

ECO208

R and RStudio

Week 5 | October 21, 2022

How you doin'? 🤘



今日もR三昧



でもその前に回帰分析、
覚えてます？

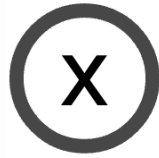
Let's review

回歸分析 is about:

A: scoring points 🏀

B: creating charts 📊

C: relationships 🧑❤️🧑



independent variable

説明変数

Pokemon's caught
捕まえたポケモン数



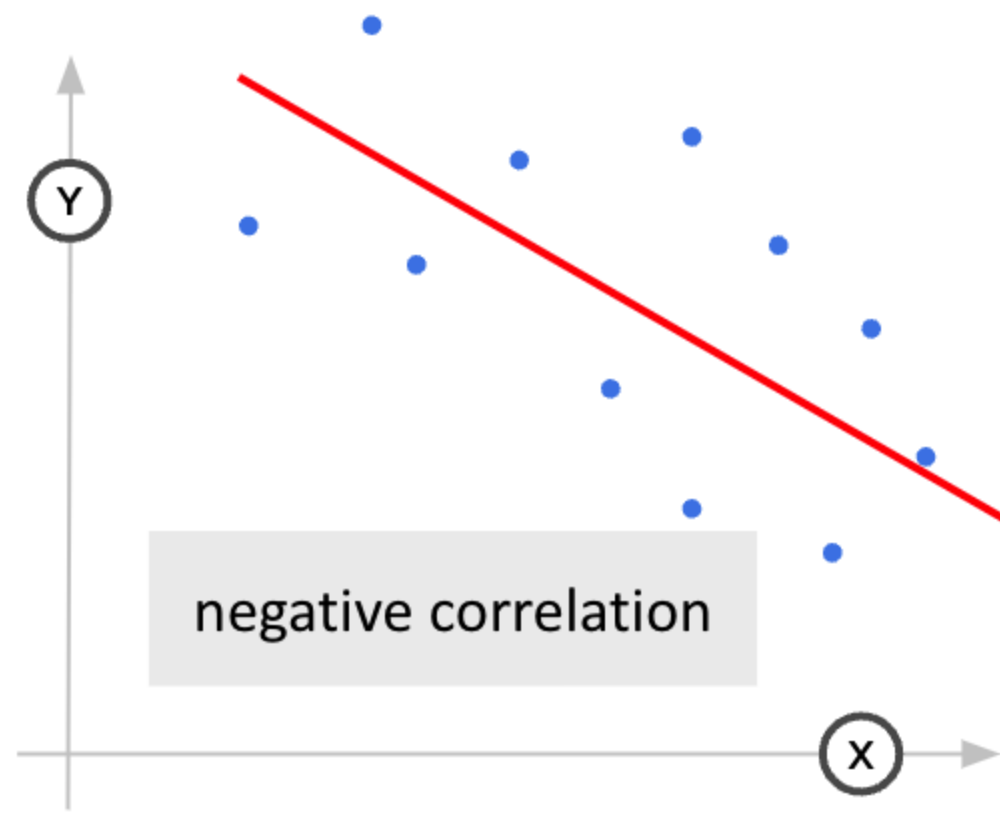
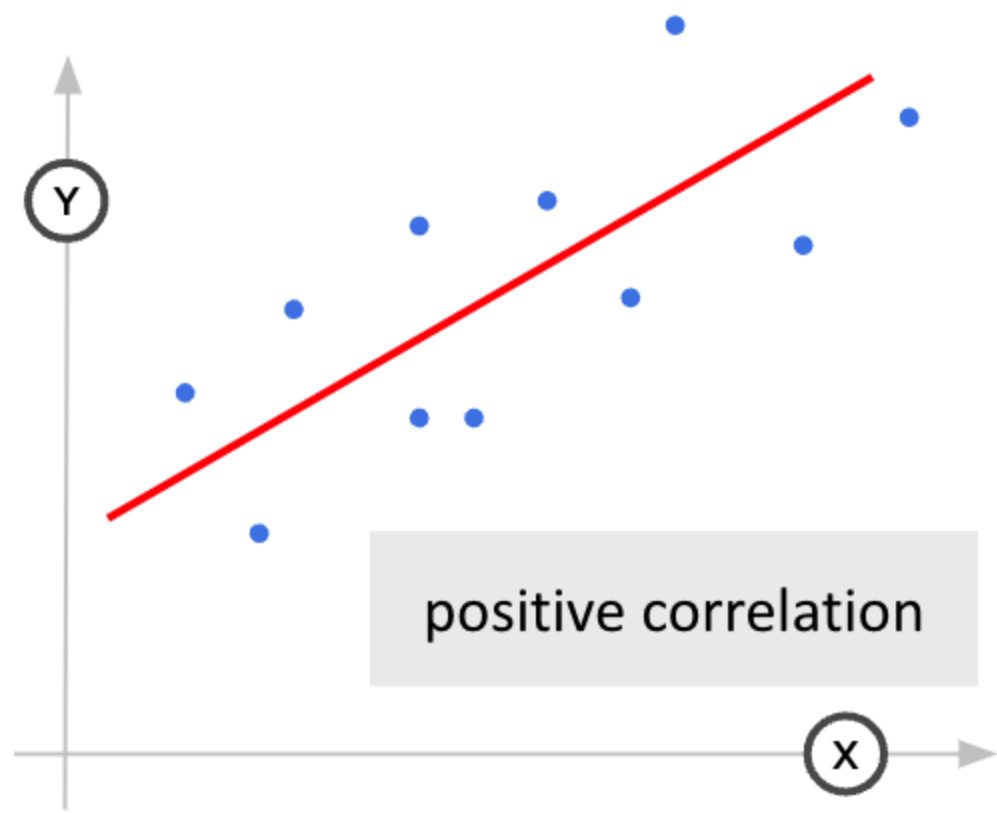
dependent variable

