データをビジネスに活用する実践アナリティクス

<第 12 回>　回帰分析の結果を解釈する

梶山昌之
株式会社ワイハット

前回までの連載で，回帰分析は Excel の表計算のみを用いて行ってきました．その方法を習得することは，回帰分析の基礎を学び，自身で回帰分析を利用したツールを開発するためには必要です．

しかしながら，表計算による方法で，回帰分析の結果として必要になる様々な値（統計量）を計算するのは大変です．

一方, Excel にはアドインツールとして各種の統計解析ができる「分析ツール」が提供されています．今回は，このツールを用いて回帰分析を実行してみることにしましょう．

ただし，Excel の導入時点では「データ分析」は利用可能な状態になっていません. Excel を起動し「データ」タブのメニューに「データ分析」がない場合は,「ファイル」タブの「オプション」にあるアドインで有効にしてください．

1. 標準偏差と標準誤差

連載の第 7 回では，ある集団から 5 名の身長のデータをサンプリングして標本平均を求めました．この標本平均は元の集団の母平均の不偏推定量になることについて解説しました．

また, 同じデータから, 標本分散を求めました．この標本分散は元の集団の母分散の不偏推定量になります．さらに，標本分散の平方根を計算したものが，標本標準偏差です．

一方，サンプリングを繰り返して，標本平均を求めると，この平均はいつでも同じ値になるわけではなく，ある分布に従う量になります．すなわち，標本平均もバラツキを持つ量ですので，その標準偏差を計算することができます. このような, 推定量の標準偏差は真の値との差（誤差）を示すものですので「標準誤差」と呼ばれます．

例えば，身長の例では，母集団の標準偏差（母標準偏差）が 5cm の場合，100 名による標本平均の標準誤差は 0.5cm (=5/√100) になります．

2. 身長と体重の分析

前回の連載では，男性 10 名の身長と体重の関係の分析を Excel の表計算で行いました（表 1).

表 1　男性 10 名の身長と体重

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

単回帰分析では 2 変量を分析対象にしますが，この場合，個々の変量の統計量を最初に分析するのが基本です．ここでは，まず体重の基本統計量を Excel の分析ツールで計算します．

まず,「データ分析」をクリックします. すると, 分析手法を選択する画面が現れますので,「基本統計量」を選択しクリックします．今度は分析対象と出力の内容を指定する画面が現れます（図 1).

ここでは，身長と体重の先頭行のラベルも含めて入力範囲とし，出力オプションは「統計情報」を指定して,「OK」をクリックします. その結果, 平均，標準偏差など，各種の統計量が得られました（表 2）.

ここで，標準誤差は,

$$[\text{標準誤差}] = \frac{[\text{標準偏差}]}{\sqrt{n}}$$

で計算されていることを確認してください．

例えば，身長の場合は，

$$[\text{標準誤差}] = \frac{[\text{標準偏差}]}{\sqrt{n}} = \frac{6.90156}{\sqrt{10}} = 2.182$$

となります．

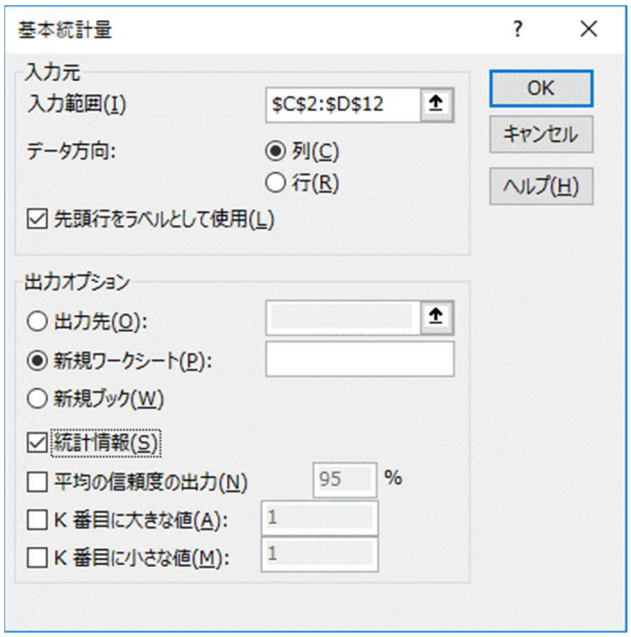

図 1　基本統計量の入力画面

表 2　身長と体重の基本統計量

身長cm		体重kg	
平均	173.26	平均	67.91
標準誤差	2.18246548	標準誤差	4.59590276
中央値 （メジアン）	172.4	中央値 （メジアン）	65.25
最頻値 （モード）	#N/A	最頻値 （モード）	#N/A
標準偏差	6.90156182	標準偏差	14.5335206
分散	47.6315556	分散	211.223222
尖度	-0.6165363	尖度	-0.2484465
歪度	0.11167938	歪度	0.5711668
範囲	22.6	範囲	47.5
最小	162.4	最小	47.7
最大	185	最大	95.2
合計	1732.6	合計	679.1
データの個数	10	データの個数	10

次に 「データ分析」の手法選択画面で，「回帰分析」を選択しクリックします．すると分析対象と出力の内容を指定する画面が現れます（図 2).

