

# Rと重回帰分析

## 1. 重回帰分析

説明変数が複数である回帰分析を重回帰分析と呼ぶ。重回帰分析も単回帰分析と同様に線形と非線形に分けられるが、特別な説明がない限り、一般的には線形重回帰分析を略して重回帰分析と言う。重回帰分析では観測データが次の式で表現できることを前提としている。

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n + \varepsilon$$

あるいは次のように定数  $a_0$  が無い式にすることもできる。

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n + \varepsilon$$

回帰分析で求める回帰式は次に示すような式である。

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$$

回帰式の係数  $a_0, a_1, a_2, \dots, a_n$  は単回帰の場合と同じく観測値  $y$  と回帰式による予測値  $\hat{y}$  との差を最小にする最小2乗法により求めることもできる。その数式展開は本稿では略する。

単回帰分析では、身長を説明変数、体重を被説明変数(目的変数)とした例を用いた。ここでは次の表1に示す、身長、ウエストを説明変数、体重を被説明変数とした重回帰分析を行ってみよう。

表1 体重、身長、ウエストデータ

体重	身長	ウエスト
50	165	65
60	170	68
65	172	70
65	175	65
70	170	80
75	172	85
80	183	78
85	187	79
90	180	95
95	185	97

まず、次のようにR上でデータセットを作成する。もちろん、単回帰で用いたデータセットに1列(ウエスト)を付け加える方法で作成することもできる。

```
>体重<-c(50,60,65,65,70,75,80,85,90,95)
>身長<-c(165,170,172,175,170,172,183,187,180,185)
>ウエスト<-c(65,68,70,65,80,85,78,79,95,97)
> taikei2<-data.frame(体重,身長,ウエスト)
> taikei2
```

```
  体重 身長  ウエスト
1   50 165     65
2   60 170     68
<後略>
```

まず、データの変数間の関係を考察するため、相関行列と対散布図を求める。相関は関数 **cor** を用いて求める。

```
> round(cor(taikei2),4)
      体重   身長  ウエスト
体重  1.0000 0.8790  0.8975
身長  0.8790 1.0000  0.5944
ウエスト 0.8975 0.5944  1.0000
```

相関係数は、相互間の線形関係に関する測度であり、非線形関係の場合は正しい相関関係を求めることができない。非線形関係をマクロ的に考察するには対散布図が便利である。

```
> pairs(taikei2,pch=21,bg="red",cex=2)
```

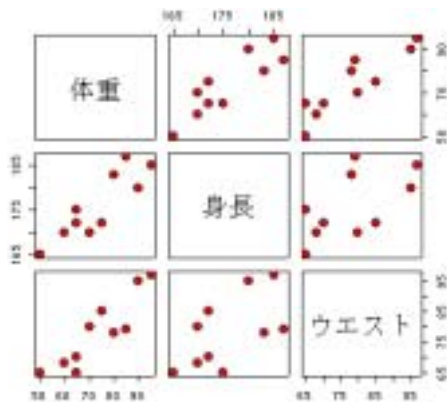


図1 taikei2 データの対散布図

相関係数と対散布図からわかるように、体重と身長、体重とウエストの間には強い線形的相関がある。また、身長とウエストの相関係数が約 0.59 で、相関関係が見られる。

ここでは体重を被説明変数(目的変数、従属変数、応答変数)とした回帰分析を行うことにする。データセットの中の体重を被説明変数、それ以外の全ての変数を説明変数とした場合は、関数 lm を用いて次のように重回帰分析行うことができる。

```
>(taikei2.lm<-lm(体重~.,data=taikei2))
Call:
lm(formula = 体重 ~ ., data = taikei2)
Coefficients:
(Intercept)      身長      ウエスト
-161.670         1.022         0.709
```

R の初心者にとって日本語文字を使用する際に特に注意すべきなことは、日本語文字以外は全て半角文字にすることである。例えば、上記の lm 関数の中の「~」「.」「,」は全て半角である。関数の中の「.」は taikei2 データの中の、被説明変数以外全ての変数を説明変数として用いることを示す。