目的変数

Distance walked
歩いた距離



プラスかマイナス？



仮設検定

帰無仮説 【null hypothesis】

H_0 = 駅までの徒歩分数は家賃に全く影響がない

coefficient 回帰係数 $\beta_1 = 0$

対立仮説 【alternative hypothesis】

H_1 = 駅までの徒歩分数は家賃に影響がある

$$\beta_1 \neq 0$$

まずは回帰式を作るう

$$Y = a + bX$$

$$\text{家賃（万円）} = a + b * \text{駅まで徒歩分数}$$

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.34046331							
5	重決定 R2	0.11591527							
6	補正 R2	0.08542959							
7	標準誤差	1.63077684							
8	観測数	31							
9									
10	分散分析表								
11		自由度	変動	分散	観測された分散	有意 F			
12	回帰	1	10.1119243	10.1119243	3.80228566	0.06090971			
13	残差	29	77.1235596	2.65943309					
14	合計	30	87.2354839						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	10.0908723	0.94352867	10.6948231	1.407E-11	8.16113945	12.0206051	8.16113945	12.0206051
18	minutes	-0.1252126	0.06421338	-1.949945	0.06090971	-0.2565437	0.00611855	-0.2565437	0.00611855

家賃（万円） = $a + b \times \text{駅まで徒歩分数}$

$$\text{家賃（万円）} = 10.09 - 0.125 \times \text{駅まで徒歩分数}$$

すなわち

南柏の駅からの徒歩分数が1分増えるごとに
家賃が1250円減る

では「a」（切片）は何？

これは「x」がゼロの時の数値。

すなわち、駅から徒歩分数がゼロの賃貸（ありますか？）の
場合、家賃は10.09万円

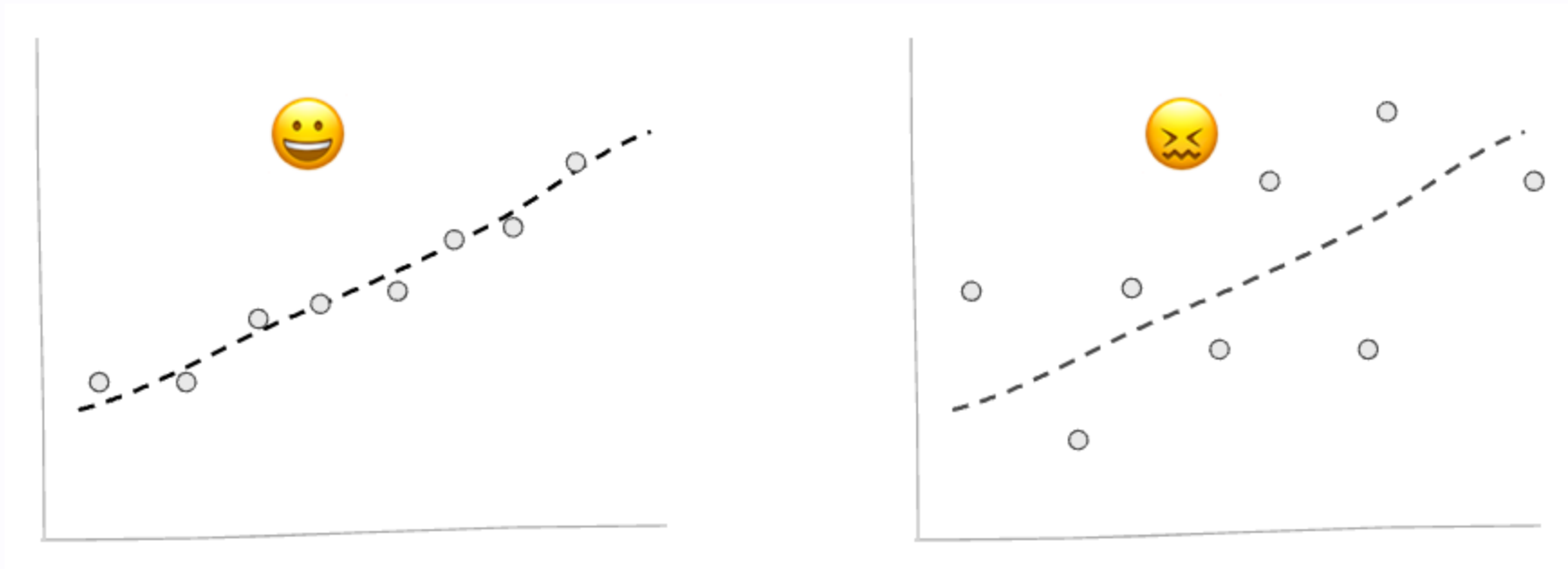
How good is this model?

