統計解析論　第2回

最小二乗法による係数推定

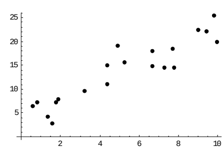
..¥homepage¥data¥consume.xls

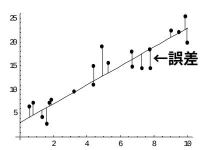
1.最小二乗法による線形回帰

- 線形回帰
  - データに<u>最も</u>フィットする直線を決める
  - 最もフィット？基準は？
- 最小二乗法
  - 基準の一つ
    - 回帰直線からのデータの隔たりの2乗の合計を最小にするという基準
  - 他の基準
    - はずれの絶対値の合計を最小化するなどさまざまな基準があるが，最小二乗法はもっともポピュラー

1．1　最小二乗基準

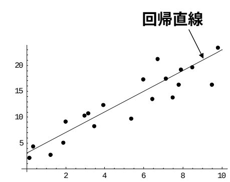
- 問題
  - 与えられたデータの組にもっとも当てはまる直線を求めよ
- 最小二乗基準
  - 「もっとも当てはまる」の意味？
  - その一つの答え
    - 誤差の二乗の合計が最小
    - 誤差とは右図の縦線の長さ





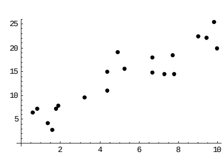
2．用語（1）

- 回帰直線
- 回帰式
  - 回帰直線を式で表したもの
  - 一般的には$Y = \alpha + \beta X$と表すことができる
  - 誤差を含めると $Y = \alpha + \beta X + 誤差項$
- 説明変数　X
  - $Y = \alpha + \beta X = \alpha \times 1 + \beta \times X$と解釈するとXと1が説明変数とも理解できる．
- 被説明変数　Y
  - X，Yは例えば所得，消費などの変数を表す



2．用語（2）

- 前のスライドは説明変数，被説明変数間の関係を扱った
- データの値を表す
  - データの値は小文字で表す
  - 何番目のデータかはその添え字で表す
  $(x_1, y_1), (x_2, y_2), \cdots$
  $y_i = \alpha + \beta x_i + 誤差項$
- データ数＝標本数　n



i		X	Y	
1	x_1	7.27	14.50	y_1
2	x_2	6.66	17.95	y_2
3	x_3	4.36	11.00	y_3
20	x_{20}	1.86	7.89	y_{20}

3．最小二乗法

- 最小二乗法
  - 最小二乗基準に従ってデータに最もフィットする回帰直線を求める．
- 問題の変形
  - 最小二乗基準→誤差の二乗の合計の最小化
  - 変化させるパラメータ＝係数→
  $Y = \alpha + \beta X + 誤差項$ のα，β
  - 誤差の二乗の和をα，βの式で表す
  $$誤差 = y_i - (\alpha + \beta x_i)$$

３．１　最小二乗法の解

- **変形された問題**
  - $\sum_{i=1}^{n} 誤差_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$を最小化するα，βを求める．
  - このようなα，βをデータの値だけの式で表す．
- **解答のための方針**
  - 最小化の二つの方向
    - 微分
    - 二次関数化

３．最小二乗法の解（2）

- **標本平均**
  - 例
    - Xのデータの標本平均 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
    - Yのデータの値の二乗の標本平均 $\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$
  - 性質

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} 1 = \sum_{i=1}^{n} x_i - n\bar{x}$$

$$= n\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}\right) = 0$$

$$\sum_{i=1}^{n} 定数(x_i - \bar{x}) = 定数\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

３．２　最小二乗法による係数推定

- 最小二乗法による係数値の決定法
  - 理論的に最小値を探す＝微分
  - 結果
    - 推定した回帰直線（式）

$$y = \hat{\alpha} + \hat{\beta} x$$

  - $\hat{\alpha}, \hat{\beta}$ で誤差の2乗の合計が最小になったとする

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

３．３　残差とその性質

- **残差**
  - 回帰直線からのデータの隔たり

$$\hat{u}_i = y_i - \left(\hat{\alpha} + \hat{\beta}\, x_i\right)$$

- **残差の性質**
  - 説明変数と残差の積の合計は0 $\quad \sum_{i=1}^{n} \hat{u}_i x_i = 0$
  - 残差の和は0 $\quad \sum_{i=1}^{n} \hat{u}_i = 0$

３．４　分散分析(1)

- RSS (Sum of Squared Residuals)　<**残差変動**>

$$RSS = \sum_{i=1}^{n} \hat{u}_i^2$$

- TSS (Total Sum of Squared Deviation) <**全変動**>

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- ESS　<**回帰変動**>
  (Explained Sum of Squared Deviation)

$$ESS = \sum_{i=1}^{n}\left\{(\hat{\alpha} + \hat{\beta} x_i) - (\hat{\alpha} + \hat{\beta} x_i)の標本平均\right\}^2$$

$$= \sum_{i=1}^{n} \hat{\beta}^2 (x_i - \bar{x})^2$$

３．５　分散分析と決定係数

- **分散分析**

$$TSS = RSS + ESS$$

- **重決定係数**
  - ESSのTSSに対する割合
  - yの変動のうちどれだけ回帰変動で説明できるか
  - 回帰式の説明力の指標

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad \left(R^2 \leq 1\right)$$

