# R による頑強な標準誤差の計算と不均一分散の検定

R による不均一分散の検定方法と最小二乗の推定量の頑強な標準誤差の計算の仕方について述べる

## 1．不均一分散の検定

### 1.1．分析の準備、データの読み込み

データは Web に載せてある makerdata03.txt を用いる。これは医薬品メーカーの売上高と経常利益のデータである。このデータを読み込む。データのあるディレクトリに作業ディレクトリを移動させ(dir()コマンドで確認して)

```
> mdata=read.table("makerdata03.txt",header=T,row.names="社名")
```

とする。読み込んだデータの最初の 5 行を確認するには

```
> head(mdata,5)
```

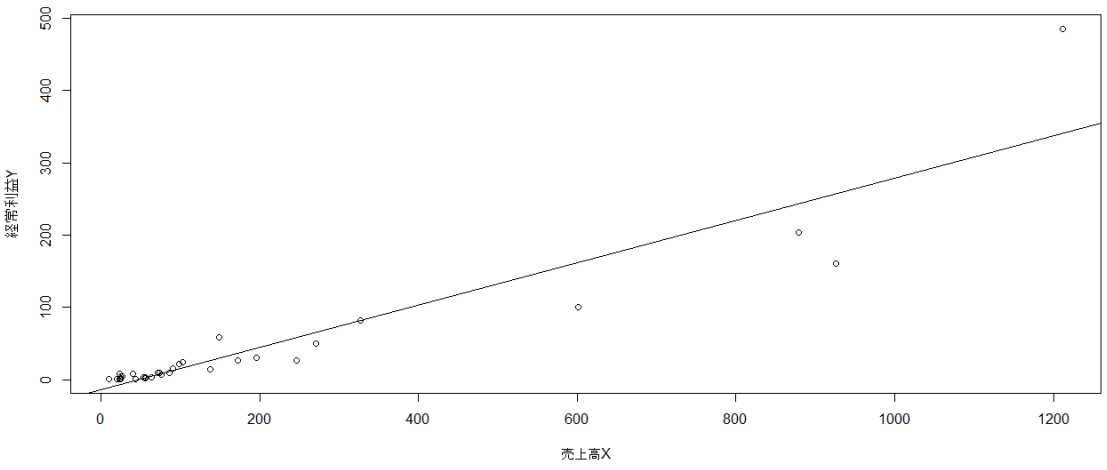
とする。以下のデータが読み込まれているのがわかる。

| | 売上高 X | 経常利益 Y |
|---|---|---|
| 1.武田薬品工業 | 1212 | 485 |
| 2.アステラス製薬 | 879 | 203 |
| 3.大日本住友製薬 | 246 | 27 |
| 4.塩野義製薬 | 196 | 30 |
| 5.田辺製薬 | 172 | 27 |

とりあえず経常利益 Y に売上高 X を線形回帰したモデルを最小二乗法によって推定し、推定した直線を経常利益 Y と売上高 X の散布図に書き入れてみよう。

```
> result = lm(経常利益 Y~売上高 X,data=mdata)
> plot(経常利益 Y~売上高 X,mdata)
> abline(result)
```

以下の図を得る。



売上高が大きくなると直線からのばらつきが大きくなっているのが見て取れる。

以下ではパッケージ lmtest にある関数 bptest を用いて、誤差項の不均一分散を検定するためのブルーシュ・ペイガン検定を計算する。

### 1.2 パッケージ lmtest のインスツール

まず R のパッケージ lmtest をインスツールする。パッケージとは通常の R には含まれていない追加的な R のコマンドの集まりのようなものである。R には追加的に 600 以上のパッケージが用意されており、それぞれ分析の目的に応じて R にパッケージを追加していくことになる。

インターネットに接続してあるパソコンで R を起動させ、「パッケージ」→「パッケージのインスツール」をクリックすると R がミラーサイトの指定を要求してくるので適当に選ぶ(特に理由がなければ Japan(TOKYO)を選んでおけばよい)。次にパッケージの名前を指定するよう要求されるので、ここで lmtest を選択する。すると R の console 上でいろいろとインスツールの途中経過が表示され、lmtest の R へのインスツールが終わる。

次にダウンロードしたパッケージを使うために R コンソール(コマンドウィンドウ)に

```
> library(lmtest)
```

と入力すると(再びコマンドウィンドウ上にいろいろと表示され)、パッケージ lmtest を使用できる様になる(代わりに「パッケージ」→「パッケージの読み込み」としてそこで lmtest を選択してもよい)。

### 1.3. bptest 関数による不均一分散の検定

bptest 関数によってブルーシュ・ペイガン検定統計量を計算するには、先ほど最小二乗推定した結果である result を用いて、以下のように入力すればよい

```
> bptest(result)
```

結果は

```
  studentized Breusch-Pagan test

  data:  result
  BP = 23.507, df = 1, p-value = 1.244e-06
```