ここでは「ラベル」にチェックをし，変数のラベルも含めてXおよびYの範囲を指定することにします．ここで，X が身長，Y が体重になります．

有意水準を指定するところがある点に注意してください．有意水準とは珍しさの基準であり，統計的慣例では，5%または 1%の値になります．

しかし，Excel のこの欄には，信頼係数（または信頼度）を入力する仕様になっています．

信頼係数から推定値の上下限の範囲（信頼区間）を計算しますが，95%を指定した場合，上下限の範囲の中に真の値（母数）が含まれる確率が 95%あるという意味になります．尚，信頼係数 95%の結果はディフォルトで出力されますので，それ以外の値，例えば 99%を指定するのがよいと思います．

その他，各種の計算結果やグラフを作成するためのチェックボックスがありますが，ここでは，それらはディフォルトの状態（指定しない）で「OK」ボタンをクリックすることにします．

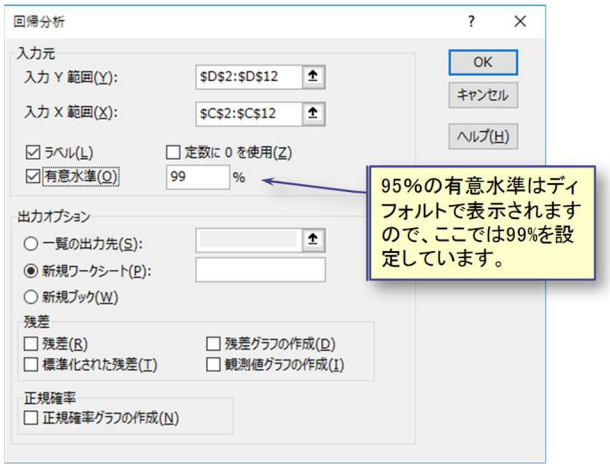

図 2　回帰分析の入力画面（Excel 回帰分析）

分析結果として，回帰統計，分散分析表，回帰係数などが得られますが，実務ではこれらの結果に表示されている数値の意味を正しく読み取れるようになる必要があります（表 3).

表 3　回帰分析の結果（Excel 回帰分析）

概要

回帰統計	
重相関 R	0.629192
重決定 R2	0.395882
補正 R2	0.320368
標準誤差	11.98141
観測数	10

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	752.5758	752.5758	5.242451869	0.051294
残差	8	1148.433	143.5542		
合計	9	1901.009			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 99.0%	上限 99.0%
切片	-161.655	100.3338	-1.61117	0.14580894	-393.025	69.71572	-498.313	175.0044
身長cm	1.324971	0.578681	2.28964	0.051293659	-0.00947	2.659412	-0.61673	3.26667

Excel の操作手順に従えば, 簡単に結果が得られるため，計算ができるようになったことで，回帰分析ができるようになったと思いがちなのですが, 出力された数字の意味を正しく解釈していない人が多いというのが現状です.

本連載では，これらの意味を正しく理解することを目標の一つとしています.

「回帰統計」では，重相関 R，重決定 R2， 補正 R2， 標準誤差，観測数の値が計算されています．Excel の「回帰分析」ツールは，説明変数が 1 つの単回帰分析のみならず，2 つ以上の重回帰分析にも対応できるツールであるため,「重」の文字が使われています．従って，単回帰分析の用語では，重相関 R は相関係数 R のことです．その 2 乗が，決定係数 R2 (寄与率 R^2) になります.

寄与率については，前回の連載では S_R/S_T の値として解説しましたが, $S_T=S_R+Se$ の関係があるので,

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

と考えることができます.

この回帰変動 S_R, 残差変動 Se を, 回帰分散 V_R ，残差分散 Ve で置き換えたものが，自由度調整済寄与率 R^{*2} です．Excel の回帰分析では，この値を「補正 R2」と表示しています.

$$R^{*2} = 1 - \frac{V_e}{V_T} = 1 - \frac{S_e}{S_T} \cdot \frac{n-1}{n-q-1}$$

ここで n はデータの個数，q は説明変数の数です. 寄与率 R^2 はモデルの選択基準として使われる値ですが，重回帰分析では説明変数の数が多いほど寄与率 R^2 は高い値になります.

