重回帰分析を用いた予測

睡眠時間を他の行動時間から予測しよう

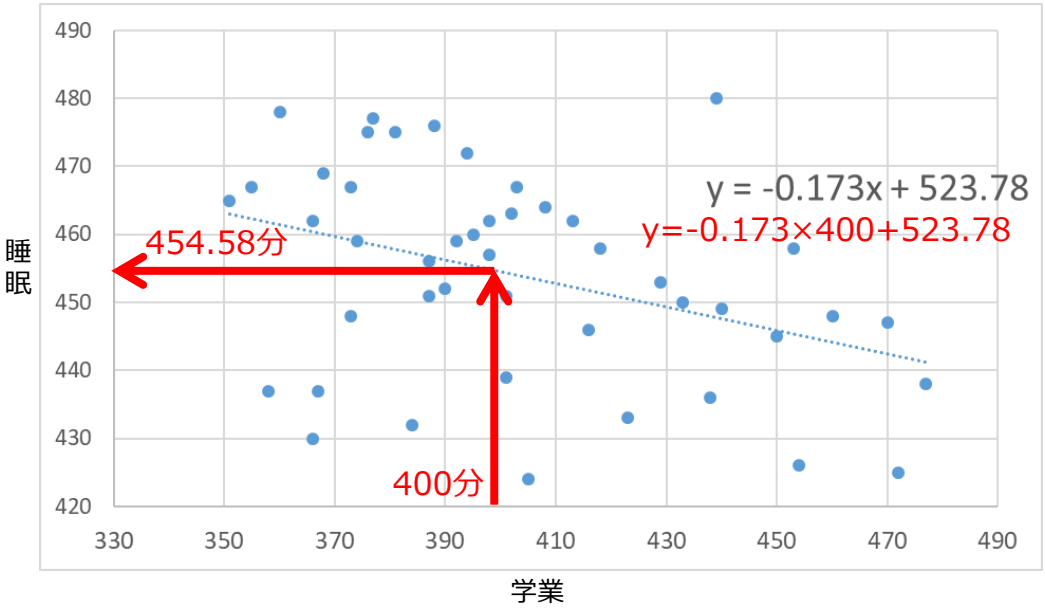
情報Ⅰで学んだこと

単回帰分析

データ

地域区分	睡眠	学業
01_北海道	467	355
02_青森県	469	368
03_岩手県	458	453
04_宮城県	448	373
05_秋田県	467	373
06_山形県	477	377
07_福島県	436	438
08_茨城県	456	387
09_栃木県	472	394

