# 第 11 回　重回帰分析で利用するその他の分析

## [1] 自由度修正済み決定係数

決定係数 $R^2$ の欠点：どんな説明変数を加えても，ふえてしまう。

⇒ **自由度修正済み決定係数** $\overline{R}^2 = 1 - \dfrac{n-1}{n-k-1}(1-R^2)$

説明変数を加える効果 $\left\{ \begin{array}{l} ①\ R^2 がふえる \\ ②\ k がふえる \end{array} \right\}$ 加えた変数が y の説明に役立つときだけ $\overline{R}^2$ はふえる

（$\overline{R}^2$ の性質）

・追加した説明変数の係数の t 値が 1 より大きいとき，$\overline{R}^2$ がふえる。

・説明変数が多いわりに $R^2$ が低いとき，$\overline{R}^2$ は負になることがある。

（例） $R^2 = 0.10$，$n = 51$，$k = 10$ のとき，$\overline{R}^2 = -0.125$

（$\overline{R}^2$ の利用法：モデルの比較）

（例 1）モデル 1： $\log(salary) = \beta_0 + \beta_1 years + \beta_2 games + \beta_3 bavg + \beta_4 hruns + u$

互いに含まない

モデル 2： $\log(salary) = \beta_0 + \beta_1 years + \beta_2 games + \beta_3 bavg + \beta_4 rb + u$

モデル 1 の $\overline{R}^2 = 0.60$，モデル 2 の $\overline{R}^2 = 0.65$　⇒　モデル 2 の方がよい。

（例 2）モデル 1： $r\&d = \beta_0 + \beta_1 \log(sales) + u$

異なる関数形

モデル 2： $r\&d = \beta_0 + \beta_1 sales + \beta_2 (sales)^2 + u$

（共通点）

sales の増加が r&d を逓減的に増やす効果を表現

モデル 1 の $\overline{R}^2 = 0.10$，モデル 2 の $\overline{R}^2 = 0.15$　⇒　モデル 2 の方がよい。

## [2] 予測

35 人の学生について，大学での成績（GPA）を sat（進学適正試験の成績：1600 点満点），hsperc（高校での成績：上位%），hsize（通学した高校の学生数：百人単位）で説明する式を最小二乗法で推定したところ，次の結果を得た（カッコ内の数字は標準誤差)。

$$\underset{}{\mathrm{GPA}} = \underset{}{1.493} + \underset{(0.0007)}{0.00149}\,\mathrm{sat} - \underset{(0.00056)}{0.01386}\,\mathrm{hsperc} - \underset{(0.01650)}{0.06088}\,\mathrm{hsize} + \underset{(0.00227)}{0.00546}\,\mathrm{hsize}^2$$
$$\underset{(0.075)}{}$$

$\overline{R}^2 = 0.277$ ,　　$\hat{\sigma} = 0.560$（$\sigma$ ：誤差項 u の分散の平方根）

・予測の問題：・ sat = 1200, hsperc = 30, hsize = 5 の学生のGPAの予測値はいくらか？

・ どのくらい正確に予測できるか？

・予測の考え方：　$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$　で，説明変数が　$x_1 = c_1, \ldots, x_k = c_k$ という定数の値をとるとき，

（考え方①）「yの平均」の予測　　⇒　定 数　$\mu = E(y)$ の予測

（例）sat = 1200, hsperc = 30, hsize = 5 である学生をたくさん集めると，彼らの平均GPAはいくらか？

（考え方②）「yそのもの」の予測　⇒　確率変数　y の予測

（例）sat = 1200, hsperc = 30, hsize = 5 である1 人の学生のGPAはいくらか？

### ① 平均 $\mu = E(y)$ の予測

$x_1 = c_1, \ldots, x_k = c_k$ のとき，仮定より，誤差項 u の期待値は 0 なので，

$$\mu = E(y) = \beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k + E(u) = \underbrace{\beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k}_{\text{予測の対象}}$$

平均 μ の点予測　$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k c_k$　（$\hat{\beta}_j$ は最小二乗推定量）

※ $\hat{\mu}$ の望ましい性質：　$E(\hat{\mu}) = \mu$　（$\hat{\mu}$ は $\mu$ の不偏推定量）

平均μの95%信頼区間　$\hat{\mu} - c \times se(\hat{\mu}) \le \mu \le \hat{\mu} + c \times se(\hat{\mu})$　（c： $t(n-k-1)$ の臨界値）

$se(\hat{\mu}) = \sqrt{\hat{Var}(\hat{\mu})}$ は $\hat{\mu}$ の標準誤差であり，次のように計算する。

$$se(\hat{\mu}) = \sqrt{\sum_{i=0}^{k} c_i^2 \hat{Var}(\hat{\beta}_i) + \sum_{i=0}^{k}\sum_{j=0}^{k} c_i c_j \hat{Cov}(\hat{\beta}_i, \hat{\beta}_j)} \quad (i \neq j,\ c_0 = 1)$$

※ 信頼区間の導出：$\hat{\beta}_j$ が正規分布にしたがうとき，その線形結合である $\hat{\mu}$ も正規分布にしたがうので，定理 4.2 と同様に考えれば，　$T = \dfrac{\hat{\mu} - \mu}{se(\hat{\mu})} \sim t(n-k-1)$

