

日本赤十字九州国際看護大学/Japanese Red
Cross Kyushu International College of
Nursing
Regression analysis

メタデータ	言語: jpn 出版者: 公開日: 2020-08-04 キーワード (Ja): キーワード (En): 作成者: メールアドレス: 所属:
URL	https://jrckicn.repo.nii.ac.jp/records/726

This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 3.0
International License.



第 13 回 回帰分析

皆さんこんにちは。今回は回帰分析についてお話しします。

二つの変数 X と Y の回帰や相関の考え方、ワークシートを用いたピアソンの積率相関係数の計算については、すでに第 4 回目の授業でお話ししました。今回の回帰分析は、目的変数 Y と一つ又はそれ以上の説明変数 X の間の関係を推定するための統計的な考え方です。最も一般的な形は線形回帰と呼ばれます。

1 回帰分析の目的

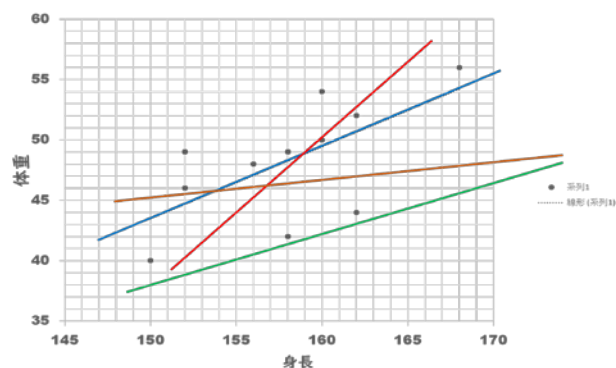
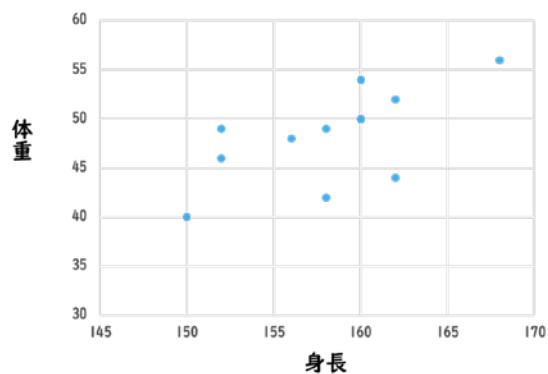
回帰分析は主に二つの目的で使われます。1 番目は予測です。線形回帰の考え方で y と x の間を数式で表すことができれば、 x から Y の値を予測できます。 Y は「予測したい変数」であり「目的変数」「従属変数」と呼びます。 x は「予測に用いる変数」であり「説明変数」「独立変数」と呼びます。学生みなさんは $m y$ 標本を用いて、たとえば身長から体重を予測するとか、勉強時間をアルバイトの時間から予測するとか、関心のあるテーマで回帰分析を行ってみてください。

2 番目の目的は、因果関係の推測です。 X と Y の間に相関関係が認められる場合、それが直接に因果関係を意味することはありません。しか相関関係は、因果関係の可能性を示します。 y を「疾病」、 x を「疾病の原因」と考えて、 x と y の関連性をより具体的に考えていくのは、皆さんが後期に学ぶ疫学の中心です。

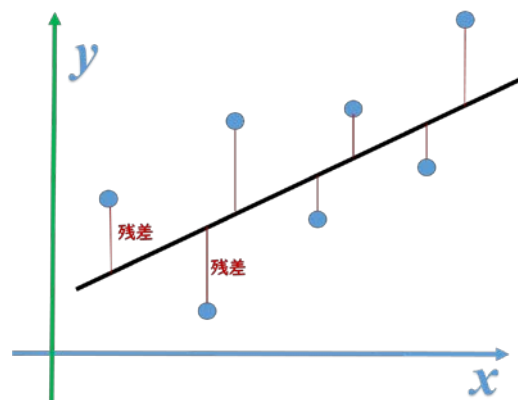
2 回帰式の求め方

さて、観測されたデータに基づいて、 x から y を予測する回帰式はどのように導いたらいいでしょうか。ある $m y$ 標本での身長と体重の関連性を図に示します。身長 X が増えると体重 Y が増える関連性が認められます。ではこの X と Y との関連性を代表する一本の直線の存在を決定するにはどうしたらよいでしょうか。このような時に使うのが最小二乗法という考え方です。最小二乗法は、回帰分析の中でも最も基本的なものです。さらに考えを進めるために、図に 5 本の直線を引いてみました。この 5 本の直線では、どの直線が最もふさわしいでしょうか。

