰分析と因果関係

相関係数は 2 つの変数の**相関関係**を測る指標であった。

回帰分析の場合、通常 2 つの変数間の**因果関係**が
あらかじめわかっていると想定して分析する。

回帰分析、最小二乗法1

- 因果関係の例

(例1)　子供の身長と親の身長
親の身長が大きければ子供の身長も大きいという
因果関係が想定される。

(例2)　身長と体重
身長が大きければ体重も大きいという因果関係が
想定される。

回帰分析、最小二乗法1

- 因果関係の例

(例3)　所得(収入)と消費(支出)
所得が増えれば、消費(支出)が増えるという因果関係が想定される。

このように、通常、回帰分析ではどちらがどちらの原因になっているかを(経済理論などをもとに)あらかじめ想定して分析する。

回帰分析、最小二乗法1

- 線形回帰分析

ここでは**線形回帰分析**について述べる。線形回帰分析とは変数間の**線形の関係**を調べる事である。

回帰分析と呼ぶ場合は通常この線形回帰分析を指している。

まず 2 変数の線形回帰分析について述べる。

2変数の回帰分析は**単回帰分析**と呼ばれる。

回帰分析、最小二乗法1

- 線形回帰分析(2変数)

2 つの変数 Y と X の間には次の関係があるとしよう。

$$Y = \alpha + \beta X + \varepsilon$$

この時 Y は**被説明変数** (もしくは**従属変数**)、
X は**説明変数** (もしくは**独立変数**)、ε を**誤差項**という。

ε は X が説明しきれない部分をまとめたものであり、
$E(\varepsilon) = 0$ の確率変数であると仮定する。

このような Y と X の間の線形の関係の事を**線形回帰モデル**という。

回帰分析、最小二乗法1

- 単回帰分析

先ほどのモデルにおいては X が Y の大きさを決定する
(X が Y の原因になっている) という因果関係がある
(例えば X は個人の所得、Y はその個人の消費である)。

説明変数 X は**確率変数ではない**と仮定する。
(実際にはこの仮定を置かなくても後述の最小二乗法の性質は成り立つが、説明の簡単化のためにこのように仮定する)。

回帰分析、最小二乗法1

- 被説明変数 Y の期待値

$E(\varepsilon) = 0$ かつ X は**確率変数ではない**ので

$$\begin{aligned} E(Y) &= E(\alpha + \beta X + \varepsilon) \\ &= E(\alpha) + E(\beta X) + E(\varepsilon) \\ &= \alpha + \beta X. \end{aligned}$$

となる事に注意。

回帰分析、最小二乗法1

- 最小二乗法

先ほどの線形回帰モデルは2 つの未知パラメータとして α と β を含んでいた。

Y と X のデータとして

$$(\{Y_1, X_1\}, \{Y_2, X_2\}, \cdots , \{Y_n, X_n\})$$

が与えられた時に、これらのデータよりこの 2 つの未知パラメータを推定する方法として、最もよく使われる方法が**最小二乗法**である。

回帰分析、最小二乗法1

- 最小二乗法

これらの観測値は線形回帰モデルより得られたデータであるので、これらのデータには次の関係がある。

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

ここで誤差項 ε_i は確率変数 ε の実現値である。

誤差項 ε_i は直接は観測されない事に注意。

回帰分析、最小二乗法1

- 最小二乗法の定義

最小二乗法 とは以下の残差平方和 (Sum of Squared Residuals; SSR) と呼ばれるものを**最小にする a と b** を未知パラメーター **α と β の推定値**とする推定法の事である (a が α の推定値、b が β の推定値)。

(残差平方和)

$$SSR(a,b) = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

このような a と b を α と β の**最小二乗推定値(量)**とよぶ。

回帰分析、最小二乗法1

- 最小二乗法の意味

今、$a + b\,X_i$ を Y_i に**フィットさせる**事を考えよう。

フィットさせるとは、X_i が与えられた時に $a + b\,X_i$ で Y_i の値を**近似する**事だとする。

$a + b\,X_i$ が Y_i に**よくフィットしているか**を測るためには何らかの基準がいる。

どのような基準が考えられるだろうか？

回帰分析、最小二乗法1

- 最小二乗法の意味: データのフィット

1 つの基準として

$$u_i = Y_i - a - b\,X_i$$

の**絶対値 $|u_i|$ の大きさをみる**というものがある。
この u_i の事を**残差(residual)** という。

この場合、絶対値が**小さいほどフィットがよい**といえる。

しかしながら、絶対値の大きさは数学的に取り扱いが難しいので $|u_i|$ の代わりに u_i^2 を使う事を考える。u_i^2 も**小さいほどフィットがよい**といえる。

