■■連載記事■

データをビジネスに活用する実践アナリティクス

<第 10 回>　回帰分析とは

梶山昌之
株式会社ワイハット

前回の連載では，商品 A を好む人は商品 B を好む傾向があるかということを示すために，相関係数を用いる方法について解説しました．ビジネスにおける分析では，2 つまたはそれ以上の変数間の関係について調べることが，課題に対する対策を検討する上で非常に重要になります．

例えば，ダイレクトメールの発送数と売上高の関係を分析し，発送数が売上高に寄与しているという結果が得られた場合は，発送数を増やすことで売上高の増加を期待できるかもしれません．今回はそのような分析の基礎となる回帰分析について解説します．

1. 親と子の身長

背の高さは遺伝するもののようです．両親共に背が高い場合，成人した子供の背も高くなる傾向があります．子供は親と同じ位の身長になるのでしょうか？

生物学者のゴルトン（1822-1911）は，　このことを検証するため，928 件の親と子の身長のデータ収集し，その関係を分析しました(Galton, 1875).

ここで，身長といっても，男性と女性の場合がありますので，女性の場合は身長を 1.08 倍した値を身長としています．ゴルトンは彼が持つデータから，男女の身長比の平均的な値として，この値を使用しました．ちなみに，日本の成人の男女身長比も概ね 1.08 になります．

ゴルトンの論文では，個々のデータを示していませんが，親の身長と子の身長の分割表を示し，各セルに該当する件数を示しています．

このデータをバブルチャートで表現するとデータの分布状況がわかります（図 1）．ここで，バブルの大きさは件数を表しています．

身長は正規分布に従いますので，中心付近のバブルが大きくなっています．例えば，親の身長 68.5（インチ）の級，子の身長 67.2（インチ）の級に該当する件数は 48 で，これが最も大きいバブルになっています．

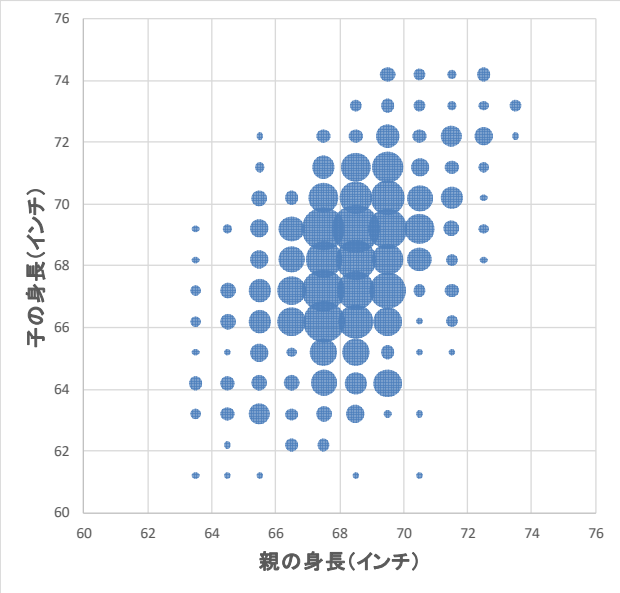

図 1　親の身長と子の身長（ゴルトン）
（注: ゴルトンのデータ(Galton, 1875)を筆者がバブルチャートを用いて表示した図です.）

2. 平均への回帰

図 1 は右肩上がりの楕円になっていますので，親の身長が高いほど子の身長も高いことがわかります．しかし，子の身長は親の身長と平均的には同じになるかどうかはよくわかりません．

そこで，親の身長の級毎に子の身長の平均値を計算できればよいのですが，論文には，平均値ではなく中央値が示されていますので，この値で代用することにします（表 1）．身長は正規分布に従うことがわかっており，平均値と中央値は近い値になるため，結果には大きな影響がないと考えら

れるためです.

親の身長（級代表値）と子の身長（中央値）との関係を可視化してみます（図 2）.

図 2 では，親の身長と子の身長が等しくなる場合を点線で示しています．また太い実践は表 1 の関係を示す回帰直線です.

