回帰分析の原理

回帰分析の型

回帰分析は，説明変数（または独立変数）と被説明変数（または従属変数）の間に何らかの関数的関係があることを想定し，その関数形を決定づけるパラメータ（回帰係数または単に係数）を，実データに基づいて求める解析方法の総称である。説明変数は，単一の場合と複数の場合とがあり，単一の場合を特に単回帰分析，複数の場合を重回帰分析と呼ぶ。

関数形については，変数の 1 次結合のみで表される場合を線形（線型）回帰と呼び，その他の場合を非線形（非線型）回帰と呼ぶ。

- 線形回帰の例
  - 単回帰： $y = \beta_0 + \beta_1 \times x$
  - 重回帰： $y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2$
- 非線形回帰の例
  - 2 次曲線（放物線）： $y = a \times x^2 + b \times x + c$
  - 対数（ただし $x > 0$）： $y = a \times \log(x) + b$
  - べき乗： $y = a \times b^x$

線形単回帰分析による計算例

最も単純な例である，線型単回帰分析は，以下のように定式化できる。

今，

データ数： n

被説明変数（従属変数）： $y_1, y_2, y_3 \cdots y_n$

説明変数（独立変数）： $x_1, x_2, x_3 \cdots y_n$

のデータがある時，これらの変数間に

$y = \beta_0 + \beta_1 \times x$ （β_0，β_1 は未知の定数）

なる関係（線形関係）があると想定されるとき（i=1,2,…,n について $y_i = \beta_0 + \beta_1 \times x_i$ がすべて成立する），β_0 と β_1 をある条件下で合理的に決定したい。

n=2 の時（n<2 では決定不能！）の求め方

$$y_1 = \beta_0 + \beta_1 \times x_1$$
$$y_2 = \beta_0 + \beta_1 \times x_2$$

を満足するためには，（簡単な連立方程式なので）

$$\beta_1 = \frac{y_2 - y_1}{x_2 - x_1}$$

$$\beta_0 = y_1 - \beta_1 \times x_1 = \frac{x_2 \times y_1 - x_1 \times y_1 - x_1 \times y_2 + x_1 \times y_1}{x_2 - x_1} = \frac{x_2 \times y_1 - x_1 \times y_2}{x_2 - x_1}$$

と求められる。具体例として，データが $(x, y) = \{(1,2),(3,5)\}$ の 2 点の集合だとすると，上式に当てはめて

$$\beta_1 = \frac{y_2 - y_1}{x_2 - x_1} = \frac{5-2}{3-1} = \frac{3}{2}$$

$$\beta_0 = y_1 - \beta_1 \times x_1 = 2 - \frac{3}{2} \times 1 = \frac{1}{2} \left(\text{または} \frac{x_2 \times y_1 - x_1 \times y_2}{x_2 - x_1} = \frac{3 \times 2 - 1 \times 5}{3-1} = \frac{1}{2}\right)$$

