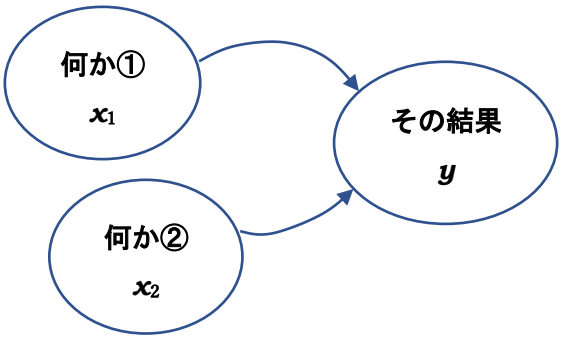
回帰分析・重回帰分析

今、**何か①(x_1)** および **何か②(x_2)** から、
その結果(y) が得られるとする。

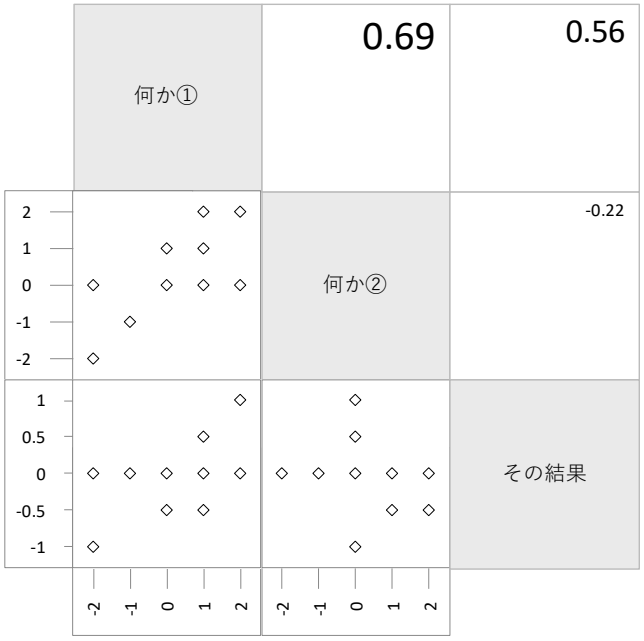


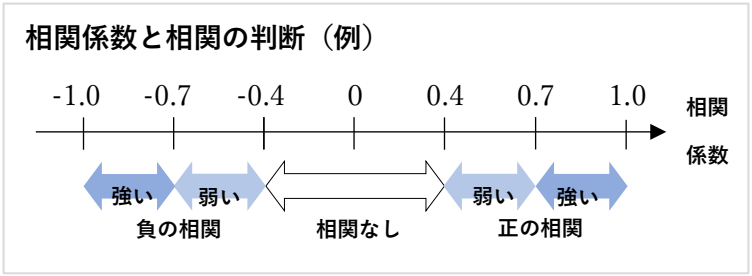
例）調味料①、②を入れる　→　味が変わる
　　エアコン①、②を入れる　→　室温が変わる。

いま、x_1, x_2, y について、次のようなデータがある。

何か①(x_1)	何か②(x_2)	その結果(y)
0	0	0
1	1	0
1	0	0.5
0	1	-0.5
2	2	0
1	2	-0.5
2	0	1
-1	-1	0
-2	0	-1
-2	-2	0

問１　このデータの散布図相関行列は次のようになる。x_1, x_2, y についての相関関係について考えてみよう。
何か①と何か②のうち、結果に影響を及ぼしているものはどれだろう？





問２　実際、もうすこしじっくりとデータを眺めてみると、x_1, x_2, y にはどのような関係があるだろう。

$$y =$$

結果にあたる変数 y(　　変数という) が 複数の変数（　　変数という）から影響を受ける場合、散布図相関行列で個別の変数の関係だけをながめてしまうと、実は相関がある変数同士を「相関なし」と判定してしまう可能性がある。

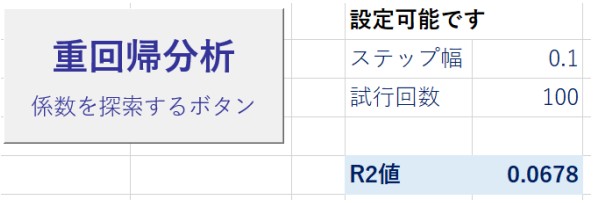
そこで、まずモデルとして例えば次のような式を仮定する。 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