しかし，説明変数の数が多いほど回帰式の安定性が低下するため，不安定な回帰式が選択されてしまう危険性があります.

それを回避するために，説明変数の多さをペナルティとして科し調整した寄与率が自由度調整済寄与率です. この式では q の値が大きいほど R^{*2} は小さくなります.

3．標準誤差（残差標準偏差）

次に回帰統計で「標準誤差」と表示されている統計量に注目します.

回帰統計の下にある分散分析表では，残差のバラツキを表す統計量である「残差分散」が計算されていますが，この平方根を取った値が残差の標準誤差 se (standard error) です.

$$\mathrm{se} = \sqrt{V_e} = \sqrt{143.554} = 11.981$$

残差の標準偏差は「残差標準偏差」とも呼ばれます．この用語の方が，残差の標準偏差であることがわかりやすいので，本連載では，残差については, 残差標準偏差の用語を用いることにします.

残差は実績値と回帰式による推定値の差のことですが，回帰分析ではこの値が正規分布に従うことを前提としています．従って，この値は回帰式の当てはまりの良さを示すと同時に，予測値がどの範囲で信頼できるかを示すために使用できます.

尚，本連載では残差標準偏差を表す記号を se と表現しました．本来は，s は小文字で e はその小文字の添え字なのですが，表記上は Se (残差平方和）に似てしまいますので，あえて se と表記しています.

残差標準偏差は線形のモデルに限らず，非線形のモデルにも計算できる量であるため，モデルの適合性の比較に使用できる重要な統計量です.

4．観測された分散比と有意 F

次に，分散分析表で計算されている，観測された分散比と有意 F の意味について説明します.

例えば，ダイエット中に体重を測り，前回の測定値は 65.8 kg だったが，今回は 65.7kg であった場合に，直ちにダイエットの効果があったと考えるでしょうか.

ダイエットをしていない場合でも，計測される体重は±0.5kg 程度は変動するものであることを,

直観的に知っていますので，効果があったとは考えないと思います.

それが 62.8kg まで体重が減少した場合は，明らかなダイエットの効果を認めると思います．すなわち，私達も日常的には，偶発的な変動の大きさと比較して効果を判断しているということになります.

その比較の手段として, バラツキの比（分散比）を考えるが自然です．このような分散の比は F 比とも言います.

「観測された分散比」は回帰分散 V_R と残差分散 Ve の比です．また，この値を F_O と表示するのが慣例です.

$$F_O = \frac{V_R}{V_e} = \frac{752.5758}{143.5542} = 5.242$$

残差の変動より回帰式で説明される変動の方が大きいほど，F_O の値は大きくなりますが，この値は回帰式が使用可能かどうかの判断に使用されます.

Excel では観測された分散比の隣に「有意 F」として値が示されていますが，この値は Excel で F.DIST.RT 関数（F 分布の右側確率を求める関数）を使って以下のように求めることができます.

$$\mathrm{F.DIST.RT}(F_O, 1, 8) = 0.0512937$$

すなわち，F 値がF_O以上となる確率を表しており，限界水準（有意水準の限界）または p 値と呼ばれている統計量と同じです．また，関数の入力パラメータとして分母と分子の自由度を与えます. ここで，1 は分母（回帰）の自由度，8 は分子（残差）の自由度です.

検定の有意水準を 5%とすれば，この例ではわずかに 5%を超えており，「回帰式は有意でない」という結論になります.

このようにして分散比を判定する方式は F 検定と呼ばれます．ちなみに，F の文字は分散比による統計を最初に開発したフィッシャー（Fisher）にちなんだものです.

連載の第 7 回では 2 つの集団の平均差を検定するために t 検定を用いましたが，考え方は F 検定も同じです．F 検定では，「2 つの集団の分散に差がないとすれば, 分散比は 1 に近いと考えられるが，1 より非常に大きい値が得られた場合には，2 つ集団の分散に差がないという仮説は間違いではないか」と考えます．この判断が誤りである確率が有意 F で計算されています.

5. 回帰係数と信頼区間

分散分析表の下に回帰係数などの計算結果が表で示されています.

紛らわしいのですが，この表は分散分析表の一部ではありません．この表では回帰係数に関する統計量が計算されています.

注目すべき点は，観測された分散比の平方根が身長の t 値 となっている点です. 前回の連載でも解説しましたが，回帰式の p 値（有意 F）は身長の回帰係数の p 値に等しい値となります.

このことは，単回帰分析の場合,「回帰式に意味がある」という判断は「説明変数の回帰係数に意味がある」という判断と同じであることを示しています.

説明変数が 2 つ以上の場合は，一部の説明変数は有意でないが，回帰式としては有意であるという状況が発生しますので，有意 F は回帰式全体に対する判断，説明変数の p 値は回帰係数に対する判断として区別することになります.