返された結果 Coefficients が回帰式の係数であり、これを用いた回帰式は次のように構築する。

$$\text{体重} = -161.670 + 1.022 \times \text{身長} + 0.709 \times \text{ウエスト}$$

回帰式の当てはまりの良さが、どの程度であるかは回帰分析の要約情報から読み取られる。

```
> summary(taikei2.lm)
Call:
lm(formula = 体重 ~ ., data = taikei2)
Residuals:
    Min       1Q   Median       3Q      Max
-3.00359 -0.56141  0.07964  1.10466  1.77918
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -161.66981   13.59310  -11.89 6.75e-06 ***
身長          1.02172    0.08960   11.40 8.95e-06 ***
ウエスト      0.70906    0.05729   12.38 5.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.6 on 7 degrees of freedom
Multiple R-Squared:  0.9901,    Adjusted R-squared:  0.9872
F-statistic: 348.7 on 2 and 7 DF,  p-value: 9.782e-08
```

返された決定係数(Multiple R-Squared)は 0.9901、調整済みの決定係数(Adjusted R-squared)は0.9872である。この値は、身長のみを説明変数とした単回帰の場合の0.7726、0.7442 より大きく増加している。

重回帰分析の回帰診断は、単回帰の場合と同じく関数 plot(あるいは plot.lm)を用いて、回帰診断図(diagnostic plots)を作成し考察を行うことができる。

```
> par(mfrow=c(2,2),oma = c(1,1,2,1),mar = c(4, 4, 2, 1))
> plot(taikei2.lm,pch=21,bg=2,col=2,cex=1.5)
```

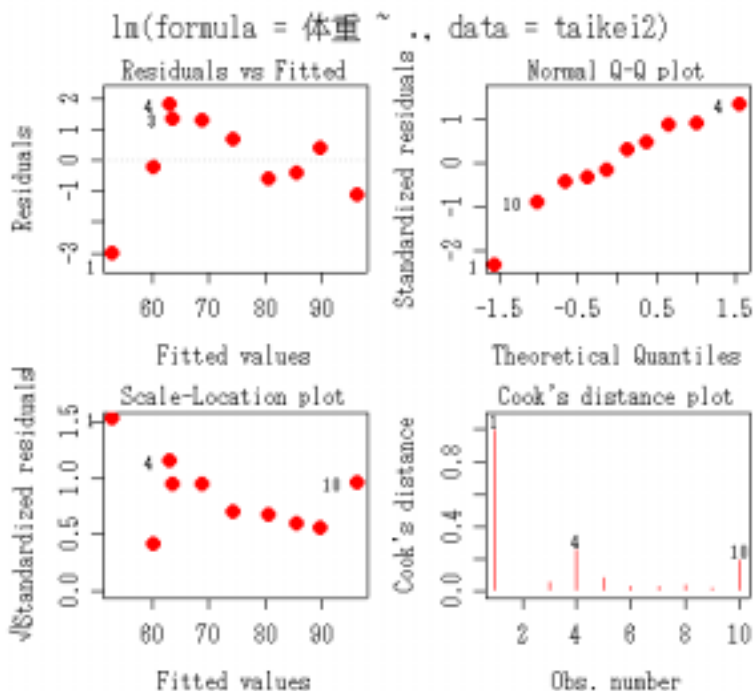


図2 回帰診断図

残差の散布図からわかるように、残差が最も大きいのは約3である。Cookの距離と残差の散布図から個体1の影響が大きいことが読み取られる。実際の問題について本格的に分析を行う際には、このような個体の影響について詳細に分析を行うことが必要である。

## 2. 相互作用モデル

taikei2のデータでは、説明変数身長とウエストの間にも相関関係がある。このような説明変数間の相関関係を相互作用(interaction)と呼ぶ。当てはまりのよいモデルを作成するためには、相互作用効果を考慮した回帰分析を行うことも可能である。説明変数が2つで

ある場合の一般式を次に示す。

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2$$

R では相互作用を考慮した回帰係数を簡単に求めることが可能である。次に相互作用を考慮した書式を示す。関数の中の「^2」が相互作用の指定である。

```
>(taikei2.lm2<-lm(体重~.^2,data=taikei2))
```

```
Call:
```

```
lm(formula = 体重 ~ .^2, data = taikei2)
```

```
Coefficients:
```