と求まる。

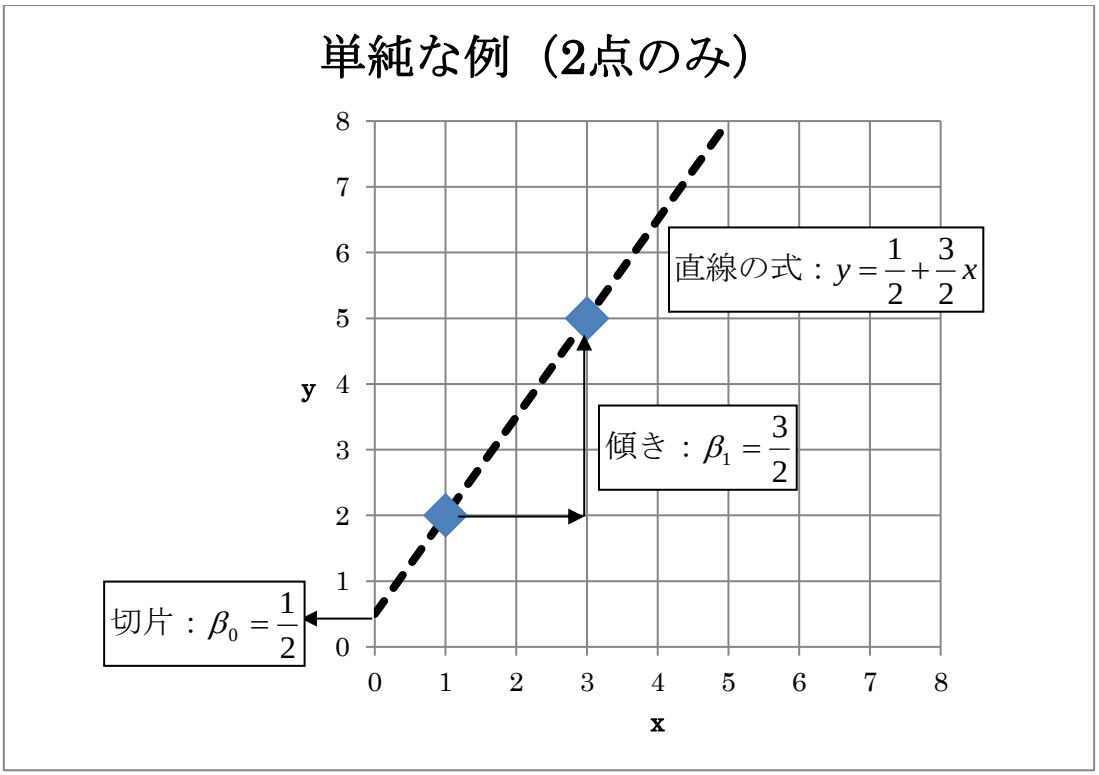

図 1 データが 2 点のみの場合の回帰直線（2 点を通る直線）

n>2 の時の求め方

一般に，データには誤差（観測誤差など）を含むので，すべての i で $y_i = \beta_0 + \beta_1 \times x_i$ を満足するとは限らない。そこで，それぞれのデータペアにおいて誤差 ε_i が含まれるものとして，モデルを

$$y = \beta_0 + \beta_1 \times x + \varepsilon$$

のように改め，各々のデータについて生じる誤差

$$\varepsilon_1 = y_1 - \beta_0 - \beta_1 \times x_1$$
$$\varepsilon_2 = y_2 - \beta_0 - \beta_1 \times x_2$$
$$\vdots$$
$$\varepsilon_n = y_n - \beta_0 - \beta_1 \times x_n$$

について考える。これらの誤差が全体で最小となるように，これらの平方和

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \times x_i)^2$$

を考えると，右辺は，式の中で未知のものが β_0 と β_1 のみなので，これらを変数とする関数 $g(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \times x_i)^2$ とみなすことができる。そこで，$g(\beta_0, \beta_1)$ の極小値を与える点を求めればよい。そのためには，β_0 と β_1 で偏微分したものを 0（傾きが 0 となる位置）と置いた式を連立方程式として解けばよい。

$$\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 \times x_i) \times (-1)$$
$$\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 \times x_i) \times (-x_i)$$

なので，結局

$$\sum_{i=1}^{n} y_i - \beta_0 \times n - \beta_1 \times \sum_{i=1}^{n} x_i = 0$$
$$\sum_{i=1}^{n} (x_i \times y_i) - \beta_0 \times \sum_{i=1}^{n} x_i - \beta_1 \times \sum_{i=1}^{n} x_i^2 = 0$$

を β_0 と β_1 に関する連立 1 次方程式として解くことになる。さらに整理して行列表現に改めると，

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} (x_i \times y_i) \end{pmatrix}$$

となるので，左辺にある正方行列の逆行列を両辺に（左側から）掛けることで，

$$\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}=\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1}\times\begin{pmatrix}\sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(x_i\times y_i)\end{pmatrix}$$

