回帰分析と因果効果分析
Regression analysis and Causal effect analysis

2018年5月1日

東京工業大学 福田研究室

鈴木　新

はじめに・目次(contents)

はじめに

- 回帰分析の理論や基本について学習.
- ところどころプログラミングで練習（R）

目次

- 回帰分析とは
- 単回帰式の導出
- プログラミング（単回帰）
- 結果の整合性
- 重回帰分析の導出
- プログラミング（重回帰）
- 練習問題

- First, we study theory and basis of Regression analysis
- And then, we try programming (R)

回帰分析とは(what is regression analysis?)

目的変数yを説明変数xでどのくらい説明できるか

・・・結果と原因の関係式を作る． Formularize relationship between result & factor

アイスがたくさん売れた

Many ice cream was sold

目的変数
Objective variable

暑かったから…？

That's because it was hot.

説明変数
Response variable

$$Y_{(アイスの売り上げ本数)} = a + bX_{(気温)} + e$$

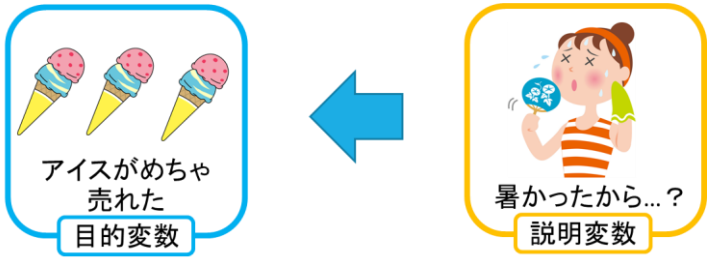
e : 誤差項
Error term

気温とアイスの売れ行きがどのくらい関係しているのか調べる

Try to estimate the relationship between # sold ice cream and the temperature

回帰分析とは(what is regression analysis?)

目的変数yを説明変数xでどのくらい説明できるか
・・・結果と原因の関係式を作る.