**② 確率変数 y の予測**

$x_1 = c_1, \ldots, x_k = c_k$ のとき，誤差項 $u^0$ に対応して $y = y^0$ が発生すると考える。

$$y^0 = \underbrace{\beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k}_{\text{予測の対象}} + u^0 \qquad （仮定： E(u^0) = 0,\ Var(u^0) = \sigma^2）$$

⇒ $\mu = E(y)$ の予測に比べると，誤差項 $u^0$ を余分に予測することになる。

$y^0$ の点予測　$\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k c_k = \hat{\mu}$　（$\hat{\beta}_j$ は最小二乗推定量）

※ $\hat{y}^0$ の望ましい性質：**予測誤差**を $\hat{e}^0 = y^0 - \hat{y}^0$ とすると，$E(\hat{e}^0) = 0$ となる。

$y^0$ の95%信頼区間　$\hat{\mu} - c \times se(\hat{e}^0) \le y^0 \le \hat{\mu} + c \times se(\hat{e}^0)$

$se(\hat{e}^0) = \sqrt{\hat{Var}(\hat{e}^0)}$ は $\hat{e}^0$ の標準誤差であり，次のように計算する。

$$se(\hat{e}^0) = \sqrt{\hat{\sigma}^2 + (se(\hat{\mu}))^2}, \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} \qquad （\hat{u}_i：残差）$$

⇒ 平均 $\mu$ の信頼区間に比べて，$\hat{\sigma}^2$ の分だけ信頼区間が広い。

※ 信頼区間の導出： $T = \dfrac{y^0 - \hat{y}^0}{se(y^0 - \hat{y}^0)} = \dfrac{\hat{e}^0}{se(\hat{e}^0)}$ ～ $t(n-k-1)$ を利用する。

（例）大学での成績(GPA)の予測

① GPA の平均の予測

$$\hat{\mu} = 1.493 + 0.00149 \times 1200 - 0.01386 \times 30 - 0.06088 \times 5 + 0.00546 \times 5^2 \approx 2.70$$

$se(\hat{\mu}) = 0.02$ が計算されたとし，自由度 30 の t 分布の臨界値 $c = 2.042$ を使うと，

$\mu$ の 95%信頼区間: $2.7 - 2.042 \times 0.02 \le \mu \le 2.7 + 2.042 \times 0.02 \quad \Rightarrow \quad 2.66 \le \mu \le 2.74$

② GPA の予測

GPA の点予測は①と同じで $\hat{GPA} = 2.70$ である。また，推定結果より $\hat{\sigma} = 0.560$ であり，$se(\hat{e}^0) = \sqrt{\hat{\sigma}^2 + (se(\hat{\mu}))^2} = \sqrt{0.56^2 + 0.02^2} \approx 0.56$ であるので，

GPA の 95%信頼区間： $2.7 - 2.042 \times 0.56 \le GPA \le 2.7 + 2.042 \times 0.56$

$$\Rightarrow 1.56 \le GPA \le 3.84$$

## [3] 残差分析

残差 $\hat{u}$ は推定した誤差項 u であるので，残差を観察すれば次の点がわかる。

・誤差項 u の仮定が適切かどうか（u の分散は一定？，u はランダム標本？）

・異常値（外れ値）の発見

（例）別紙の図の左側の x と y に関するデータ A ～ E はすべて次の関係を満たす。

$$n = 11, \quad \bar{x} = 9, \quad \bar{y} = 7.5, \quad S_{xx} = 110, \quad S_{yy} = 41.25, \quad S_{xy} = 55$$

よって，データ A～E を使って単回帰分析をすれば，次の同じ結果がえられる。

$$\hat{y} = 3 + 0.5\,x, \quad R^2 = 0.667$$

⇒ 決定係数や係数の t 値だけを観察して分析を終えるのは危険

各データの問題点は，別紙の図の右側の「残差プロット」が示している。

サンプル A：問題なし

サンプル B：負の残差，正の残差が連続する　⇒ 関数形の選択ミス，誤差の相関

サンプル C：一つだけ残差が外れている　　⇒ 異常値（外れ値）

サンプル D：x がふえるとき，残差の現れ方が広がる ⇒ 不均一分散

サンプル E：x の変動がほとんどない

＜p.1 の例 1＞
Jeffrey Wooldridge
Introductory Econometrics: A Modern Approach (3rd ed.)
South-Western, Division of Thomson Learning の p.209 の本文を引用
＜p.1 の例 2＞
Jeffrey Wooldridge
Introductory Econometrics: A Modern Approach (3rd ed.)
South-Western, Division of Thomson Learning の p.210 の本文を引用
＜p.2 の冒頭の例＞
Jeffrey Wooldridge
Introductory Econometrics: A Modern Approach (3rd ed.)
South-Western, Division of Thomson Learning の p.215 の EXAMPLE 6.5 を一部変更して利用
＜p.4 の例＞
Jeffrey Wooldridge
Introductory Econometrics: A Modern Approach (3rd ed.)
South-Western, Division of Thomson Learning の p.217 の EXAMPLE 6.6 を引用