回帰分析、最小二乗法1

ここで u_i^2 というのは変数 Y_i と $a + b\,X_i$ という i 番目の組のみのフィットのよさを見る基準なので、$i = 1,\ldots, n$ という**全体のフィットのよさ**を見るために、u_i^2 $(i = 1,\ldots,n)$ を全て足して

$$SSR(a,b) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

という基準を考える。これは先ほどの残差平方和である。つまり残差平方和とは $a + b\,X_i$ が Y_i を全体的によく近似しているかどうかを見るための1つの指標であるということがわかる。

回帰分析、最小二乗法1

SSRの値が**小さいほど**、全体的に Y_i が $a + b\,X_i$でよくフィットできていると考えられる。

最小二乗法とはこの SSR を最小にするような a とb を α と β を推定値とする、つまり Y_i と $a + b\,X_i$ の**フィットが全体的に最もよくなる**ような a と b を推定値とするという事である。

回帰分析、最小二乗法1

- 最小二乗法による推定

最小二乗法によって推定してみよう。

$$SSR(a, b) = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

という目的関数を変数 a と b について最小化する。

最小化のための1階の条件は…

回帰分析、最小二乗法1

$$\frac{\partial SSR(a,b)}{\partial a} = 0 \Leftrightarrow -\sum_{i=1}^{n} 2(Y_i - a - bX_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} Y_i - an - b\sum_{i=1}^{n} X_i = 0$$

および

$$\frac{\partial SSR(a,b)}{\partial b} = 0 \Leftrightarrow -\sum_{i=1}^{n} 2X_i(Y_i - a - bX_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} X_i Y_i - a\sum_{i=1}^{n} X_i - b\sum_{i=1}^{n} X_i^2 = 0$$

となる。これら a と b に関する連立方程式を解くと…

回帰分析、最小二乗法1

$$b = \hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2},$$

および

$$a = \hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X}$$

となる。ここで、$\overline{X}$と$\overline{Y}$はそれぞれX_iとY_iの標本平均を表す。

回帰分析、最小二乗法1

- 直線のあてはめ

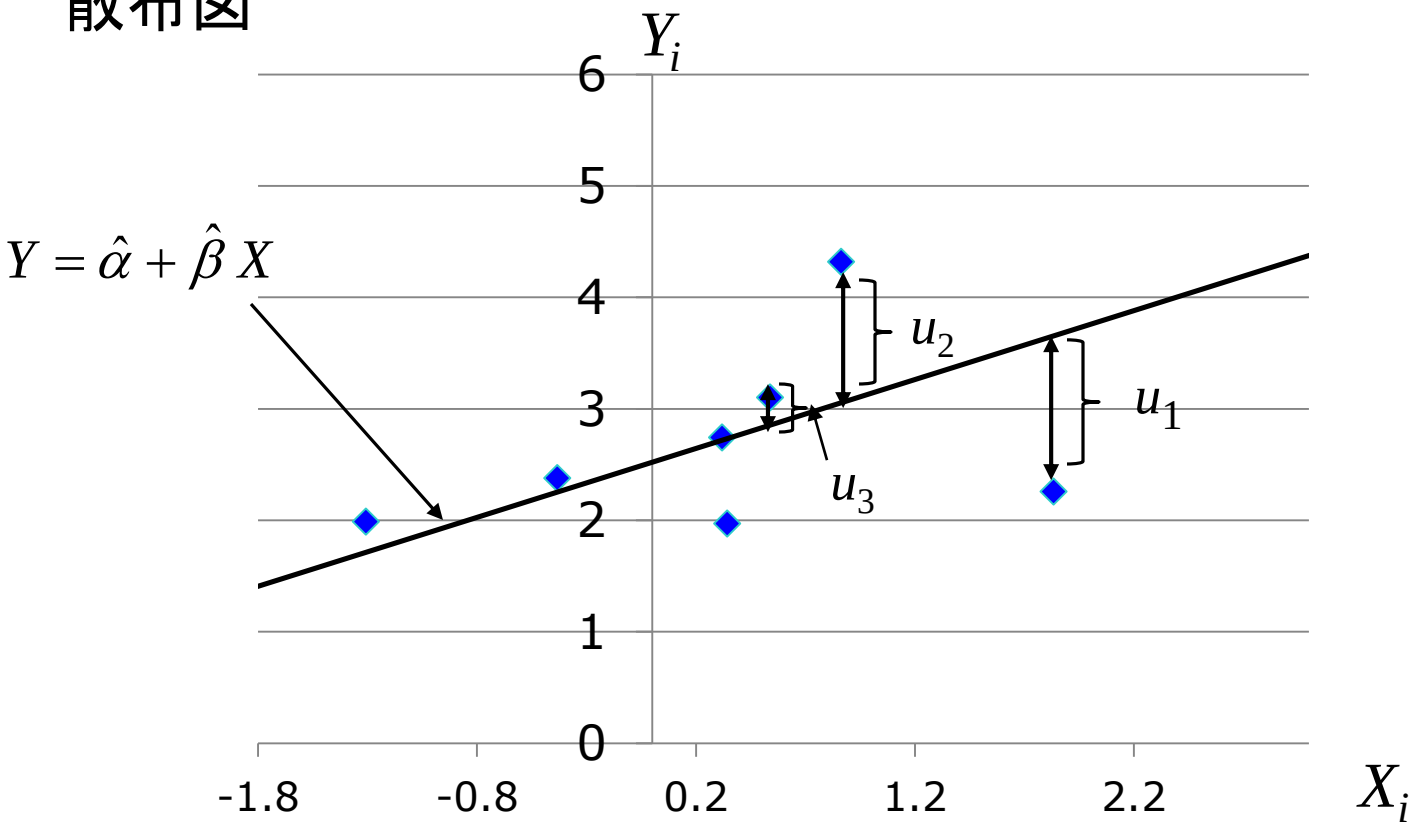
最小二乗推定法は$\{Y_i,\ X_i\}, i = 1,\ldots, n$ の散布図に良く当てはまる直線を引いているとイメージするとわかりやすい。