最も身長が高い級で，子の身長が親の身長に近い値（外れ値）となっていますが，これはデータ数が少ないためのバラツキの影響であり除外可能です．すると，全体的には，子の身長（中央値）は親の身長と同じ値を表す線（破線）に沿った値にはならず，親の身長が低い範囲では，子の身長はその親よりは高くなり，親の身長が高い範囲では，この身長はその親よりは低くなるという結果が得られました.

表 1　親の身長（級代表値）と子の身長（中央値）

No.	親の身長(インチ)	子の身長(インチ)
1	72.5	72.2
2	71.5	69.9
3	70.5	69.5
4	69.5	68.9
5	68.5	68.2
6	67.5	67.6
7	66.5	67.2
8	65.5	66.7
9	64.5	65.8

（注: ゴルトンのデータ(Galton, 1875)より引用）

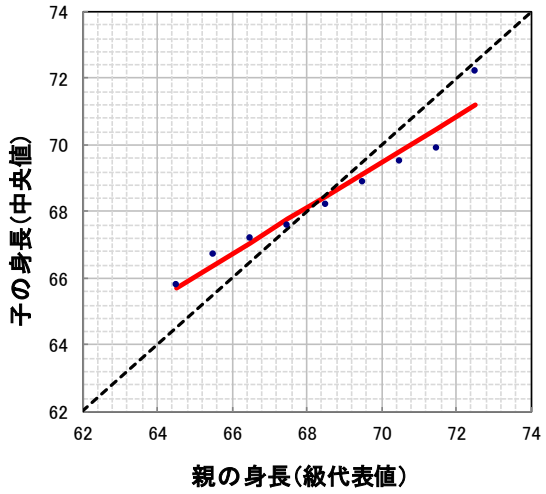

図 2　親の身長（級代表値）と子の身長（中央値）
（注: ゴルトンのデータ(Galton, 1875)を用いて
筆者が作図した図です.）

何故，このような結果になるのでしょう．このことは，もし子の身長の平均が親の身長に等しくなるならば何が起きるかを考えればわかります．例えば，180 センチの親の子の半数は親より背が高くなるとすれば，その子からはさらに背の高い子が生まれることになります．同様に背が低い親からは，さらに背が低い子が生まれることになり，これを繰り返せば，人類は次第に巨人族と小人族に分かれてしまうことになります.

そこで，ゴルトンは，自然の摂理により，親よりは平均に近い子供が生まれると考えました.

しかし，その後の研究で，自然の摂理とは無関係に，統計学的な現象としても平均への回帰（regression）が発生することを発見しました．値がある傾向へ帰っていくのです．これが，今日，予測ための分析手法として，最も活用されている回帰分析の語源です.

回帰分析の手法を利用した報告が間違っている場合がありますが，最も多いのは，この平均への回帰という現象があることを知らないまま，報告しているケースではないかと思います.

この現象は対象とするデータのバラツキが大きい時に発生しますので，親と子の身長の分析でも，自然の摂理として平均的な子が生まれるということ以外にもバラツキの影響を受けて平均への回帰が発生していると考えるべきということになります.

統計的現象として平均への回帰が発生するということのイメージは，まだ掴めないと思いますが，回帰分析の手法について具体的な計算方法を学んだ後に解説の予定です.

3. 身長と体重

回帰分析は 2 つまたはそれ以上の量の関係を調べる手法ですが，2 つの量の関係を調べる場合が，単回帰分布とよばれます.

最初は身近な例で考えると理解しやすいので，ここでは身長と体重の関係について調べましょう.

日本の 20 代の成人男性 100 名のサンプリング調査を行ったとします．得られたデータを散布図にすると身長が高いほど体重が増える傾向があることは明らかです（図 3）．しかし，個人差もかなりあるようです.

太りすぎ，痩せすぎは成人病の発生率が高くなりますので，健康診断では肥満度の指標の一つある BMI（Body Mass Index）を計算します.

$$\mathrm{BMI} = \frac{[\text{体重 kg}]}{[\text{身長 m}]^2}$$

この値が 17.6 以下の場合は痩せ過ぎ，26.5 以上の場合は太り過ぎとなります．この上下の境界値を散布図上に点線でプロットしています．皆さんは上下限の範囲に入っているでしょうか．

