第 14 回：一般化線形回帰モデル
【教科書第 11 章】

北村　友宏

2021 年 1 月 21 日

本日の内容

1. 一般化線形回帰モデルの推定
2. gretl を用いた実習

古典的線形回帰モデル

y_i を x_i に単回帰することを考える．これまでに登場した線形回帰モデル

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + u_i, \\
E(u_i \mid x_i) &= 0, \\
E(u_i u_j \mid x_i) &= 0 \quad (i \neq j), \\
V(u_i \mid x_i) &= \sigma^2, \\
i &= 1, 2, \cdots, n
\end{aligned}
$$

は，古典的線形回帰モデル（Classical Linear Regression Model）．

$$\Downarrow$$

誤差項 u_i の条件付き分散が一定．

一般化線形回帰モデル

誤差項 u_i の条件付き分散が，個体によって異なる（一定でない）と仮定すると，一般化線形回帰モデル（Generalized Linear Regression Model）となる．

単回帰の場合，$V(u_i \mid x_i)$ が一定でないとした一般化線形回帰モデルは，

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i, \\ E(u_i \mid x_i) &= 0, \\ E(u_i u_j \mid x_i) &= 0 \quad (i \neq j), \\ V(u_i \mid x_i) &= \sigma_i^2, \\ i &= 1, 2, \cdots, n \end{aligned}$$

不均一分散

- $V(u_i \mid x_i)$ が一定でないことを（条件付き）不均一分散（heteroskedasticity）という．
- 不均一分散があっても，（古典的線形回帰モデルを推定して）頑健標準誤差（robust standard error）を計算すれば，標準誤差をより厳密に求め，仮説検定をより厳密に行うことができる．
- 他の方法として，モデル推定の段階で誤差項の不均一分散を取り除き，古典的線形回帰モデルに変換することが考えられる．

一般化最小二乗法

$$V(u_i \mid x_i) = \sigma_i^2,$$

とした一般化線形回帰モデル

$$y_i = \beta_0 + \beta_1 x_i + u_i, \tag{1}$$

を，古典的線形回帰モデル（誤差項の条件付き分散が一定）に変換する．

そのために，(1) の両辺を σ_i で割ると，

$$\begin{aligned} \frac{y_i}{\sigma_i} &= \frac{\beta_0 + \beta_1 x_i + u_i}{\sigma_i} \\ &= \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \cdot \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i}. \end{aligned} \tag{2}$$

- 被説明変数：$\dfrac{y_i}{\sigma_i}$
- 説明変数：$\dfrac{1}{\sigma_i}, \dfrac{x_i}{\sigma_i}$
- 誤差項：$\dfrac{u_i}{\sigma_i}$

$\Rightarrow$ (2) では定数項がなくなったが，$\dfrac{1}{\sigma_i}$ の係数 β_0 が元の定数項．

(2) の，説明変数を所与とした誤差項の条件付き分散を確認する．σ_i は確率変数でないので，

$$
\begin{aligned}
V\left(\frac{u_i}{\sigma_i} \mid \frac{1}{\sigma_i}, \frac{x_i}{\sigma_i}\right) &= V\left(\frac{u_i}{\sigma_i} \mid x_i\right) \\
&= \left(\frac{1}{\sigma_i}\right)^2 \underbrace{V(u_i \mid x_i)}_{=\sigma_i^2} \\
&= \frac{1}{\sigma_i^2} \cdot \sigma_i^2 \\
&= 1.
\end{aligned}
$$

よって，(2) では，説明変数を所与とした誤差項の条件付き分散は一定．

$$E\left(\frac{u_i}{\sigma_i} \mid \frac{1}{\sigma_i}, \frac{x_i}{\sigma_i}\right) = 0,$$

$$E\left(\frac{u_i}{\sigma_i} \cdot \frac{u_j}{\sigma_j} \mid \frac{1}{\sigma_i}, \frac{x_i}{\sigma_i}\right) = 0 \quad (i \neq j),$$

も満たされる（証明は省略）.

$$\Downarrow$$

$$\frac{y_i}{\sigma_i} = \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \cdot \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i}, \tag{2}$$

は古典的線形回帰モデルになっている.