(Intercept)	身長	ウエスト	身長:ウエスト
-372.94598	2.21610	3.46816	-0.01555

```
>summary(taikei2.lm2)
```

```
Call:
```

```
lm(formula = 体重 ~ .^2, data = taikei2)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-1.3925	-0.7506	0.1545	0.5282	1.5525

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-372.94597	85.83156	-4.345	0.00485	**
身長	2.21610	0.48648	4.555	0.00387	**
ウエスト	3.46816	1.11361	3.114	0.02073	*
身長:ウエスト	-0.01555	0.00627	-2.480	0.04784	*

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.214 on 6 degrees of freedom
```

```
Multiple R-Squared: 0.9951, Adjusted R-squared: 0.9926
```

```
F-statistic: 405.5 on 3 and 6 DF, p-value: 2.582e-07
```

決定係数は 0.9951、調整済みの決定係数は 0.9926 で、相互作用効果を考慮していない場合の 0.9901、0.9872 より増加している。調整済みの決定係数が 0.9926 で 1 に近いことからこのモデルがデータに非常によく当てはまっていると言えるであろう。次に回帰診断図を示す。

```
> plot(taikei2.lm2,pch=21,bg=2,col=2,cex=1.5)
```

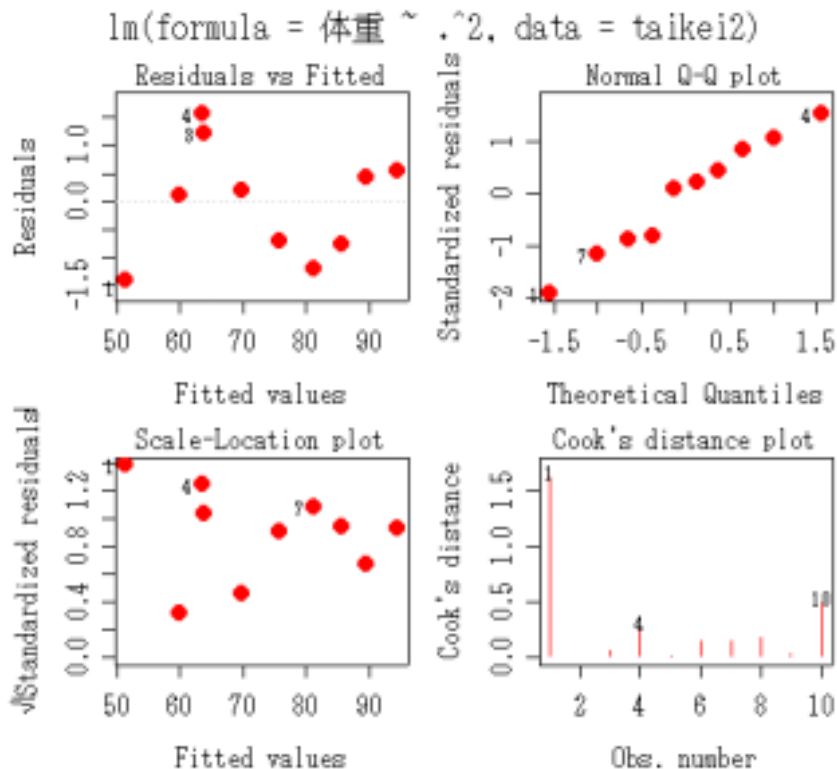


図3 taikei2.lm2 の回帰診断図

最大の残差が約 1.5 であることから相互作用を考慮していないモデルより予測の精度がよいことがわかる。Cook の距離で個体 1 の値が 1.5 を超えており、個体 10 の値も約 0.5 前後になっている。しかし、今問題のようなデータの数が少ない場合は、Cook の距離が大きい個体を外して、回帰分析を行うことに大きい意味があるとは言い切れない。

この例では 2 つの説明変数全てが回帰モデルの構築に非常に役に立っている。しかし、説明変数が多い場合は、説明変数が回帰モデルの構築に役に立たない場合もある。そのときには、どのように回帰モデルに役立つ説明変数を選択するかを考えなければならない。

### 3 . 変数・モデルの選択

変数の選択は、変数を入れ換えながら回帰モデルを構築し、そのモデルを比較してより当てはまりがよいモデルを採択する。よって、変数を選択することは、モデルを選択することに等しい。

