データ解析基礎

第１０回 回帰分析（その２）

回帰分析の復習

- 二つの変数 X とY の関係を調べる．
- 相関係数 （ $r = S_{XY} / S_X S_Y$ ）
  - X とY のあいだの1次の関係を表す指標
  - 1と-1のあいだの値をとり，0のとき無相関．
- 回帰直線 （ $Y = a + b X$ ）
  - データ Yと X から予測されるY の平均2乗誤差を最小にする直線

$$b = \frac{S_{XY}}{S_X^2} = \frac{N \Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{N(\Sigma X_i^2) - (\Sigma X_i)^2} \qquad a = \overline{Y} - b\overline{X}$$

相関係数が0であるかどうかの検定

- 例題
  - 数学と化学の成績に相関があるか否かを知りたい．
- 設定する仮説
  - 帰無仮説 H0：$\rho = 0$
  - 対立仮説 H1：$\rho \neq 0$

    ρ: 真の相関係数

    r : データの相関係数

相関係数が0であるかどうかの検定

- 帰無仮説の下で標本（つまりデータ）の相関係数 r がどのような分布に従うかがわかれば，いままでと同じように棄却域が設定できる．
- 天下り的であるが，帰無仮説の下で，次の量がt 分布に従うことが知られている．

$$t = \frac{r}{\mathrm{sqrt}((1 - r^2) / (N - 2))}$$

（自由度は$N - 2$）

  - あとは，いままで同じやり方．

例題：

- 66人の学生について調査したところ，化学と数学の成績の相関係数は0.60であった．相関の有無を有意水準5%で検定せよ．
- step1：
  - 前ページの式に値を代入すると，検定統計量は

$$t = 0.60 / \mathrm{sqrt}((1 - 0.60^2) / (66 - 2))$$

$$= 0.60 / \mathrm{sqrt}(0.64 / 64) = 6$$

例題の解答

- step 2:
  - 有意水準5%の場合，自由度64の t 分布の限界点はおよそ±2．
- step 3:
  - 上で求めた統計量は棄却域に含まれるので，帰無仮説は棄却される．
  - つまり，「相関がある」という結論が得られる．

相関係数がcに等しいかどうかの検定

- 帰無仮説が $r = c$ となる検定はどうやればよいか？
- FisherのZ変換と呼ばれる方法がある．
  - 詳細は省略．おおまかなステップは以下の通り．
    1. Z変換の表を使って，r から Zr に変換する．
    2. 同様にして，c をZc に変換する．
    3. 次の値が標準正規分布に従うことを使って検定，推定を行なう．

$$z = \frac{Zr - Zc}{1 / \mathrm{sqrt}(N - 3)}$$

二つの母集団における相関係数の比較

- 関連のない標本の場合：
  -数学と化学の成績の相関係数は男女の間で違うか？
- 次の量が標準正規分布に従うことを使って検定する．

$$z = (Zr_1 - Zr_2) / \mathrm{sqrt}\left(\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3} \right)$$

- Zr_1, Zr_2は，相関係数 r_1 とr_2 にZ変換を施したもの．
- N_1, N_2は，r_1 とr_2を計算するのに用いたデータ数．

相関係数の比較

- 関連のある標本の場合
  - 「数学と物理の成績の相関」と「数学と化学の成績の相関」が等しいかどうかを共通の学生に対して調べる．
- 三つの変数 X, Y, Zについて「XとYの相関係数ρ_{XY}」と「XとZの相関係数ρ_{XZ}」が等しいかどうかは，次の量が自由度 N - 3 の t 分布に従うことを使って検定する．

$$t = \frac{(r_{XY} - r_{XZ}) \, \mathrm{sqrt}((N - 3) (1 + r_{YZ}))}{\mathrm{sqrt}(2 (1 + 2 r_{YZ} r_{XY} r_{XZ} - r_{YZ}^2 - r_{XY}^2 - r_{XZ}^2))}$$

例題：

- 「数学と物理」「数学と化学」の成績の相関係数を比較したい．100人の学生について三科目の成績の相関係数を求めたところ，数学と物理は0.5, 数学と化学は0.25，物理と化学は0.1であった．
- 以下の仮説について有意水準5%の検定を行なえ．
  - 帰無仮説：二つの相関係数は等しい．
  - 対立仮説：「数学と物理」の方が大きい．
- 略解：
  - 検定統計量は2.18．限界点は約1.66．よって棄却．

