統計解析論　2

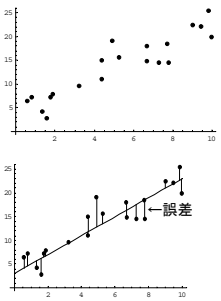
単回帰
第2章
..¥homepage¥data¥consume.xls

2－1　線形回帰式

- 線形回帰
  - データに**最も**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

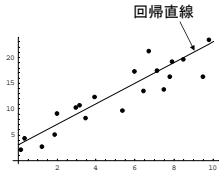
2－1－1　最小二乗基準

- 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



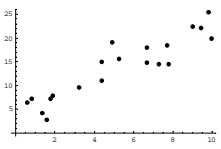
2－2－2　用語（1）

- 回帰直線
- 線形回帰式
  - 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



2－2－2　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＝標本サイズ　n



i	x_i	X	Y	y_i
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}

2－2．回帰式の推定

- 最小二乗法（OLS）
  - 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 $= y_i - (\alpha + \beta x_i)$

２－２－１　最小二乗法の解（1）

- 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$\alpha$，$\beta$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

7

２－２－１　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

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

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

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

8

２－２－２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

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}} = r_{xy}\frac{s_y}{s_x} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

9

２－２－３　残差

- 残差
  - 推定した回帰直線からのデータの隔たり

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

- 残差分散
  - 誤差項の分散の推定値（回帰直線とデータとの間の隔たりの大きさの2乗の平均）

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

10

２－３　決定係数
２－３－１　RSS, TSS, ESS

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

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

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

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

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

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

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

11

２－３－２　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

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

12

２－３－３（重）相関係数と重決定係数

- （重）相関係数（復習）

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

- 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

13

２－３－４　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***最小二乗法***で推定すると相関係数が求まる

$\hat{\beta}=S_{XY}/S_X^2$

$\hat{\beta}S_X/S_Y=(S_{XY}/S_X^2)\cdot(S_X/S_Y)=S_{XY}/(S_XS_Y)=r$

14

２－３－５　トレンド回帰

- 説明変数を時間変数にとる

$y_i=\alpha+\beta t+u_t$(誤差項)

- 例えばtとして西暦を用いる
- tの代わりに皇紀t^*を利用する場合

$y_i=\alpha+\beta t^*+u_t$(誤差項)

$t^*=t+660$

$\hat{\beta}$ はかわらない．$\hat{\alpha}$　はかわる．

15

２－３－６　変数変換と回帰係数

- 変数変換

$y_i=\alpha+\beta x_i+$誤差項

$z_i=cy_i+hx_i+o$

$w_i=px_i+q$

$z_i=\gamma+\delta w_i+$誤差項

- 係数推定値

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

16

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（β Xが残るように変形）

$\alpha+\beta X=\alpha+(3\alpha+5\beta+1)(X/5)+(-3\alpha/5-1/5)X$

    - のこりの係数をαがある部分だけまとめる．

$=\alpha-3\alpha X/5+(3\alpha+5\beta+1)(X/5)-1/5X$

    - α，βに関わらない変数を左辺＝被説明変数に移行

$Y+1/5X=\alpha(1-3X/5)+(3\alpha+5\beta+1)(X/5)$

17

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

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

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

18

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

19

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

$Y = \alpha + \beta X$

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

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

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

20

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

21

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。

22

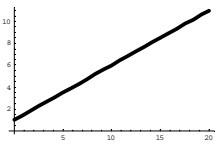
被説明，説明変数の単位変換

- 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

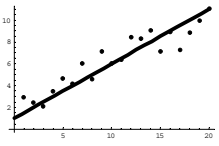
23

データの成り立ちについての想定

- データの成り立ち
  - 回帰直線を想定



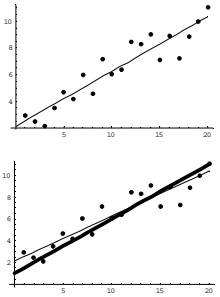
  - 回帰直線＋確率的に決まる誤差



24

推定誤差

- 最小二乗法による推定
- データの成り立ちの回帰直線と推定された回帰直線の比較



25

2－4　推定結果に関する検定

- 標準誤差（係数推定値の）
  - 係数推定値の変動の指標
    - 係数値の真の値からの隔たりの標準偏差を推定（＊）

$$\hat{\beta}: s.e.(\hat{\beta}) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad \hat{\alpha}: s.e.(\hat{\alpha}) = s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

