■■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 11 回>　信頼できる回帰式とは

梶山昌之
株式会社ワイハット

前回の連載では，10 名の男性の身長と体重のデータから，身長と体重の関係を表す回帰直線を求めました．

その結果，身長が高いほど体重が重いという結果が得られました．しかし，身長が低いが体重が重い人と，身長が高いが体重が軽い人が，サンプルとして選ばれた場合，回帰直線の傾きが負の値になり，「身長が高いほど体重が軽い」という結論になるかもしれません．

この例からわかるように，回帰式は常に真の関係を表しているとは言えないことを理解することは，とても重要です．

身長と体重の場合は，「身長が高いほど体重が軽い」という結論が誤りであることは経験的にわかりますが，通常は分析の対象とする変数の間にどのような関係があるかを知ることはできません．

従って，回帰分析を正しく使うためには，回帰式が信頼できるのかどうかを判断するための知識が必要になります．

1. バラツキはばらせ

回帰分析は目的変数の変動を，説明変数で説明される量と，説明変数では説明できない量に分解する手法であるともいえます．

身長は体重を説明する（予測する）ための一つの変数ですが，身長では説明できない要因（食生活や運動量など）の影響を受けて体重は変動します．

例えば，10 名の男性の身長と体重の例（図 1）では，体重の平均は 68.0kg ですが，その中の一人の男性は，身長 185cm，体重 95.2kg でした．この男性の平均からの増加分は 27.2kg（=95.2-68.0）です．

回帰式によれば，体重の予測値は 83.5kg ですので，平均からの 15.5kg (=83.5-68.0)　の増加は，身長が高いことで説明できます．それ以上の増加分 11.7kg (=95.2-83.5) は，身長以外の要因で増加したものと考えることができます．

このように全体の変動 27.2kg を回帰式により説明できる変動 15.5kg と，回帰式では説明できない変動（残差）11.7kg のように分解できたと考えることができます．

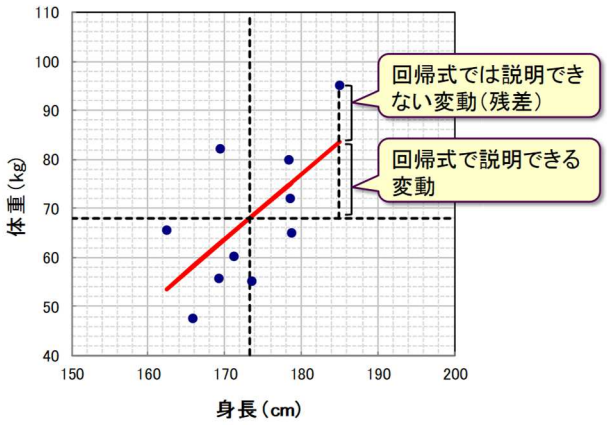

図 1　身長と体重（男性 10 名）

ここで，任意のデータにおける体重の平均 $\bar{y}$ からの変動を $y_i - \bar{y}$ とし，全体としての変動を考えることにします．

平均からの変動には正負があり，合計すると 0 になってしまいますので，平方して合計した値を計算するものとし，これを全平方和 S_T と呼びます．

回帰式によって説明できる変動，残差についても同様に考えて，以下の量を求めます（表 1）．

全平方和　：$S_T = \sum(y_i - \bar{y})^2 = 1901.009$
回帰平方和：$S_R = \sum(\hat{y}_i - \bar{y})^2 = 752.576$
残差平方和：$S_e = \sum(y_i - \hat{y}_i)^2 = 1148.433$

$\hat{y}_i$は前回の連載で求めた回帰式

$$\hat{y} = a + bx = -161.7 + 1.32x$$

による予測値です.

ここで，回帰平方和と残差平方和の合計を計算してみましょう．すると，不思議なことにその値 1901.009 (= 752.576+1148.433) は全平方和に一致します．すなわち,

$$S_T = S_R + S_e$$

という関係があります．このことは，全体の変動を回帰によって説明できる変動と説明できない変動に分けることができたことを意味します．これにより，説明変数が現象を説明できる程度を知ることができるようになりました.

