統計解析論　その4

多変数回帰

1．多変数回帰の概念

- 概念
  - 複数の変数を説明変数にした線形回帰
    - 定数項も説明変数とする

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

普通は $x_{1i} = 1$

- 例：
  - コブ－ダグラス型生産関数の推定

$$Y_i = \gamma + \alpha K_i + \beta L_i + \varepsilon_i$$

- さらにこの多変数回帰を一般的な回帰と考える
  - 単回帰はこの多変数回帰に含まれている

2．係数値の推定

- 正規方程式（＊）
  - 係数決定のために解くべき方程式

$$\sum_{i=1}^{n} \hat{u}_i x_{1i} = 0,\ \sum_{i=1}^{n} \hat{u}_i x_{2i} = 0,\ \cdots\ ,\ \sum_{i=1}^{n} \hat{u}_i x_{Ki} = 0$$

ただし　$\hat{u}_i = y_i - \left(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_K x_{Ki}\right)$

- <u>つまり，すべての説明変数と残差の積和が0ということ（統計屋的に言うと説明変数と残差が直交する）</u>
- 正規方程式を解くのは，Excelなど統計ソフトが連立方程式を解いてくれる．

3．多変数回帰からみた単回帰

- 係数推定値の公式（＊）

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} x_{ji}^* y_i^*}{\sum_{i=1}^{n} \left(x_{ji}^*\right)^2} = \frac{\sum_{i=1}^{n} x_{ji}^* y_i}{\sum_{i=1}^{n} \left(x_{ji}^*\right)^2} = \frac{\sum_{i=1}^{n} x_{ji} y_i^*}{\sum_{i=1}^{n} \left(x_{ji}^*\right)^2}$$

ただし，

$x_{ji}^* = x_{ji}$をx_j以外の全ての説明変数に回帰した時の残差

$y_i^* = y_i$をx_j以外の全ての説明変数に回帰した時の残差

つまり，$y_i^* = \beta_j x_{ji}^* + u_i$の回帰をしたときのβjの最小二乗推定量

3．多変数回帰からみた単回帰（続）

- 単回帰の公式との類似

$x_{1i} = 1,\quad x_{2i} = x_i,\quad \beta_1 = \alpha,\quad \beta_2 = \beta$ とおくと単回帰

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_{2i}^* y_i^*}{\sum_{i=1}^{n} \left(x_{2i}^*\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \hat{\beta}$$

ということは，$x_{2i}^* = x_i - \bar{x} = x_i^*,\quad y_i^* = y_i - \bar{y}$

3．1　類似性の種明かし

- まず $x_{2i}^* = x_i$をx以外の全ての説明変数に回帰した時の残差

  $= x_i$を1に回帰した時の残差

  $= x_i - \hat{\mu} \times 1$

  ここで係数推定値を$\hat{\mu}$とした．
- ところで，残差 x_{2i}^*と説明変数1との積和は0である
  ここから，$\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0$ となり$\sum_{i=1}^{n} x_i - n\hat{\mu} = 0$であるから

  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$
- yについても同様

4．1 多変数回帰の特殊例(1)

- 定数(1)だけに回帰する

$$y_i = \mu + \varepsilon_i \qquad \hat{\mu} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} 1^2 = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

  - $E[\varepsilon_i] = 0$ なので，$E[y_i] = \mu$であるから，y_iの期待値を推定している．
  - $V[y_i] = V[\varepsilon_i] = \sigma^2$なので，$\sigma^2$ は y_i の分散である．
  - 残差と説明変数の積和が0は前のスライド参照

$$RSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad ESS = 0$$

  - σ^2の推定量である残差分散がy_iの不偏分散推定量

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

4.2 多変数回帰の特殊例(2)

- 定数項を含まない回帰

$$y_i = \beta x_i + \varepsilon_i$$

$\hat{\beta} = \sum_{i=1}^{n} y_i x_i / \sum_{i=1}^{n} x_i^2$ (1変数の場合の一般型)

理由：残差と説明変数の積和が0なので

$$\sum_{i=1}^{n} \hat{u}_i x_i = \sum_{i=1}^{n} (y_i - \hat{\beta} x_i) x_i = \sum_{i=1}^{n} y_i x_i - \hat{\beta} \sum_{i=1}^{n} x_i^2 = 0$$

