第 8 章: 重回帰分析：不均一分散

Jeffrey Wooldridge (2016).
Introductory Econometrics: A Modern Approach
Seventh Edition. Cengage Learning.

2026-03-08

準備

必要なパッケージの読み込み

- wooldridge パッケージの読み込み

```
library(wooldridge)
```

- sandwich パッケージの読み込み

```
library(sandwich)
```

- lmtest パッケージの読み込み

```
library(lmtest)
```

8-1 不均一分散の含意

不均一分散 (Heteroskedasticity) とは

- これまでのガウス＝マルコフ仮定（MLR.5）では、説明変数 x の値に関わらず、誤差項 u の分散が一定であることを仮定していた。
- $Var(u|x) = \sigma^2$（等分散性, Homoskedasticity）
- これが満たされない場合（$Var(u|x) = \sigma^2(x)$）、**不均一分散**と呼ぶ。

不均一分散が OLS に与える影響

1. **不偏性・一致性には影響しない。**OLS 推定量は依然として不偏かつ一致である。
2. **OLS はもはや最良線形不偏推定量 (BLUE) ではない。**より効率的な推定量が存在し得る。
3. **標準誤差の推定が正しくなくなる。**通常の t 統計量や F 統計量を用いた仮説検定が妥当ではなくなる。

8-2 不均一分散に頑健な推論

頑健な標準誤差 (Robust Standard Errors)

- 不均一分散が存在しても、適切な修正を加えることで妥当な推論（t 検定など）が可能になる。
- ホワイトの頑健な標準誤差（White standard errors）や、不均一分散に頑健な標準誤差（Heteroskedasticity-robust standard errors）と呼ばれる。
- 大規模標本において、誤差項の分散の形を知らなくても正当化される。

Rでの実装例 (gpa3)

```r
data(gpa3)
res <- lm(cumgpa ~ sat + hsperc + tothrs + female + black + white,
          data = gpa3, subset = (spring == 1))

# 通常の標準誤差
# summary(res)$coefficients[1:3, 1:2]

# 頑健な標準誤差 (sandwich パッケージを使用)
coeftest(res, vcov = vcovHC(res, type = "HC1"))[1:3, 1:2]
```

```
##                 Estimate   Std. Error
## (Intercept)  1.470064766 0.2206802109
## sat          0.001140728 0.0001915317
## hsperc      -0.008566358 0.0014179275
```

8-3 不均一分散の検定

なぜ検定するのか？

- 不均一分散が疑われる場合、まず統計的に確認することが重要。
- **検定の役割：**「等分散という帰無仮説を棄却できるか」を判断する。
- 主な検定方法：
  - **ブロイシュ＝ペーガン (BP) 検定**：分散が説明変数の線形関数である場合に強力。
  - **ホワイト (White) 検定**：より一般的な不均一分散（非線形など）も検出できる。
- p 値が小さい（例：0.05 未満）なら、等分散の仮定を棄却し、頑健な標準誤差や WLS で対応する。

ブロイシュ＝ペーガン検定 (Breusch-Pagan Test)

- **帰無仮説** $H_0 : Var(u|x_1, \ldots, x_k) = \sigma^2$（等分散、不均一分散なし）
- **対立仮説** H_1：分散が少なくとも 1 つの x_j に依存する（不均一分散あり）
- **直感：**もし分散が x_j に無関係なら、$\hat{u}^2$ を x_j で回帰しても $R^2 \approx 0$ になるはず。
- **手順：**
  1. OLS 残差 $\hat{u}$ を求める。
  2. 残差の 2 乗 $\hat{u}^2$ を独立変数 $x_1, \ldots, x_k$ で**補助回帰**する。
  3. $LM = n \cdot R^2$ を計算し、自由度 k の χ^2 分布で p 値を求める。

BP 検定の実装例（gpa3 データ、8-2 の res を使用）

```
# bptest() は lmtest パッケージの関数
# 引数に OLS 結果オブジェクト (res) を渡すだけで自動的に BP 検定を実行
bptest(res)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  res
## BP = 44.557, df = 6, p-value = 5.732e-08
```

```
# 出力の見方：
#   BP = LM 統計量 (n * R^2 補助回帰)
#   df = 補助回帰の説明変数の数（= 元モデルの k）
#   p-value が小さい → H0（等分散）を棄却 → 不均一分散の証拠あり
```