表 1　平方和の計算（男性 10 名の身長と体重）

No	X	Y	$\hat{Y}$	S_T	S_R	Se
1	178.4	80.0	74.72	146.168	46.381	27.875
2	169.3	55.8	62.66	146.652	27.530	47.102
3	178.8	65.0	75.25	8.468	53.881	105.069
4	178.5	72.0	74.85	16.728	48.203	8.139
5	173.6	55.3	68.36	159.012	0.203	170.576
6	162.4	65.5	53.52	5.808	207.049	143.501
7	165.9	47.7	58.16	408.444	95.097	109.374
8	169.5	82.3	62.93	207.072	24.819	375.270
9	171.2	60.3	65.18	57.912	7.450	23.820
10	185.0	95.2	83.47	744.744	241.963	137.706
合計	1732.6	679.1	679.10	1901.009	752.576	1148.433

単回帰分析では一つの要因の効果を示しますが，複数の要因が現象に関与する場合は，要因毎にその大きさを把握することができます．これが「バラツキはばらせ」の意味です.

2. 寄与率とは

全平方和に対する回帰平方和の割合 S_R/S_T は説明変数が目的変数を説明できる程度を表す指標となり，これを寄与率 R^2 と呼びます.

寄与率 $R^2 = S_R/S_T$

また，この寄与率は経済学などでは決定係数とも呼ばれます．10 名の身長と体重の例では，寄与率 0.40 となり身長で体重の約 40%は説明できるという言い方で説明変数の効果を示すことができます.

尚，この寄与率の平方根は相関係数 r になります．相関係数を計算すると 0.63 (= 0.40 の平方根) となり「相関がある」と言うことができます.

ただし，寄与率が現象を説明できる割合という明確な意味があるのに対して，相関係数はその数値の意味が曖昧です．そこで，実務では寄与率（または決定係数）で説明することを推奨します.

寄与率は回帰式が使用できる場合に，どの程度の説明力があるかを示す数値であり，回帰式が信頼できるかどうかを判定する数値ではありません．この点は注意する必要があります.

例えば，寄与率が 0.9 以上であっても，回帰式が信頼できないケースもあるということですが，その意味については後ほど解説します.

3. 分散分析の方法と解釈の仕方

測定値の変動を各要因（因子）に起因する変動と誤差に起因する変動に分解する手法が分散分析です.

表 1 では個々のデータ毎に変動を計算しそれを集計して S_T，S_R，S_eを算出しましたが，通常は，前回の連載で解説した偏差平方和 S_{xx}，S_{yy} および偏差積和 S_{xy} の計算結果を用いて，以下の公式で計算します.

$$S_T = S_{yy} = 1901.009$$

$$S_R = {S_{xy}}^2/S_{xx} = 567.994^2/428.684 = 752.576$$

$$S_e = S_T - S_R = 1901.009 - 752.576 = 1148.433$$

次に各変動の自由度について考えます.

全変動 S_T の計算には 10 個の y_i が計算に使われていますが，$\bar{y}$ は その 10 個のデータから計算されていますので，全変動の自由度 f_T は 10 より 1 少ない値の 9 になります.

回帰変動 S_R の自由度 f_R については，単回帰分析では説明変数が 1 つだけなので，自由度は 1 になると考えてください.

残差変動 S_e の自由度 f_e については，e_i を求めるために，回帰直線のパラメータである a と b の 2 つを推定しなければならないので，自由度は 2 少ない値の 8 になります.

その結果，各変動の自由度についても，平方和の場合と同様の以下の関係が成立しています.

$$f_T = f_R + f_e = 1 + 8 = 9$$

回帰分析では誤差によるバラツキより回帰によ

るバラツキの方が十分に大きければ回帰式は意味があると考えます.

このバラツキ（分散）は平方和を自由度で割って求めます. 単回帰分析の場合, 回帰の自由度は 1 ですので，回帰分散は回帰平方和に等しくなります．残差分散は残差平方和を残差の自由度で割った値になります.

以上の関係を表にまとめたものが回帰の分散分析表（ANOVA: ANalysis Of VAriance）です（表 2）.