実行不可能な一般化最小二乗法

(2) の残差二乗和

$$\sum_{i=1}^{n}\left(\frac{y_i}{\sigma_i}-\hat{\beta}_{0,\mathrm{GLS}}\cdot\frac{1}{\sigma_i}-\hat{\beta}_{1,\mathrm{GLS}}\cdot\frac{x_i}{\sigma_i}\right)^2,$$

が最小になるような $\hat{\beta}_{0,\mathrm{GLS}}$ と $\hat{\beta}_{1,\mathrm{GLS}}$ を求めることを，（実行不可能な） 一般化最小二乗法（Generalized Least Squares, GLS）という．これで，元のモデル

$$y_i = \beta_0 + \beta_1 x_i + u_i, \tag{1}$$

の定数項 β_0 と x_i の係数 β_1 を推定できる．

- σ_i は未知なので，この方法は実行不可能．
  ➡ まずは σ_i を推定する必要がある．

実行可能な一般化最小二乗法

そこで，以下の手順で β_0 と β_1 を推定する（gretl の場合）．

1. y_i を x_i に OLS 回帰し，残差

$$e_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

を得る．

2. $\ln e_i^2$ を x_i（と x_i^2）に OLS 回帰し，$\ln e_i^2$ の予測値

$$\widehat{\ln e_i^2}$$

を得る．

    - 誤差項の条件付き分散が「個体ごとに異なる＝説明変数の値によって変化する」と仮定している．

3. x_i を所与とした誤差項 u_i の条件付き標準偏差 σ_i の推定値を

$$\hat{\sigma}_i = \sqrt{\exp\left(\widehat{\ln e_i^2}\right)},$$

とし，それを用いて $\dfrac{y_i}{\hat{\sigma}_i}$ を $\dfrac{1}{\hat{\sigma}_i}$ と $\dfrac{x_i}{\hat{\sigma}_i}$ に回帰する．すなわち，

$$\frac{y_i}{\hat{\sigma}_i} = \beta_0 \cdot \frac{1}{\hat{\sigma}_i} + \beta_1 \cdot \frac{x_i}{\hat{\sigma}_i} + \frac{u_i}{\hat{\sigma}_i},$$

を OLS で推定する．

これは，元のモデルの残差から誤差項の条件付き標準偏差を推定し，それを用いて

$$\sum_{i=1}^{n}\left(\frac{y_i}{\hat{\sigma}_i}-\hat{\beta}_{0,\mathrm{FGLS}}\cdot\frac{1}{\hat{\sigma}_i}-\hat{\beta}_{1,\mathrm{FGLS}}\cdot\frac{x_i}{\hat{\sigma}_i}\right)^2,$$

が最小になるような $\hat{\beta}_{0,\mathrm{FGLS}}$ と $\hat{\beta}_{1,\mathrm{FGLS}}$ を求めていることになる．この方法を，実行可能な一般化最小二乗法（Feasible Generalized Least Squares, FGLS）という．

- 誤差項の不均一分散を取り除いているので，標準誤差の修正（頑健標準誤差の計算）は不要．

「OLS + 頑健標準誤差」と「FGLS」

- OLS で推定し，頑健標準誤差を計算する方法
  - 誤差項に不均一分散があるという前提で推定を行い，標準誤差を修正している．
  - 実証分析では，こちらがよく使われる．
- FGLS で推定する方法
  - 誤差項の不均一分散を取り除いた状態にして推定を行っている．
  - 実証分析ではあまり使われないが，観測値数が十分に大きいとき OLS 推定量よりも分散の小さな推定量が得られるという長所がある．

注：元のモデルの誤差項の条件付き分散が均一でない限り，OLS 推定量と FGLS 推定量は異なった値になる．

gretl での FGLS 推定

一般化線形回帰モデルを仮定し，第 11 回授業で推定した，ワンルームダミーを含むモデル

$$price_i = \beta_0 + \beta_1 minutes_i + \beta_2 age_i + \beta_3 area_i + \beta_4 d_i + u_i,$$

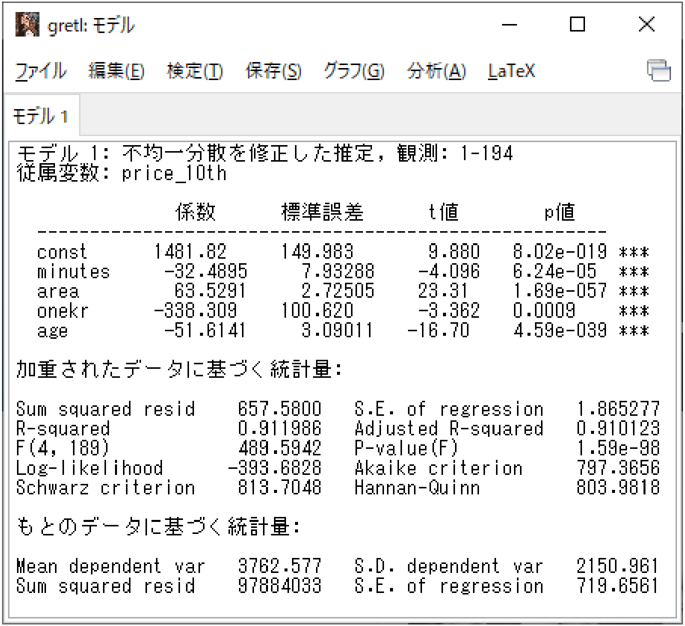
- $price_i$：中古マンション価格（万円）
- $minutes_i$：最寄り駅までの所要時間（分）
- age_i：築年数（年）
- $area_i$：面積（m^2）
- d_i：ワンルームダミー
- i：中古マンション番号