説明変数1個・・・単回帰分析
Single regression analysis

$$Y = a + bX + e$$

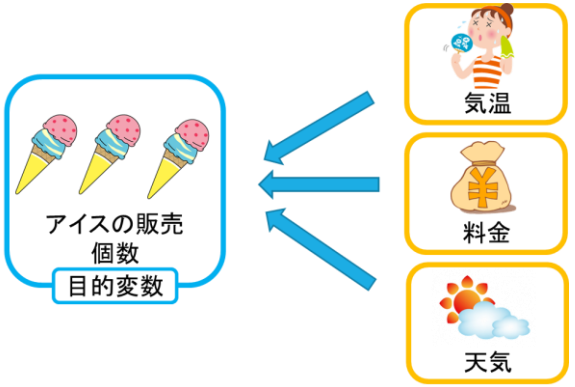
Constant term 定数項
（気温に関係なく
売れるアイスの本数）

回帰係数 Regression coefficient
（気温がアイスの売り上げに
どのくらい影響与えるのか）

説明変数2個以上・・・重回帰分析
Multiple regression analysis

$$Y = a + bX_1 + cX_2 + dX_3 + e$$

気温 temperature　料金 price　日照時間 hours of sunlight

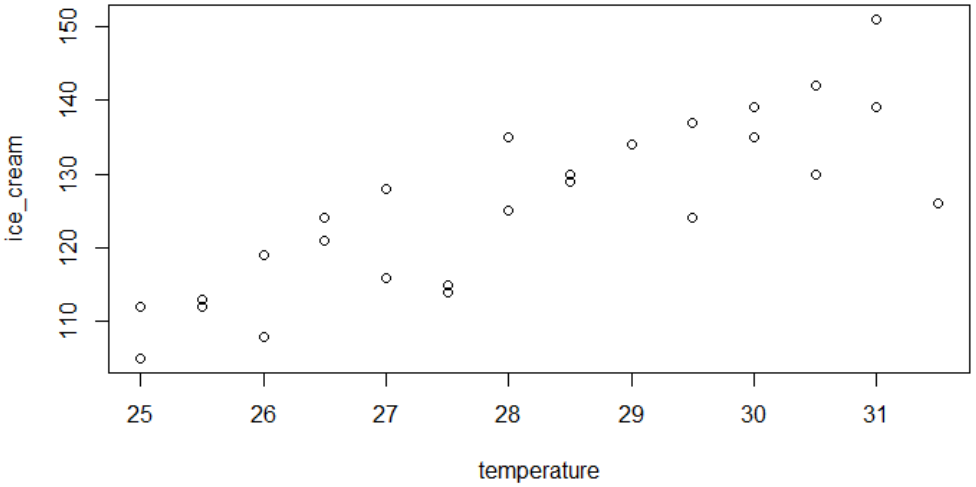


どのようにして回帰式を求めるのか？

単回帰式の導出(derivation of regression model)

アイスの売り上げ本数とその日の気温の関係図をプロット

Write Scatter plot between # sold ice cream & temperature.



目的変数yと説明変数xの間に線形の関係があると仮定

Make assumption “liner relationship” Between objective variable & response variable

$Y = a + bX + e$ 単回帰モデル

$\hat{Y} = a + bX$ 単回帰式

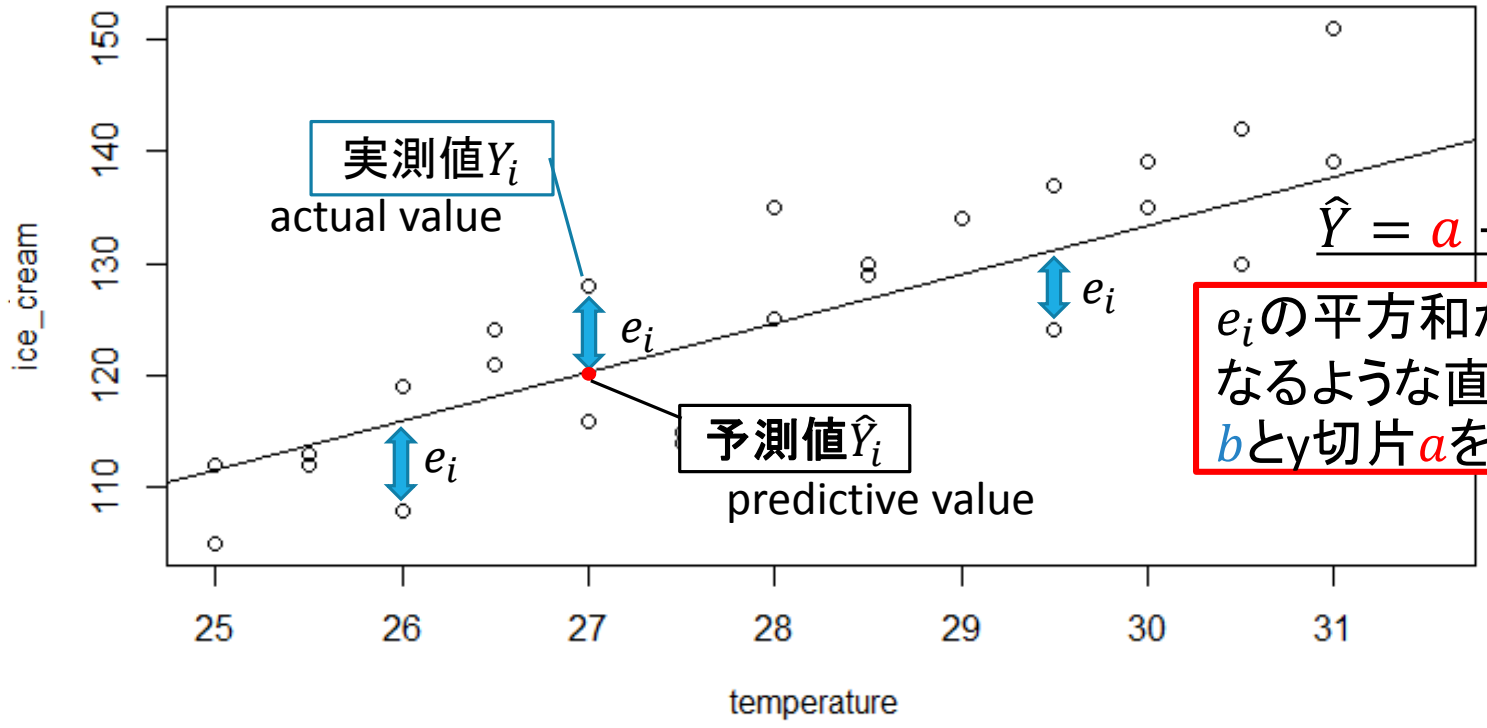
単回帰式の導出(derivation of regression model)

