データ解析基礎

第2回 二つの変数のあいだの関係

本日の予定

- 二つの変数のあいだの関係
- 相関係数
- 線形回帰
- 線形回帰の良さ
- 誤差の分解

二つの変数の関係

3

- 例：入学試験と期末試験の成績の関係

case1：

Student	Entrance	Achievement
A	10	4
B	8	3.5
C	7	3
D	3	1.5
E	1	1

- 入試の成績から期末試験の成績を予測できるか？
- 二つの成績のあいだには関係があるか？

二つの変数の関係

4

case 2:

Student	Entrance	Achievement
A	10	1.0
B	8	1.5
C	7	3.0
D	3	3.5
E	1	4.0

case 3:

Student	Entrance	Achievement
A	10	1.0
B	8	4.0
C	7	3.5
D	3	3.0
E	1	1.5

二つの変数のあいだの系統的な関係

5

- **正の線形関係** (positive systematic linear relationship)：

  X **が大きくなれば** Y **も大きくなり，** X **が小さくなれば** Y **も小さくなるような関係．**

- **負の線形関係** (negative systematic linear relationship)

  X **が大きくなれば** Y **が小さくなり，** X **が小さくなれば** Y **は大きくなるような関係．**

- **線形関係が弱い，ない．**

  X **が大きくなったり，小さくなったりするのにかかわらず，** Y **が大きくなったり小さくなったりする．**

散布図を用いたデータの記述

6

- **散布図** (scatter plot)
  - **データの組を**2**次元のグラフ上に表すことで，両者の関係を視覚的に捉える．**

- **例題**
  - **前の三つのケースについて，それぞれ散布図をかいてみる．**

二つの変数の関係を表す指標

7

- 散布図により変数間の関係は視覚的に把握できる．
- データ数が少ない場合は，変数間の関係をデータから直接発見できる．
  - データ数が多いときはより系統的な方法が必要．
- 変数の関係を「定量的に」評価することも重要．

→　二つのデータの関係を示す指数を使う．

相関係数の考え方

8

- データの相関関係

  データ X_i が平均値 $\overline{X}$ からどれだけ離れているか？

  データ Y_i が平均値 $\overline{Y}$ からどれだけ離れているか？
- 二つの「離れ具合い」の符号をみると
  - 同符号ならば正の関係，逆符号ならば負の関係．
- 二つの「離れ具合い」の積の平均値をとる．

$$S_{XY} = \frac{1}{N} \Sigma_i (X_i - \overline{X}) (Y_i - \overline{Y})$$　　**共分散** (covariance)

相関係数の考え方

9

- Pearsonの積率相関係数（correlation coefficient)
  - 共分散のままでは，ばらつきの大きさに依存して係数が変動してしまう．
  - そこで，X，Y の標準偏差を用いて標準化する．

$$r_{XY} = \frac{1}{N} \Sigma_i \frac{(X_i - \overline{X})}{S_X} \frac{(Y_i - \overline{Y})}{S_Y}$$

$$= \frac{S_{XY}}{S_X S_Y}$$

  - 完全な線形関係では＋１あるいは-１になる．

相関係数の計算の簡略化

10

- 計算上の便法：
  - データから相関係数を直接求めるときは，共分散や標準偏差を経由する必要はなく，次式によって計算の手間を減らすことができる．

$$r_{XY} = \frac{N \Sigma_i X_i Y_i - (\Sigma_i X_i)(\Sigma_i Y_i)}{\sqrt{N(\Sigma_i X_i^2) - (\Sigma_i X_i)^2} \quad \sqrt{N(\Sigma_i Y_i^2) - (\Sigma_i Y_i)^2}}$$

相関係数の計算の簡略化

11

- 前ページの式の導出
  →演習問題
- 例題
  最初のケースについて計算してみる．
- 注意：
  - この式が「公式」として書かれている本もあるが，式を憶える必要はない．定義から随時導けばよい．
  - 分散計算での数値計算の丸め誤差には要注意．

相関係数の例

12

- 相関係数を実際に計算してみる．
  - MS Excel には，相関係数を計算する関数(Pearson)がある．
- 「相関係数が0である」と「無関係」であるとは違う．
  - 相関係数が0であっても強い関係がある場合はある．
  - 相関係数はあくまで1次の（つまり直線で表せる）関係だけを表している．

相関係数の幾何学的解釈

13

(数式の定義ではすっきりしない人のために．．．)

- XとYのデータをそれぞれ棒グラフで書いてみる．
  - 平均値分だけ引いて，原点に対して対称形にする．
  - できた波形をそれぞれ一つのベクトルとみなす．
  - 二つのベクトルのなす角度を計算する．

$$\cos\theta = \frac{\boldsymbol{x}\,\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} = \frac{x_1y_1 + x_2y_2 + \ldots + x_Ny_N}{\sqrt{x_1x_1 + \ldots + x_Nx_N}\ \sqrt{y_1y_1 + \ldots + y_Ny_N}}$$

→　この量が相関係数と一致する．

相関係数の解釈の仕方

14

- 相関係数の大きさだけで「関係の強弱」を論じるのは危険である．

  そのデータがもつ本来の関係（すなわち文脈）に応じて相対的に論じるべきである．

- 相関係数が大きくても，二つの変数間に直接的な関係があるとは限らない．

  二つの変数に同時に影響を与える第三の変数があるかもしれない．データ間の因果関係を考える重要性．

- データにノイズが多ければ，相関係数は低下する．

  データのノイズの大きさで相関係数の意味が変わる．

相関係数の特殊なケース

15

- **ファイ係数：**
  - **二つの変数が共に2値しかとらない場合．**
- point bi-serial correlation coefficient：
  - **一方の変数が2値で，他方が間隔尺度の場合**
- Spearman rank correlation coefficient
  - **二つの変数が「順位」として与えられている場合**