表 2　分散分析表（男性 10 名の身長と体重）

変動因	自由度 f	平方和 S	分散 V	観測された分散比 Fo	有意F (5%)	p値
回帰 R	1	752.576	752.576	5.242	5.318	0.051
残差 e	8	1148.433	143.554			
全体 T	9	1901.009				

分散分析表で，回帰分散と残差分散の比 F_O (=V_R/V_e) を計算していますが，この F_O が回帰式を使用可能かどうか判定するための指標になります．F_O は収集したデータから計算される量ですので，「観測された分散比」とも呼ばれます.

偶然の誤差（残差）による変動に対して，回帰式で説明される変動が大きくなれば，F_O も大きくなります．従って，F_O が大きいほど回帰式に意味があると考えることができます.

F_O は F 分布と呼ばれる分布に従うことがわかっており，これを利用して得られた回帰式に意味があるかどうかを判定します.

F 分布についてはまだ説明していませんので，ここでは, F_O が大きくなるほど残差の変動と回帰による変動の大きさに違いがあることを見誤らない確率が少なくなり，F_O がある値以上になる確率が 5%の時のF_O の値が「有意 F(5%)」として計算されているものと理解してください.

男性 10 名の身長と体重のケースでは，F_O = 5.242 に対して，有意 F=5.318 ですので，F_O は有意 F より小さく，回帰式に意味があるとは言えないという結論になります．すなわち，偶然の作用で右上がりの直線が得られた可能性を捨てきれないということです.

この様に「身長が高いほど体重が重くなる」ということが真実であったとしても，分析対象のデータからはそのことを証明できない場合があります．データ数が少なくバラツキが大きい場合に，そのような結果になると言えます.

ここで，表 2 の一番右の列に計算されている p 値を見てください．この値は，もし 2 変量の関係が無相関であった場合に，b=0 ではない値が観測される確率を示しています.

身長と体重のケースで p 値は 0.051 であり，わずかに 0.050（5%）を超えたために，回帰式は有意でないと判定したことになります.

この 5%は有意水準を判定するために, 統計的慣例として使用されている基準ですが，実務的には p 値をみて，それが許容される値であるかどうかで，回帰式を使用するかどうかを決めることを推奨します.

この例では，身長から体重を予測する式を作りたいが，無相関の場合でも b=1.32 となる確率が 5.1% ということになります. その誤りの確率が許容できるならば，回帰式を使用すればよいということになります.

4. ダイレクトメールと成約数

前章では，身長と体重の関係について分析しました．身長が高いほど体重が重くなるのは自明の知見ですが，実務では相関があるかどうかも不明な場合があります.

例えば，あなたは不動産会社の営業担当者であるとします．新しい物件はホームページや SNS でも広告していますが，それに加えて，ダイレクトメール（DM）を発送して成約に努めています.

ところが，5 回分の DM を発送した時点で，500 通の DM を発送したときの成約数が, 300 通の DM を発送したときの成約数より少ないケースもあることがわかりました（図 2）.

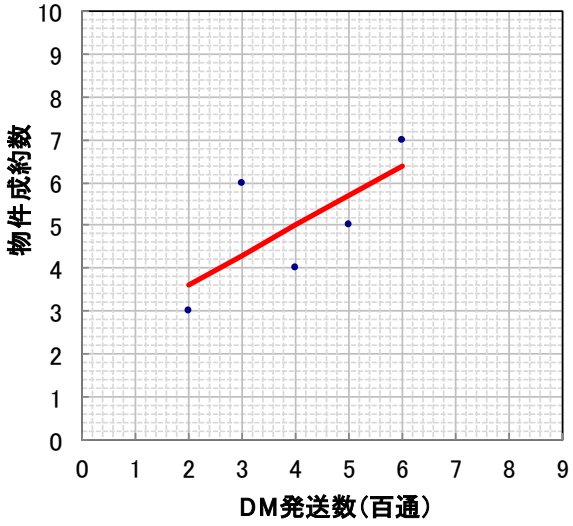

図 2　DM 発送数と成約数（5 回分）

身長と体重の場合と同様に Excel で集計を行い．回帰係数を公式で求めた結果は以下の通りです．

$$b = \frac{S_{xy}}{S_{xx}} = \frac{7.000}{10.000} = 0.700$$

$$a = \bar{y} - b\bar{x} = 5.000 - 0.700 \cdot 4.000 = 2.200$$

また，分散分析表は以下の通りです（表 3）．

表 3　分散分析表（DM 発送数と成約数）

変動因	平方和S	自由度f	分散V	分散比Fo	有意F	p値
回帰 R	4.9	1	4.9	2.882	10.128	0.188
残差 e	5.1	3	1.7			
合計 T	10.0	4	2.5			

