実証分析の基礎：
回帰分析とその応用（6）
重回帰分析

松井啓之
京都大学　経営管理大学院
2018年9月10日

重回帰分析（1）

- 一つの目的変数を、複数の説明変数で予測する事を考える。たとえば3つの独立変数がある場合、重回帰式は $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ となる。それぞれの独立変数にかかっている係数を「偏回帰係数」と呼ぶ。

$$\Sigma e_i^2 = \Sigma (y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}))^2 \rightarrow Min$$

なので、α、β_1、β_2、β_3で偏微分して0とおくことで正規方程式を作成し、それを解いて求めれば良い

  - 単回帰分析と同様に、モデルの当てはまりの良さは（自由度修正済み）決定係数で確認可能
  - ただし、単回帰と違い様々な要因が絡み難易度は上がる。
  - 当然、偏回帰係数についての統計的検定が必要！

重回帰分析（2）

- $Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i$
  - iはi番目、Y_iは被説明変数、X_{ji}は説明変数j、u_iは誤差項
  - $i=1$からnまで、この式が成立するとすれば、

$$Y_1 = \alpha + \beta_1 X_{11} + \cdots + \beta_k X_{k1} + u_1$$
$$\vdots$$
$$Y_n = \alpha + \beta_1 X_{1n} + \cdots + \beta_k X_{kn} + u_n$$

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

とすれば、この重回帰モデルは

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$$

と表現できる

重回帰分析（3）

$\boldsymbol{\beta}$の通常の最小二乗推定量を$\widehat{\boldsymbol{\beta}}$とすると、
モデルからの$\boldsymbol{y}$の推定値は

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

であり、$\boldsymbol{y}$と$\widehat{\boldsymbol{y}}$との差は残差

$$\widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

である。これを整理すると最小二乗法の正規方程式

$$(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

を得るので、この正規方程式を$\widehat{\boldsymbol{\beta}}$について解くと

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

が得られる。また残差$\widehat{\boldsymbol{u}}$は以下の性を持つ。

$$(1)\boldsymbol{X}'\widehat{\boldsymbol{u}} = \boldsymbol{0}, (2)\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{y}} = 0, (3)E(\widehat{\boldsymbol{u}}) = 0, (4)Var(\widehat{\boldsymbol{u}}) = \sigma^2\boldsymbol{M}$$

ただし、$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

重回帰分析（4）

- 線形回帰モデルの古典的仮定
  - 仮定0：真のモデルは線形
  - 仮定1：説明変数は非確率変数（＝$X_1,\ldots,X_n$は定数）
  - 仮定2：誤差項の期待値は（全てのiについて）0
    - $E(\boldsymbol{u})=\mathbf{0}$
  - 仮定3：分散均一性＝誤差項の分散は全てのiについて等しい
    - $E(\boldsymbol{uu}')=\sigma^2\boldsymbol{I}$ （$\boldsymbol{I}$は$n\times n$の単位行列）
  - 仮定4：誤差項は互いに独立⇒相関は存在しない。
  - 仮定5：誤差項の確率分布は正規分布に従う
    - この仮定と仮定1、仮定2、仮定3をあわせると

    $u_i\sim N(0,\sigma^2)$ i.i.d. ：誤差項は互いに独立で、同一の正規分布$N(0,\sigma^2)$に従う

重回帰分析（5）

- これまでと同じ議論から、個々の観測Y_iは正規分布に従う

$$Y_i\sim N(\alpha+\beta_1X_{1i}+\ldots+\beta_kX_{ki},\sigma^2)$$
$$E(Y_i)=\alpha+\beta_1X_{1i}+\ldots+\beta_kX_{ki}$$
$$Var(Y_i)=\sigma^2$$

- ガウス・マルコフの定理
  - 古典的仮定1～4が成立するならば、最小二乗推定の期待値と分散は

$$\mathrm{E}(\hat{\beta}_i)=\beta_i,\mathrm{Var}(\hat{\beta}_i)=\frac{\sigma^2}{S_{jj}(1-R_j^2)},j=1,2,\cdots,k$$

  - ここで、S_{jj}は、第j説明変数X_{ji}の偏差2乗和、R_j^2は、X_{ji}をそれ以外のk-1個の説明変数に重回帰した際の決定係数。上式の分散は、線形不偏推定量の中で最小となる。