を，FGLS で推定する．

実習 1

1. gretl を起動.
2. 「ファイル」→「データを開く」→「ユーザー・ファイル」と操作.
3. setagayaapartment.gdt を選択し,「開く」をクリック.
4. gretl のメニューバーから「モデル」→「その他の線形モデル」→「不均一分散を修正した推定」と操作.
5. 出てきたウィンドウ左側の変数リストにある price_10th をクリックし，3 つの矢印のうち上の青い右向き矢印をクリック.
    - 推定式の左辺の変数（被説明変数，従属変数）が price_10th（万円単位の中古マンション価格）となる.

6. ウィンドウ左側の変数リストにある minutes をクリックした後，Ctrl キーを押しながら，area, onekr, age をクリックし，3 つの矢印のうち真ん中の緑の右向き矢印をクリック．
   - 推定式の右辺の変数（説明変数，独立変数）が minutes（最寄り駅までの所要時間），area（面積），onekr（ワンルームダミー），age（築年数）の 4 つとなる．
   - 最初から説明変数リストに入っている const は推定式の切片（定数項）のこと．
7. 「分散の推定式に説明変数の二乗項を含む」にチェック．
8. 「OK」をクリックすると，結果が新しいウィンドウに表示される．



このような画面が表示されれば成功.

9. 表示された「gretl: モデル 1」のウィンドウのメニューバーから「ファイル」→「名前を付けて保存」と操作。
10. 「標準テキスト」を選び，「OK」をクリック。
11. results20210121.txt という名前で 2020microdatag フォルダに保存．すると，表示された推定結果をそのままテキストファイルで保存できる．

モデル推定結果

- 最寄り駅所要時間の係数
  - −32.4895（符号は負）
  - 有意水準 1%で係数ゼロの H_0 棄却.
    ➥ 最寄り駅までの所要時間はマンションの価格と統計的に有意に相関している.
    ➥ 築年数，面積，ワンルームかどうかを一定とした上で，最寄り駅までの所要時間が 1 分長くなると，マンションの市場価値が 32.4895 万円（324,895 円）安くなる傾向がある.

- 築年数の係数
  - −51.6141（符号は負）
  - 有意水準 1%で係数ゼロの H_0 棄却.
    ➡ 築年数はマンションの価格と統計的に有意に相関している.
    ➡ 最寄り駅までの所要時間，面積，ワンルームかどうかを一定とした上で，築年数が 1 年長くなると，マンションの市場価値が 51.6141 万円（516,141 円）安くなる傾向がある.
- 面積の係数
  - 63.5291（符号は正）
  - 有意水準 1%で係数ゼロの H_0 棄却.
    ➡ 面積はマンションの価格と統計的に有意に相関している.
    ➡ 最寄り駅までの所要時間，築年数，ワンルームかどうかを一定とした上で，面積が 1m^2 広くなると，マンションの市場価値が 63.5291 万円（635,291 円）安くなる傾向がある.

- ワンルームダミーの係数
  - −338.309（符号は負）
  - 有意水準 1%で係数ゼロの H_0 棄却．
    ➥ ワンルームであるかどうかはマンションの価格と統計的に有意に相関している．
    ➥ 最寄り駅までの所要時間，築年数，面積を一定とした上で，ワンルームマンションはそれ以外の種類のマンションに比べ，市場価値が 338.309 万円（3,383,090 円）安い傾向がある．

- 定数項
  - 1481.82（符号は正）
  - 有意水準 1%で，係数ゼロの H_0 棄却.
    ➥ 定数項は統計的に有意にゼロと異なる.
- 自由度修正済み決定係数
  - $\bar{R}^2 = 0.910123.$
    ➥ 最寄り駅までの所要時間，築年数，面積，ワンルームかどうかの違いで，「価格」のバラつきが約91%説明できている.

本日の作業はここまで.
今回は gretl のデータセットに変更を加えていないので，gretl のデータセット（setagayaapartment.gdt）を上書き保存する必要はない.