R², t-statistic, p-value


R-Squared: How well does it fit?

$R^2 = 1$ (perfect fit 😊)

$R^2 = 0$ (bad fit 😞)



ではこの分析のR2乗は？

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.34046331							
5	重決定 R2	0.11591527							
6	補正 R2	0.08542959							
7	標準誤差	1.63077684							
8	観測数	31							
9									
10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	1	10.1119243	10.1119243	3.80228566	0.06090971			
13	残差	29	77.1235596	2.65943309					
14	合計	30	87.2354839						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	10.0908723	0.94352867	10.6948231	1.407E-11	8.16113945	12.0206051	8.16113945	12.0206051
18	minutes	-0.1252126	0.06421338	-1.949945	0.06090971	-0.2565437	0.00611855	-0.2565437	0.00611855
19									

「目的変数である家賃は値変動を説明変数である徒歩分数は 1
1.59%しか説明できていない。」

What about the P value?

P値で説明変数（徒歩分数）が目的変数（家賃）に対して関係があるかどうかを確認する

P値が優位水準0.05未満であれば、
「説明変数が目的変数に有意に影響している」と判断ができる。

要するに、この関係性はランダムではないので、帰無仮説
【null hypothesis】をrejectすることができる。

でも...

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.34046331							
5	重決定 R2	0.11591527							
6	補正 R2	0.08542959							
7	標準誤差	1.63077684							
8	観測数	31							
9									
10	分散分析表								
11		自由度	変動	分散	観測された分散	有意 F			
12	回帰	1	10.1119243	10.1119243	3.80228566	0.06090971			
13	残差	29	77.1235596	2.65943309					
14	合計	30	87.2354839						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	10.0908723	0.94352867	10.6948231	1.407E-11	8.16113945	12.0206051	8.16113945	12.0206051
18	minutes	-0.1252126	0.06421338	-1.949945	0.06090971	-0.2565437	0.00611855	-0.2565437	0.00611855

うわ！ 6.09%！

微妙～



微妙でもオッケーの判断はあなた次第。
有意水準を上げてもいい。

要するに：

「p値の有意水準を0.1だと、説明変数（徒歩分数）が目的変数（家賃）に対して有意に影響していることが言える。」

At 94% level of confidence that relationship is not due to random chance. That relationship actually exists in the housing market.