下の散布図を見ると、直線が異なると残差の大きさが異なる事がわかる。

どのような直線がもっとも当てはまりがよいかの1 つの客観的な基準が残差平方和である。

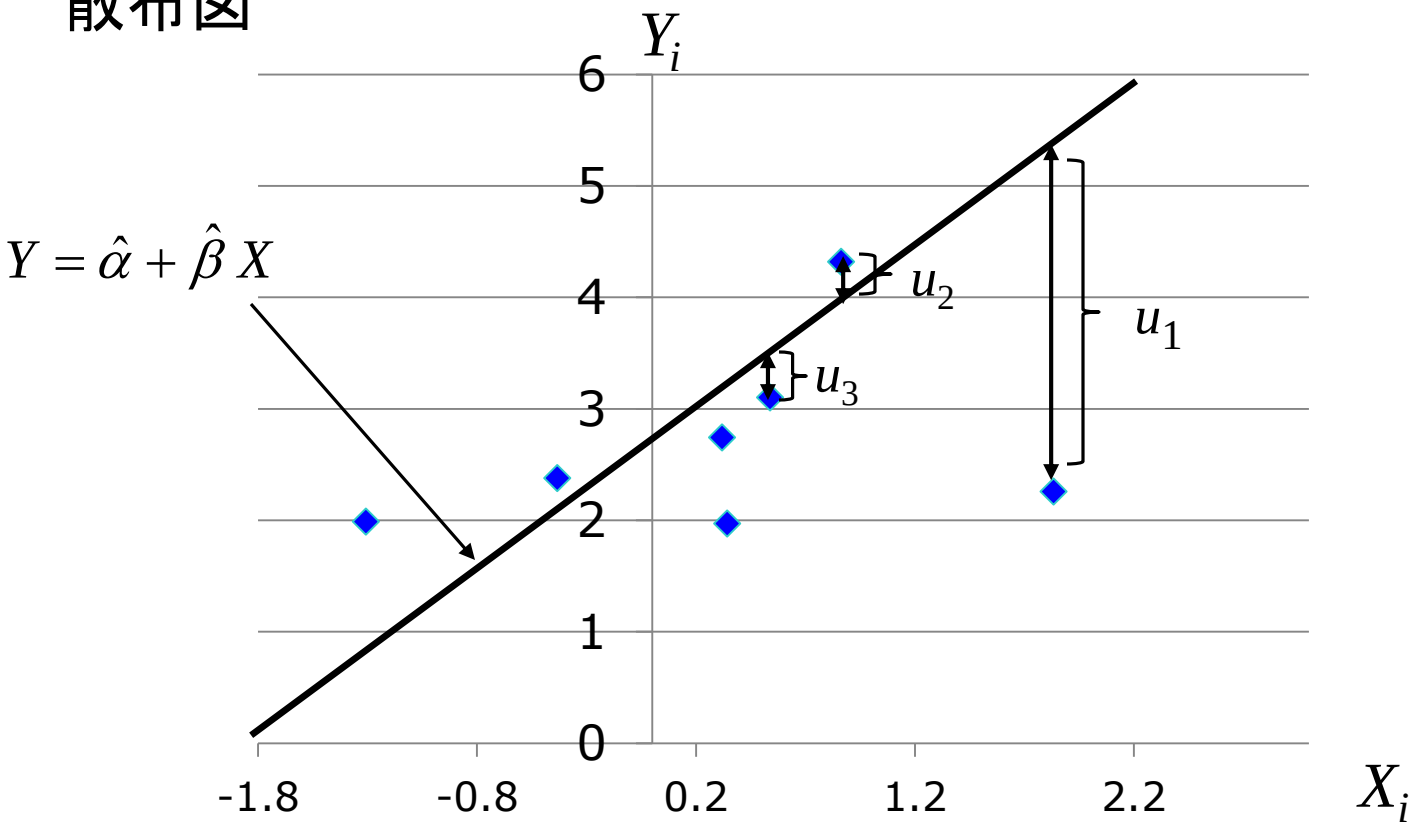
回帰分析、最小二乗法1

散布図



回帰分析、最小二乗法1

散布図



回帰分析、最小二乗法1

例題1

先ほどの OLS 推定量

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\,\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

は

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

と書き換えられる事を示しなさい。

回帰分析、最小二乗法1

- 決定係数

決定係数とは推定された回帰直線が実際のデータの変動をどの程度説明できているかを表す尺度である。

モデルが定数項(切片)をもっている場合、決定係数は以下のように定義される。決定係数はR^2 (アールスクエアとよむ) と表記される。

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

ここで $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ (**あてはめ値**と呼ばれる)である。

回帰分析、最小二乗法1

決定係数の分子は推定した回帰直線の平均からの変動、分母は実際のデータの平均からの変動を表している。

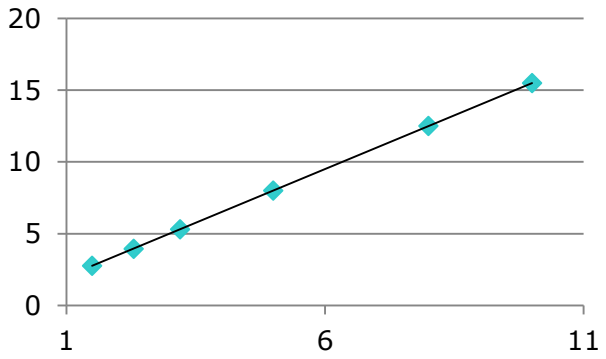
決定係数のとりうる範囲は0以上1以下である。

もし回帰直線がデータを完全に説明できている(すべてのデータが回帰直線上に乗っている) のであれば $\hat{Y}_i = Y_i$ となるので、分子と分母が完全に一致し、$R^2 = 1$ となる(このような事は実際のデータにおいてはまずないが)。