2変数の確率分布

- 二つの変数の組の確率分布
  - 2次元の空間の上に確率密度関数 $p(x, y)$ を描く．
    xyが作る平面上に高さをもったグラフをかく．
    したがって，体積が確率を表すことになる．

2次元正規分布

- 2次元正規分布の確率密度関数は次式で与えられる．

$$p(\boldsymbol{x}) = \frac{1}{2\pi\,|D|^{1/2}} \exp\left(-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} D^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- 2次元正規分布の性質
  - X, Yの分布（周辺分布）はそれぞれ正規分布．
  - Xを固定したときのYの条件付き分布が正規分布．
  - Yを固定したときのXの条件付き分布が正規分布．
  - 条件付き分布の分散は X (or Y)の値によらず一定．
  - Y (or X）の条件付き分布の平均値は直線上に並ぶ．

母集団の線形回帰式

- 第2回の講義での回帰分析はデータに基づいていた．
- データに関する回帰直線の方程式 $Y = a + bX$ は

$$\hat{Y} = b\,(X - \bar{X}) + \bar{Y}$$

と書き直せる．

ただし，$b = \dfrac{S_{XY}}{S_X^2}$

- これを2次元正規分布の立場から見直してみる．

母集団の線形回帰式

- 前ページの回帰直線に対応する母集団の回帰直線は，

$$Y = B\,(X - \mu_X) + \mu_Y$$

となる．

- Bは母集団に対する相関係数×SD比($\rho_{XY}\ \dfrac{\sigma_Y}{\sigma_X}$)
- μ_X, μ_Y は母集団 X, Yの平均．

母集団の線形回帰

- 前ページの二つの式を比較したとき，
  - $\bar{X}$ と $\bar{Y}$ は，それぞれμ_X，μ_Yの不偏推定量，
  - b は，B の不偏推定量，

    という関係にある．
- したがって，データから求めた回帰式は「母集団の回帰式の不偏推定量」である，といえる．

回帰直線の傾きが0であるかどうかの検定

- B が0であれば，Y の予測値はX によらずμ_Y になる．
  - それは X とY が無相関であることを示す．
  - 実際，$B = \rho_{XY}\,\sigma_y / \sigma_X$ より，$B = 0$ と$\rho_{XY} = 0$ は同値．
- したがって，$B = 0$ を検定するには $\rho_{XY} = 0$ を検定すればよい．
  - 相関係数が0であるかどうかの検定は既出．

予測誤差の不偏推定量

- データに関する平均2乗予測誤差の定義（復習）

$$D = \mathrm{sqrt}(\Sigma (Yi - \hat{Yi})^2 / N \quad)$$

（「データと予測値の差の2乗」の平均の平方根）

- では，母集団の予測誤差の不偏推定量はどうなるか？
  - X given という条件でのY の2乗誤差と解釈できる．「X が一定時のY のばらつき」にほかならない．
  - 平均2乗誤差は「Xが与えられた条件の下での Yの標準偏差」であるといえる．

条件付き分布の標準偏差

- (X, Y) が2次元正規分布に従うという仮定の下では，Yの条件付き標準偏差は Xの値に依存しない．
- その値を$\sigma_{Y|X}$とかけば，それは母集団全体に共通する予測誤差として使える．
- 誤差の標準偏差の不偏推定量は

$$\sigma_{Y|X} = \mathrm{sqrt}(\frac{1}{N - 2} \Sigma (Yi - \hat{Yi})^2 \quad)$$

で与えられる．

  - この値を使うとデータの信頼区間を求められる．

例題

- 2次元正規分布に従っている母集団から，ランダムに15対のデータを取得したところ以下の値が得られた．

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	2	3	3	5	5	5	6	6	6	7	7	7	7	8	10
Y	3	3	4	7	3	5	7	6	9	5	7	8	10	9	8

1. データの線形回帰式を求めよ．
2. 係数Bが0であるかどうかを有意水準5%で検定せよ．
3. $\sigma_{Y|X}$を推定せよ．
4. X=6に対するYの予測値を求めよ．
5. 4.で求めたYの予測値の上下2の範囲に含まれる近似的な確率を求めよ．

例題の解答 1.