回帰係数は 0.7 であり，DM を多く発送するほど成約数が見込めることになるのですが，果たしてこの数字は信頼できるのでしょうか．

もし，DM の発送数は成約数とは無関係である場合も，成約数は他の要因で変化します．その結果，偶然の作用により，正の相関が観測されただけかもしれません．

次回の売り出しで DM を発送した方がいいのか，またはその予算を他の広告に使用した方がいいのかについて悩んでいる状況です．

5. 単回帰モデル

我々は，収集したデータ（サンプル）からしか，真の関係を推し量ることができません．ここで，真の関係とサンプリングされたデータの関係のイメージを示します（図 3）．

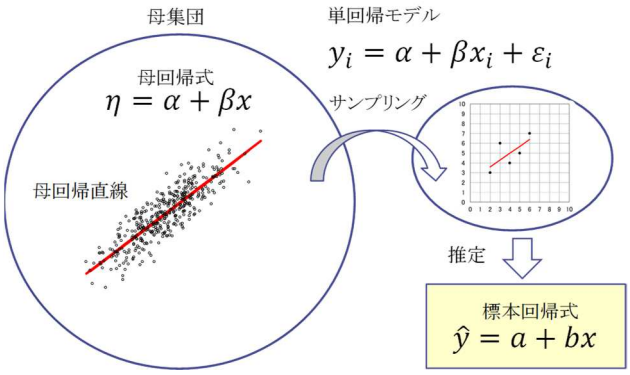

図 3　母回帰直線と標本回帰式

ここで真の回帰定数および回帰係数をα，βとしましょう．これらは真の値であるため，母回帰定数，母回帰係数と呼ばれます．何故，ギリシャ文字をつかうのかについては連載の第 7 回で解説していますが，統計学では母数をギリシャ文字で表す慣例になっているためです．母数は確率変数ではなく定数（未知定数）です．

左側の円内にある無数の点は分析対象となるすべてのデータを示しています．また，この円内に描かれている直線は，分析対象のすべてのデータで計算された回帰直線ですので，母回帰直線と呼ばれます．

DM 発送数と成約数の例でいえば，今後の売り出しで毎回 DM の発送を繰り返し，その結果得られる成約数を記録したとすれば，回帰直線はある一つの直線に収束すると考えられます．その直線が母回帰直線です．また，その直線を表す回帰式は母回帰式と呼ばれます．

6. 母回帰式の推定

母回帰直線は神のみぞ知るものなので，我々は，サンプリングしたデータでこの直線の係数を推定しようとしています．

図 3 では母集団として右肩上がりの点の集合が描かれており，母回帰直線も右肩上がりの直線になっていますが，DM 発送数と成約数が無関係ならば，円状または水平の帯状に散らばった分布になり，母回帰直線の傾きも 0 になります．

一方，右側の円内にはサンプリングされたデータが散布図にプロットされています．このデータから計算された回帰式は標本回帰式と呼びます．

ここで，回帰係数 b に対して，Ho: $\beta=0$ の検定を行うことを考えます．そのためには，b の期待値と分散を求める必要があります．

統計の理論より，b は正規分布に従い，期待値 E[b] と分散 V[b]は

$E[b]=\beta$
$V[b]=\sigma^2/Sxx$

となることを示すことができます．ここで，σ^2 は残差分散です．

この式の導出はやや複雑なのですが，b が β の不偏推定量になっており，b のバラツキは残差（予測と実績の差）が小さく x の変域が広いほど小さくなると理解してください．σ は未知の母数ですので，σ^2 の代わりに，その不偏推定量 Ve を使

用することにします.

ここで，b の標準偏差を s(b) とすれば

$$\mathrm{s(b)} = \sqrt{V[b]} = \sqrt{Ve/Sxx} = \sqrt{1.7/10} = 0.4123$$

となります. ここからは, 連載第 7 回で解説した t 検定と同じ内容になり b を標準化した

$$\mathrm{t} = (b - \beta)/\mathrm{s(b)}$$

は自由度 n-2 の t 分布に従います.