    - 大きいほど最小二乗法による推定値は当てにならない．小さいほど信頼できる．

26

2－4－1　t統計量（1）

- t統計量
  - 理論，仮説が示す値を検証するためのツール
  - 理論や仮説で与えられている値を α_0, β_0 とする
  - β のt統計量 $t_\beta = \dfrac{\hat{\beta}-\beta_0}{s.e.(\hat{\beta})} = \dfrac{\hat{\beta}-\beta_0}{\dfrac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}$
    - この値は自由度 $n-2$ のt分布に従って変動する（＊）
      - これは標本数が大きいと標準正規分布（平均0，分散1）とほぼ同じ分布
      - この絶対値が大きい（目安としては1.96以上，検定で述べる）と理論，仮説ははずれている可能性が高いか，回帰式が不適切かのいずれかである

27

2－4－1　t統計量（2）

- α に関するt統計量（自由度n-2のt分布）

$$t_\alpha = \frac{\hat{\alpha}-\alpha_0}{s.e.(\hat{\alpha})} = \frac{\hat{\alpha}-\alpha_0}{S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}$$

- 通常出力されるのは，$\alpha_0 = 0, \beta_0 = 0$ としたときのt統計量
  - Excel出力を変換すれば任意の α，β に対してt統計量が計算できる

28

2－4－2　信頼区間（1）

- □%信頼区間
  - 本当の係数値 β_0 がその中に入っている確率を□%にするように設定した区間
  - □としては95%，99%の順でよく使われる
  - 「この区間に真の係数値が入る確率は95%だから，はずれは5%なので，この範囲に備えておけば大丈夫」といった考え方
    - これに似た考え方としてVaR（ヴァリュー　アット　リスク）がある

29

2－4－2　信頼区間（2）

- β のk%信頼区間

$$\left[\hat{\beta} - \text{自由度}n-2\text{のｔ分布の}(50+k/2)\%\text{点} \times s.e(\hat{\beta}),\right.$$
$$\left.\hat{\beta} + \text{自由度}n-2\text{のｔ分布の}(50+k/2)\%\text{点} \times s.e(\hat{\beta})\right]$$

  - Excelによる自由度n-2のt分布のm%点の計算の仕方

    =TINV((1-m/100)*2,標本サイズ-2)

- 有意水準k%の検定で棄却される＝（100－k）%信頼区間の外側に α_0, β_0 がある．

30

２－５ 最小二乗法の諸性質
２－５－１ 残差とその性質

- 残差の性質
  - 説明変数と残差の積の合計は0 $\sum_{i=1}^{n} \hat{u}_i x_i = 0$
  - 残差の和は0（定数項を含む場合） $\sum_{i=1}^{n} \hat{u}_i = 0$

31

２－５－２ 予測

- 予測値 $\hat{y}_{n+1} = \hat{\alpha} + \hat{\beta} x_{n+1}$
  - 過去のデータを用いて回帰係数を推定
  - 説明変数に将来値 x_{n+1} を代入
- 予測誤差の分散

$$V(\hat{y}_{n+1}) = s^2 \left\{ 1 + 1/n + (x_{n+1} - \bar{x})^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}$$

- k%予測信頼区間

$$\left[\hat{y}_{n+1} - \text{自由度} n-2 \text{の t 分布の} (50 + k/2)\text{%点} \times \sqrt{V(\hat{y}_{n+1})}, \right.$$
$$\left. \hat{y}_{n+1} + \text{自由度} n-2 \text{の t 分布の} (50 + k/2)\text{%点} \times \sqrt{V(\hat{y}_{n+1})} \right]$$

32