線形関係にあるデータにおける予測

16

- **散布図や相関係数**
  - **与えられたデータの性質を分析する．**
- **線形関係であることがわかった場合**
  - **その関係を用いて未知のケースについて予測する．**
    X からY，あるいは Y から X を求める「公式」を得る．
  - **データから得られた「法則性」を利用する．**
  - **データを説明する「線形モデル」を作る．**

どのような直線を求めればよいか？

17

- 回帰直線の方程式

$$Y = a + b\,X$$

- 「誤差を最小にする」という考え方

  直線から得られる予測値 $\tilde{Y}_i$

  データの値 Y_i

  これらの差ができるだけ小さい直線が望ましい．

  すべてのデータについての影響を考慮する．

最小2乗法 (least squared error)

18

- 誤差の2乗の和を最小化する（最小2乗法)

$$E = \Sigma_i \, (Y_i - \tilde{Y}_i)^2$$

  を評価関数とし，これを最小にするa, bを求める．

- 平均2乗誤差 (mean squared error)の最小化と考えても同じことである．

$$D^2 = \frac{1}{N} \Sigma_i \, (Y_i - \tilde{Y}_i)^2$$

最適な線形方程式

19

- 2乗誤差を最小化するパラメータの決定法
  - 微分を用いて求める方法
  - 2次式の和への展開による方法
  - 幾何学的な考え方による方法（一般化逆行列）
- 最終的に得られる答え

$$b = \frac{S_{XY}}{S_X^2} = \frac{N \Sigma_i X_i Y_i - (\Sigma_i X_i)(\Sigma_i Y_i)}{N(\Sigma_i X_i^2) - (\Sigma_i X_i)^2}$$

$$a = \overline{Y} - b\overline{X}$$

線形回帰のよさを示す指標

20

- 平均2乗誤差と相関係数の関係

$$D^2 = S_Y^2 (1 - r_{XY}^2) = S_Y^2 - \frac{S_{XY}^2}{S_X^2} = S_Y^2 - \tilde{S}_Y^2$$

平均2乗誤差は，Y の分散から $\tilde{Y}$ の分散（つまり回帰で説明できる部分の分散）を引いたものに等しい．

- 決定係数（0から1の値をとる）

変数 Y の分散のうち回帰で説明できる部分の割合

$$R^2 = \frac{\tilde{S}_Y^2}{S_Y^2} \quad (= r_{XY}^2 \text{ ：1次の場合は相関係数の2乗に等しい})$$

誤差の分解

- 「誤差」の観点からもう一度見直してみよう．

$$S_Y^2 = \frac{S_{XY}^2}{S_X^2} + D^2 = \tilde{S}_Y^2 + D^2$$

- 両辺を N 倍すると

（左辺）＝（データYの全変動）

（右辺第1項）＝（回帰で説明される変動の平方和）

（右辺第2項）＝（回帰で説明できない誤差平方和）

- 個々のデータに対する誤差の関係と対比せよ．

線形回帰分析の幾何学的解釈 1

- XとYのデータを並べそれぞれ棒グラフで書いてみる．
  - 平均値分だけ引いて，原点に対して対称形にする．
  - できた波形をそれぞれ一つのベクトルとみなす．
    これらをそれぞれ $\boldsymbol{X}$, $\boldsymbol{Y}$ と表すことにする．
- このベクトルを用いると
  - Xの分散：　$S_X^2 = \| \boldsymbol{X} \|^2 / n$
  - Yの分散：　$S_Y^2 = \| \boldsymbol{Y} \|^2 / n$
  - XとYの共分散：$S_{XY} = (\boldsymbol{X} \cdot \boldsymbol{Y}) / n$

線形回帰分析の幾何学的解釈 2

23

- ベクトルYをベクトルXの張る直線に正射影してみる

$$Y' = [\, Y \cdot (X \,/\, \| X \|) \,] \; (X \,/\, \| X \|)$$

(かっこの中は単位ベクトル)

- この式を変形すると

$$Y' = Y \cdot (X \,/\, \| X \|) \; (X \,/\, \| X \|)$$

$$= [\, (X \cdot Y) \,/\, \| X \|^2 \,] \; X$$

$$= \frac{S_{XY}}{S_X{}^2} X = b X$$

線形回帰分析の幾何学的解釈 3

24

- Yの予測値は，YからXの直線に対して下ろした垂線の足に対応している．
- 平均2乗誤差は，垂線の長さに対応している．
  - 垂線が短いほど，YとY'が近いことになる．
- 誤差の分解式は，ピタゴラスの定理に対応している．

もう一度分布をみてみよう

25

- 線形関係にあることが大前提である.
  - この関係が成立しているかどうかを目で確認するのが大事.
  - 異常値を検出することができる.

より高度な話題

26

- 回帰分析の結果に対する評価
  - 相互関係に関する検定
- 重回帰分析
  - 説明変数が複数ある場合
- AIC（赤池情報量）
  - 説明変数をどうやって選択すればよいか？
  - 「経験誤差」と「予測誤差」