を得る。右辺を展開して整理すると

$$\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}=\frac{1}{n\sum_{i=1}^{n} x_i^2-(\sum_{i=1}^{n} x_i)^2}\begin{pmatrix}\sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n\end{pmatrix}\times\begin{pmatrix}\sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(x_i\times y_i)\end{pmatrix}$$

$$=\frac{1}{n\sum_{i=1}^{n} x_i^2-(\sum_{i=1}^{n} x_i)^2}\begin{pmatrix}\sum_{i=1}^{n} x_i^2\times\sum_{i=1}^{n} y_i-\sum_{i=1}^{n} x_i\times\sum_{i=1}^{n}(x_i\times y_i) \\ -\sum_{i=1}^{n} x_i\times\sum_{i=1}^{n} y_i+n\sum_{i=1}^{n}(x_i\times y_i)\end{pmatrix}$$

となる[1]。

n=3 で，データが $(x,y)=\{(1,2),(3,5),(1.8,4.6)\}$ である場合を実際に計算してみよう。この時,

$$\sum_{i=1}^{3} x_i=1+3+1.5=5.8$$

$$\sum_{i=1}^{3} y_i=2+5+4.6=11.6$$

$$\sum_{i=1}^{3} x_i^2=1^2+3^2+1.8^2=1+9+3.24=13.24$$

$$\sum_{i=1}^{3}(x_i\times y_i)=1\times 2+3\times 5+1.8\times 4.6=2+15+8.28=25.28$$

---

[1] さらに，xの平均 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$ や yの平均 $\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$，$x$の分散

$s_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n} x_i^2-\bar{x}^2$，および xと yの共分散

$s_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\sum_{i=1}^{n}(x_i y_i)-\bar{x}\bar{y}$ であることを利用すると

$\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}=\begin{pmatrix}\bar{y}-\frac{s_{xy}}{s_x^2}\bar{x} \\ \frac{s_{xy}}{s_x^2}\end{pmatrix}$ と，極めてシンプルな式に改めることができる（導出は各自で確認してみよ）。

とそれぞれ求まるので,

$$\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}=\frac{1}{3\times13.24-5.8^2}\begin{pmatrix}13.24\times11.6-5.8\times25.28\\ -5.8\times11.6+3\times25.28\end{pmatrix}=\frac{1}{6.08}\begin{pmatrix}6.96\\ 8.56\end{pmatrix}\fallingdotseq\begin{pmatrix}1.145\\ 1.408\end{pmatrix}$$

