



第7回 重回帰分析

問題解決に役立つSQC手法（統計的品質管理の手法）の解説の第7回目です。

今回は、結果系の1つの特性値（長さ、重さ、など）に着目したときにその原因の可能性があると思われる複数の特性値について、相関関係がありそうかどうか、あるとするとどのような関係式で表せるかを導く、重回帰分析について解説します。

結果系の特性値 y に対して、原因系の候補が x_1 と1つしかない場合は、第6回で紹介した単回帰分析を行うことになります。しかし、原因系の候補が例えば温度だけではなく、湿度や圧力など複数の候補が考えられ、且つそれらのいずれもが関係していると考えられる場合があります。仮に、それぞれ要因の1つである、温度と y 、湿度と y 、圧力と y を個々に単回帰分析を行っても、同時にどのように関連しているかを説明することは困難です。

このようなときに役立つSQC手法が重回帰分析です。特に、加工工程や作業工程における測定データの記録など、多くのデータを取得済みの場合に威力を発揮します。

ここではイメージしやすいように、考えられる要因が、温度 x_1 、湿度 x_2 、圧力 x_3 の3つしかないものとします。そして、それらの各組合せにおいて、結果である特性 y の測定データがあるものとします。この場合のデータは、下の表7.1のようになります。

表7.1 重回帰分析対象データのイメージ

追番	x_1 : 温度(°C)	x_2 : 湿度(%)	x_3 : 圧力(kPa)	y : 結果
1	209	74.5	317	116.6
2	216	76.9	320	120.1
3	208	72.4	318	116.2
4	214	76.6	327	119.6
⋮	⋮	⋮	⋮	⋮

重回帰分析では、要因の候補となる x_1 、 x_2 、 x_3 を説明変数と呼び、これに対して結果系の y を目的変数といいます。そして、説明変数と目的変数 y との関係を決めるような1次式で表します。

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

β_0 は定数項で切片、 $\beta_1 \sim \beta_3$ はそれぞれの説明変数についての傾きで回帰係数といいます。

上の式だけを見ると、変数の2乗の項を考慮すべき場合には適用できないようにも感じてしまいますが、そのようなことはありません。2乗の項を考慮すべき場合には、例えば $x_2 = x_1^2$ のような変数変換を行います。もしも対数の項を考慮すべき場合には、例えば $x_2 = \log(x_1)$ のような変数変換を行います。このような変数変換を行うことによって、重回帰分析でもかなり複雑な式まで対応することが可能です。

さて、重回帰分析によって得られる重回帰式 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ をどのようにして求めるかという、実測値 y_i と予測値 $\hat{y}_i$ との差 $y_i - \hat{y}_i$ 、これを残差といいます。残差の2乗を全部足したもの ($\sum (y_i - \hat{y}_i)^2$) が、最小となるように計算して求めます。この計算方法は、一般に最小二乗法といわれています。データ数が多く、変数も複数あるため、最小二乗法を手計算で行うのは厳しいものがあります。このため、通常は専用のアプリケーションソフトを用います。因みに(一財)日本科学技術連盟のSQCセミナーでは、(株)日本科学技術研修所製のStatWorks®を使用しています。

ここまで読んでくると、「パソコンとソフトがあれば答えが見つかるのか。意外と簡単そうだな。」と思われるかもしれません。

しかし、実際の重回帰分析では、それぞれの説明変数を重回帰式に加えるか否かを、固有技術も加味して検討しなくてははいけません。

また、解析対象のデータは通常既存データであり、実験計画法に基づいて得たデータと異なります。このようなデータは、様々な未知の変動要素が加わっているため、回帰係数が妥当なものかどうかを回帰診断によってチェックする必要があります。事実、説明変数間の間には、その相関の強弱はともかく、相関関係が存在することが少なくありません。変数間に強い相関関係が存在すると、例えば β_2 本来の符号がプラスでなければならないにも関わらず、得られた $\hat{\beta}_2$ の符号がマイナスになることが起きます。符号が変わらなくても、その絶対値がずれることは当たり前に起こります。よって、得られた回帰係数の推定値 $\hat{\beta}_1 \sim \hat{\beta}_3$ は偏回帰係数と呼ばれます。

ここでは、回帰診断のうち、イメージしやすいものをごく簡単に紹介しておきます。

まずは、残差のヒストグラムです。実測値 y_i に対して予測値 $\hat{y}_i$ とのズレが少なければ、残差 $y_i - \hat{y}_i$ の平均値が0の正規分布となることが期待されます。よって、正規分布といえそうかどうかをチェックします。

次に残差の時系列グラフです。意図的に傾向を持たせるように採ったデータであればともかく、そうでなければ時系列グラフはランダムに変動する筈です。よって、ランダムに変動しているかどうかをチェックします。

次に説明変数と残差の散布図です。説明変数に特に問題がなければ、説明変数と残差の散布図は無相関になります。もしも説明変数の2乗の項を考慮すべき場合は、散布図の全体的な形状が凸あるいは凹のような、無相関とはいえない形状を示します。よって、無相関といえそうかどうかをチェックします。

以上、回帰診断のうち比較的イメージしやすいものを3つ紹介いたしました。回帰診断のチェックで引っかかってくる場合は、未知の変数を見逃していたり、適切な層別を行っていないなど、何らかの問題点があることを示しています。

このようにして、重回帰分析を行って得られた結果として、StatWorks® の出力例とその重回帰式を示しておきます。

表7.2 重回帰分析結果例

目的変数名	重相関係数	寄与率R^2	R^2	R**^2	
y	0.851	0.724	0.715	0.707	
	残差自由度	残差標準偏差			
	96	0.855			
説明変数名	分散比	P値（上側）	偏回帰係数	標準偏回帰	トレランス
定数項	0.9325	0.337	-9.077		
x1：温度	81.5796	0.000	0.333	0.524	0.855
x2：湿度	24.8585	0.000	0.194	0.303	0.776
x3：圧力	23.1524	0.000	0.130	0.290	0.792

この分析結果の場合の重回帰式は、

$$y = -9.077 + 0.333x_1 + 0.194x_2 + 0.130x_3$$

となります。

重回帰分析を適用する状況はいろいろと考えられますが、ここでは次の3つを紹介しておきます。1つは、どの因子をどのようなレベルの値にすれば期待する結果 y が得られそうかという、制御を目的とする場合です。1つは、材料の特性などを含む製造条件が予め得られている場合において、最終の特性値 y がどのような値になるかを事前に予測しておきたい場合です。1つは、製造工程で採られたデータが多くある状況において、不具合が生じたときに何に起因して不具合が生じているかの絞り込みを行いたい場合です。

いずれの場合でも、目的に応じた解析を注意深く行う必要がありますが、費用および手間がかかる実験を行わなくとも、PCによる解析だけである程度の目処をつけられるかもしれない点は相当のメリットです。

既存データがあるのであれば、実験を行う前にまずは重回帰分析を行っていただき、メリットを享受していただければと思います。

著者紹介

吉富 公彦（よしとみ きみひこ）

1986年新日本無線(株)入社。赤外発光ダイオード、チューナーモジュールの生産技術業務に15年間従事。その後、社内SQC手法教育および品質マネジメントシステム業務に従事（ISO9001およびIATF16949の内部監査員）。元VDA6.3 Process Auditor。

現在、東京理科大学、東京情報大学非常勤講師。

（一財）日本科学技術連盟において企業向け講師派遣型研修SQCベーシックコース講師の他、品質管理セミナーベーシックコースおよび品質管理セミナー入門コースで講師をつとめる。（一財）日本規格協会 通信講座品質管理中級コース教材作成検討委員。