最小二乗法にて，残差平方和が最小となる回帰直線を算出

Calculate regression line by least squares method “minimizing residual sum of squares”

$Y = a + bX + e$ 単回帰モデル　$\hat{Y} = a + bX$　単回帰式

$$min \sum_i {e_i}^2 = min \sum_i (Y_i - \hat{Y}_i)^2$$



単回帰式の導出(プログラミングで)

統計ソフトRでやってみましょう （try programing R）

●ソースコードを開く Open source code by clicking source file

・・・コードファイルをダブルクリックかCtrl+Oでファイル選択

・データの読み込み Read csvfile

```
1  setwd("C:/Users/Owner/Desktop/Dropbox (fukudalab-tokyotech)/基礎ゼミ2018")
2  data.ice = read.csv("ice.csv")
```

CSVファイルの読み込み

ファイルが保存されているフォルダに
（もしくはsessionから選択）

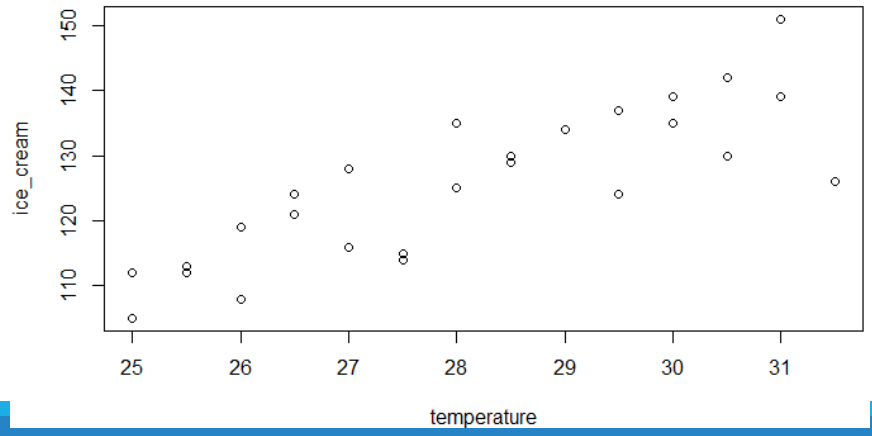
・散布図の記入

```
3  plot(ice_cream ~ temperature, data = data.ice)
```

縦軸　　横軸

名前はcsv
ファイルの
ここに対応

	A	B	C	D
1	ice_cream	temperature		
2	105	25		
3	112	25		
4	112	25.5		



単回帰式の導出(プログラミングで)