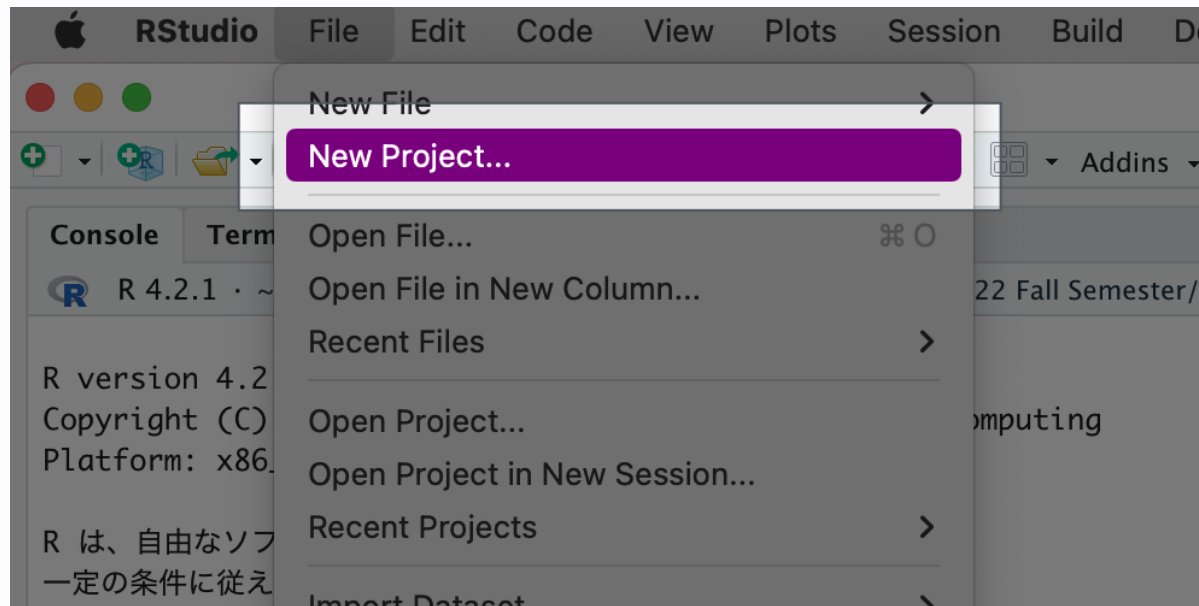
では、これをRStudioで
やってみよう！



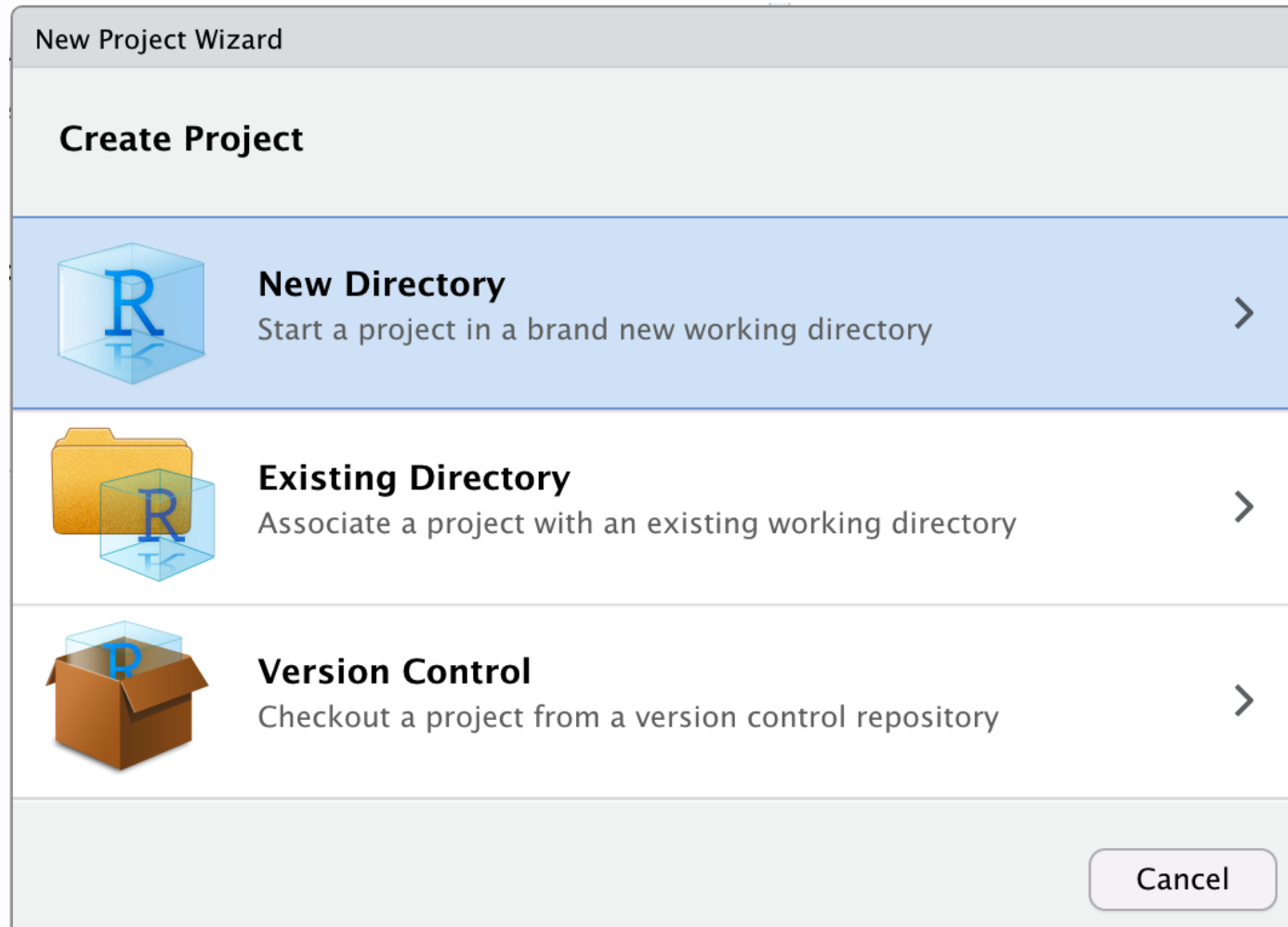
