**回帰分析・重回帰分析**

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 |

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。

**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？



０

0.4

0.7

1.0

-0.4

-0.7

-1.0

**相関**

**係数**

**相関なし**

**弱い**

**弱い**

**負の相関**

**正の相関**

**相関係数と相関の判断（例）**

**強い**

**強い**

**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* =

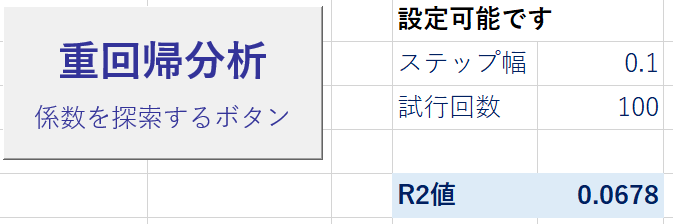
結果にあたる変数y(　　　　　　　変数という)　が　複数の変数（　　　　　　変数という）から影響を受ける場合、散布図相関行列で個別の変数の関係だけをながめてしまうと、実は相関がある変数同士を「相関なし」と判定してしまう可能性がある。

　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。

　※*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（β0、β1、β2、…）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。R2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、R2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…？



(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R2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、（R2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 |
| β1: | β2: | β3: | β4: | β0: |

R2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 |
| β1: | β2: | β3: | β4: | β5: | β0: |

**（【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。

　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。

**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。



**①**

**②**

**③**

**④**

**⑤**

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。