ORTHO MEDICO
Industry News Topics

〒112-0002
東京都文京区小石川1丁目4番1号
住友不動産後楽園ビル2階

2020年11月

相関分析と回帰分析【第32回生物統計学】

1　概要

ある変数と別の変数の間の関連性を要約する手法には相関係数(correlation coefficient)を中心にした相関分析と、回帰直線(regression line)を中心にした回帰分析があります。これらの手法は広く利用されていますが、しばしば間違って用いられます。その原因は計算原理と深く関わっているので、原理をよく理解して正しく利用するようにしましょう。

2　相関分析とは

2つ以上の変数が存在する場合に、ある変数(x)が変化した時に、他方の変数(y)もその変化に応じて変化する関係のことを相関関係といいます。そして、この関係を統計的に分析することを相関分析と言います。相関関係は2種類のデータがお互いに影響を与え合っている相互関連性のことであるので、因果関係ではないことに注意しましょう。

相関分析を行うにあたり、散布図が便利になります。散布図は変数間の関係を視覚的に表現した図で、その図から相関係数(r)が求められます。相関係数とは2つの変数間の類似性の高さを表す指標です。相関分析では散布図を描き、外れ値など異常値がないか確認し、相関分析の結果として相関係数が求められます。そして、以下の式で相関係数が算出されます。

$$r = \frac{S_{xy}}{S_x \times S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

r: x と y の相関
s_{xy}: x と y の共分散
s_x: x の標準偏差
s_y: y の標準偏差
n: データの総数
$\bar{x}$: x の平均
$\bar{y}$: y の平均

相関係数は「-1～1」の範囲で表され、正であれば、一方の変数が大きくなるにつれてもう一方の変数も大きくなる右肩上がりの関係になり、絶対値が1に近づく程その関係が強いことを表しています。また、負は正の逆の関係になります。

ORTHO MEDICO

お問い合わせ先　株式会社オルトメディコ　企画部
〒112-0002 東京都文京区小石川1丁目4番1号　住友不動産後楽園ビル2階
TEL: 03-3818-0610　FAX: 03-3818-0617　Mail: planning-department@orthomedico.jp

〒112-0002
東京都文京区小石川1丁目4番1号
住友不動産後楽園ビル2階

3　回帰分析とは

回帰分析はある変数(x)から、もう一方の変数(y)を予測するために用いられます。つまり、x が原因で y がその結果という因果関係を、回帰直線として求める分析になります。このとき、2 種類の変数の関係を示す当てはまりのいい直線を引くために、定数 a および b について、最小二乗法を用いて算出します。

最小二乗法とは、回帰式 $y = ax + b$ を想定した場合に、ある点 x における観測値を y として、その時の計算値 Y との差（$\varepsilon = y - Y$）の二乗の和が最小になるように、定数 a および b を決定する方法です。なお、a は直線の傾きを表すが、回帰では回帰係数と呼びます。式で表すと以下のようになります。

$$Q = \sum \varepsilon^2 = \sum_{i=1}^{n} (y_i - Y_i)^2 = \sum_{i=1}^{n} (y_i - ax_i + b)^2 \tag{2}$$

この式からもわかるように、y軸方向の距離だけを最小にする、これが最小二乗法の本質だと言えます。そして、Qの最小値を求めるためには、a および b で偏微分して、それぞれを 0 と置いた連立方程式を解けば求められます。具体的には、

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^{n} x_i \{y_i - (ax_i + b)\} \tag{3}$$

$$\frac{\partial Q}{\partial b} = -2 \sum \{y_i - (ax_i + b)\} \tag{4}$$

それぞれ =0 と置くと、

$$\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} {x_i}^2 + b \sum_{i=1}^{n} x_i = 0 \tag{5}$$

$$\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i + b\, n = 0 \tag{6}$$

これを、まず a について解くと、

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} {x_i}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x} \tag{7}$$

となります。

お問い合わせ先　株式会社オルトメディコ　企画部
〒112-0002 東京都文京区小石川1丁目4番1号　住友不動産後楽園ビル2階
TEL: 03-3818-0610　FAX: 03-3818-0617　Mail: planning-department@orthomedico.jp

〒112-0002
東京都文京区小石川1丁目4番1号
住友不動産後楽園ビル2階

b については、

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} = \bar{y} - a\bar{x} \tag{8}$$

となります。

つまり、回帰式 $y = ax + b$ は

$$y = \frac{s_{xy}}{s_x}x + (\bar{y} - \frac{s_{xy}}{s_x}\bar{x})$$

という式で求めることができます。

4　回帰係数と相関係数の違い

では、回帰係数と相関係数はどのように違うのかを解説します。まず、式(7)の分母の計算に注意してください。x のみで決まり、y は全く関係しないことが分かります。これは x の平均からの変化に伴って、y がその平均からどのように変化するかを表した式と言えます。例えば、x と y を入れ替えたデータからできる回帰直線は、改めて y 軸方向の残差が最小になるようにして得られる直線であることから、元の直線と一致しないことがわかります。一方、相関係数を求める式(1)では、x と y に関して対称式であり、x と y を入れ替えても全く同じ式になります。つまり、回帰分析はある変数から、もう一方の変数を予測するために用いられ、相関分析はある変数とある変数の関係の強さを数量で表すものです。回帰式の回帰係数を相関係数のように扱うのは本来の目的ではないので、混同して用いないように注意してください。

5　まとめ

今回は、相関分析と回帰分析について、それぞれの特徴をとらえながら紹介しました。相関係数も回帰分析も、2 つの変数の関係性を形にする点では共通しています。しかし、研究のデザインを考える上で、この 2 つは明確に異なるものなので、相関係数を提示するべき時に、回帰係数を提示したり、予測式を立てる時に、散布図と相関係数を提示したりしないように気をつけて下さい。

お問い合わせ先　株式会社オルトメディコ　企画部
〒112-0002 東京都文京区小石川1丁目4番1号　住友不動産後楽園ビル2階
TEL: 03-3818-0610　FAX: 03-3818-0617　Mail: planning-department@orthomedico.jp