となる。

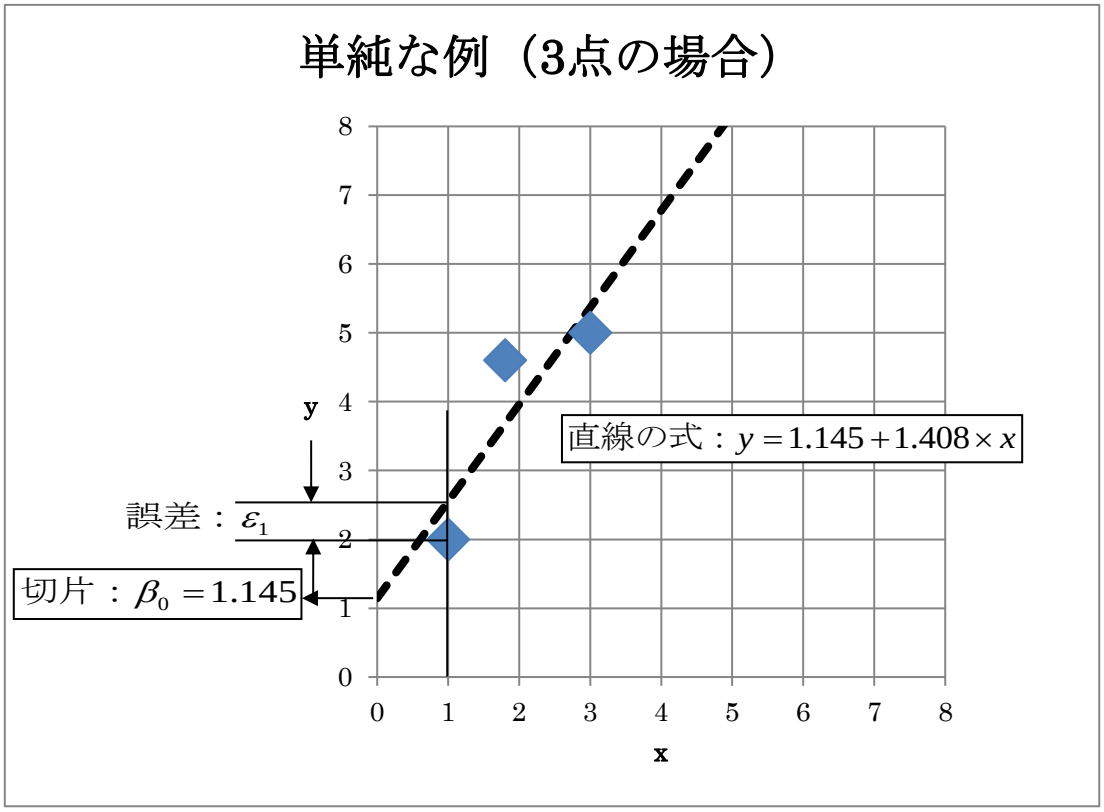

図 2 データが 3 点の場合の回帰直線

なお，この時,

$$g(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1\times x_i)^2$$

$$=\sum_{i=1}^{n}(y_i^2+\beta_0^2+\beta_1^2\times x_i^2-2\beta_0\times y_i+2\beta_0\times\beta_1\times x_i-2\beta_1\times x_i\times y_i)$$

$$=\sum_{i=1}^{n}y_i^2+n\times\beta_0^2+\beta_1^2\times\sum_{i=1}^{n}x_i^2-2\beta_0\times\sum_{i=1}^{n}y_i+2\beta_0\times\beta_1\times\sum_{i=1}^{n}x_i-2\beta_1\times\sum_{i=1}^{n}(x_i\times y_i)$$

なので，実際のデータを当てはめた曲線,

$$g(\beta_0,\beta_1)=50.16+3\times\beta_0^2+\beta_1^2\times13.24-2\beta_0\times11.6+2\beta_0\times\beta_1\times5.8-2\beta_1\times25.28$$

$$=50.16+3\times\beta_0^2+13.24\times\beta_1^2-23.2\times\beta_0+11.6\times\beta_0\times\beta_1-50.56\times\beta_1$$

