Rによる重回帰分析(最小二乗法)

1. 相関行列、分散共分散行列、平均の計算

2. 最小二乗法によって重回帰モデルの係数を推定する

について簡単に説明する。

Rによる単回帰分析を最小二乗法で行うやり方は説明した。ここではRによる重回帰分析のやり方を説明する。重回帰分析とは説明変数の数が2個以上ある場合の事である。ここでは data01.txt にあるデータを用いて説明する。data01.txt を開くと

```
# 北海道・東北・関東14都道県の農業粗生産額、農家数、耕地面積および専業農家の割合(2004年)

# 農林水産省「農業構造動態調査報吉書」
# 農林水産省「生産農業所得統計」
# 都道府県名 農業粗生産額(億円) 農家数(10戸) 耕地面積(100ha) 専業農家の割合(%)
          Y    X1    X2    X3
北海道 10942  5799 11720  49.2
青森    2953  5544  1597  17.7
岩手    2619  7087  1571  13.4
宮城    2101  6758  1386  10.2
              ・・・・・
```

のようになっている。

データの読み込み

単回帰の場合と同様、まず作業ディレクトリをデータの置いてあるディレクトリに移動する。移動したらそのディレクトリにデータあるかどうかを

```
> dir()
```

で確認する。次に `read.table()` によってdata01.txtのデータを読み込みdata01という名前を付ける。

```
> data01=read.table("data01.txt",header=T,skip=4)
```

データが読み込まれた事を確認するには

```
> data01
```

と入力して Enterーキーを押すと

```
           Y    X1    X2    X3
北海道 10942 5799 11720  49.2
青森    2953 5544  1597  17.7
岩手    2619 7087  1571  13.4
宮城    2101 6758  1386  10.2
          …
```

と出力され、データが読み込まれていることがわかる。

1. 相関行列、分散共分散行列、平均の計算

data01 にある Y, X1, X2, X3 という 4 つの系列の相関行列を計算するには

```
> cor(data01)
```

と入力する。すると

```
            Y          X1         X2          X3
Y  1.0000000  0.3448498  0.9506409   0.8153449
X1 0.3448498  1.0000000  0.1656063  -0.1898360
X2 0.9506409  0.1656063  1.0000000   0.8282634
X3 0.8153449 -0.1898360  0.8282634   1.0000000
```

と出力される。相関行列なので、例えば Y の行と X1 の行がクロスする部分にある数字 0.3448498 というのは Y と X1 の相関係数を表す。

同様に Y, X1, X2, X3 の分散共分散行列を計算するには

```
> var(data01)
```

と入力する。すると

```
              Y           X1           X2            X3
Y  6407652.09  2030947.445  6894622.26   20405.15000
X1 2030947.45 5413004.863   1103927.36   -4366.63462
X2 6894622.26 1103927.357   8208977.61   23461.83462
X3   20405.15   -4366.635     23461.83     97.74577
```

と出力される。今回は分散共分散行列なので Y の行と X1 の行がクロスする部分にある数字は Y と X1 の共分散を表す。

またそれぞれの系列の平均を計算するには mean() 関数を用いる事ができる。例えば Y の平均を計算したい場合は

```
> mean(data01$Y)
```

と入力すればよい。すると

```
[1] 2972.643
```

のように出力される。またそれぞれの列のデータにおける平均を一度に計算するには colMeans() 関数を用い

```
> colMeans(data01)
```

と入力すると

```
         Y        X1       X2       X3
  2972.643 5789.357 1926.071   18.850
```

と出力される。それぞれの系列名の下にある数字がそれぞれの系列の標本平均である。

2. 最小二乗法によって重回帰モデルの係数を推定する

次に最小二乗法によって重回帰モデルの係数を推定するやり方を説明する。モデルは

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \varepsilon$$

というモデルであり、推定しようとしているのは係数の β_0、β_1、β_2、β_3 である。この推定を最小二乗法を使って推定するには単回帰の時と同様 lm( )を用いる。推定結果に results という名前を付けるとすると

```
> results=lm(Y~X1+X2+X3, data=data01)
```