ここでは，β=0（DM 発送数は成約数と無関係）ではないことを示したいので, 帰無仮説 Ho: β=0 となり，検定統計量 t_o は

$$t_o = \frac{\mathrm{b}}{\mathrm{s(b)}} = 0.7/0.4123 = 1.698$$

となります. t 分布表で自由度 3（=5-2) の両側確率 0.05 の値は 3.182 ですが，t_o はこれより小さな値となっており，回帰係数は有意ではないという判定になります.

7. ダイレクトメールの効果は？

前章の分析で回帰係数は有意ではないという判定になりましたが，この結果は，DM に効果がないことを積極的に証明しているわけではなく，データ数が不足しているため，残念ながら DM の効果を証明できなかったと解釈すべきです.

回帰係数は正の値であり，DM の効果を期待できる状況であるため，引き続き，DM を継続することになります.

次の売り出しでは 700 通の DM を発送し, 成約数が 8 件となりました（図 4）.

その結果，$t_o = 2.960$ となります． t 分布表で自由度 4（=6-2) の両側確率 0.05 の値は 2.776 ですので，t_o はこれより大きな値となっており，有意水準 5% で DM は効果があることを示すことができました.

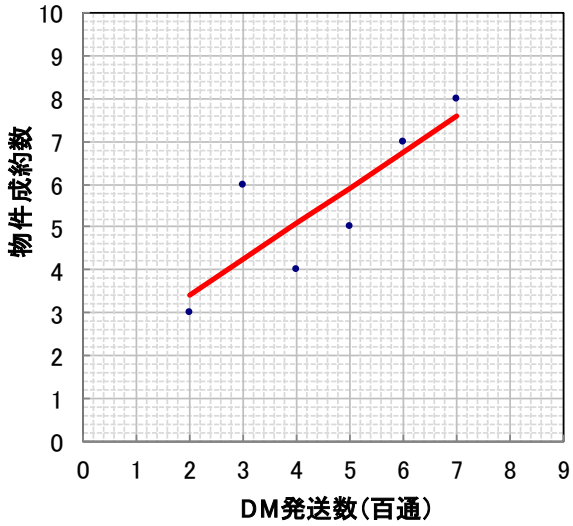

図 4　DM 発送数と成約数（6 回分）

単回帰分析では，説明変数が 1 つですので，その回帰係数が有意となった場合は，回帰式に意味があると考えます.

説明変数が 2 つ以上の場合は，最初の変数は有意だが，2 番目の変数は有意でないというケースも考えられますので，回帰式が使用可能かどうかという判断と回帰係数に意味があるかどうかという判断は異なります.

回帰式に意味があるかどうか，言い換えればその回帰式を使用可能かどうかの判定は分散分析表の分散比 F_o を使い F 分布を用いた検定を行います.

単回帰分析の場合は

$$t_o = \sqrt{F_o}$$

の関係があります.

例えば，表 3 の分散分析表では $F_o = 2.882$ ですが，その平方根は 1.698 となり，t_o に一致しています．すなわち，単回帰分析の場合に限り，この 2 つの量は，同じ判断を表す統計量です.

F 分布と F 検定については説明変数が 2 つ以上の回帰分析（重回帰分析）でその知識が必要となりますが，詳細は後日の連載で解説の予定です.

次回は引き続き回帰分析で理解すべき各種の統計量について解説します.

実務では予測と判断のために使われる手法としては，回帰分析が最も活用されているため，本連載では回帰分析の活用に必要な知識については詳

細に解説の予定です.

今回は回帰分析の計算を Excel の表計算で行いましたが，統計分析のツールを使えば簡単に回帰式が得られるため，通常は，そのようなツールを使用して計算することになると思います.

しかし，回帰分析を十分に理解しないまま使用した結果，誤った結論になってしまっている事例をよく目にします.

回帰分析の正しい利用の仕方については，ここに解説したこと以外にも，いろいろと考慮すべき事項があります．本連載では，それらの事例を示しながら，できるだけわかりやすく解説する予定ですのでご期待ください.

参考文献

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.