をグラフに描くと，図 3 のようになる。

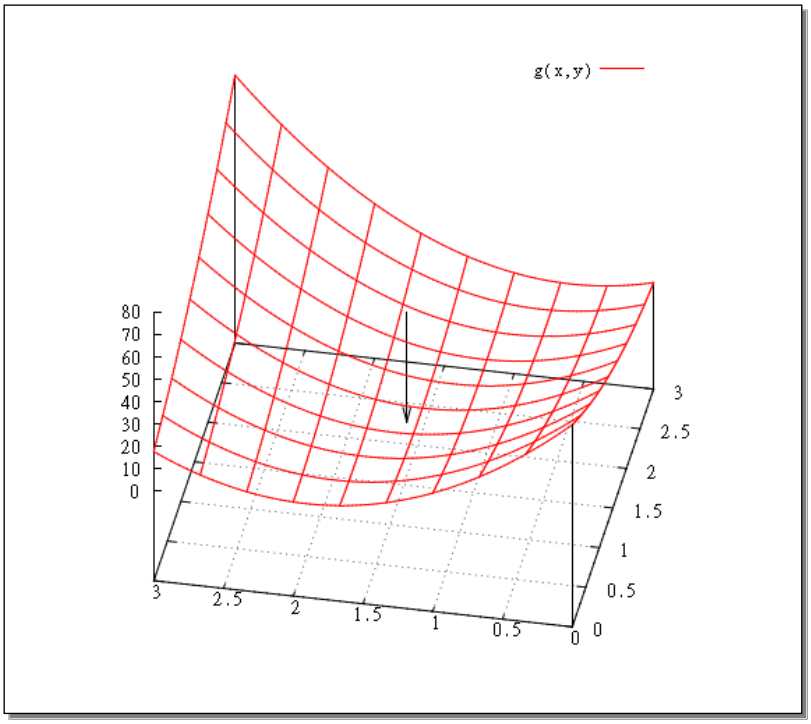

図 3 誤差の平方和を示す関数 $g(\beta_0,\beta_1)$ と極小点の位置

分析結果の判断指標

回帰分析の結果の判断は，説明力と信頼性の 2 点について検討する必要がある。

説明力の指標としては，従属変数（被説明変数）の挙動を，独立変数（説明変数）と得られた係数で求めた理論値の挙動でどの程度説明できているかを示す決定係数（しばしば記号として R^2 が用いられる）が代表的である。

従属変数の挙動は，その平均 $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ からのばらつき度合い，つまり，分散 $s_y^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$ を用いて示す[2]。これに対して，得られた係数（β_0 および β_1）と独立変数 x_i で得られる理論値（回帰直線上の値）を $\hat{y}_i$ で表すこととすると，次の性質を持つことが分かる。定義より，$\hat{y}_i=\beta_0+\beta_1\times x_i=\bar{y}-\frac{s_{xy}}{s_x^2}\bar{x}+\frac{s_{xy}}{s_x^2}x_i=\frac{s_{xy}}{s_x^2}(x_i-\bar{x})+\bar{y}$ となるので，その平均は，

[2] もしくはデータ数で除する前の平方和 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ で示す。

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\frac{s_{xy}}{s_x^2}(x_i - \bar{x}) + \bar{y}) = \frac{1}{n}\times\frac{s_{xy}}{s_x^2}\sum_{i=1}^{n}x_i - \frac{1}{n}\times\frac{s_{xy}}{s_x^2}\times n\times\bar{x} + \frac{1}{n}\times n\times\bar{y} = \bar{y}$$

から，従属変数の平均と一致する。よって，理論値の挙動をその平均からのばらつき度合い，すなわち，分散で示すと，

$$s_{\hat{y}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{s_{xy}}{s_x^2}(x_i - \bar{x}))^2 = \frac{1}{n}\times\left(\frac{s_{xy}}{s_x^2}\right)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{s_{xy}}{s_x^2}\right)^2\times s_x^2 = \frac{s_{xy}^2}{s_x^2}$$

となる。決定係数R^2は，従属変数の実際の分散と理論値の分散の比の形で示して，

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{s_{xy}^2 \big/ s_x^2}{s_y^2} = \frac{s_{xy}^2}{s_x^2\times s_y^2}$$

と示すことができる。なお，$R = \frac{s_{xy}}{\sqrt{s_x^2\times s_y^2}}$は，（Pearson の）相関係数と呼ばれる指標でもあり，決定係数はその平方値の形となっているので慣習的にR^2を用いて表している。実際に n=3 の場合の例で計算してみると，

$$s_{xy} = \frac{1}{3}\sum_{i=1}^{3}(x_i - \frac{5.8}{3})\times(y_i - \frac{11.6}{3})$$
$$= \frac{1}{3}\{(1-\frac{5.8}{3})\times(2-\frac{11.6}{3}) + (3-\frac{5.8}{3})\times(5-\frac{11.6}{3}) + (1.8-\frac{5.8}{3})\times(4.6-\frac{11.6}{3})\}$$
$$= \frac{1}{27}\{(3-5.8)\times(6-11.6) + (9-5.8)\times(15-11.6) + (5.4-5.8)\times(13.8-11.6)\}$$
$$= \frac{1}{27}(-2.8\times(-5.6) + 3.2\times3.4 + (-0.4)\times2.2) = \frac{1}{27}(15.68 + 10.88 - 0.88) = \frac{25.68}{27}$$