Let's get started

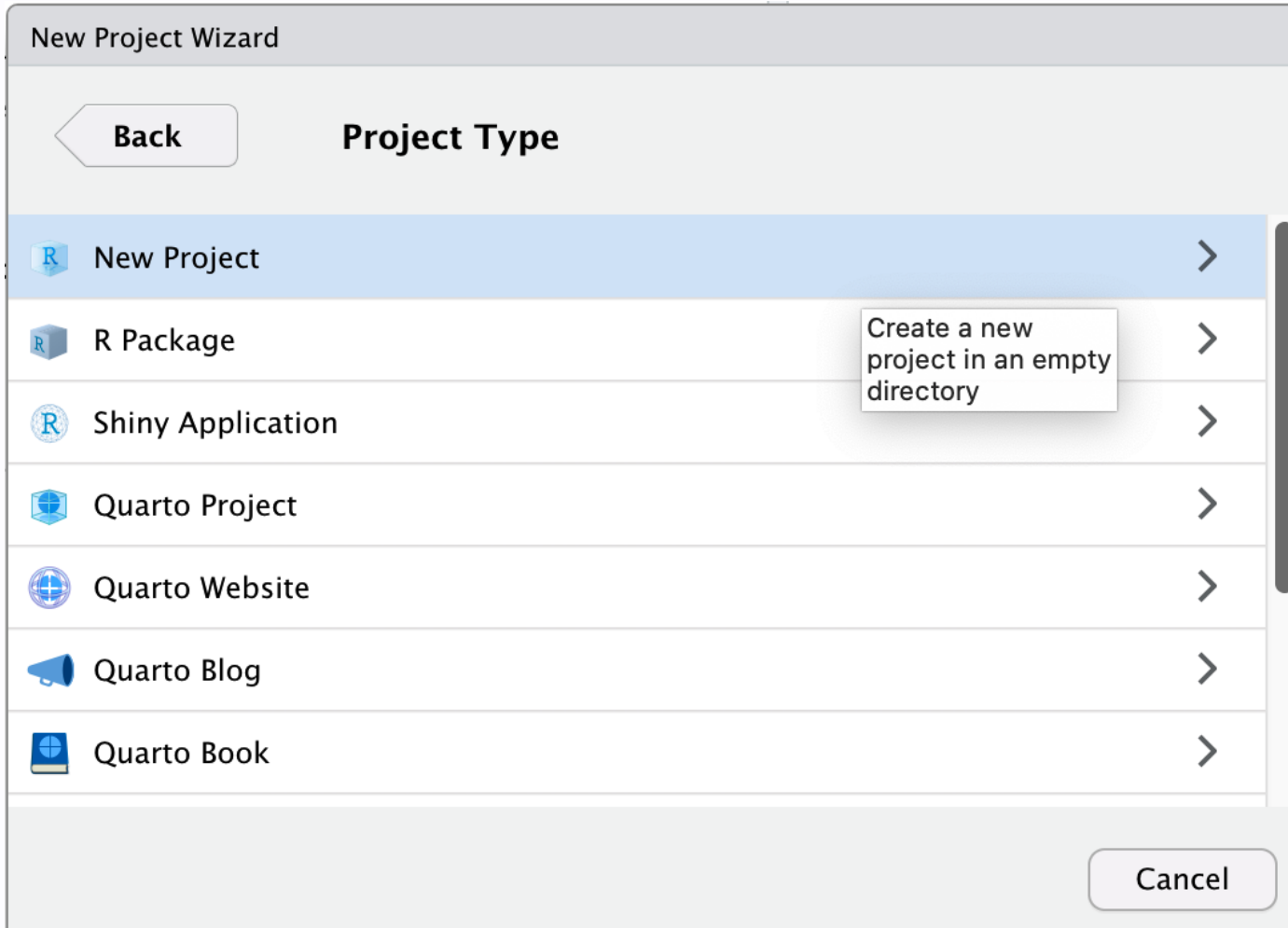
Launch R Studio (RStudioを使おう)