- X, Yの平均，分散，共分散を求めると，

$$\overline{X} = 5.8,\ \overline{Y} = 6.27,\ S_X^2 = 4.03,\ S_Y^2 = 5.13,\ S_{XY} = 3.39$$

- したがって，
  - 傾き $b = 3.39 / 4.03 = 0.841$，
  - 切片 $a = 6.27 - 0.841 * 5.8 = 1.39$

  となる．

例題の解答 2.

- 仮説 $B = 0$を検定するには，相関係数について
  - 帰無仮説：$\rho = 0$
  - 対立仮説：$\rho \neq 0$

  を検定すればよい．
- 相関係数は，

$$r = S_{XY} / S_X S_Y = 3.39 / \mathrm{sqrt}(4.03 \times 5.13) = 0.745$$

例題の解答 2.（続き）

- 帰無仮説の下で，以下の統計量が自由度13 (=15 - 2) の t 分布に従うことから，

$$t = r / \mathrm{sqrt}((1 - r^2) / (N - 2))$$

$$= 0.745 / \mathrm{sqrt}(0.445 / 13) = 4.03$$

- 自由度13の t 分布の数表から5%の限界点は2.16．
  - したがって，帰無仮説は棄却される．
- すなわち，傾きBは0ではないと結論される．

例題の解答 3.

- データのXのそれぞれに対して回帰式をあてはめてYの予測値を求め，予測誤差を計算すると，以下の表が得られる．

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	2	3	3	5	5	5	6	6	6	7	7	7	7	8	10
Y	3	3	4	7	3	5	7	6	9	5	7	8	10	9	8
$\underline{Y}$	3.07	3.91	3.91	5.59	5.59	5.59	6.43	6.43	6.43	7.28	7.28	7.28	7.28	8.12	9.80
Error	-0.07	-0.91	0.09	1.41	-2.59	-0.59	0.57	-0.43	2.57	-2.28	-0.28	0.72	2.72	0.88	-1.80

- これより，予測誤差の不偏推定量は

$$\sigma_{Y|X} = \mathrm{sqrt}(\text{（誤差の2乗和）} / (N - 2))$$

$$= \mathrm{sqrt}(34.21 / 13) = 1.62$$

例題の解答 4. および 5.

- 上の表より $X = 6$ に対するYの予測値は 6.43．
- 確率を計算するため，もとのデータを標準正規分布の空間に変換する必要がある．
- 先に求めた標準偏差を用いて標準化すると，

$$z = (2 - 0) / 1.62 = 1.23$$

- 1.23より上側の部分の累積確率を標準正規分布を数表から読みとると，0.10935 である．

例題の解答（続き）

- したがって，求める確率は

  P = 1 - 0.10935 x 2 = 0.7813.

- 以上より，
  - $X = 6$のときには，約78%の確率で Y の値は，6.43 ± 2 の区間に含まれることになる．
  - ここでの議論は，厳密な信頼区間ではなく，あくまで近似的に計算しているにすぎないことに注意．

重回帰分析

- 重回帰分析
  - 説明変数が二つ以上ある場合の回帰分析のこと
  - 複数の変数を$X_1, \ldots, X_m$ を使って，別の変数Yを以下のようにして予測する．

    $$Y = b_0 + b_1 X_1 + \ \ldots + b_m X_m$$

- 以下で議論する問題
  - 係数 b_i をどのようにして推定するか？
  - 決定係数はどのようにして求めるか？
  - 予測誤差の分散不偏推定量はどうやって求めるか？

係数の決定方法

- 基本的考え方は単純回帰の場合と同じ．
  - データと予測値の2乗誤差を最小化する．
    + 微係数が0になる点を探す
    + 幾何学的に考える．
- 幾何学的な方法
  - n 個のデータについて，それぞれ回帰式を表すと，

$$Y_1 = b_0 + b_1 X_{11} + \ \ldots + b_m X_{1m} + e_1$$
$$Y_2 = b_0 + b_1 X_{21} + \ \ldots + b_m X_{2m} + e_2$$
$$Y_n = b_0 + b_1 X_{n1} + \ \ldots + b_m X_{nm} + e_n$$