重回帰分析の検定（1）

- 母分散s^2の不偏推定量

$$s^2=\frac{1}{n-(k+1)}\sum\hat{u}_i^2,E(s^2)=\sigma^2$$

- s^2より、各bの標準誤差

$$s.e.(\hat{\beta}_j)=\frac{s}{\sqrt{S_{jj}(1-R_j)}},j=1,2,\cdots,k$$

を得る。さらに、標準誤差を使えば、t統計量となる。

$$t_j=\frac{\hat{\beta}_j-\beta_j}{S/\sqrt{S_{jj}(1-R_j)}}=\frac{\hat{\beta}_j-\beta_j}{s.e.(\hat{\beta}_j)}\sim T(n-(1+k))$$

- 係数β_jに関する仮説検定 H_0:$\beta_j=\beta_{j*}$が可能に

重回帰分析の検定（2）

- $Y_i=\alpha+\beta_1X_{1i}+\ldots+\beta_kX_{ki}$という回帰関係を仮定し、個々の偏回帰係数$\beta_j$＝0という仮説を検定する。
- 回帰係数β_j（説明変数x_i）の有意性検定

$$H_0:\ \beta_j=0$$

- 帰無仮説が真⇒偏回帰係数に意味がない
- 帰無仮説を棄却

⇒対立仮説を採択：xとyの間に統計的な関係がある

「β_jは（x_jは）統計的に有意である」

「β_jは有意にゼロと異なる」　　と表現される

  - 統計ソフトで最小二乗検定を実施する場合に、t検定（t値が計算）されるのは、優位性検定を実施しているため

重回帰分析の検定（3）

- 回帰モデル全体の評価を考える。
- $Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$という回帰関係を仮定し、全ての偏回帰係数$\beta_1$、…、$\beta_k$が0という仮説を検定する。

  帰無仮説　H_0: $\beta_1 = \ldots = \beta_k = 0$

  対立仮説　H_1: H_0でない
  - 帰無仮説が真⇒偏回帰係数に意味がない
- 最小二乗法で推定した決定係数R^2を用いてF値を求めると

$$F = \frac{R^2/K - 1}{(1-R^2)/(n-K)} \sim F(K-1, n-K)$$

- F値が棄却域に入らなければ、帰無仮説を棄却できず、説明変数は全て説明力を持たない＝モデルが意味を持たない。

自由度修正済み決定係数

- 決定係数R^2の問題
  - 一般の回帰モデル（重回帰モデル）では複数の説明変数が用いられる⇒説明変数が増えるとR^2は必ず値が大きくなる。

    ⇔モデルを複雑にすればするほど（説明変数が多いモデルほど）説明力の高いモデル
- モデルの複雑度に左右されないためには

  ⇒自由度修正済み決定係数 $\overline{R}^2$

$$\overline{R}^2 = 1 - \frac{RSS/(n-(k+1))}{TSS/(n-1)} = 1 - \frac{\sum_{i=1}^{n} e_i^2 / (n-(k+1))}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)} = 1 - (1-R^2)\frac{n-1}{n-k}$$

なお、kは変数の個数

最適なモデルの決定

- F検定
  - nested modelの場合
- Adjusted R^2を用いる方法
- AIC基準　(Akaike Information Criteria)