統計ソフトRでやってみましょう （try programing R）

・回帰分析～lm()～ Regression Analysis
線形モデルによる回帰を行う

使用するデータ

```
4  result <- lm(ice_cream ~ temperature, data = data.ice)
```

目的変数 Objective variable

説明変数 Response variable

$$Y_{(アイスの本数)} = a + bX_{(気温)} + e$$

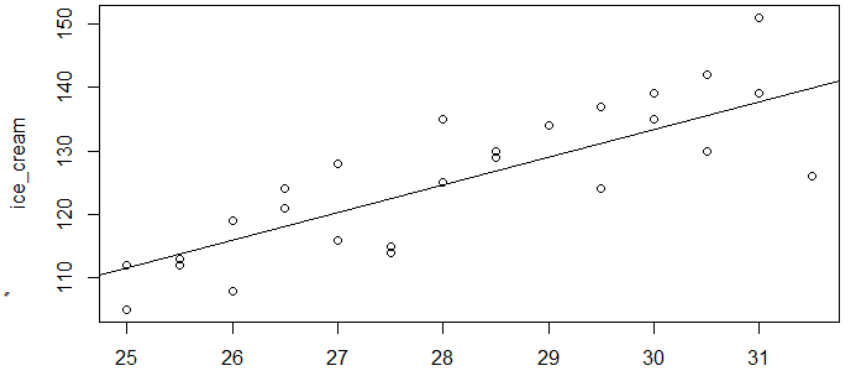
分析結果をresultの中に収納

名前はここに対応

	A	B	C	D
1	ice_cream	temperature		
2	105	25		
3	112	25		
4	112	25.5		

単回帰式の導出(プログラミングで)

・結果の出力
プロット図に結果を記入・・・abline()
結果の詳細・・・summary()




```
5  abline(result) # 推定回帰直線を描く
6  summary(result)
```

切片，回帰係数の結果
Estimate・・・推定値
（interceptがy切片a）
（temperatureが傾きb）

```
> summary(result)

Call:
lm(formula = ice_cream ~ temperature, data = data.ice)

Residuals:
    Min      1Q  Median      3Q     Max
-13.957  -5.877   1.392   5.079  13.218

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.950     18.750   0.157    0.876
temperature     4.349      0.662   6.570 5.75e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.061 on 26 degrees of freedom
Multiple R-squared:  0.6241,    Adjusted R-squared:  0.6096
F-statistic: 43.16 on 1 and 26 DF,  p-value: 5.754e-07
```

$\hat{Y} = 2.95 + 4.349X$ という単回帰式が得られた！

結果の整合性

- **算出した式はどのくらいの精度で予測しているのか，**

⇒決定係数R^2（「予測の当てはまりの良さ」を表す指標）

- **気温はどのくらいアイスの売り上げに影響しているのか**

⇒t値t-value
（説明変数の影響力）

```
> summary(result)

Call:
lm(formula = ice_cream ~ temperature, data = data.ice)

Residuals:
    Min      1Q  Median      3Q     Max
-13.957  -5.877   1.392   5.079  13.218

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.950     18.750   0.157    0.876
temperature    4.349      0.662   6.570 5.75e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.061 on 26 degrees of freedom
Multiple R-squared:  0.6241,    Adjusted R-squared:  0.6096
F-statistic: 43.16 on 1 and 26 DF,  p-value: 5.754e-07
```

tvalue･･･t値

決定係数

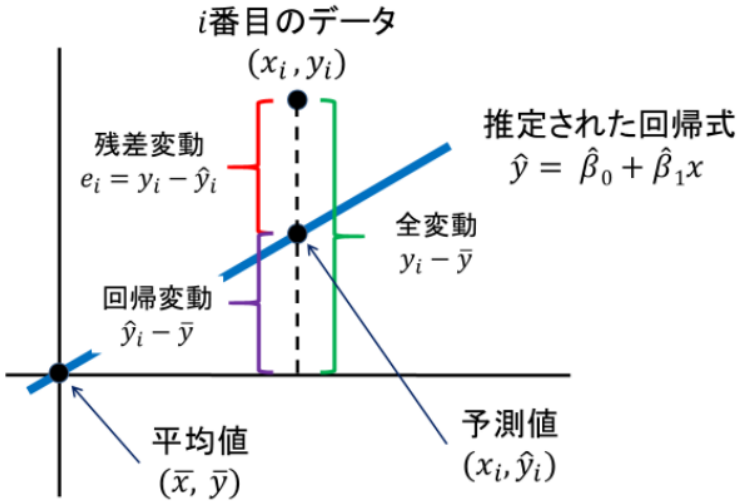
結果の整合性

●決定係数（R-square, R^2）
・・・「予測の当てはまりの良さ」を表す指標.
0~1の値をとり，1に近ければ近いほどモデルの当てはまりが良い
（何%説明変数で説明できてるのか）