モデルの選択とは、真のモデルがあって、それに近似する複数のモデルが構築された場合、複数の近似モデルの中から真のモデルに最も近いモデルを見つけ出すことである。

調整済みの決定係数を回帰モデルの選択基準とすることもできるが、R ではモデルの選

択基準として広く知られている、元統計数理研究所長赤池弘次氏が提案した AIC(Akaike's Information Criterion)を用いて回帰モデルを選択する関数が用意されている。AIC は次の式で定義されている。

$$AIC = -2 \times (\text{モデルの最大対数尤度}) + 2 \times (\text{モデルのパラメータ数})$$

モデル選択の場合は、AIC の値が小さいモデルがよいモデルであると評価する。

ここでは R の中にあるデータ attitude を用いて説明することにする。データ attitude は無作為に抽出した 30 部門に所属している 35 人の事務員のアンケート回答から得られたデータである。

データセットの中の数値は各質問項目に対する好意的な反応の割合(パーセンテージ)である。データは 30 行 7 列(変数)のデータフレームである。その 7 つの変数を次に示す。

rating : 総合評価  
complaints : 従業員の苦情の取り扱い  
privileges : 特別な特権は許さない  
learning : 学習の機会  
raises : 能力に基づいた昇給  
critical : 加重  
advancel : 昇進

```
>data(attitude)
>attitude[1,]
  rating complaints privileges learning raises critical advance
1     43          51         30      39      61      92      45
```

ここでは、従業員が会社を評価した総合評価(rating)を被説明変数とした回帰モデルを作成し、総合評価がどの項目の影響を大きく受けているかを分析する。

回帰分析をする前に、まず 7 つの変数の相関関係を考察してみよう。相関関係は、相関係数と対散布図で考察できる。まず相関係数を次に示す。

```
> round(cor(attitude),2)
      rating complaints privileges learning raises critical advance
rating    1.00      0.83      0.43      0.62      0.59      0.16      0.16
complaints 0.83      1.00      0.56      0.60      0.67      0.19      0.22
privileges 0.43      0.56      1.00      0.49      0.45      0.15      0.34
learning   0.62      0.60      0.49      1.00      0.64      0.12      0.53
```



対散布図から見られるように **rating** と **complaints**、**learning**、**raises** との相関関係は線形的であると言えよう。

そこで、まず **rating** を被説明変数と残りに全ての変数を説明変数とした回帰分析の情報を求めてみる。

```
>attitude.lm1 <- lm(rating ~ ., data = attitude)
>summary(attitude.lm1 )
Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
    Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708    11.58926   0.931 0.361634
complaints   0.61319     0.16098   3.809 0.000903 ***
privileges  -0.07305     0.13572  -0.538 0.595594
learning     0.32033     0.16852   1.901 0.069925 .
raises       0.08173     0.22148   0.369 0.715480
critical     0.03838     0.14700   0.261 0.796334
advance     -0.21706     0.17821  -1.218 0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-Squared:  0.7326,    Adjusted R-squared:  0.6628
F-statistic: 10.5 on 6 and 23 DF,  p-value: 1.240e-05
```

返された  $\Pr(>|t|)$  は **complaints**、**learning** の項目を除くと、いずれも大きい値になっている。このような場合では、変数を選択して用いる必要がる。変数の優先候補は  $\Pr(>|t|)$  値が小さい順である。変数の選択は、変数増加法、変数減少法などの方法がある。

変数増加法とは、 $\Pr(>|t|)$  値が最も小さい変数から順次に変数を加える方法で、最適のモデルが選択されるまで変数を付け加える操作を繰り返す。変数減少法とは  $\Pr(>|t|)$  値が大きい変数から順次変数を除去する方法で、最適モデルが選択されるまで、変数を削除する操作を繰り返す。