Open RStudio, go to File → New Project



Create a new project






Change "W3" to "W4"

New Project Wizard

Back Create New Project

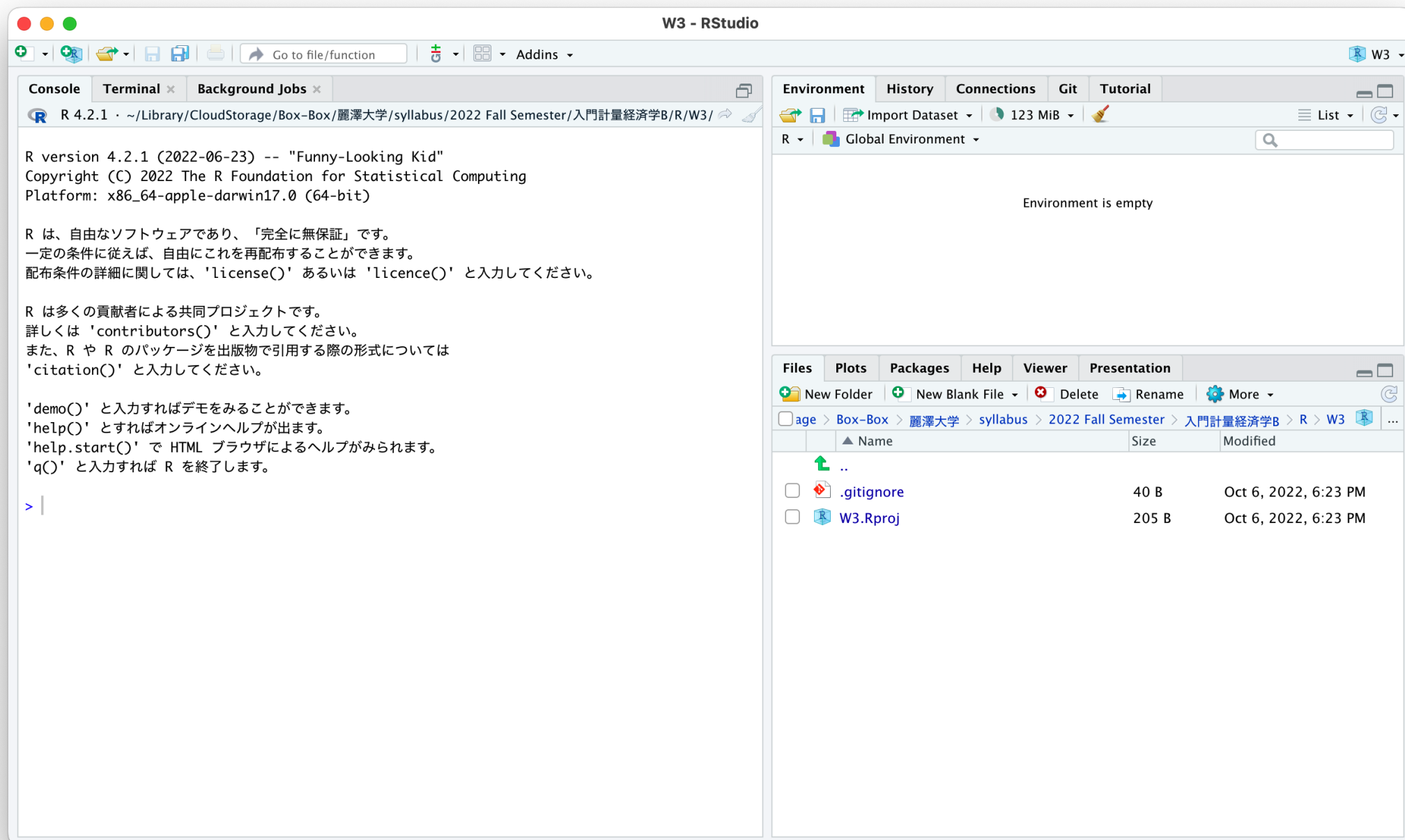
 Directory name:

Create project as subdirectory of:

☒ Create a git repository
☐ Use renv with this project

ここは大事！！

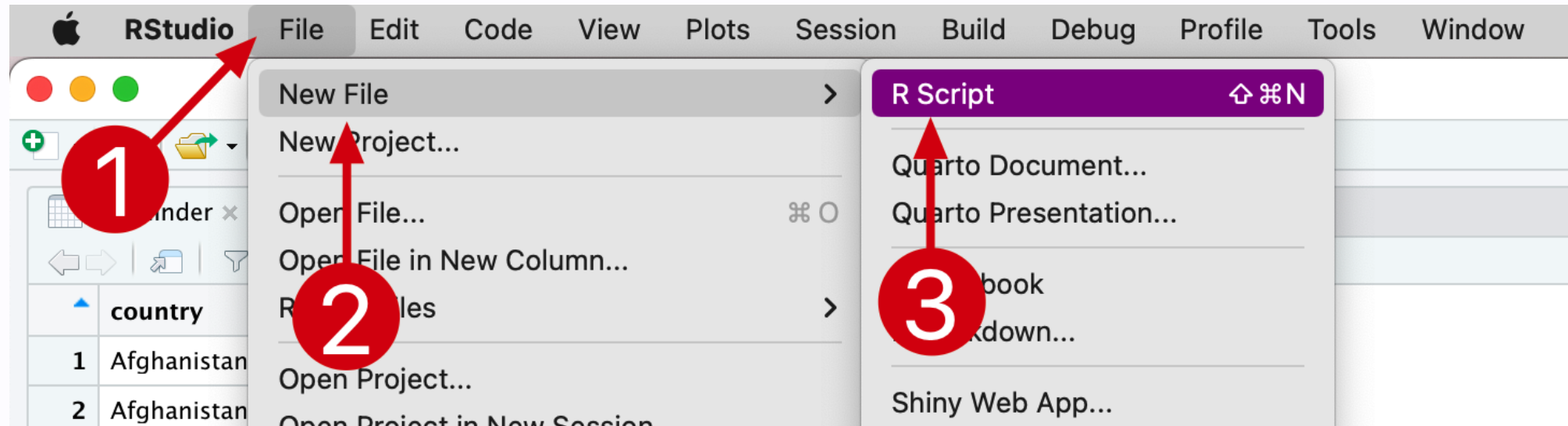
☐ Open in new session



Pause: Are we all here?

Create an R script file

R Scriptファイルを作成



Rで回帰分析

Load the data

```
# データを取得  
chiba <- read.csv("data/chiba_rent.csv")  
  
# データを表示  
head(chiba)  
  
# データの統計  
summary(chiba)
```

データはGoogle Classroomからダウンロード

attach the data

```
# attach the dataset  
attach(chiba)  
  
# 散布図  
plot(minutes,rent) # x,y
```

散布図に近似直線(回帰直線)を付ける

```
abline(lm(rent~minutes), col="red")
```

*注意！ `lm(y~x)` ではyは目的変数、xは説明変数なので、`plot(x,y)` とは逆

式は？

lm 関数を使おう

```
lm(y~x, data = dataset) #yは目的変数、xは説明変数
```

```
# 回帰分析を実行  
lm(rent~minutes)
```


lm output

```
> lm(rent~minutes)
```

Call:

```
lm(formula = rent ~ minutes)
```

Coefficients:

(Intercept)

10.0909

minutes

-0.1252

家賃 (万円) = $a + b \times$ 駅まで徒歩分数

回帰式

$$\text{家賃（万円）} = 10.0909 - 0.1252 \times \text{駅まで徒歩分数}$$

すなわち

南柏の駅からの徒歩分数が1分増えるごとに
家賃が1250円減る

回帰分析の `summary()`

これだけじゃ足りないので...

```
# 変数に入れる  
slr = lm(rent~minutes)  
  
# 回帰分析のsummary結果  
summary(slr)
```