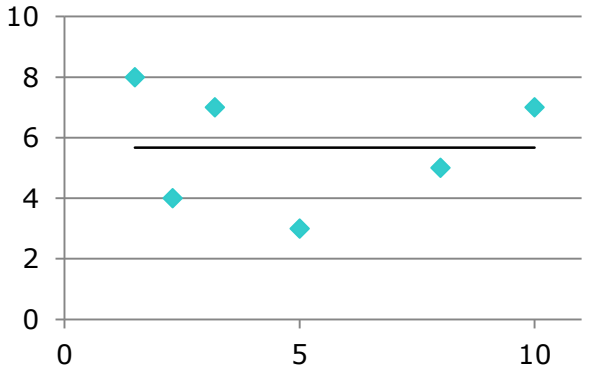
回帰分析、最小二乗法1

また、$R^2 = 0$ となるような状況とは(これは推定された回帰直線がY_i の変動を全く説明していないような状況) は分子が 0 になる様な時、すなわち、すべての i について $\hat{Y}_i = \bar{Y}$ となるような状況である。これは $\hat{\beta} = 0$ である時に起こる。

($R^2 = 1$)



($R^2 = 0$)



回帰分析、最小二乗法1

- 決定係数直観的な解釈

通常、決定係数は 0 と 1 の間の数値をとり、

決定係数が1に近い
→ 推定された回帰直線はデータの変動をよく説明
している。

決定係数が0に近い
→ 推定された回帰直線はデータの変動をあまりよく
説明していない。

と解釈される。

回帰分析、最小二乗法1

- 決定係数

決定係数は変数 X と Y の間の**線形関係**が強いほど高くなる。

別の言い方をすると、決定係数が低いという事は変数 X と Y の間に強い線形関係は存在しないという事である。

これは、変数 X と Y の間にはそもそも関係がない場合や、変数 X と Y の関係は線形ではなく**非線形**である場合のような状況である。

演習問題

問題 1

r_{xy} を変数 X_i と Y_i の間の相関係数としよう。この時、単回帰分析において、R^2 は r_{xy} の2乗、すなわち

$$R^2 = r_{xy}^{\,2}$$

に等しいことを示しなさい。

ヒント: OLS 推定量

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{と} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\,\bar{X},$$

をあてはめ値に代入して R^2 を計算しなさい。