  $AIC=-2\ln(L)+2k$

  $\ln(L)$: 対数尤度，k: パラメータの数（説明変数の数）

  AICを最小にするようなモデルを選ぶ
- BIC基準 ベイジアン情報量規準

  $BIC=T*\log(s^2)+K*\log(T)$

  Tは標本の大きさ、s^2はモデルの誤差項の分散推定量で、Kはモデルに含まれる係数の数

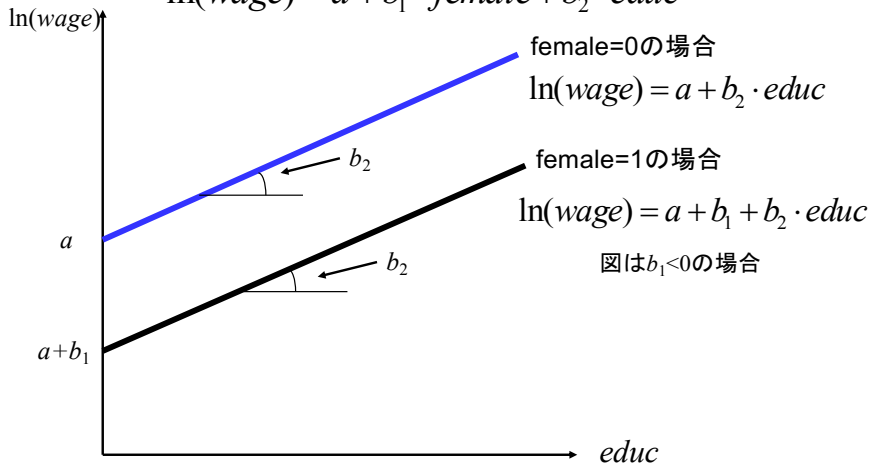
  BICを最小にするようなモデルを選ぶ
- 変数増減法(stepwise regression)

ダミー変数（1）

- 2つのグループの所属を区別するための質的変数
  - $D_i = \begin{cases} 0 & i\text{がグループに該当しない場合} \\ 1 & i\text{がグループに該当する場合} \end{cases}$
  - 複数のグループの場合は、「グループ数－1」のダミー変数を用意
    - 中卒、高卒、大卒の3区分→$D_{高卒}$、$D_{大卒}$の2つのダミー変数
      - 中卒：$D_{高卒}=0$、$D_{大卒}=0$　高卒：$D_{高卒}=1$、$D_{大卒}=0$
      - 大卒：$D_{高卒}=1$、$D_{大卒}=1$
- $Y=\alpha+\beta X$
  - $Y=\alpha+\beta_1 X+\beta_2 D$ ←切片パラメータのみに影響
  - $Y=\alpha+\beta_1 X+\beta_2 (X \times D)u$　←交差項：傾き（回帰）パラメータにも影響

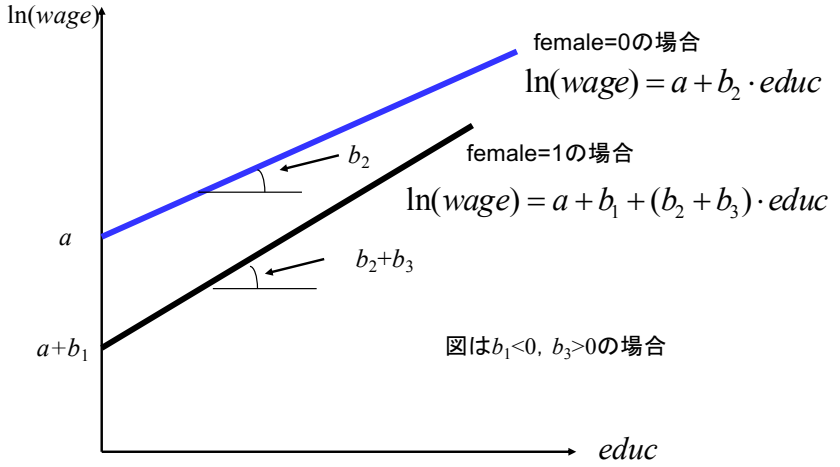
定数項ダミー

$$\ln(wage) = a + b_1 \cdot female + b_2 \cdot educ$$



傾きのダミー

$$\ln(wage) = a + b_1 \cdot female + b_2 \cdot educ + b_3 \cdot (female \times educ)$$



ダミー変数（2）

- 重回帰モデルの構造変化が生じる
  ⇔グループ間での重回帰パラメータ全てが異なる
  ⇔交差項まで含めたF検定＝大変！
- チョウ（Chow）検定
  - 全体の残差二乗和：SSR
  - 2つのグループの残差二乗和：SSR_1、SSR_2

$$F = \frac{SSR - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \times \frac{n - 2(k+1)}{k+1} \sim F(2, n-2k-2)$$