※x_1 と x_2 がそれぞれ y に影響すると考えているところがポイント。

そして、実際の結果をうまく説明／予想できるような、x の係数 β のセット（β_0、β_1、β_2、…）を、β をああでもない、こうでもないと変化させ、あてはめながら決めていく。(実際には、残差（実際と理論の差）を小さくするような β の値のセットを、β を変えながら決めていく。) この分析手法を（　　）回帰分析という。

以下、Excel ファイル（説明変数 6 つまでに対応）を開いて、実際にやってみましょう。

【実習①】サンプルデータ①が入力されていることを確認したら、ステップ幅（ベータをどれくらい変化させて試すか）を 0.1、試行回数を 100 として、重回帰分析ボタンを押してみましょう。R^2 値*が 1 に近ければ近いほど、実際の値をよく説明する（誤差の少ない）モデルとなっています。データ①については、R^2 値が 0.999 くらいになるまで繰り返してみましょう。



【実習②】家賃のデータ

駅からの距離、築年数、部屋の広さなどによって、アパートの家賃は違いますよね。皆さんがアパートを建てるとしたら、どれくらいの家賃を設定しますか？あるいは、アパートを借りるときに、割安なアパートをみつけることはできるでしょうか…？

β1 ○する	β2 ○する	β3 ○する	β4 ○する	β5 ×しない	β6 ×しない	β0 ×しない
0	0	0	0	0	0	0
変数 X1	変数 X2	変数 X3	変数 X4	変数 X5	変数 X6	目的変数 y
駅からの距離 徒歩(分)	築年数(年)	部屋の広さ (m2)	立地エリアの 人気度			家賃(円)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

(1)データのクリア消去　→　サンプルデータ②家賃ボタンを押して、データを読み込みましょう。

(2)係数のセル（B3～H3）に、それぞれの説明変数の効果がどの程度か予想して、係数の初期値（その値から、変化させてセットを決める種のようなもの）を決めます。この値がいい感じだと、少ない試行回数で R^2 値が 1 に近づきます。データ②の場合はその値が 1 増えたとき、家賃がいくら上がるか、下がるかを想像してみるといいと思います。

(3)重回帰分析をやってみる。このデータは、(R^2 値 0.98 くらいまでは狙えるはず。) それぞれの係数はいくらになったか。表に書き入れましょう。

駅からの距離	築年数	部屋の広さ	立地エリアの人気	切片
β_1:	β_2:	β_3:	β_4:	β_0:

R^2 値*　…　（1 －（残差 ^{2} の合計）／（y^2 の合計））で計算できる。y の値に対し、残差が小さいと 1 に近づく。

【実習③】テストの点数

テストでいい点を取りたい！と思ったら、まずは勉強する。そして、睡眠も大事な気がする。スマホは使いすぎるとよくない…さて、1日の生活をどうマネジメントしたらいいでしょう。**それぞれの項目は、テストの点数をおよそ何点分押し上げるでしょうか。**

データ番号	学習時間（時間/日）	睡眠時間（時間/日）	スマホ利用時間（時間/日）	朝食習慣の有無	余暇の時間（時間/日）	テストの点数
1	1.69	5.16	4.21	0	2.01	24
2	4.28	8.18	1.42	1	3.43	96
3	3.29	6.57	1.81	1	2.63	47
4	2.69	7.54	5.49	1	0.65	48
5	0.7	9.54	4.03	0	0.28	39
6	0.7	6.25	1.05	0	2.57	16
7	0.26	7.05	1.51	0	0.11	40
8	3.9	8.78	4.32	1	2.34	75
9	2.71	6.14	1.03	1	3.76	54
…	…	…	…	…	…	…

データのクリア消去　→　サンプルデータ③テストの点数ボタンを押すところからやってみよう。

変数 x_1	変数 x_2	変数 x_3	変数 x_4	変数 x_5	目的変数 y
学習時間	睡眠時間	スマホ利用時間	朝食習慣の有無	余暇の時間	テストの点数
β_1:	β_2:	β_3:	β_4:	β_5:	β_0:

（【実習④】実際の探究で…）

総合的な探究の時間や、大学での研究などでデータを分析する1つのツールとして、重回帰分析を活用してください。単回帰分析だけを使うよりも面白い発見がありそうです。今回は、x_1 や x_2 は1乗であることを仮定しましたが、ひょっとすると2乗や3乗のものもあるかもしれませんよね。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 {x_2}^2 + \beta_3 x_3 + \beta_4 {x_4}^3$$