ホワイト検定 (White Test)

- **帰無仮説** H_0：等分散（BP 検定と同じ）
- **BP 検定との違い：**補助回帰に説明変数の **2 乗項・交差項**も含める。
  - 分散が非線形な形で x_j に依存する場合も検出できる。
- **問題点：**変数が多いと交差項が膨大になり自由度が急減する。
- **簡略版（推奨）：**OLS 予測値 $\hat{y}$ とその 2 乗 $\hat{y}^2$ を補助回帰に使う。
  - $\hat{y}$ は全説明変数の線形結合なので、2 乗項・交差項の情報を集約している。
  - 自由度の消費は常に 2（$\hat{y}$ と $\hat{y}^2$ の 2 変数）。

ホワイト検定の実装例（簡略版）

```r
# ステップ 1：OLS 残差の 2 乗と予測値を取得
u2    <- resid(res)^2      # 残差の 2 乗（被説明変数）
yhat <- fitted(res)        # OLS 予測値（= 説明変数の線形結合）

# ステップ 2：補助回帰 — 残差 2 乗 を yhat と yhat^2 で説明
# I(yhat^2) は R の数式内で yhat の 2 乗を計算する記法
white_reg <- lm(u2 ~ yhat + I(yhat^2))
```

ホワイト検定：LM 統計量の計算

```
# LM統計量 = n * R~2（補助回帰の R~2）
n         <- nobs(res)                                  # サンプルサイズ
lm_stat   <- n * summary(white_reg)$r.squared   # LM統計量
# 簡略版ホワイト検定の自由度 = 2（yhat と yhat~2 の 2変数）
p_val     <- 1 - pchisq(lm_stat, df = 2)        # χ~2(2) の p値
data.frame(LM = lm_stat, df = 2, p_value = p_val)
```

```
##          LM df   p_value
## 1 1.264452  2 0.5314075
```

```
# p値が小さい → H0（等分散）を棄却 → 何らかの不均一分散が存在する
```

BP 検定 vs. ホワイト検定：使い分け

	BP 検定	ホワイト検定（簡略版）
補助回帰の説明変数	元の x_j のみ	$\hat{y}$, $\hat{y}^2$
自由度の消費	k（変数の数）	常に 2
検出できる不均一分散	線形な依存	線形・非線形
推奨される場面	変数数が少ない場合	変数数が多い場合

8-4 加重最小二乗法 (WLS)

加重最小二乗法 (Weighted Least Squares)

- 不均一分散の構造 $Var(u|\mathbf{x}) = \sigma^2 h(\mathbf{x})$ が既知である場合、OLS よりも効率的な推定が可能。
- 各観測値を $1/\sqrt{h_i}$ で割って変換し、変換モデルに OLS を適用 →これが WLS。
- R では `lm()` の `weights` 引数に $1/h_i$ を指定するだけでよい。
- これを**一般化最小二乗法 (GLS)** の一種と呼ぶ。

WLS の実装例：Example 8.6（金融資産方程式）

- データ：`k401ksubs`（単身世帯 `fsize == 1`）
- モデル：

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 (age - 25)^2 + \beta_3 male + \beta_4 e401k + u$$

- 仮定：$Var(u|inc) = \sigma^2 \cdot inc$（分散が所得に比例）
- 重み：$1/h_i = 1/inc_i$

WLS の実装例: Example 8.6（続き）

```
data(k401ksubs)
# 単身世帯のみを抽出（fsize == 1）
k401_s <- subset(k401ksubs, fsize == 1)

# OLS 推定（比較用）
ols_w <- lm(nettfa ~ inc + I((age-25)^2) + male + e401k, data = k401_s)

# WLS：Var(u|inc) = σ²*inc を仮定
# weights に 1/h_i = 1/inc を指定 → 所得が大きい観測は信頼度が低いとみなす
wls_w <- lm(nettfa ~ inc + I((age-25)^2) + male + e401k,
            data = k401_s, weights = 1/inc)
```

WLS の実装例: Example 8.6（続き）

```
# OLS と WLS の係数・標準誤差を比較
round(cbind(OLS_coef  = coef(ols_w),
            WLS_coef  = coef(wls_w)), 3)
```

```
##                  OLS_coef WLS_coef
## (Intercept)       -20.985  -16.703
## inc                 0.771    0.740
## I((age - 25)^2)     0.025    0.018
## male                2.478    1.841
## e401k               6.886    5.188
```