最小二乗法では、測定データ Y は、モデル関数と誤差（残差）の和で表わされます。モデル関数が測定データにどのくらい適合するかは、残差で判断できます。



残差は測定データからモデル関数の値を引いたものです。残差は正、または負の値を取るため、負の値を取らないように残差を二乗し、残差二乗（残差平方）とした上で、その合計、つまり残差二乗和が最小になるように連立方程式を解いてモデル関数を決定するのが、最小二乗法です。



3 回帰分析の歴史

最小二乗法は回帰分析の原型とされる重要な計算方法です。18 世紀の間に天文学や数学の分野で行われた進歩がこの方法に結集しました。最小二乗法の論文は 1805 年にフランスの数学者ルジャンドルが公開し、それより前に最小二乗法を見つけたというドイツの数学者ガウスと論争になりました。ガウスはさらに考えを発展させて確率の原理に結びつけ、またガウスはこの過程で正規分布を発明しました 19 世紀の初めに話題になっていた小惑星セレスの軌道をガウスはこの最小 2 乗法で正確に予測したことが知られています。

19 世紀にフランシス・ゴルトンがスイートピーの親種と子種の直径を比較する研究から見出した傾向を表すために、回帰、regression という用語を用いたことは、すでに第 4 回目の授業でお話ししました。この回帰の概念がその後発展し、最小二乗法と結びつき、回帰分析として知られるようになりました。

回帰分析は 20 世紀半ばまでは膨大な計算を必要とするため、コンピューターが一般化する前、1970 年代以前は機械式計算機で一つの結果を得るまでに、24 時間かかる場合もあったとされます。その後コンピューターの普及とともに回帰分析は発展し、今では経済学・天文学・医学など様々な分野で使われています。

4 エクセルでの計算

計算の事例はあなたの m y 標本から選んでください。ここでは m y 標本から取った以下のデータを用います；

Y; ネットテレビ時間、一日当たりのネットやテレビ視聴にかかる時間（分）

X; 予習復習時間、一日当たりの予習復習にかかる時間（分）

予習復習	ネットテレビ
60	90
70	60
0	120
60	60
80	70
30	100

一次回帰の予測式は $Y = \alpha + \beta x$ です。Y は「目的変数（従属変数）」です。x は「説明変数（独立変数）」です。α と β を回帰分析で求めます。

- ・すでに設定したエクセルの分析ツールを使うためには、まずエクセルの画面上部にあるメニューからデータのタブを選びます。
- ・すると上の右端に分析というタブが現れるので、それをクリックします。
- ・すると分析ツールのボックスが現れるので、メニューから「回帰分析」を選びOKを押します。
- ・すると回帰分析のボックスが現れるので「入力 Y 範囲」および「入力 X 範囲」を指定します。
- ・入力 Y 範囲とは、回帰分析で予測したい y、今回の例ではネットテレビ時間が入っているカラムです。入力 x 範囲とは、回帰分析で予測に用いる X、予習復習時間が入っているカラムです。
- ・両方を指定してOKを押すと直ぐに計算結果が表示されます。結果は三つの表に分かれて示されます。

1) 最初の表、回帰統計では相関係数が 0.8907 と示されています。相関係数の前に「重」という文字がついています。その理由は、回帰分析で説明変数の数が複数になると、相関係数も複数の変数から計算することになり、その際は単に相関といわず、重相関というからです。回帰分析を行う前に、説明変数 x と目的変数 Y との間に相関があることを確認するのは重要です。X と Y の間に相関が認められなければ、回帰式を立てることが意味を持ちません。

2) 二番目の表は分散分析です。前回の授業で学んだ分散分析は、実はこのように、回帰分析の中にも組み込まれています。データ Y の全変動を示す偏差二乗和 2933.33 に対し、回帰式で説明できる変動（偏差二乗和）が 2327.27、それを引き算した残差の変動が 606.06、そして「観測された分散比」である F 値が 15.36、この F 値に対応する p 値（有意確率）が 0.01726 です。有意水準を 5% (0.05) と設定した場合、p 値はそれより低い値をとるため、帰無仮説は棄却されます。棄却される帰無仮説とは「回帰式の予測による誤差（残差）の減少は無い」です。帰無仮説を棄却すると、「回帰式による予測が誤差を減少させる」と認めることになり、回帰式の予測が意味あるものと位置付けられます。