$$s_x^2 = \frac{1}{3}\sum_{i=1}^{3}(x_i - \frac{5.8}{3})^2 = \frac{1}{3}\{(1-\frac{5.8}{3})^2 + (3-\frac{5.8}{3})^2 + (1.8-\frac{5.8}{3})^2\}$$
$$= \frac{1}{27}\{(3-5.8)^2 + (9-5.8)^2 + (5.4-5.8)^2\} = \frac{1}{27}(2.8^2 + 3.2^2 + 0.4^2)$$
$$= \frac{1}{27}(7.84 + 10.24 + 0.16) = \frac{18.24}{27}$$

$$s_y^2 = \frac{1}{3}\sum_{i=1}^{3}(y_i - \frac{11.6}{3})^2 = \frac{1}{3}\{(2-\frac{11.6}{3})^2 + (5-\frac{11.6}{3})^2 + (4.6-\frac{11.6}{3})^2\}$$
$$= \frac{1}{27}\{(6-11.6)^2 + (15-11.6)^2 + (13.8-11.6)^2\} = \frac{1}{27}(5.6^2 + 3.4^2 + 2.2^2)$$
$$= \frac{1}{27}(31.36 + 11.56 + 4.84) = \frac{47.76}{27}$$

なので,

$$R^2 = \frac{\left(25.68/27\right)^2}{18.24/27 \times 47.76/27} = \frac{25.68^2}{18.24 \times 47.76} \fallingdotseq 0.7570$$

となる。

ところで，決定係数は，データ数 n や独立変数の数 p によって影響を受けるので，その影響を補正した（自由度を調整した）自由度調整済み決定係数 $R^2_{adjusted}$ と呼ばれるものを用いることがある。この時,

$$1 - R^2_{adjusted} = \frac{n-1}{n-p-1}(1-R^2)$$

という関係があるので,

$$R^2_{adjusted} = 1 - \frac{n-1}{n-p-1}(1-R^2) = \frac{n-p-1-(n-1)+(n-1)R^2}{n-p-1} = \frac{(n-1)R^2-p}{n-p-1}$$

で求めたものを用いる。上の例では,

$$R^2_{adjusted} = \frac{(3-1)\times 0.7570 - 1}{3-1-1} = 0.5140$$

となる。一般に，p>1 の時（重回帰分析）は自由度調整済み決定係数によって判定するのが良いとされている。また，決定係数は当てはまりの良さの目安であるので，対象によって異なるが社会科学系のデータおおむね 0.6 を超えると「良好」と判定してよいとされている[3]。

回帰分析のもう 1 つの指標は信頼性に関するものである。これは得られた回帰係数，特に独立変数の乗数として得られた係数（傾き）に関しての検定であり，回帰係数が 0 とみなせるかどうかの検定結果となっている。回帰係数が 0 と等しいとみなせる場合は，従属変数は独立変数とは無関係である，ということを意味してしまうので，そもそもモデル式が意味を失ってしまうこととなる。導出の詳細は省くが，検定はモデル全体にわたるものと，回帰係数ごとに行わるものとがあり，前者は F 検定，後者は t 検定で行われる。

F 検定は,「すべての回帰係数が 0 とみなせる」ことの仮説検定として行われ，回帰分析を行うソフトウェアでは，その検定量（F 値）とともに確率値（すべての回帰係数を 0 とみなしてよい確率の値）として表示される。確率値が小さい値（0.1 未満）を取ると,「すべての回帰係数を 0 とみなしてよい」確率は小さいので,「少なくとも 1 つの回帰係数は 0 ではない」と判定できる。

同様に，t 検定は,「当該の回帰係数が 0 とみなせる」ことの仮説検定として行われ，その検定量（t 値）とともに確率値が表示される。解釈についても同様で，確率値が小さい時