実際に説明変数が 2 つ以上の場合の分析（重回帰分析）を行ってみないとイメージが掴めないと思いますので，後日の連載で解説の予定です.

さらに,「P-値」の右側には，回帰係数の信頼区間が計算されています．最初に 95%の信頼区間の上下限の値が示されています．身長の回帰係数の下限は -0.00947，上限は 2.65941 となりました．回帰係数の 95%はこの範囲に入るということを示しています.

この上下限の値は，回帰係数の推定値 1.324971 と標準誤差 0.578681 から求めることができます. 例えば上限の値は，Excel の T.INV 関数（t 分布の左側逆関数）を使って，以下の様に求めることができます.

$$b_u = b + s(b) \cdot T.INV(0.975, 8)$$

$$= 1.3250 + 0.5787 \cdot 2.3060 = 2.659$$

ここで，b_u は回帰係数の上限，b は回帰係数の推定値，s(b)は b の標準偏差です.

6. 変動係数

Excel の回帰分析では計算結果を表示していませ

んが，変動係数（CV: Coefficient of　Variation）

CV = [標準偏差]/[平均]

は，異なるデータ同士でバラツキを比較するために重要な統計量ですので，ここで説明しておきます.

標準偏差はバラツキを表す量ですが，バラツキの大きさを比較できる量ではありません. 例えば，40 代男性の体重の標準偏差は 9.9kg，同年代の女性の標準偏差は 7.9kg ですが，この結果で，男性の方がバラツキが大きいとはいえません．何故なら，男性の方が重いため，標準偏差も大きくなるためです.

そこで，平均体重に対する標準偏差の割合を考えると，両者の比較が可能になります．40 代の体重のケースでは男性の平均体重は 68.9kg，女性の平均体重は 53.6kg ですが，これらから変動係数を計算すると，男性は 14.4%，女性は 14.7% になります．その結果，40 代の場合は，女性の方がバラツキが大きいということになります.

余談になりますが，20 代の場合は女性の方が変動係数が小さくなります.

尚，変動係数は無名数（単位のない値）であるため，身長と体重など単位の異なる量の比較も可能です.

例えば，表 1 の身長と体重の例では，表 2 で平均と標準偏差が計算されていますので，これを用いて変動係数を計算すると，

[身長の CV] = 6.902/173.26 = 0.040 = 4%
[体重の CV] = 14.534/67.91 = 0.215 = 21.5%

となり，体重は身長に比べて約 5 倍バラツキが大きいことがわかります.

回帰分析でも予測値のバラツキの大きさを評価するために変動係数が使われています．回帰分析における変動係数は残差標準偏差を目的変数の平均で割ったものです。

CV = [残差標準偏差]/[目的変数の平均]

この値は予測値のバラツキを標準化したものと考えることができます.

例えば，身長から体重を予測する回帰式における変動係数は，

CV = [残差標準偏差]/[体重の平均]
= 11.98141/67.91 = 0.176

すなわち，変動係数は約 18% になります.

回帰分析はプロジェクトのコストを予測するためにも活用されています．今，2 種類の見積り手法があり，その予測精度を比較したいとします.

しかし，予測誤差の大きさはプロジェクトの規模に依存しますので，残差標準偏差では比較できません．この場合も，変動係数を用います．コスト見積りの知識体系として CEBoK（Cost Estimating Body of Knowledge）がありますが，この体系では，コスト予測を行う場合は，変動係数が15%未満であることが望ましいとしています.

本連載の第 10 回では表計算で回帰分析の計算を行いましたが，Excel の分析ツールを使えば，簡単に結果が得られることが確認できたかと思います. しかしながら，得られた結果の解釈には統計学の知識が必要でした.

7. Excel と R の使い分け

今回は Excel の分析ツールによる回帰分析の方法を学びました. Excel は多くの方が利用しているソフトウェアですので，分析結果を共有するためには必須のツールと言えます．また，現場で活用するためのツールを開発し，それを配布する場合にも最適です.

しかしながら，高度な統計解析を行う場合は，市販ツールまたは統計解析用のフリーソフトを使うのがよいでしょう.

特に，統計解析フリーソフトとして知られている R は学術論文でも多用されており，信頼性のあるソフトと言えます．そこで，本連載でも，R を用いた分析を解説の予定ですのでご期待ください.

Excel と R は，それぞれ良い点がありますので，実務では使い分けながら活用することを推奨します.

参考文献

今泉忠，田村義保，中西寛子，美添泰人 (2015). 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.

末吉正成，末吉美喜 (2017). EXCEL ビジネス統計分析［ビジテク］第 3 版. 翔泳社.