回帰分析により身長から体重を予測する式を作ることができます．100 名のデータから得られた予測式は，以下の通りです．

$$[\text{体重}] = -112.0 + 1.036\,[\text{身長}]$$

図 3 の直線（実線）は上記の予測式を表す直線です．

平均身長は 171.3[cm], 平均体重 65.5[kg]ですが，この値を示す点が散布図の重心になります．回帰分析により求められた直線（回帰直線）はこの重心を通る直線になります．

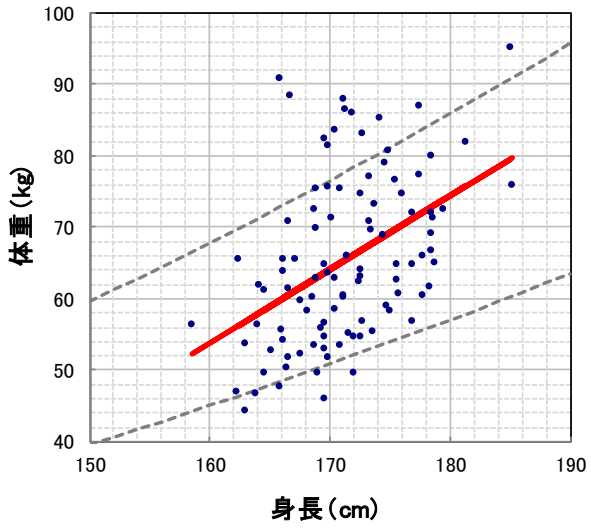

図 3　成人男性の身長と体重（100 名）
（注: 統計局の集計結果から得られる成人男性の身長と体重の，平均値と標準偏差を用いて，シミュレーションにより生成したデータです．）

4. 説明変数と目的変数

回帰分析では独立に変動する変数 x と，x に対応して従属的に定まる変数 y の関係を分析します．x と y の関係は必ずしも，直線的な関係ではありませんが，直線的な関係があるものとしてモデル化するのが線形回帰分析です．

それでは曲線的な関係は分析できないのかというとそうではなく，適切な変数変換により直線的な関係になれば，線形回帰分析の問題として扱うことができます．例えば，$y=a+bx^2$ は曲線的な関係ですが，$u=x^2$ とすれば，$y=a+bu$ となり，u と y は直線的な関係になります．

前述の親と子の身長の関係，身長と体重の関係は，いずれも x により y を説明する（予測する）関係になっていますので，x は説明変数，y は目的変数と呼ばれます．

説明変数（独立変数）と目的変数（従属変数）の関係を図示する場合は，説明変数を横軸，目的変数を縦軸に取るようにしましょう．これを逆にすると誤解や混乱のもとになります．

親の身長と子の身長の関係のように x が原因となって y の結果となる因果関係が成り立つ場合には，原因を横軸（x 軸），結果を縦軸（y 軸）にします．親と子の身長の場合は，親の身長が高いので，（遺伝により）子の身長が高いといえますので，親の身長が横軸になります．また，x と y の文字も x を説明変数，y を目的変数に対応させるのが慣例です．これを逆にするとわかりにくくなってしまいます．

尚，説明変数には制御できる変数と制御できない変数があります．例えば，成人男性の身長と体重の関係では，サンプリングの結果，身長の分布は正規分布になりますが，身長は制御できる変数ではありません．

それに対して，説明変数の値を変化させながら，効果を確認する実験の場合のように，制御できる変数の場合があります．

回帰分析はいずれのタイプの説明変数にも適用可能です．

5. 最小 2 乗法

回帰分析では散布図上の点に最も当てはまる直線を見つけることが課題ですが，最も当てはまるとはどういうことかについて考える必要があります．

当てはまりのよい直線とは直線からの差（残差）が小さいことですが，残差には正と負がありますので，合計をとっても全体的なバラツキの大きさを表しません．

そこで，残差の絶対値を取ってその合計が最も小さい直線を求める方法が考えられます．直観的にはこの方式はわかりやすいかもしれません．

もう一つの方法は，残差の 2 乗を取って，正の値にし，その合計が最も少なくなる直線を求めることです．

2 つの方法は異なった結果になりますが，どちらが優れているというものではありません．しかし，数学的には絶対値を取る方法は取り扱いが難

しくなるため，回帰分析では 2 乗の和を最小にする方法が使用されています（図 4）．

ここで，説明変数を x，目的変数を y とし，個々のデータを (x_i,y_i) （i=1,2…n）で表すことにします．ここで，x_i における予測値を

$$\hat{y}_i = a + bx_i$$

とします．a は回帰定数，b は回帰係数です．

回帰式により予測された値は，実績値と区別するため，文字に山形の記号（ハット）を冠し，y の予測値はワイハットと読みます．統計学ではハット記号は推定値であることを表します．