と出力される。ここで BP がブルーシュ・ペイガン検定統計量の値、p-value がその検定の p 値である。このように bptest 関数を用いた場合、bptest はブルーシュ・ペイガン検定の対立仮説のもとでの分散モデルの説明変数として result を計算するときに用いた全ての説明変数を自動的に用いている(今回の場合であれば「売上高 X」)。その他の変数を用いたい場合は、引数として varformula=~X+Z(X や Z が使用したい変数のデータ)と入力する。例えば、被説明変数の「経常利益 Y」を用いたい場合は

```
> bptest(result, varformula=~経常利益 Y,data=mdata)
       studentized Breusch-Pagan test

data:  result
BP = 28.537, df = 1, p-value = 9.193e-08
```

とする(経常利益 Y の前に 〝 ~ 〝 がついていることに注意。また、data=mdata がないと経常利益 Y が R に認識されないことにも注意)。「経常利益 Y」と「売上高 X」の両方を用いたいときは。

```
> bptest(result, varformula=~経常利益 Y+売上高 X,data=mdata)
        studentized Breusch-Pagan test

data:  result
BP = 28.593, df = 2, p-value = 6.181e-07
```

とすればよい。いずれの場合も均一分散の帰無仮説は棄却される。

ここで結果が studentized Breusch-Pagan test と表示されていることに注意しよう。実はここで行われた検定は、正確に言うと Breusch-Pagan 検定そのものではなく、その改良版(student 化されたもの)である。もともと提案された(student 化されていない)オリジナルの Breusch-Pagan 検定を計算するには

```
> bptest (result,studentize=FALSE)
```

とする。結果は

```
        Breusch-Pagan test

data:  result
BP = 122.36, df = 1, p-value < 2.2e-16
```

のようになる。結果の表示から studentized が取れていることに気づくだろう。

## 2. 頑強な標準誤差の計算

次にホワイトの不均一分散に頑強な標準誤差の計算の仕方を説明する。そのために lmtest に加えて、もう一つパッケージを読み込む。読み込むパッケージは sandwich であり、その中の vcovHC 関数を用いる。

lmtest の時と同様に sandwich パッケージを読み込む。読み込んだら、ブルーシュ・ペイガン検定の時と同様、推定結果 result を使って

```
> vcovHC(result,type="HC0")
```

と入力すると

```
             (Intercept)      売上高 X
(Intercept)  43.1384490 -0.380722301
売上高 X       -0.3807223  0.003849222
```

と出力される。これがホワイトの方法により計算した(係数の推定量の)「分散共分散行列」である(標準誤差ではないことに注意)。標準誤差は対角成分(分散)の平方根を取ることにより計算できる。例えば売上高 X の係数の推定量の標準誤差は

```
> sqrt(0.003849222)
[1] 0.0620421
```

と計算できる。引数 type により様々な不均一分散に頑強な推定量の分散共分散行列を計算することができる。選択できるものは HC0,HC1, HC2, HC3, HC4 および const がある。

const は均一分散を仮定した時の分散共分散行列（つまり lm でも計算される通常のものと同じ）を出力する。HC0 はホワイトの方法、HC1 ~ HC4 はその改良版である。シミュレーション実験などによると HC0 ~ HC3 の中では HC3 のパフォーマンスが小標本では一番良いとされている。さらに HC4 は HC3 の改良版なので、さらによいパフォーマンスをすると考えられる。

この分散共分散行列から標準誤差を計算してそれを用いて t 値を計算して不均一分散に頑強な t 検定をすることができる。これをしてくれる関数が coeftest 関数である（これ自体は lmtest パッケージに含まれている）。これは t 検定を行う関数だが、そこでどのような標準誤差（分散共分散行列）を用いるか指定することができる。例えば標準誤差の計算に HC0 を用いて t 統計量を計算し、t 統計量が漸近的に標準正規分布に従うことを用いて、t 検定を行うには

```
> coeftest(result, df=Inf, vcov=vcovHC(result,type="HC0"))
```

と入力する。結果は

```
z test of coefficients:

               Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -13.349673   6.567987  -2.0325    0.0421 *
売上高 X       0.292538   0.062042    4.7152 2.415e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

となる。Std. Error の列が標準誤差を表している。売上高 X の所を見ると先ほど計算した標準誤差と同じであることがわかる。引数 df は検定に用いる t 分布の自由度を指定する。何も入力しなければ result で使用されたものと同じ自由度を使用する。ここでは標準正規分布を用いるので自由度を無限大に設定する（t 分布は自由度が無限大の時に標準正規分布になる）。Inf は無限大を表す。vcov で統計量を計算するための分散共分散行列を指定している。

**練習問題**

**問題１**　不均一分散を解消するための一つの方法として、対数変換をするという方法がある。makerdata03.txt のデータに対して

$$\log(\text{経常利益 } Y_i) = \alpha + \beta \log(\text{売上高 } X_i) + u_i$$

という回帰モデルを推定し、不均一分散を検定しなさい。結果はどのようになるか？

**問題２**　頑強な分散共分散行列として vcovHC で type= "HC3"として売上高 x の係数の推定量について t 検定を行いなさい。また type= "HC4" として同様に検定を行いなさい。結果はどのようになるか？