と入力する。2 つ目の引数として『data=data01』とあるが、これは Y、X1、X2、X3 という系列が data01 というデータにあるというのを R に教えるためである。『data=』の部分は省略して

```
> results=lm(Y~X1+X2+X3, data01)
```

と入力しても同じ推定を行う。推定結果は

```
> summary(results)
```

で確認する。

```
Call:
lm(formula=Y ~ X1 + X2 + X3, data = data01)

Residuals:
   Min     1Q Median     3Q    Max
-450.5 -303.8  -16.6  286.3  604.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.230e+03  6.120e+02  -3.644 0.004507 **
X1           3.728e-01  6.059e-02   6.153  0.000108 ***
X2           4.635e-01  8.621e-02   5.376  0.000312 ***
X3           1.142e+02  2.509e+01   4.549  0.001059 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 404.1 on 10 degrees of freedom
Multiple R-squared: 0.9804,    Adjusted R-squared: 0.9745
F-statistic: 166.7 on 3 and 10 DF,  p-value: 7.765e-09
```

と出力される。Estimate の列の数値が上から順に β_0、β_1、β_2、β_3 の推定値である。Pr(>|t|) は t 統計量の P 値であるが、全て 0.01 より低い、つまり全て有意水準 1%で真の係数のが値が 0 (つまりその説明変数は Y に何の影響も与えない)という帰無仮説は棄却される事がわかる。一般にこの数字が小さいほどその説明変数は説明力があると解釈できる。

推定結果のうち係数の推定値だけを見たい場合は summary()ではなく coefficients()を用いる。

```
> coefficients(results)

(Intercept)            X1            X2            X3
-2230.1641197     0.3727748     0.4634662   114.1650777
```

(coefficients とは英語で係数という意味である)。 また当てはめ値や残差を見るには predict()と residuals()を用いる。これらは

```
> predict(results)

      北海道       青森          岩手       宮城        秋田       山形       福島
10980.3026  2597.3766  2669.6081  2095.8957  2238.4560  1775.7402  2989.5213
      茨城       栃木          群馬       埼玉        千葉       東京       神奈川
4191.3815  2251.6449  2723.6874  2308.9831  3619.3319   221.6502   953.4205

> residuals(results)

      北海道       青森          岩手       宮城        秋田       山形       福島
-38.302643  355.623424 -50.608074  5.104321 -450.455974  364.259827 -421.521333
      茨城       栃木          群馬       埼玉        千葉       東京       神奈川
11.618542  517.355146 -442.687398 -340.983143  604.668054  78.349789 -192.420537
```

のように出力される(また『 prevalue = predict(results)』などのようにすれば当てはめ値に prevalue というような名前を付けることもできる。以後は prevalue と入力すれば同じ結果が出力される)。

説明変数が多くなってくると全て書き込むのに手間がかかる場合がある。その場合

```
> results=lm(Y~., data=data01)
```

と入力すればよい(『.』で説明変数として data01 にある系列で Y 以外の全てを使うという事を意味する)。また逆に、説明変数として Y 以外の全部を用いず、例えば X2 と X3 だけ用いたいような場合は

```
> results=lm(Y~X2+X3, data=data01)
```

もしくは

```
> results=lm(Y~.-X1, data=data01)
```

のように入力すればよい。後者は X1 を除いてすべての変数を用いるというコマンドになる。また例えば $Y = \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \varepsilon$ のように切片を除いたモデルの係数を推定するには

```
> results=lm(Y=X1+X2+X3-1,data=data01)
```

のように最後に『-1』を付ければよい。

練習問題

ファイル data02.txt にあるデータについて先ほどと同じことをやってみる。

補足: data02.txt をそのまま上記のやり方で読み込んで分析すると実はうまくいかない。もとのテ

キストファイルを開いて country の文字列を消すか、

```
> data02=read.table("data02.txt",header=T,skip=6,row.names="country")
```

のように打ち込んで読み込むとうまくいく。