[3] R^2 は理論上，$0 \le R^2 \le 1$ の範囲に収まる（1 が最大で 100%～誤差が全くないことを示す）。

は，その回帰係数は 0 ではなく，意味のある数値（統計学の言葉では，有意（ゆういと表現する）とみなしてよい，と判定できる。

これらの検定量は，手計算では求められないので，たとえば Excel のアドインツール（「分析ツール」の「回帰分析」）や，統計処理用のプログラム（SPSS や R，TSP，e-Views など）を用いて行なう。R 言語では，次のようなプログラム（自分で入力する部分を青字で示す）で，先の例題に対する結果を求めることができる。

```
> x = c(1, 3, 1.8)    # 独立変数 x の定義
> y = c(2, 5, 4.6)    # 従属変数 y の定義
> lm.result = lm(y~x)    # プログラム lm=Linear Model を使って回帰分析を実行
> summary(lm.result)    # 結果の要約表示

Call:
lm(formula = y ~ x) ←モデル式（係数や定数項は省いて表現）

Residuals:
      1       2       3 ←残差（実際の y と理論値の差）
-0.5526 -0.3684  0.9211

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1447     1.6757   0.683    0.618
x             1.4079     0.7977   1.765    0.328

Residual standard error: 1.136 on 1 degrees of freedom
Multiple R-squared:  0.757,    Adjusted R-squared:  0.514
F-statistic: 3.115 on 1 and 1 DF,  p-value: 0.3282

>
```

定数項の値（β_0）

回帰係数ごとの t 検定結果

自由度調整済み決定係数

モデル全体の F 検定結果

傾きの値（β_1）

回帰係数と決定係数は，手計算の場合と一致していることが分かる。自由度調整済み決定係数については，0.6 を下回っているので，当てはまりは「良くはない」と言える。また，F 検定と t 検定の結果からは，（0 とみなせる確率が）0.328（32.8%）で，10%を超えているため，モデル全体，回帰係数ごとの信頼性はともに「十分ではない」と言える。

R 言語では，回帰分析以外の統計的手法も利用できるので，使い方を一通り覚えておくとよいだろう（オープンソースで，かつ無料で利用でき，最近では参考書も数多い）。

重回帰分析

p>1 となる場合，すなわち独立変数（説明変数が）が複数の場合にも，同様の計算手順（誤差の平方和を最小化する方法～最小二乗法）となる。

モデル式は，

$$y = \beta_0 + \beta_1 \times x_1 + \cdots + \beta_p \times x_p$$

のように拡張され，最小二乗法によって解くべき連立方程式は，

$$\begin{pmatrix} n & \sum x_1 & \sum x_2 & \cdots & \sum x_p \\ \sum x_1 & \sum x_1^2 & \sum x_2 \times x_1 & & \sum x_p \times x_1 \\ \sum x_2 & \sum x_1 \times x_2 & \sum x_2^2 & & \sum x_p \times x_2 \\ \vdots & & & \ddots & \vdots \\ \sum x_p & \sum x_1 \times x_p & \sum x_2 \times x_p & \cdots & \sum x_p^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \sum y \\ \sum x_1 \times y \\ \sum x_2 \times y \\ \vdots \\ \sum x_p \times y \end{pmatrix}$$

となる（煩わしくなるので，$\sum_{i=1}^{n} x_{j,i}$ などは$\sum x_j$ の形で略記してある)。解き方も左辺の正方行列の逆行列を両辺の「左側」から掛けて解けばよいが，解析的に表現すると煩雑になる上，あまり得るところがないので，統計用プログラムでは「数値的に」解を求めていることだけ知っておけば実用上は十分であろう。

非線形回帰分析

モデル式が線形でない場合は，回帰係数を連立方程式の形で求められないことも多く（極値を与える点を関数形で表すことが困難であるから)，回帰係数に適当な初期値を与えてプログラムを使って最適化計算（目標となる計算値が最小あるいは最大になる組合せを試行）によって行わせる。

変数変換による非線型の回帰分析

非線形のモデル式による場合は，一般的に，適当な変数変換を行うことで，多くの場合，線形回帰分析として扱うことができる。