統計解析論　第3回

最小二乗法の統計学

1．問題設定（1）

- 係数推定量の信頼性を知る
  - 係数推定量のばらつきの評価（標準誤差）
    - そのための仮定

$$y_i = \alpha + \beta x_i + \text{撹乱項（誤差項）}$$

    - 撹乱項ε_iに確率的仮定をする
      - » 平均0，分散σ^2（一定）でiが違うと独立
        （つまり，i毎にサイコロをふって撹乱項の値がきまる）
      - » σ^2　や，α，βは我々には分からないが，確かにあるとする．（真のパラメータと呼ぶ）
    - パラメータの推定とそのばらつき＝分散の評価

2．問題設定（2）

- 回帰自身に意味があるかの検討
  - R^2の統計的性質はあまり明確ではない
  - 統計的な意味づけをもった指標
  - F統計量

3．残差分散

- 回帰直線からの乖離の程度を測る
  - 残差分散

$$S^2 = \frac{RSS}{\text{標本数} - \text{説明変数の数（定数項含む）}} = \frac{RSS}{n-K}$$

    - この値が回帰直線$y_i = \alpha + \beta x_i$からの乖離＝撹乱項　ε_i　の分散σ^2の一つの推定量となる．
    - 不偏推定量
      - この推定量S^2の確率的期待値が実はσ^2　　（∗）

4．標準化残差による回帰のチェック

- 残差標準偏差　$S = \sqrt{\frac{RSS}{n-K}}$
- 標準化残差　$\frac{\hat{u}_i}{S}$
  - 残差を残差の標準偏差で割るのでその標本分散は1である．この絶対値が2以上になる確率は少ない．（チェビシェフの不等式）
  - 2以上があまりに多い（標本数の5％よりかない多い），2．5以上の値がある（標本数の1％よりかなり多い）場合決定係数がいくら大きくてもさらなるチェックが必要
  - S^2はそのままでは残差の分散とならない．（教科書参照）従って，標準化残差によるチェックはおおまかなもの．

5．回帰分散と残差分散

- 回帰分散　$\frac{ESS}{K-1}$
  - 回帰係数推定値の変動（推定誤差）による予測値$\hat{\alpha} + \hat{\beta} x_i$の変動の指標
  - もし，$\beta = 0$で撹乱項が正規分布なら　$\frac{ESS}{\sigma}$　は自由度K-1のカイ二乗分布に従って変動する　　（∗）
- 残差分散　$\frac{RSS}{n-K}$
  - 　$\frac{RSS}{\sigma}$　は自由度$n-K$のカイ二乗分布に従って変動する　　（∗）
  - 残差分散と回帰分散は独立に変動する

6．F統計量による回帰のチェック

- F統計量＝回帰分散／残差分散
- F統計量は $\beta = 0$ のとき，分子自由度K－1，分母自由度n－KのF分布に従う（正規分布の仮定に依存）　（＊）
- $\beta = 0$ のチェックに使用できる．（検定参照）
- こちらのほうが決定係数より統計学的根拠をもつ
- 大きいと回帰変数の一部または全部が意味をもち，小さいと回帰は意味を持たない．

7．係数値の信頼性

- 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（＊）

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

    - 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．
  - 信頼区間が計算される（＊）

8．t統計量

- 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, β とする
  - βのt統計量 $t_{\hat{\alpha}} = \dfrac{\hat{\beta} - \beta}{s.e.(\hat{\beta})} = \dfrac{\hat{\beta} - \beta}{\dfrac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$
    - この値は自由度 $n - K$ 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

続き

- αに関するt統計量（自由度n-Kのt分布）

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

- 通常出力されるのは，$\alpha = 0$，$\beta = 0$ 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