３．６　(重) 相関係数と重決定係数

- (重) 相関係数 (復習)

$$r=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}\qquad(-1\le r\le 1)$$

- 重決定係数との関係

$$R^2=r^2$$

  - 相関係数の絶対値は線形関係の当てはまりの強さを示す

４．　変数変換と回帰係数

- 変数変換

$$y_i=\alpha+\beta x_i+\text{誤差項}$$
$$z_i=cy_i+hx_i+o$$
$$w_i=px_i+q$$
$$z_i=\gamma+\delta w_i+\text{誤差項}$$

- 係数推定値

$$\hat{\gamma}=\frac{c\hat{\beta}+h}{p}\qquad \hat{\gamma}=c\hat{\alpha}+o-\frac{c\hat{\beta}+h}{p}q$$

係数の変換

- 係数の一次式を係数に持つ新たな回帰式をつくる
  - $Y=\alpha+\beta X$ を変形してＸの係数が$3\alpha+5\beta+1$となる回帰式を作る。
  - 方針
    - 左辺を変形 (βＸが残るように変形)
      $\alpha+\beta X=\alpha+(3\alpha+5\beta+1)(X/5)+(-3/5\alpha-1/5)X$
    - のこりの係数をαがある部分だけまとめる．
      $=\alpha-3/5\alpha X+(3\alpha+5\beta+1)(X/5)-1/5X$
    - α，βに関わらない変数を左辺＝被説明変数に移行
      $Y+1/5X=\alpha(1-3/5X)+(3\alpha+5\beta+1)(X/5)$

説明変数の変換

- 回帰式 $Y=\alpha+\beta X$ から説明変数が５Ｘ＋３である回帰式を求める．つまり，回帰式

  $Y=c+(\ \ )(5X+3)$

  の５Ｘ＋３の係数を求める．
- やりかた

  $Y=\alpha+\beta X=\alpha+(\beta/5)(5X+3)-(3/5\beta)$
  $=(\alpha-3/5\beta)+(\beta/5)(5X+3)$

説明変数の単位変換

- 説明変数の単位変換
  - １００億円単位のデータを１０兆円に換える
    - データの桁数が大きいので3桁分へらす
    - 説明変数を１／１０００倍する
- 数式で表現
  - $Y=\alpha+\beta X$ において $X\to cX$
    $Y=\alpha+(\beta/c)(cX)$
    なのでc倍された変数の回帰係数は1/c倍
- ３桁減らすとその回帰係数のみ３桁増える

被説明変数の変換

- 回帰式$Y=\alpha+\beta X$から被説明変数が-２Y+５Ｘ＋３である回帰式を求める．つまり，回帰式

  $-2Y+5X+3=(\ \ )+(\ \ )X$

  の()内の係数を求める
- やりかた

$Y=\alpha+\beta X$
$-2Y=(-2\alpha)+(-2\beta)X$
$-2Y+5X=(-2\alpha)+(-2\beta)X+5X=(-2\alpha)+(-2\beta+5)X$
$-2Y+5X+3=(-2\alpha)+3+(-2\beta+5)X=(-2\alpha+3)+(-2\beta+5)X$

被説明変数の単位変換

- **被説明変数の単位変換**
  - １００億円単位のデータを１０兆円に換える
    - データの桁数が大きいので3桁分へらす
    - 被説明変数を１／１０００倍する
- **数式で表現**
  - $Y = \alpha + \beta X +$ 誤差 において $Y \to cY$
    $$cY = c\alpha + (c\beta)X$$
    なので回帰係数はc倍される
- **3桁減らすと回帰係数は全て3桁減る**

説明変数，被説明変数の変換の組み合わせ

- **まず説明変数を変換し，被説明変数を変換する。**

被説明，説明変数の単位変換

- **被説明変数と説明変数の両方をc倍したら？**
  - 例
    - 両方を3桁減らす
    - C＝１／１０００
- **式変形**
  $$cY = c\alpha + (c\beta)X$$
  $$cY = c\alpha + (c\beta / c)(cX)$$
- **定数項はc倍，傾きは同じになる**
  - 定数項は3桁減る，傾きはそのまま

5．相関係数再訪

- ＹをＸに回帰する
  $$Y = \alpha + \beta X$$
- Y / S_Y を X / S_X に回帰する
  $$Y = \alpha + \beta X = \alpha + (\beta S_X)(X / S_X)$$
  $$Y / S_Y = (\alpha / S_Y) + (\beta S_X / S_Y)(X / S_X)$$
- **回帰係数を*<u>最小二乗法</u>*で推定すると相関係数が求まる**
  $$\hat{\beta} = S_{XY} / S_X^2$$
  $$\hat{\beta} S_X / S_Y = (S_{XY} / S_X^2) \cdot (S_X / S_Y) = S_{XY} / (S_X S_Y) = r$$