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y_i})^2$$

全変動　回帰変動　残差変動

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y_i}-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y_i})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$



https://bellcurve.jp/statistics/course/9706.html

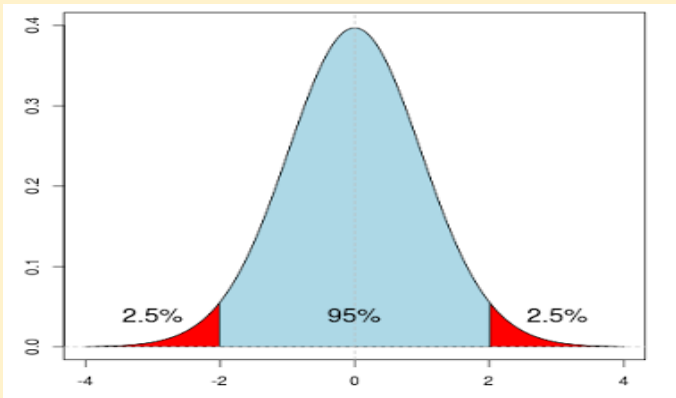
●自由度調整済み決定係数（adjusted R-square, Adjusted R^2）
・・・サンプルサイズや説明変数の数に左右されないよう修正した決定係数

結果の整合性

t値・・・各説明変数が目的変数にどのくらい影響を与えるのかを示す.
大きさ1.96以上なら優位水準5%を満たす（右図の赤い部分内）

$$t\text{値} = \frac{b}{\sigma}$$

パラメーターbを標準誤差σで割ったもの,
t値が大きいほど標準誤差が小さい



http://www.geisya.or.jp/~mwm48961/statistics/bunsan1.htm

Ex)パラメーターb =4，標準誤差σ = 2の時

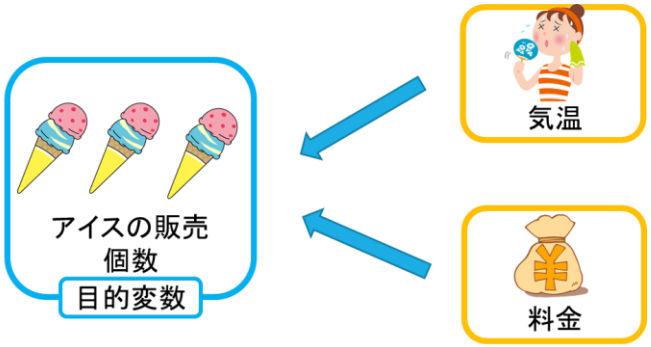
パラメーターは$b \pm \sigma$の値（2~6）をとる可能性95%以上
⇒ 0や負の値になる可能性もなくはないけどほぼない