  - F値が棄却域に入れば、帰無仮説「グループ間で違いがない」を棄却する。＝構造変化が起った
    - 時系列データでも利用。その場合には、何時生じたかも確認

多重共線性

- 説明変数間で線形関係があるとき、正規方程式をで解を一意に定めることが出来ない状態となる。変数間に高い相関がある場合も同様。
  - 指定されたパラメータの符号が理論と合わない。説明変数のt値が小さい反面、決定係数が高い。データ数が変動すると、推定値が大きく変動する。新しい説明変数を追加すると、パラメータの符号や値が大きく変わる。
- 相関の高い場合には、どちらかの変数をモデルから、取り除く必要がある。
  - VIF（Variance Inflation Factor）の値が10以上
  - ⇔相関係数が0.95以上
- 変数の変更の際には、適切なモデル化どうか、自由度済み決定係数やAIC、BICを用いて確認。

多重共線性の検出

- OLSにおいて説明変数x_jの係数の分散は次の通りになる。

$$\mathrm{var}(b_j) = \frac{\sigma^2}{S_j(1-R_j^2)}$$

σ^2: 誤差項の分散，　S_j : 説明変数x_jの平均値の回りの平方和，　R_j^2: 説明変数x_jを 他の説明変数に回帰した場合のR^2（決定係数）

多重共線性→R_j^2が高い→b_jの分散が大きくなる

- VIF (Variance inflation factor　分散増幅因子)

$$VIF(b_j) = \frac{1}{1-R_j^2}$$

均一分散が成立しない場合の対応

- 古典的仮定の定理3が成立しない
  ⇒$Var(\hat{\beta})$を推定できない→標準誤差を求められない
- 不均一分散を確認するために
  ⇒ホワイト検定、ブルーシュ・ベーガン検定など
  - 補助回帰（残差を二乗した系列で回帰）の決定係数（nR^2）がカイ二乗分布に従うことを利用
  - 帰無仮説：不均一分散は生じていない
- 不均一分散への対応
  ⇒（ホワイトの）頑健標準誤差を用いる
  - 一般に、分析ソフトのOLSのオプションで用意されている
  ⇒分散の構造が分かっている場合ウェイト付き回帰（WLS）を用いる＝推定の精度を向上させる

分散不均一性の検出

- Whiteのテスト
  - 残差の平方 e^2 を被説明変数
  - 説明変数：x_j, x_jの平方，x_jとx_hの交差項
  - これらの説明変数の係数が全て0という仮説を検定する
- Breusch and Paganのテスト

estimate: $y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + u_i$

save: $e_i = y_i - a - b_1 x_{1,i} - b_2 x_{2,i} - \cdots - b_k x_{k,i}$

compure e_i^2

estimate : $e_i^2 = \delta_0 + \delta_1 x_{1,i} + \delta_2 x_{2,i} + \cdots + \delta_k x_{k,i} + v_i$

test $H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0$

$$\frac{(RSS - TSS)/k}{RSS/(n-(k+1))} = \frac{ESS/k}{RSS/(n-(k+1))} \sim F(k, n-(k+1))$$

加重最小二乗法　Weighted Least Square

- 不均一性のテストは検出のみ
  - どのような方法で対処すべきかは教えてくれない
  - 推定する方程式の関数型を変えることで解決する場合もある
- 誤差項の分散がある変数に比例していることがわかっている場合
  → Weighted Least Square 加重最小二乗法
- WLS : 次の式を最小化するように係数を決定

$$\sum_{i=1}^{n} w_i \left(y_i - a - b_1 x_{1,i} - \cdots b_k x_{k,i}\right)^2$$

w_i ： weight

Weighted Least Square

次のモデルを考える

$$y_i = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} + u_i \qquad (1)$$

ただし, $\mathrm{var}(u_i) = \sigma_i^2 = h_i\sigma^2$ (誤差項の分散が変数hに比例している→ 分散不均一性)。このとき次のように式変換すれば

$$\frac{y_i}{\sqrt{h_i}} = \alpha\frac{1}{\sqrt{h_i}} + \beta_1\frac{x_{1,i}}{\sqrt{h_i}} + \cdots + \beta_k\frac{x_{k,i}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}} \qquad (2)$$

$\mathrm{var}\left(\frac{u_i}{\sqrt{h_i}}\right) = \sigma^2$ → 分散は均一

Weighted Least Square (2)