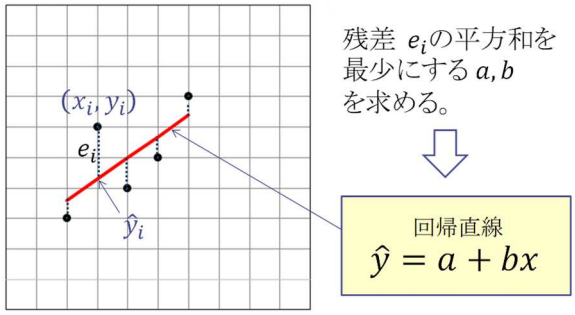

図 4　最小 2 乗法

実績値と予測値の差（残差） e_i は

$$e_i = y_i - \hat{y}_i$$

であり，e_i の平方和（残差平方和）Se は

$$Se = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$
$$= \sum (y_i - a - bx_i)^2$$

となりますが，この値はデータが直線から乖離する量を表すと考えることができます．最小 2 乗法はこの残差平方和 Se が最小になるように係数 a, b を求める方法です．

Se は a と b の関数ですので，a, b で Se を偏微分すると正規方程式と呼ばれる連立方程式が得られます．その連立方程式を解いて，a, b を求める公式が得られます．これらの計算内容は理解していなくても大丈夫ですが，計算の結果得られる以下の公式は重要です．

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Sxx は偏差平方和，Sxy は偏差積和と呼ばれる量です．変数にバー（－）の記号を冠した文字はその変数の平均を意味し，x の平均はエックスバーと読みます．

6. 回帰式の計算

実務での回帰分析の計算は，Excel の分析ツールや，統計解析ツールを使うことになると思います．

しかし，回帰分析の学習の最初の段階では，簡単な例を使って，具体的な計算方法を習得しておきましょう．それにより計算のメカニズムが理解でき，Excel などで回帰分析を組み込んだツールを作る場合にも，その知識が役立つことになります．

ここでは，100 名の身長と体重のデータから最初の 10 名のデータを用いて回帰式の計算方法を示すことにします（表 2）．

表 2　身長と体重（成人男性）

No	身長cm	体重kg
1	178.4	80.0
2	169.3	55.8
3	178.8	65.0
4	178.5	72.0
5	173.6	55.3
6	162.4	65.5
7	165.9	47.7
8	169.5	82.3
9	171.2	60.3
10	185.0	95.2

身長 x，体重 y の数値そのものを使うと扱う桁が大きくなりますので，計算のテクニックとして，仮平均を身長 170cm，体重 65kg として元のデータを変換し，変換後のデータを u, v とします(表 3).

表 3　身長と体重（変数変換後）
(u=x-170, v=y-65)

No	u	v	u^2	v^2	uv
1	8.4	15.0	70.56	225.00	126.00
2	−0.7	−9.2	0.49	84.64	6.44
3	8.8	0.0	77.44	0.00	0.00
4	8.5	7.0	72.25	49.00	59.50
5	3.6	−9.7	12.96	94.09	−34.92
6	−7.6	0.5	57.76	0.25	−3.80
7	−4.1	−17.3	16.81	299.29	70.93
8	−0.5	17.3	0.25	299.29	−8.65
9	1.2	−4.7	1.44	22.09	−5.64
10	15.0	30.2	225.00	912.04	453.00
合計	32.6	29.1	534.96	1985.69	662.86
平均	3.26	2.91	53.496	198.569	66.286

表 3 で計算した合計および平均の値を使って，以下の計算を行い，元のデータの平均，平方和，回帰係数などを求めます．

平均:

$\bar{x} = 170 + \bar{u} = 170 + 3.26 = 173.26$

$\bar{y} = 65 + \bar{v} = 65 + 2.91 = 67.91$

偏差平方和:

$$S_{xx} = \sum {u_i}^2 - \frac{(\sum u_i)^2}{n} = 534.96 - \frac{32.6^2}{10} = 428.684$$

$$S_{yy} = \sum {v_i}^2 - \frac{(\sum v_i)^2}{n} = 1985.69 - \frac{29.1^2}{10} = 1901.009$$

偏差積和:

$$S_{xy} = \sum u_i v_i - \frac{\sum u_i \sum v_i}{n}$$

$$= 662.86 - 32.6 \cdot 29.1/10 = 567.994$$

回帰係数:

$$\mathrm{b} = \frac{S_{xy}}{S_{xx}} = \frac{567.994}{428.684} = 1.32497$$

回帰定数:

$$\mathrm{a} = \bar{y} - b\bar{x} = 67.91 - 1.32497 \cdot 173.26 = -161.65$$

以上の計算の結果，回帰式は

$$\hat{y} = -161.7 + 1.325x$$

となります（図 5）．

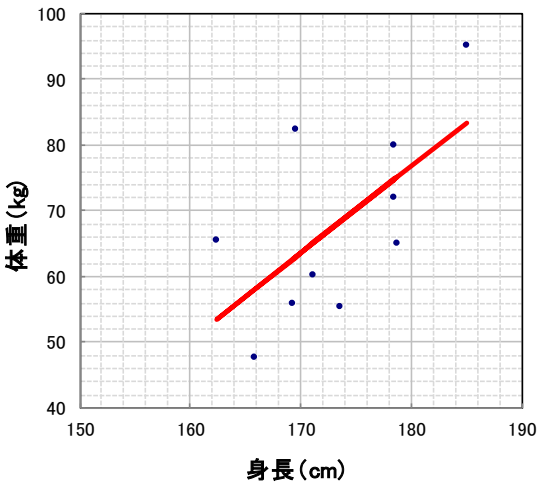

図 5 成人男性の身長と体重（10 名）

手順通りに計算すれば，回帰式を求めること自体は難しくないことがわかりました．

しかし，このケースでもわかるように，得られたパラメータ a, b の値は 100 名の場合とかなり異なっています．

今回は最初の 10 名のデータで計算しましたが，次の 10 名で計算すると，また異なった結果が得られることでしょう．このように，推定量 a, b はバラツキを持つ量です．

7. 回帰式の適用範囲

回帰分析による予測値は，説明変数がその平均から離れるほど信頼度が低くなります．成人男性 10 名の例では，身長 120cm の人の体重の予測値は -2.7kg と負の値になってしまいます．

従って，回帰式は説明変数のどの範囲でも使えるものではありません．基本的には説明変数のデータが存在する範囲内で使用します．この範囲内の予測を内挿とよびます．

しかし，実務ではこの範囲を超えて予測（外挿）したい場合も多いと思います．例えば，見積りでは，今までに受注した規模より大きい案件の見積りを行わなければならないことがあります．

適切な外挿の範囲は分析対象のデータの特性により異なると考えられるため，一般的な理論はないのですが，観測した説明変数の範囲の半分までとするのが無難です．ただし，外挿が妥当性を持つためには，説明変数の説明力が高いなどの条件が必要です．

成人男性 10 名の例では，観測された身長は 162.4～185.0 ですので，範囲は 22.6（=185.0-162.4）

であり，その半分 11.3（=22.6/2）まで拡張した範囲（151.1～196.3) を予測の範囲とします.

今回は手順に従い回帰式を求めるところまでを解説しましたが，得られた回帰式をそのまま信頼できるとは限りません．なぜなら，2 つの変数に関係がない場合でも，偶然の作用で正または負の相関が観測される場合があるためです.

次回は回帰式が信頼できるのかどうかを判定するための理論と手順をわかりやすく解説したいと思いますので，ご期待ください.

最後に本稿の理解に役立つ文献を以下に挙げます.

参考文献

Galton, F. (1875). *Anthropological Miscellanea*. Galton's Data (Diagramme de correlation cree par Francis Galton en 1875). http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf, (accessed 2017-5-16).

今泉忠，田村義保，中西寛子，美添泰人 (2012). 日本統計学会公式認定 統計検定 2 級対応 統計学基礎. 東京図書.