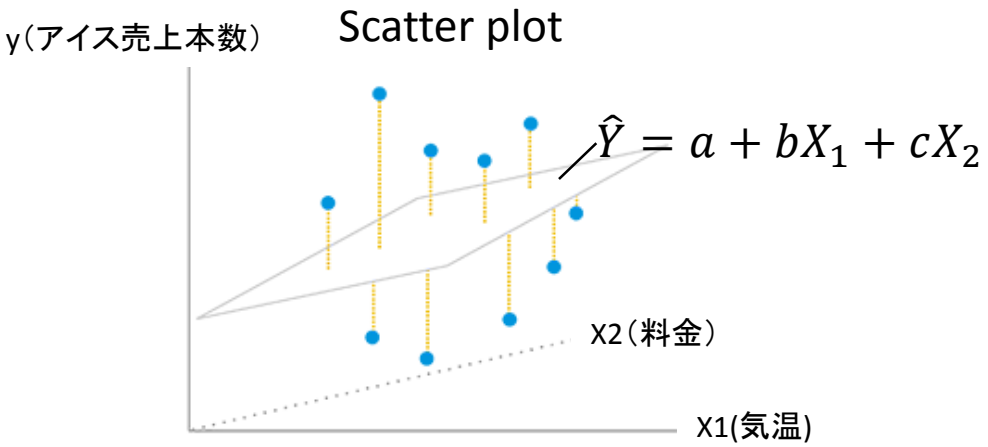
Ex)パラメーターb = 4，標準誤差σ = 3の時

パラメーターは$b \pm \sigma$の値（1~7）をとる可能性約95%以下
⇒ 0や負の値になる可能性高い

重回帰式の導出

アイスの売り上げ本数はその日の気温とアイスの価格に関係するのでは…?





http://xica.net/magellan/marketing-idea/stats/about-coefficient-of-determination/

最小二乗法によって残差の二乗和が最も小さくなる回帰式を求める

$\hat{Y} = a + bX_1 + cX_2$　重回帰式

重回帰式の導出(プログラミングで)

統計ソフトRでやってみましょう （try programing R）

・データの読み込み

```
13  #重回帰分析
14  data.ice2 = read.csv("ice2.csv")
```

使用するデータ

	A	B	C
1	ice_cream	temperature	price
2	105	25	160
3	112	25	130
4	112	25.5	140
5	113	25.5	130
6	119	26	140

・重回帰分析 Regression Analysis

```
15  result2 <- lm(ice_cream ~ temperature+price, data = data.ice2)
```

目的変数
Objective variable

説明変数
Response variable

・結果の出力
グラフに記入・・・abline()
結果の詳細・・・summary()

```
5   abline(result) # 推定回帰直線を描く
6   summary(result)
```

```
> summary(result)

Call:
lm(formula = ice_cream ~ temperature + price, data = data.ice)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1935 -4.2656 -0.8731  3.7920 14.4696

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.48134   27.37538   2.940 0.006973 **
temperature  2.88370    0.69836   4.129 0.000355 ***
price       -0.27788    0.08032  -3.460 0.001952 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.921 on 25 degrees of freedom
Multiple R-squared:  0.7458,    Adjusted R-squared:  0.7255
F-statistic: 36.67 on 2 and 25 DF,  p-value: 3.671e-08
```

重回帰式の導出(プログラミングで)

・結果の出力
結果の詳細・・・summary()

```
16  summary(result2)
17
```

```
Call:
lm(formula = ice_cream ~ temperature + price, data = data.ice)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1935 -4.2656 -0.8731  3.7920 14.4696

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.48134   27.37538   2.940 0.006973 **
temperature  2.88370    0.69836   4.129 0.000355 ***
price       -0.27788    0.08032  -3.460 0.001952 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.921 on 25 degrees of freedom
Multiple R-squared:  0.7458,    Adjusted R-squared:  0.7255
F-statistic: 36.67 on 2 and 25 DF,  p-value: 3.671e-08
```

切片，回帰係数の結果
Estimate・・・推定値
Std.Error・・・標準偏差
tvalue・・・t値

決定係数

$\hat{Y} = 80.48 + 2.88X_1 - 0.278X_2$という重回帰式が得られた！

「temperature」のt値の絶対値＞「price」のt値の絶対値
⇒価格よりも気温の方がアイスの売り上げに与える影響は大きい

練習問題

知名度のない東工大，東工大の人気を上げるためにはどうすればいいのか！？

⇒回帰分析を使って調べてみよ～

練習問題

・使用するデータ・・・大学別〇〇のデータ（2018）

Applicant：**志願者数（人気度，千人）**
Prof_student：**教員一人当たりの学生数（面倒見の良さ，人）**
Area：**大学敷地面積**（*1000m^2）