係数の決定方法

- これを行列とベクトルを使って書き直すと

$$\boldsymbol{y} = \mathrm{X}\,\boldsymbol{b} + \boldsymbol{e}$$

となる．ここで，$|\boldsymbol{e}|^2$ を最小化すればよい．

- 射影変換の知識を使うと，$|\boldsymbol{e}|^2$ を最小化する $\boldsymbol{b}$ の値は，

$$\boldsymbol{b} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \boldsymbol{y}$$

として求められる．

決定係数と誤差の分解

- 決定係数 R^2：
  - Yの変動のうち，$X_1, \ldots, X_m$の変動で説明できる部分の割合

$$R^2 = \frac{\Sigma (\hat{Y}i - \overline{\hat{Y}i})^2}{\Sigma (Yi - \overline{Y})^2} = 1 - \frac{\Sigma (Yi - \hat{Y}i)^2}{\Sigma (Yi - \overline{Y})^2}$$

- この式からわかるように，Yの変動の平方和は

  （説明される変動の平方和）＋（誤差の平方和）

  という形に分解できる（ピタゴラスの定理）.

予測誤差の不偏推定量

- 予測誤差の不偏推定量は次のように与えられる.

$$\sigma_{Y|X} = \mathrm{sqrt}\left(\frac{1}{N - p} \Sigma (\hat{Y}i - Yi)^2 \right)$$

  p は係数 bi の数（説明変数の数 + 1)である.

- 例：説明変数が一つの場合は，
  - 定数項 a と傾き b の二つのパラメータがあるので，

$$\sigma_{Y|X} = \mathrm{sqrt}\left(\frac{1}{N - 2} \Sigma (\hat{Y}i - Yi)^2 \right)$$ （復習）

例題：

- システムの入出力関係の同定

  二つの入力を与えると一つの出力を出すシステムを考える．いま，8組の入力を与えたときの出力を調べたところ，次の結果が得られた．入力から出力を推定する回帰式を求めよ．

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

#	1	2	3	4	5	6	7	8
X1	2	4	6	3	5	1	4	5
X2	1	5	3	2	4	2	3	1
Y	1.54	-6.68	3.14	0.08	-2.01	-4.33	-1.28	7.04

例題の解答

- ベクトルと行列の設定

$$\boldsymbol{y} = \begin{bmatrix} 1.54 \\ -6.68 \\ 3.14 \\ 0.08 \\ -2.01 \\ -4.33 \\ -1.28 \\ 7.04 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 4 & 5 \\ 1 & 6 & 3 \\ 1 & 3 & 2 \\ 1 & 5 & 4 \\ 1 & 1 & 2 \\ 1 & 4 & 3 \\ 1 & 5 & 1 \end{bmatrix}$$

例題の解答

- 行列の計算

$$X^{\mathrm{T}}X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 3 & 5 & 1 & 4 & 5 \\ 1 & 5 & 3 & 2 & 4 & 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 4 & 5 \\ 1 & 6 & 3 \\ 1 & 3 & 2 \\ 1 & 5 & 4 \\ 1 & 1 & 2 \\ 1 & 4 & 3 \\ 1 & 5 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 8 & 30 & 21 \\ 30 & 132 & 85 \\ 21 & 85 & 69 \end{bmatrix}$$

例題の解答

- 行列の計算

$$(X^{\mathrm{T}}X)^{-1} = \begin{bmatrix} 8 & 30 & 21 \\ 30 & 132 & 85 \\ 21 & 85 & 69 \end{bmatrix}^{-1} = \begin{bmatrix} 1.0167 & -0.1539 & -0.1199 \\ -0.1539 & 0.0599 & -0.0270 \\ -0.1199 & -0.0270 & 0.0842 \end{bmatrix}$$

$$X^{\mathrm{T}}\boldsymbol{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 3 & 5 & 1 & 4 & 5 \\ 1 & 5 & 3 & 2 & 4 & 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1.54 \\ -6.68 \\ 3.14 \\ 0.08 \\ -2.01 \\ -4.33 \\ -1.28 \\ 7.04 \end{bmatrix} = \begin{bmatrix} -2.50 \\ 11.14 \\ -35.78 \end{bmatrix}$$

