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()を用い，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 = データ名)**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 =は省略可能で，データ名だけを書いても 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う構造になります。

関数 lm()を用いる場合，偏回帰係数と標準偏回帰係数（β）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。なお，同時に計算してくるパッケージもありますが，それは最後に紹介します。

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 という名前で読み込み，変数をまとめて標準化したデータ（xs）を作ります。

```
label_n <- 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[label_n]
x1 <- scale(x0)
xs <- 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，

```
xs <- 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，xs)
summary(sr.sd)
```

このように，違いはデータ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，xs)
summary(mr.sd)
```

このように，独立変数は「+」で結びます。もちろん，両者の違いはデータ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, xs)
> 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 = xs)

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=.495 であり，$F(7,154)=21.56$, $p<.01$ 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」は１％水準で有意，「家にある本」は有意ではなく，知識欲は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，ここだけ取り出したい(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 という名前で入っていますので，round(mr.sd$coefficients, 3)というように丸めて取り出しておくと便利かもしれません。

ちなみに Intercept は切片のことです。

```
> round(mr.sd$coefficients, 3)
   (Intercept)        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
         0.000           0.261           0.009           0.390          -0.161          -0.264          -0.030
      本好き
         0.436
```

もう少し整えられた，きれいな形で結果を出力したいなら，また回帰係数の 95%信頼区間も得たいなら，stargazer というパッケージをインストールして，summary にかえて以下のようにして表示させることもできます。なお，カッコの中が偏回帰係数（この例では mr.sd を使っているので標準偏回帰係数）の 95%信頼区間，上下限です。注意は，アスタリスクのつけ方が独特なところ。通例なら，5%でアスタリスク１つ，２つで 1%です。

```
stargazer(mr.sd, type = "text", single.row=TRUE, ci = TRUE)
```

```
> stargazer(mr.sd, type = "text", single.row=TRUE, ci = TRUE)

================================================
                        Dependent variable:
                    ---------------------------
                              読書習慣
------------------------------------------------
親の様子                 0.261*** (0.091, 0.431)
家にある本                 0.009 (-0.158, 0.175)
知識欲                  0.390*** (0.241, 0.539)
不可欠さ                -0.161** (-0.301, -0.021)
インターネット           -0.264*** (-0.393, -0.135)
マンガ                  -0.030 (-0.154, 0.093)
本好き                  0.436*** (0.305, 0.567)
Constant              -0.000 (-0.112, 0.112)
------------------------------------------------
Observations                    162
R2                             0.495
Adjusted R2                    0.472
Residual Std. Error       0.727 (df = 154)
F Statistic           21.561*** (df = 7; 154)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

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，おかしな結果になる…）。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）を使って判断する場合が多いようです。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，　10 とか 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。もし大きな値が見られるようだったら，その独立変数をどうするかを考えなければならないでしょう。

最後に，ここまで触れてきた指標を一気に計算してくれるパッケージ，関数を紹介しておきます。標準化したデータセットを用意しなくても，標準偏回帰係数（beta）を算出してくれますし，VIF も計算してくれます。パッケージは pequod，関数は lmres です。構造は lm()と同じです。

```
library(pequod)
mr2 <- lmres(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(mr2)
```

出力をこれまで計算してきたものと比べることで，結果の読み取りは問題なく理解できると思います。

今回は，回帰分析の基本的な部分のみを説明してきました。例として行った重回帰分析の変数の選択や投入にはかなり無理があります…。あくまでも例ととっておいてください。

変数の選出方法や投入手順などに関連しては様々な方法がありますので，その情報も知っておかないと実際に分析を行うのは難しいかもしれません。また最近では交互作用項の影響を重回帰分析を使って行う研究もあります。R は様々な回帰分析に対応できますし，web を検索すると様々な情報や使い方が得られます。いろいろと勉強して，使いこなせるようになってください。

本日はここまでにします。