回帰直線 → 予測



学習時間以外のデータも使えたら・・・

睡眠時間を予測するために、
学業の時間だけでなく他の時間も使うことはできないか？

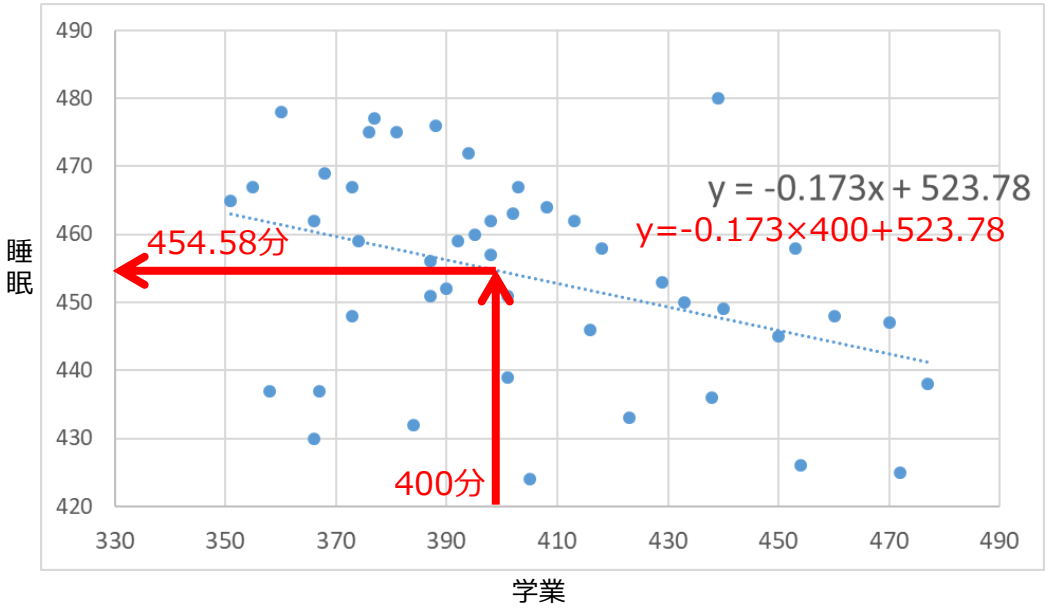
例）通学時間、買い物、休息・・・など

令和３年社会生活基本調査　生活時間－地域（調査票Ａ）
第６５－１表　曜日，男女，スマートフォン・パソコンなどの使用時間，年齢，行動の種類別総平均時間（10歳以上）－全国，都道府県

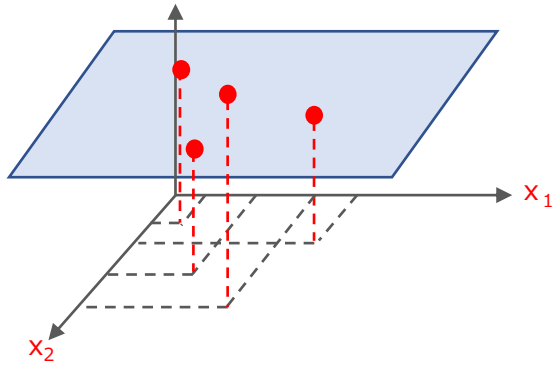
					総平均時間 行動の種類	総平均時間 行動の種類	総平均時間 行動の種類	総平均時間 行動の種類	総平均時間 行動の種類	総平均時間 行動の種類	総平均時間 行動の種
					01_睡眠 (分)	02_身の回りの用事 (分)	03_食事 (分)	04_通勤・通学 (分)	05_仕事 (分)	06_学業 (分)	07_家事 (分)
曜日	地域区分	男女	スマートフ	年齢							
2_平日	01_北海道	0_総数	0_総数	02_15～19歳	467	72	74	62	41	355	
2_平日	02_青森県	0_総数	0_総数	02_15～19歳	469	71	90	53	34	368	
2_平日	03_岩手県	0_総数	0_総数	02_15～19歳	458	69	76	53	47	453	
2_平日	04_宮城県	0_総数	0_総数	02_15～19歳	448	75	80	72	44	373	
2_平日	05_秋田県	0_総数	0_総数	02_15～19歳	467	61	84	45	29	373	
2_平日	06_山形県	0_総数	0_総数	02_15～19歳	477	74	83	52	43	377	
2_平日	07_福島県	0_総数	0_総数	02_15～19歳	436	60	87	64	41	438	
2_平日	08_茨城県	0_総数	0_総数	02_15～19歳	456	88	76	71	31	387	
2_平日	09_栃木県	0_総数	0_総数	02_15～19歳	472	66	82	69	46	394	
2_平日	10_群馬県	0_総数	0_総数	02_15～19歳	457	74	78	69	41	398	

2つの値を使って予測する

1つの値を使って予測



2つの値を使って予測



$$y=a_1x_1 + a_2x_2 + b$$

睡眠　学業　通勤･通学

さらに多くの値を使って予測する

図として表すことはできないけれど、
同じような考え方を使って予測をする。

	1	2	3	4	5	6	7	8	9	10	11
1	地域区分	睡眠	身の回りの	食事	通勤・通学	仕事	学業	家事	買い物	移動(通勤	テレビ
2	01_北海道	467	72	74	62	41	355	5	3	19	
3	02_青森県	469	71	90	53	34	368	0	0	13	
4	03_岩手県	458	69	76	53	47	453	5	3	8	
5	04_宮城県	448	75	80	72	44	373	3	3	7	
6	05_秋田県	467	61	84	45	29	373	4	6	13	
7	06_山形県	477	74	83	52	43	377	4	5	6	
8	07_福島県	436	60	87	64	41	438	7	2	15	
9	08_茨城県	456	88	76	71	31	387	6	6	18	
10	09_栃木県	472	66	82	69	46	394	13	12	9	
11	10_群馬県	457	74	78	69	41	398	7	7	12	
12	11_埼玉県	439	83	91	80	31	401	5	3	12	