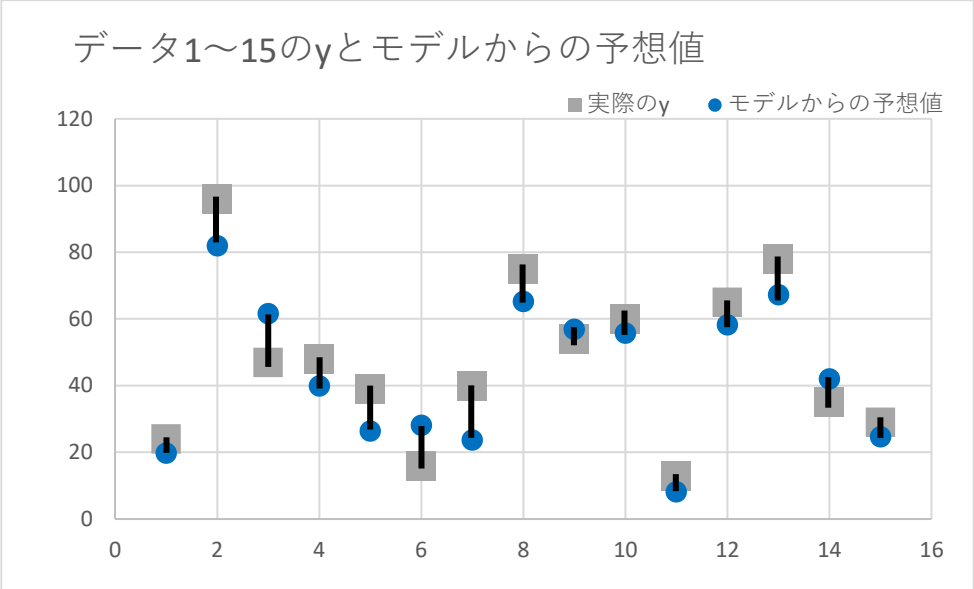
あるいは、何乗かわからないときは、それもランダムで変えながら調べる必要があるかもしれません。

$$y = \beta_0 + \beta_1 {x_1}^{\alpha_1} + \beta_2 {x_2}^{\alpha_2} + \beta_3 {x_3}^{\alpha_3} + \beta_4 {x_4}^{\alpha_4}$$

重回帰分析と Excel シートの仕組み

重回帰分析、というとそれこそ重々しく感じますが、原理はそんなに難しくありません。

例えば、次のグラフの四角いマークが実際の値、丸いマークがモデルから計算される予想値、黒い直線がその差分を表しています。



要するに、四角と丸ができるだけ重なるようになると、それだけそのモデルは、現実のデータをうまく言い当てている（つまり、素敵で優秀なモデル）ということになります。

つまり、黒い直線の長さを計算して、それを全部合計したものができるだけ短くなる「理想の β のセット」を探すことが、重回帰分析の原理です。

$\beta 1$ ○する	$\beta 2$ ○する	$\beta 3$ ○する	$\beta 4$ ○する	$\beta 5$ ○する	$\beta 6$ ×しない	$\beta 0$ ○する		$y = \beta_0 + \beta_1$
12	4	① -5	5	0	0	0	切片	とモデル化されると仮
変数 x_1	変数 x_2	変数 x_3	変数 x_4	変数 x_5	変数 x_6	目的変数 y		⑤ 9770.4536
学習時間	睡眠時間	スマホ利用時間	朝食習慣の有無	余暇の時間		テストの点数	モデルからの予測値	残差	残差の2乗
1.69	5.16	4.21	0	2.01		24	19.87	4.13	17.0569
4.28	8.18	1.42	1	3.43		96	81.98	14.02	196.5604
3.29	6.57	1.81	1	2.63		47	② 61.71	③ -14.71	④ 216.3841

ワークシートのつくりも案外単純です。

重回帰分析のボタンを押すと、①の β のセットの値をランダムで少しずつ変化させ、それに基づいて②の予想値を各データから計算します。③は実際の y の値（今の場合はテストの点数）と、モデルから計算された予想値の差分です。このまま合計するとプラスマイナスを含むので、それを足し合わせても合計の長さにはなりません。そこで、③を 2 乗して、④を計算してから、④の全データの合計である⑤を計算しています。（だから、正確には黒い線の長さの 2 乗の和を計算しています。）例えば試行回数を 1000 回とすると、1000 回 β をさまざま変えた中で、最も⑤の値が小さくなった β のセットを、最適なモデルとして記録します。