(2)式をもとに係数を推定→次の式の最小化

$$\sum_{i=1}^{n}\left[\frac{y_i}{\sqrt{h_i}} - a\frac{1}{\sqrt{h_i}} - b_1\frac{x_{1,i}}{\sqrt{h_i}} - \cdots - b_k\frac{x_{k,i}}{\sqrt{h_i}}\right]^2$$
$$= \sum_{i=1}^{n}\frac{1}{h_i}\left[y_i - a - b_1x_{1,i} - \cdots - b_kx_{k,i}\right]^2$$
$$= \sum_{i=1}^{n} w_i\left(y_i - a - b_1x_{1,i} - \cdots b_kx_{k,i}\right)^2$$

元のモデルの誤差項の分散がhに比例する→weight変数を$1/h$にする

非線形の回帰分析(1)

- 線形回帰モデル:非説明変数と説明変数に線形線形を仮定している。ただし、線形回帰モデルのOLS推定が通用するための要件は、Y_iと回帰係数の線形性であり、X_iに関しては、適切な変数変換で対応可能
- 例:多項式モデル
  - $Y = \alpha + \beta_1X + \beta_2X^2 + \ldots + \beta_kX^k$、$X \rightarrow X_1$、$X^2 \rightarrow X_2$、…、$X^k \rightarrow X_k$、と変数変換することで対応可能
- 例:対数線形モデル
  - 対数変換することで、掛け算→足し算へ変換可能
  - $Y = \alpha + \beta X_1 \cdot X_2 \cdot \ldots \cdot X_k$

$\rightarrow Log(Y) = \mathrm{Log}(\alpha) + \beta\{Log(X_1) + Log(X_2) + \ldots Log(X_k)\}$

対数変換したモデルの解釈

応答変数	説明変数	係数 b の意味
無変換	無変換	説明変数が 1 単位増えると、応答変数は b だけ増える
無変換	自然対数	説明変数が 1%増えると、応答変数が b だけ増える
自然対数	無変換	説明変数が 1 単位増えると、応答変数が $100b$%増える
自然対数	自然対数	説明変数が 1%増えると、応答変数が $100b$%増える(弾力性)

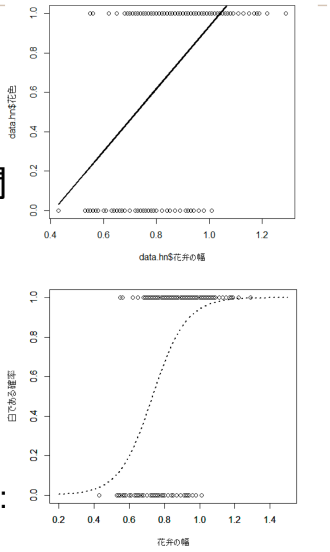
注:b が 0 に近くないときは $\exp(b) - 1$ を計算する必要がある

非線形の回帰分（2）

- 被説明変数が「０－１」になるような変数の場合に、線形回帰ではなく条件付き確率として割り当てる
- プロビット：正規分布の累積分布関数を割り当てる。
- ロジット：ロジステック関数を割り当てる。

$$Y = \frac{1}{1 + e^{-(\alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})}}$$

  - 均一分散を満たさないので、OLS推定は出来ないので、最尤推定量（ML: Maximum Likelihood estimator）を求める。→ML推定。



スコアリング

- 個々の見込み客が持つ、自社への将来的な「価値」を予測し、その価値に準じて順位をつけること
- 具体的には、ロジステック回帰分析で推定したロジステック関数を用いて確率を計算する。
  - 確率の予測は、医療分野では疾病リスク、金融分野では貸し倒れリスク、マーケティング分野では見込み客が財・サービスを利用する確率といった形で、多くの分野で応用
  - 例：ある妊婦が未熟児を生まれる確率を求めたい
    - birthwt「体重2.5kg未満の新生児が出産される原因となる因子」を探索するために調査されたデータ(1986年マサチューセッツ州スプリングフィールドにある湾岸州医療センター)
    - low:2.5kg未満の出生体重の指標。age:母親の年齢。lwt:最後の月経時の母親の体重（ポンド）。race:母親の人種（1=白、2=黒、3 =他）。smoke：妊娠中の喫煙状態。ptl：早産の経験。 ht:高血圧の病歴。ui:子宮過敏性の存在。ftv：最初の妊娠中の医師の診察数。bwt:出生体重（グラム）。