例題の解答

・以上より，求めるべき係数ベクトルは

$$\boldsymbol{b} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \boldsymbol{y} = \begin{bmatrix} 1.0167 & -0.1539 & -0.1199 \\ -0.1539 & 0.0599 & -0.0270 \\ -0.1199 & -0.0270 & 0.0842 \end{bmatrix} \begin{bmatrix} -2.50 \\ 11.14 \\ -35.78 \end{bmatrix}$$

$$= \begin{bmatrix} 0.0328 \\ 2.0184 \\ -3.0149 \end{bmatrix}$$

つまり，$b_0 = 0.0328, \quad b_1 = 2.0184, \quad b_2 = -3.0149$

という結果が得られた．

誤差の評価

・データと誤差を比較してみると，

#	1	2	3	4	5	6	7	8
Y	1.54	-6.68	3.14	0.08	-2.01	-4.33	-1.28	7.04
$\underline{Y}$	1.05	-6.97	3.1	0.06	-1.93	-3.98	-0.94	7.11
Error	0.49	0.29	0.04	0.02	-0.08	-0.35	-0.34	-0.07

- これより，平均２乗誤差は　$D = 0.267$

・データYの分散は 16.258であるから，決定係数は，

$R^2 = 1 - 0.267^2 / 16.258 = 0.996$

・予測データの分散16.187を使っても同じ結果になる．

$R^2 = 16.187 / 16.258 = 0.996$

重回帰分析の難しさ

- 現実の問題では教科書どおりにうまく処理ができないことが多い．
- しばしば生じる問題として以下のものが挙げられる．
  - 理論との乖離
    1. 誤差が無相関でない．
    2. 誤差の分散が一定でない．
  - 手続き上の問題
    1. 異常値をどのようにして取り除くか？
    2. 説明変数をどのようにして選ぶか？

説明変数の選び方

- 説明変数の組み合わせ
  - 説明変数の候補が10個あれば，説明変数の選び方は全部で1023通りある．
  - これらすべてを比較するのは非効率である．
- 先験的な情報による絞り込み
  - 確実に関係している変数を選び，選択肢を減らす．
  - 例えば，上の場合で，変数3個を事前に固定すれば，選び方は127通りに減る．

説明変数の選び方

- 変数の順序付けの利用
  - 変数に順序がついている場合は，少ない数から初めて順次増やし，一定の条件で増加を停止させる．
- 「説明力」による変数の追加
  - 説明力が大きな変数を順次追加していく．
  - 説明力の小さな変数を減らしていく方法もある．

説明変数の選び方

- 予備検定の利用
  - 係数 $Bi = 0$ という帰無仮説を検定し，それが棄却されれば説明変数として採用する．
- 基準量に基づくモデルの選択
  - 一定の評価関数を設け，その値を最大化（最小化）するモデルを選ぶ．

基準量の基づく説明変数の選択

- 基準量として決定係数を用いる場合
  - これはうまくいかない．
    説明変数を増やせば，決定係数は必ず増加する．
- 自由度を修正した決定係数の利用
  - 自由度が増えた分，決定係数の増加分を割り引く．
- 情報量基準の利用
  - AIC, MDL などの情報量基準を用いる．

情報量基準に基づくモデル選択

- 赤池情報量(AIC: Akaike's Information Criterion)
  - 説明される変数Y の真の確率分布とモデルの確率分布との距離(KL-divergence)の漸近的不偏推定量

$$\mathrm{AIC} = -2 \log L(Y \mid \hat{\theta}) + 2p$$

L: 尤度関数
$\hat{\theta}$: 最尤推定量
p: パラメータ数

- AICが最小になるモデルを選択すれば，真の確率分布に最も近いモデルを選ぶことができる．

回帰分析におけるAIC基準

- 回帰分析の場合
  - 回帰分析の仮定を用いると，最終的に

$$\mathrm{AIC'} = n \log (D^2) + 2\, m$$

を評価関数として用いればよいことになる．

n ：データ数

m：パラメータ数

D^2：平均2乗誤差

変数選択に関わるその他の問題

- 変数の多重共線性
  - 説明変数が相互に関係している場合
    （極端な例として，$X_2 = 2\, X_1$の場合）
  - 行列式の値が0に近づき，推定が不安定になる.
    → 一方の変数を除去する．
- 非線形関係の取り扱い
  - X と Yの間の関係が線形でない場合
    → 変数変換を行ない線形性を向上させる．