R のパッケージ **stats** には変数を選択する関数 **step** が用意されている。関数 **step** では、モデルの選択は AIC を基準としている。次に関数 **step** の使用例を示す。

```

>attitude.lm2<-step(attitude.lm1)
Start: AIC= 123.36 #全ての変数を用いた場合
rating ~ complaints + privileges + learning + raises + critical + advance
< 中略 >

Step: AIC= 121.45 #1 つの変数を除去した場合
rating ~ complaints + privileges + learning + raises + advance
< 中略 >

Step: AIC= 119.73 #2 つの変数を除去した場合
rating ~ complaints + privileges + learning + advance
< 中略 >

Step: AIC= 118.14 #3 つの変数を除去した場合
rating ~ complaints + learning + advance
< 中略 >

Step: AIC= 118 #4 つの変数を除去した場合
rating ~ complaints + learning
      Df Sum of Sq    RSS    AIC
<none>          1254.65 118.00
- learning     1    114.73 1369.38 118.63
- complaints   1   1370.91 2625.56 138.16

```

このデータの場合は  $\Pr(>|t|)$  が大きいも 4 つの変数を除去した段階での線形モデルが、当てはまりが最も良いと判断され、変数の選択が終了した。その結果、説明変数 complaints、learning を用いたモデルが選択された。次に回帰分析の要約を出力する。

```

> summary(attitude.lm2)
Call:
lm(formula = rating ~ complaints + learning, data = attitude)
Residuals:
    Min       1Q   Median       3Q      Max
-11.5568  -5.7331   0.6701   6.5341  10.3610
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8709     7.0612   1.398   0.174
complaints    0.6435     0.1185   5.432 9.57e-06 ***
learning      0.2112     0.1344   1.571   0.128
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 6.817 on 27 degrees of freedom  
Multiple R-Squared: 0.708, Adjusted R-squared: 0.6864  
F-statistic: 32.74 on 2 and 27 DF, p-value: 6.058e-08

上記の情報をを用いた回帰式を次に示す。

$$\text{rating} = 9.8709 + 0.6435 \times \text{complaints} + 0.2112 \times \text{earning}$$

```
>plot(attitude.lm2)
>par(temp.par)
```

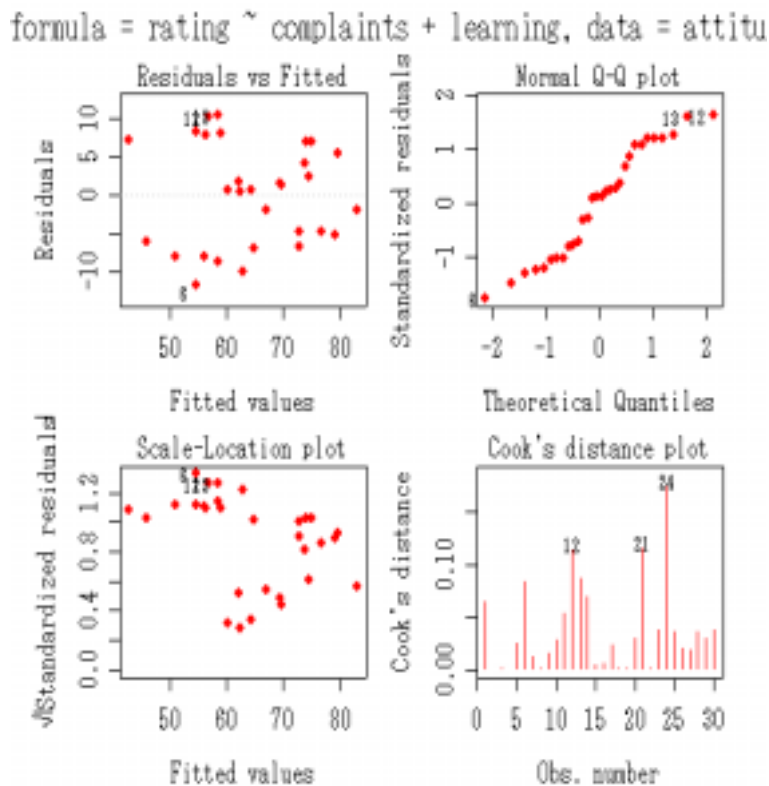


図5 回帰診断図

回帰診断図の中の残差の散布図から作成した回帰式による予測値の最大の誤差が約10ポイントを超えていることがわかる。

アンケート調査のデータとしてはデータが少なすぎるので、この回帰分析の結果で断言することは危険であるが、このデータで読み取られたのは、会社に対する総合評価(rating)は、従業員の苦情の取り扱いに対する満足度(complaints)と学習の機会(learning)が主な要因であること言えよう。