Circles：**サークル数(個)**
Girlsratio：**女子率(%)**
Age：**大学の歴史の長さ(年)**

	A	B	C	D	E	F	G
1		applicant	Prof_stud	area	circles	girlsratio	age
2	東工大	4.69	9.08	46.9887	51	13	137
3	東大	9.534	7.12	97.0543	351	9	141
4	慶応	44.845	15.9	88.7588	366	36	160
5	早稲田	114.983	30.09	57.6115	631	40	136
6	明治	113.507	31.96	26.6718	414	35	127
7	一橋	4.484	19.57	40.7282	80	28	143
8	青山	60.966	33.04	22.9828	275	50	144
9	立教	62.691	34.18	17.5617	255	54	144
10	上智	29.277	26.05	23.7953	204	58	105

練習問題

・重回帰式（今回は面倒見の良さ，敷地面積，女子率を採用）

$$Y_{(Applicant)} = a + bX_{1(Prof\_student)} + cX_{2(Area)} + dX_{3(girlsratio)}$$

```
20  #練習問題
21  "########"
22  data.enshu = %%%("enshu.csv")#read csv file
23  result3 <- %%%(%%%~Prof_student+area+girlsratio, data = %%%)
24  %%%%%%%%(result3) #detail of the result
25
```

%%%に文字を入れて，プログラムを回してみてください

```
Residuals:
      1       2       3       4       5       6       7       8       9
 18.679 -13.728   7.084  12.675  18.893 -26.992 -16.208 -11.743  11.341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -100.5041    42.2613  -2.378  0.06331 .
Prof_student    6.9310     1.5503   4.471  0.00657 **
area            0.9003     0.3971   2.268  0.07266 .
girlsratio     -1.4403     0.8348  -1.725  0.14509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.74 on 5 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.8398,    Adjusted R-squared:  0.7437
F-statistic: 8.739 on 3 and 5 DF,  p-value: 0.01968
```

←うまくいけばこんな結果が出てきます

練習問題

```
28  #練習問題
29  "########"
30  data.enshu = read.csv("enshu.csv")#read csv file
31  result3 <- lm(applicant~Prof_student+area+girlsratio, data = data.enshu)
32  summary(result3) #detail of the result
```

```
Residuals:
     1      2      3      4      5      6      7      8      9
18.679 -13.728   7.084  12.675  18.893 -26.992 -16.208 -11.743  11.341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -100.5041    42.2613  -2.378  0.06331 .
Prof_student    6.9310     1.5503   4.471  0.00657 **
area            0.9003     0.3971   2.268  0.07266 .
girlsratio     -1.4403     0.8348  -1.725  0.14509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.74 on 5 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.8398,    Adjusted R-squared:  0.7437
F-statistic: 8.739 on 3 and 5 DF,  p-value: 0.01968
```

t値から分かったこと

→敷地面積を広くして，教員一人当たりの学生を増やせば（放任気味），
理論上大学の人気は上がる（大学の女子率はそんな関係ない←）

最後に

説明変数変えたりすると結果もいろいろ変わるので
時間あるときにやってみてください

参考文献

実証分析のための計量経済学

http://blog.livedoor.jp/pcclgk2/archives/52195523.html

http://www.keinet.ne.jp/dnj/result/ippan/2279.html

https://gakucir.com/search/?university=653

https://bellcurve.jp/statistics/course/9700.html

https://atarimae.biz/archives/12142

http://r-office-room.jugem.jp/?eid=124

http://cogpsy.educ.kyoto-u.ac.jp/personal/Kusumi/datasem13/shinya.pdf

www.eco.nihon-u.ac.jp/eco_kyouin/go/Go_seminar//2010/3.pps

http://sorb.co.jp/blog/1654/