$y = a_1 \times$ 学業 $+ a_2 \times$ 通勤・通学 $+ a_3 \times$ 食事 $+ \cdots + b$（分）

Pythonを使って重回帰分析をしてみよう

重回帰分析の流れ

1. 予測に必要なデータを収集する
2. 分析できるようにデータを整形する
3. 重回帰分析をして、モデルを作成する
4. 作成したモデルを用いて予測する

プログラムの解説

```
import pandas as pd
```

pandasというライブラリを使えるようにする

```
df = pd.read_csv('shakai.csv')
```

引数のファイル名のcsvファイルをpandasで読み込む

```
df.head()
```

変数dfに格納されているデータの最初の5件を表示する

プログラムの解説

```
X = df[['学業']]
X.head()
```

説明変数として「学業」を設定し、最初の5件を表示する

```
y=df[['睡眠']]
y.head()
```

目的変数として「睡眠」を設定し、最初の5件を表示する

プログラムの解説

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
```

線形回帰モデルを使えるようにする

```
model.fit(X, y)
```

変数Xを説明変数、変数yを目的変数として、回帰分析を行う

```
print(model.intercept_, model.coef_)
```

回帰分析の結果として得られる

切片（定数項）と説明変数の係数を出力する

プログラムの解説

```
X_yosoku = pd.DataFrame([[400]], columns = ['学業'])
model.predict(X_yosoku)
```

学業の時間を400分として、回帰式にあてはめて睡眠時間を予測する

回帰式で求めると

$y = -0.17300592 \times 400 + 523.77740922$

$y = 454.57504168$[分]

プログラムの解説

```
X2=df[['学業', '通勤・通学', '休養・くつろぎ','趣味・娯楽']]
X2.head()
```

説明変数として「学業」、「通勤・通学」、「休養・くつろぎ」、「趣味・娯楽」を設定し、最初の5件を表示する

```
y=df[['睡眠']]
y.head()
```

目的変数として「睡眠」を設定し、最初の5件を表示する

プログラムの解説

```
from sklearn.linear_model import LinearRegression
model2 = LinearRegression()
model2.fit(X2, y)
print(model2.intercept_, model2.coef_)
```

単回帰分析：説明変数X （1つの項目）
重回帰分析：説明変数X2（複数の項目）

とした以外は同じ

係数の読み取り

```
print(model2.intercept_, model2.coef_)
```

[612.39863457]：定数項（切片）

```
[[-0.23689623 -0.54526488
  -0.16213081 -0.10581234]]
```

「学業」が1分増えると「睡眠」が「0.23689623」分減る

「通勤・通学」、「休養・くつろぎ」、「趣味・娯楽」についても同様に変化する

プログラムの解説

```
X2_yosoku = pd.DataFrame([[400, 70, 135, 40]],
    columns = ['学業', '通勤・通学', '休養・くつろぎ','趣味・娯楽'])
model.predict(X2_yosoku)
```

学業を400分、通勤・通学を70分、休養・くつろぎを135分、
趣味・娯楽を40分として、回帰式にあてはめて睡眠時間を予測する

回帰式で求めると

$$y = -0.237 \times 400 - 0.545 \times 70 - 0.162 \times 135 - 0.106 \times 40 + 612.399$$

$$y \fallingdotseq 453.3\text{[分]}$$

係数から傾向を読み取ることは注意

重回帰分析をしたときに求まるcoef_の値（係数）

```
[[-0.23689623 -0.54526488
  -0.16213081 -0.10581234]]
```

多重共線性に注意！

説明変数間に相関が強いものがあるときに起き、係数が安定しなくなる

質的データも使って予測するには

ワンホットエンコーディングという方法を使う

「スマートフォン・パソコンなどの使用時間」

列を追加「1～3時間」「3～6時間」「6～12時間」
該当するデータに対して「1」を割り当てる
12時間以上は、追加した項目の値を「0」にする

地域区分	スマートフォン・パソコンなどの使用時間	スマートフォン1～3時間	スマートフォン3～6時間	スマートフォン6～12時間	睡眠	……
01_北海道	22_1～3時間未満	1	0	0	480	
01_北海道	23_3～6時間未満	0	1	0	439	
01_北海道	24_6～12時間未満	0	0	1	450	
01_北海道	25_12時間以上	0	0	0	389	

こんなことに応用できます

スポーツの記録を予測

100m走のタイム、ボール投げ、幅跳びの記録などを使って予測

農作物の生産量の予測

気温、湿度、日照時間などを使って予測

アパートの家賃を予測

広さ、築年数、駅からの時間などを使って予測