3) 三番目の表では、回帰式の α と β を具体的な数値として求め、またその区間推定値も示します。最終的に得られる予測式は $Y=119.697-0.7273X$ となります。以上がエクセルによる回帰分析です。エクセルではこの他に散布図や回帰直線を描くこともできます。その方法をお話しします。

- ・まず先ほど回帰分析を行ったデータのあるセルをもう一度範囲選択します
- ・エクセルの画面上部にある挿入のタブを選びグラフのグループにある散布図のボタンをクリックします。すると散布図が描けます
- ・散布図を描いたらエクセルの画面上部「デザイン」のタブを選び、上部左のクイックレイアウトを選ぶと散布図に様々な情報を付け加えられる選択肢が画像ボタンで表示されます。
- ・一番上の行の 3 番目にある画像ボタン（直線を描く）を押すと、先ほど散布図に回帰分析で得た直線を描き込むことができます。

5 Casio、Linear regression Calculator

エクセル以外の計算方法として今回はCasio、Linear regression Calculatorを紹介します。

<https://keisan.casio.com/exec/system/14059929550941>

- ・サイトにアクセスするとすぐに現れる表にはすでに X と Y の値が例として書き込まれています。
- ・まずこれを下のクリアキーを押して消去します。
- ・空白の表にしたら、そこに X と Y の値を入力します。

- 入力が終わったら、計算実行を意味する EXECUTE のキーを押します。
- そうすると X と Y それぞれの平均値、相関係数そして回帰式のアルファとベータの値が表で表示されます。
- Excel の場合とほぼ同様の式 $Y=119.6970-0.7273x$ が得られたことを確認してください。このサイトでは同時に直線のグラフも描いてくれます。

6 まとめ

さて今回の授業までで、学生の皆さんにお伝えしたい統計学の基本的な考え方は、ほぼお話ししたように思います。統計学の教科書を見ると、私がこれまでにお話しした以外にも様々な方法が示されています。様々な方法の名前を知ることも無駄ではありません。でも統計学は暗記科目ではありません。統計は、皆さんがこれから日々の生活や仕事や研究の中で、それらを実際に使って物の見方考え方を広げていくための道具です。

これまで皆さんにお話しした基本的な方法、特にクロス集計表とカイ二乗検定、二つの平均の差の t 検定、分散分析と F 検定、回帰分析などに親しみ、それらを用い始めれば、まだお話ししていない他の統計の方法についても、徐々に理解し使いこなせるようになると思います。

私が自身の仕事の中で、これまで最もよく用いていたのは回帰分析です。「あることから別なことを予測する」というのは興味深く、特に間の発育に関連して回帰分析を使ってきました。人は子ども時代から思春期に入った時に、身長や体重が突然増加速度を速めたり、また思春期を過ぎて逆に増加速度が遅くなったりします。そのような時に、1年前の身長や体重の増加から1年後の増加をどのくらい予測できるかなどに、関心を持ちました。また人の代謝に関連した基本的な物質としてクレアチニンがあります。小児において、尿中のクレアチニン排泄量を身長や体重で予測することも、私がかって関心を持ったテーマです。皆さんも、皆さんらしい課題を見つけ、回帰分析を使ってみてください。

演習問題

1. あなたはどのような課題に回帰分析を使ってみたいですか。50字以内で書いてください。
2. エクセルを用い、my 標本で何か回帰分析を行い、結果を 50 字以内で書いてください。（my 標本ではなく、150 名データを用いても構いません。）エクセルが利用できない場合は、動画中のエクセルの説明を見て、思ったことを 50 字以内で書いてください。（登校禁止が解除されたら、ぜひ情報処理室でエクセルに触れてください。）
3. カシオ Linear regression Calculator で何か回帰分析を行ってください。
すでにエクセルで計算済みであっても、こちらでも計算を試みてください。異なる手段で計算してみることで、手段の特徴や限界を理解できます。結果や気づいた点を 50 字以内で書いてください。