WLS 結果の解釈

```
# 標準誤差の比較
round(cbind(OLS_se = sqrt(diag(vcov(ols_w))),
            WLS_se = sqrt(diag(vcov(wls_w)))), 3)
```

```
##                   OLS_se WLS_se
## (Intercept)        2.472  1.958
## inc                0.061  0.064
## I((age - 25)^2)    0.003  0.002
## male               2.048  1.564
## e401k              2.123  1.703
```

- inc の係数：OLS $\approx$ 0.82 $\rightarrow$ WLS $\approx$ 0.79（高所得層に少ない重みをつけた結果）
- WLS の標準誤差は OLS より小さい $\rightarrow$分散の構造を正しく利用すれば効率が上がる。

実行可能な一般化最小二乗法 (FGLS)

- $h(\mathbf{x})$ の形が未知の場合、データから推定する → **FGLS（実行可能 GLS）**。
- 分散関数のモデル化：
  $Var(u|\mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k)$
  - 指数関数を使うので推定された分散は常に正。
- **FGLS の手順：**
  1. OLS を実行し残差 $\hat{u}$ を取得。
  2. $\log(\hat{u}^2)$ を独立変数で回帰し、予測値 $\hat{g}_i$ を得る。
  3. $\hat{h}_i = \exp(\hat{g}_i)$ を計算（分散の推定値）。
  4. 重み $1/\hat{h}_i$ で WLS を実行。

FGLS の実装例：Example 8.7（タバコ需要関数）

- データ：smoke
- モデル：$cigs = \beta_0 + \beta_1 \log(income) + \beta_2 \log(cigpric) + \beta_3 educ + \beta_4 age + \beta_5 age^2 + \beta_6 restaurn + u$

```
data(smoke)

# ステップ 1：OLS 推定（残差を取得するため）
ols_c <- lm(cigs ~ log(income) + log(cigpric) + educ +
              age + I(age^2) + restaurn, data = smoke)
```

```r
# ステップ 2：残差の 2 乗を対数変換 → log(û²)
# log(0) を避けるため resid()^2 が 0 になる場合は注意（通常は問題ない）
log_u2 <- log(resid(ols_c)^2)

# ステップ 3：log(û²) を元の説明変数で回帰（分散関数を推定）
# 予測値 ghat = log(h_i) の推定値
var_reg <- lm(log_u2 ~ log(income) + log(cigpric) + educ +
                age + I(age^2) + restaurn, data = smoke)

# ステップ 4：hhat = exp(ghat) → 各観測の分散推定値（常に正）
hhat <- exp(fitted(var_reg))
```

```r
# ステップ 5：WLS（重み = 1/hhat）で本来のモデルを推定
# 分散が大きいと推定された観測には小さな重みがつく
fgls_c <- lm(cigs ~ log(income) + log(cigpric) + educ +
               age + I(age^2) + restaurn,
             data = smoke, weights = 1/hhat)

# OLS と FGLS の係数比較（テキスト式 (8.35)・(8.36) に対応）
round(cbind(OLS  = coef(ols_c),
            FGLS = coef(fgls_c)), 3)
```

```
##                   OLS   FGLS
## (Intercept)  -3.640  5.635
## log(income)   0.880  1.295
## log(cigpric) -0.751 -2.940
## educ         -0.501 -0.463
## age           0.771  0.482
## I(age^2)     -0.009 -0.006
## restaurn     -2.825 -3.461
```

FGLS 結果の解釈

- OLS の `log(income)` 係数 $\approx$ 0.880（有意でない）$\rightarrow$ FGLS では $\approx$ 1.30（有意）
- BP 検定で強い不均一分散を確認済み（$LM \approx 32.3$, $p < 0.001$）。
- FGLS は分散の大きい観測の影響を抑えることで、係数推定を改善。
- OLS と FGLS の係数が大きく異なる場合は、モデルの関数形の誤特定の可能性もある点に注意。

まとめ

- 不均一分散下では、OLS 推定量は不偏だが効率的ではなく、通常の検定が使えない。
- **解決策 1:** 頑健な標準誤差を用いる（最も一般的）。
- **解決策 2:** 不均一分散の検定を行い、必要に応じて WLS/FGLS で効率性を高める。
- 実証分析では、まずは頑健な標準誤差で結果を確認するのが定石である。