

## 20日目：回帰分析

本日は、単回帰分析と重回帰分析を取り上げます。

回帰分析は、関数 `lm()` を使い、説明される変数（従属変数）と、それを説明する変数（独立変数）を指定することでできます。基本的には、`lm(従属変数 ~ 独立変数, data = ファイル名)` という構造です。独立変数が複数ある重回帰分析の場合、複数の独立変数を「+」でつなぎます。また `data =` は省略可のうで、ファイル名だけを書いても OK です。

なお、この `lm()` 関数の場合も、いったん何かに保存しておいて、`summary` で結果を表示させるといった構造になります。

SPSS では、偏回帰係数と標準偏回帰係数 ( $\beta$ ) は同時に計算されますが、R では、それぞれ別に計算します。データそのままを利用して計算すると、偏回帰係数が計算されます。標準偏回帰係数が欲しい場合には、前もってデータを標準化しておいてから計算します。命令自体は同じなので、ちょっと注意しておく必要があるでしょう。

偏回帰係数を計算する場合は、特に必要はありませんが、標準偏回帰係数を求める場合は、従属変数と独立変数をまとめておいて、それを標準化したデータを新しく準備しておく方が便利でしょう。今回は、`no` 以外のすべての変数を使いますので、これらをまとめて標準化したデータを準備します。

サンプルデータを `x` という名前を読み込み、変数をまとめて標準化したデータを作ります。

```
xn <- c("親の様子", "家にある本", "知識欲", "不可欠さ", "インターネット", "マンガ", "本好き", "読書習慣")
x0 <- x[xn]
x1 <- scale(x0)
x1 <- data.frame(x1)
```

上の2行は以前にもやった、指定した変数のみを取り出す命令です。3行目の `scale` は標準化（平均値0、標準偏差1に変換）を行う命令です。これで標準化が行われた `x1` というファイルができます。しかし、これは行列型のデータとして作られるので、4行目でデータフレームに変換します。

なお、以上の内容は、

```
x1 <- data.frame(scale(x[c("親の様子", "家にある本", "知識欲", "不可欠さ", "インターネット", "マンガ", "本好き", "読書習慣")]))
```

と書いても同じですね。 `x1 <- data.frame(x1)` の `x1` は、`scale(x0)` のことなので `x1 <- data.frame(scale(x0))` と書いても同じで、その `x0` の部分は…とやっていくと、4行を

1行にすることができます。ただし、わかりにくくなりますが…

### ●単回帰

偏回帰係数どうしを比較することはないので、標準偏回帰係数を算出する意味はないでしょうが、とりあえず両方の計算をやってみます。

「読書習慣」を独立変数、「家にある本」を従属変数とした、単回帰分析（偏回帰係数を求める）。

```
sr <- lm(読書習慣 ~ 家にある本, x)
summary(sr)
```

「読書習慣」を独立変数、「家にある本」を従属変数とした、単回帰分析（標準偏回帰係数を求める）。

```
sr.sd <- lm(読書習慣 ~ 家にある本, x1)
summary(sr.sd)
```

このように、違いはファイル名の指定の部分だけです。

### ●重回帰

「親の様子」「家にある本」「知識欲」「不可欠さ」「インターネット」「マンガ」「本好き」を独立変数、「読書習慣」を従属変数とした、重回帰分析（偏回帰係数を求める）。

```
mr <- lm(読書習慣 ~ 親の様子 + 家にある本 + 知識欲 + 不可欠さ + インターネット +
マンガ + 本好き, x)
summary(mr)
```

「親の様子」「家にある本」「知識欲」「不可欠さ」「インターネット」「マンガ」「本好き」を独立変数、「読書習慣」を従属変数とした、重回帰分析（標準偏回帰係数を求める）。

```
mr.sd <- lm(読書習慣 ~ 親の様子 + 家にある本 + 知識欲 + 不可欠さ + インターネット +
マンガ + 本好き, x1)
summary(mr.sd)
```

このように、独立変数は「+」で結びます。もちろん、両者の違いはファイル名の指定の

部分だけです。

さて、結果の読み取りですが、最後の命令の出力を使って説明します。

```
> mr.sd <- lm(読書習慣 ~ 親の様子 + 家にある本 + 知識欲 + 不可欠さ + インターネット + マンガ + 本好き, x1)
> summary(mr.sd)

Call:
lm(formula = 読書習慣 ~ 親の様子 + 家にある本 +
    知識欲 + 不可欠さ + インターネット + マンガ +
    本好き, data = x1)

Residuals:
    Min       1Q   Median       3Q      Max
-1.47123 -0.50470 -0.02041  0.46925  1.78444

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.136e-16  5.709e-02   0.000  1.00000
親の様子      2.610e-01  8.682e-02   3.006  0.00309 **
家にある本    8.947e-03  8.493e-02   0.105  0.91623
知識欲        3.901e-01  7.594e-02   5.137  8.35e-07 ***
不可欠さ     -1.610e-01  7.161e-02  -2.248  0.02600 *
インターネット -2.642e-01  6.574e-02  -4.019  9.12e-05 ***
マンガ       -3.036e-02  6.294e-02  -0.482  0.63027
本好き        4.361e-01  6.694e-02   6.515  9.79e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7266 on 154 degrees of freedom
Multiple R-squared:  0.495, Adjusted R-squared:  0.472
F-statistic: 21.56 on 7 and 154 DF, p-value: < 2.2e-16
```

上から、**Residuals** の部分は、回帰式の誤差の部分の情報です。

次に下から2行が重要になります。これが決定係数 ( $R^2$ ) の検定結果です。今回の場合は、 $R^2=0.495$  であり、 $F(7,154)=21.56$ ,  $p<0.01$  で有意なものであることがわかります。

これが有意であることが確認できたら、**Coefficients** に移ります。**Estimate** が標準偏回帰係数（もちろん、これは標準化されたデータを使っているため）、**t value** と **Pr(>|t|)** が、t 値と t 検定の結果になります。「親の様子」の  $\beta$  は 0.261 で 1% 水準で有意、「家にある本」の  $\beta$  は 0.009 で有意ではなく、知識欲は 0.390 で 1% 水準で有意、などということがわかります。なお、偏回帰係数は論文にも記載することもよくありますので、こっだけ取り出したい (e- の表示で間違えそうだし...) です。これは **lm** の出力の中に、**coefficients** という名前が入っていますので、`round(mr.sd$coefficients, 3)` というように丸めて取り出しておくとも便利かもしれません。

ちなみに **Intercept** は切片のことです。

さて、重回帰分析の場合、多重共線性が問題になることがあります（相互相関の高い変数を独立変数として採用すると、おかしい結果になる…）。相関行列の結果と重回帰分析の結果を比べたりすると、ある程度予測はつきますが、最近ではVIF（variance inflation factor；分散拡大係数）を使って判断するケースが多いようです。VIFは、値が大きくなるほど多重共線性が疑われます。いろいろと資料を探しても、どの値を基準にすべきかという明確な記載はないようですが、一般的に10もしくは4が目安として考えられているようです。

このVIFを求めるためには、`car`というパッケージを使います。

`car`をとってきておいて…

`library(car)`

`vif(mr.sd)`

これで計算してくれます。`vif()`のカッコ内には、重回帰分析の結果のファイルを指定します。ちなみに、標準化したものでもしていないものでも、同じ結果になります。

今回の場合は、いずれもVIFは4未満であり、特に多重共線性の問題はないようです。

本日はここまでにします。