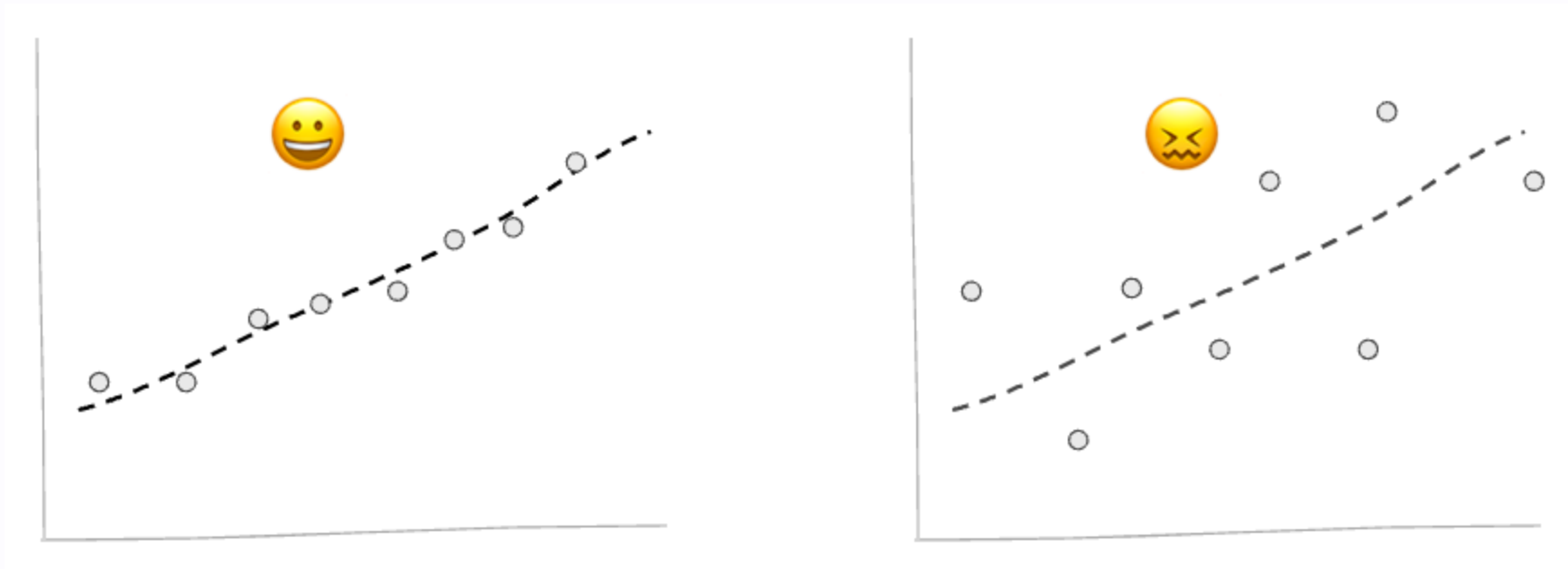
summary output

回帰に用いた式	→	Call: lm(formula = rent ~ minutes)
残差の統計量 残差は実績値と予測値の差で定義される	→	Residuals: Min 1Q Median 3Q Max -2.9631 -0.8883 -0.3379 0.4495 2.9134
回帰係数 (Intercept)が回帰直線の切片minutesが傾き	→	Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 10.09087 0.94353 10.70 1.41e-11 *** minutes -0.12521 0.06421 -1.95 0.0609 . ---
残差の標準誤差とその自由度	↘	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
決定係数R2	→	Residual standard error: 1.631 on 29 degrees of freedom Multiple R-squared: 0.1159, Adjusted R-squared: 0.08
分散分析のF-値とその自由度	↗	F-statistic: 3.802 on 1 and 29 DF, p-value: 0.06091

R-Squared: How well does it fit?

$R^2 = 1$ (perfect fit 😊)

$R^2 = 0$ (bad fit 😞)



ではこの分析のR2乗は？

```
> summary(slr)
```

Call:

```
lm(formula = rent ~ minutes)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.9631	-0.8883	-0.3379	0.4495	2.9134

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.09087	0.94353	10.70	1.41e-11 ***
minutes	-0.12521	0.06421	-1.95	0.0609 .

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

Residual standard error: 1.631 on 29 degrees of freedom

Multiple R-squared: 0.1159, Adjusted R-squared: 0.08543

F-statistic: 3.802 on 1 and 29 DF, p-value: 0.06091



すなわち

「目的変数である家賃は値変動を説明変数である徒歩分数は 1
1.59%しか説明できていない。」

P値は？

```
Call:
lm(formula = rent ~ minutes)

Residuals:
    Min       1Q   Median       3Q      Max
-2.9631 -0.8883 -0.3379  0.4495  2.9134

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.09087    0.94353   10.70 1.41e-11 ***
minutes     -0.12521    0.06421   -1.95  0.0609 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.631 on 29 degrees of freedom
Multiple R-squared:  0.1159,    Adjusted R-squared:  0.08543
F-statistic: 3.802 on 1 and 29 DF,  p-value: 0.06091
```

単回帰分析の場合はこの値は変わらない

要するに：

「p値の有意水準を 0.1 で設定すると、説明変数（徒歩分数）が目的変数（家賃）に対して有意に影響していることが言える。」

```
# データを取得
chiba <- read.csv("data/chiba_rent.csv")

# attach する
attach(chiba)

# 散布図
plot(minutes,rent)

# 散布図に近似直線(回帰直線)を付ける
abline(lm(rent~minutes), col="red")

# 回帰分析を実行
lm(rent~minutes)

# 変数に入れる
result = lm(rent~minutes)

# 回帰分析のsummary結果
summary(result)
```

では、 it's group time!