  - $\sum_{i=1}^{n} \hat{u}_i = 0$が一般的には成立しない．
  - TSS=RSS+ESSは成立しないので，R^2が負になることがある．($\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (\hat{\beta} x_i)^2 + RSS$ が成立)

5．多変数回帰における分散分析

- 説明変数に定数項が入っている場合に限り TSS=ESS+RSSがなりたつ．(＊)
- その時に限り重決定係数は意味を持つ．
- 重決定係数の平方根値はもはや重相関係数と関係がない．
- 回帰式 $y_i^* = \hat{\beta}_j x_{ji}^* + \hat{u}_i$における決定係数 R^2 は重要(説明変数に定数項がある場合)
  - 他の変数の影響を取り除いた上で被説明変数の変動のうちでどれだけをj番目の説明変数の変動で説明できるかを示す．

5．1　偏相関係数

- y_i^*とx_{ji}^*との相関係数　$r_{y,x_j|x_j\text{以外の説明変数}}$

$$\frac{\sum_{i=1}^{n} x_i^* y_i^*}{\sqrt{\sum_{i=1}^{n} (x_i^*)^2}\sqrt{\sum_{i=1}^{n} (y_i^*)^2}}$$

- 偏相関係数の二乗は被説明変数の変動のうちでどれだけをj番目の説明変数の変動で説明できるかを示す．ただし，予め他の説明変数の影響を取り除いてある．
- t値から計算できる．　$r^2 = \dfrac{t^2}{n-K+t^2}$

6．係数推定値の分散とt値

- 係数推定値の分散．その平方根が標準誤差

$$V[\hat{\beta}_j] = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ji}^*)^2}$$

- t値

$$t_{\beta_j} = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\dfrac{S}{\sqrt{\sum_{i=1}^{n} (x_{ji}^*)^2}}} \qquad S^2 = \frac{RSS}{n-K}$$

7．自由度(1)

- 変動に貢献する変数の個数
  - 逆に言うと変動を決める方程式の個数分減る．
- TSS ($y_i - \bar{y}$の二乗和)
  - 標本数nを自由に決めることができる．
  - 式 $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$ によって $\bar{y}$ が決まり，その結果 $y_i - \bar{y}$が決まる
  - 式1本なのでTSSの自由度は1下がっている．
  - よって自由度はn－1

７．自由度（その２）

- ＲＳＳ（ $\hat{u}_i$ の二乗和）
  - 正規方程式（係数の個数ある）によって係数推定値が決まり，$\hat{u}_i$ が決まる．
  - 式が係数の個数Ｋあるので，自由度はＫ下がり，結果として自由度はｎ－Ｋ
- ＥＳＳ（ $\hat{\beta}_1(x_{1i}-\bar{x}_1)+\cdots+\hat{\beta}_K(x_{Ki}-\bar{x}_K)$ の二乗和）
  - $x_{1i},\cdots,x_{Ki}$ は所与なので最初は自由度は０
  - $x_{1i}=1$ なので $\bar{x}_1=1$ となり，$\hat{\beta}_1(x_{1i}-\bar{x}_1)=0$ である．したがって，$\hat{\beta}_2,\cdots,\hat{\beta}_K$ のみが変動に貢献する．
  - よって自由度はＫ－１

８．モデル選択

- 変数を増やすと重決定係数は必ず増える．従って回帰モデルに変数を入れるかどうかの選択には重決定係数は無意味．
- 自由度修正付き決定係数

$$\bar{R}^2 = 1-\frac{RSS/(n-K)}{TSS/(n-1)}$$

cf. $R^2 = 1-\frac{RSS}{TSS}$

- ある変数を１つ回帰に新たに加えたときに $\bar{R}^2$ が増えるのはt値の絶対値が１以上の場合

９．マルチ･コリニアリティ（多重共線性）

- 説明変数の間に線形の関係，または，それに近い関係がある場合（その例）
  - 国民所得＝消費＋投資であるが，それを考えずに説明変数に国民所得，消費，投資を入れてしまった場合
  - 定数項のある回帰で説明変数１＋説明変数２＝定数の場合
- 統計ソフトは結果を出さないか異常な結果を出す．（例あり）
  - 係数値の符号が理論とあわない．また，t値が異常に小さい．
- 対策
  - 不要な係数を取り除く（相関係数が大きいもの同士を回帰に入れない）