医学統計勉強会

東北大学病院循環器内科・東北大学臨床研究推進センター　共催

東北大学大学院医学系研究科 EBM 開発学寄附講座

宮田　敏

“Data! data! data!” he cried impatiently. “I can't make bricks without clay.”
From *The Adventure of the Copper Beeches, The Adventure of Sherlock Holmes.*

「データ！データ！データ！」ホームズはいらいらして叫んだ。「粘土が無ければレンガは作れない」

第 2 回　回帰分析

1.　線形回帰モデル

第 1 回で取り上げた「基本統計量」は、単独の変数の持つ特徴、傾向、分布を解析するものでした。しかし、自然科学、社会科学において取り扱われる現象の多くは、複数の要因が相互に依存しあって成り立っています。今回取り上げる**回帰分析（regression analysis）**は、多数の変数の間の関係を解析する**多変量解析（multivariate analysis）**と呼ばれる手法の一つで、一つの連続数（実数）の値を複数の変数によって説明、予測する統計モデルになります。

多数の変数の間の関係を解析するのが回帰分析の目的ですが、最初は 2 変数の間の関係を解析することから始めます。

1．1　二変量データの解析

いま、x と y、2 つの変数の組が n 組得られたとします。

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

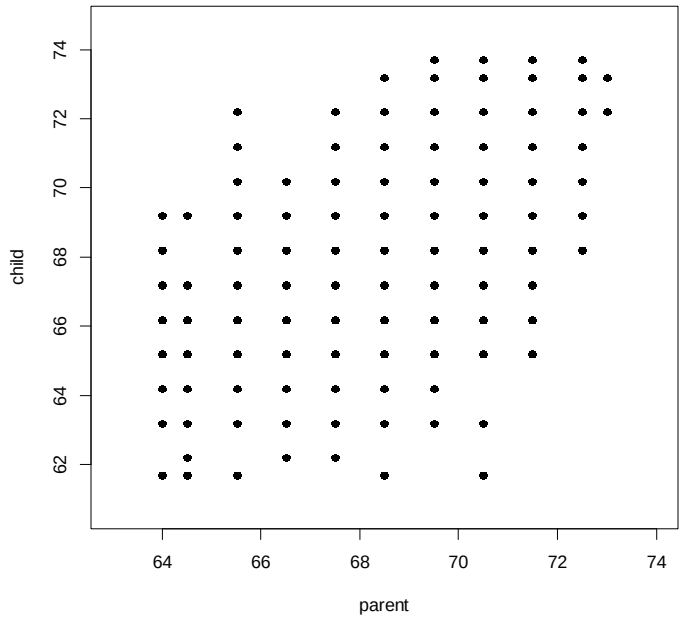
このとき、x と y の関係を推測することを目的とします。例として、以下のものを考えます。

Example 1: Galton の親子の身長のデータ

図 1.1 は、205 組の夫婦の平均身長（インチ単位）と、彼らから生まれた 928 人の成人した子供の身長の関係を図示したものです。データは 0.1 インチ単位に丸められているため、格子点上に分布しています。また、そのために多くのサンプルが同じ位置にプロットされています。図 1.1 から明らかな通り、身長の高い両親からは身長の高い子が生まれる傾向があり、親子の身長の間には正の相関関係があることがわかります。（「相関」という概念については、すぐ後で詳述します）

　なおこのデータは、Galton, F. (1886) で取り上げられたものですが、筆者の Francis Galton は回帰分析や相関係数の概念を提唱した人物として知られており、この親子の身長のデータは回帰分析の歴史のごく初期の例として有名です。

図 1. 1



Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature *Journal of the Anthropological Institute*, 15, 246-263

1. 2　二変量データの要約

基本統計量の解説で強調した通り、データ解析の第一歩はデータを数値的、視覚的に要約し、データの持つ特徴、傾向を把握することにあります。

二変量データを視覚的に要約するもっとも簡単な方法は、図 1. 1 にあるように二つの変数の値を二次元平面にプロットしたもので、これを**散布図 (scatter plot)** と呼びます。

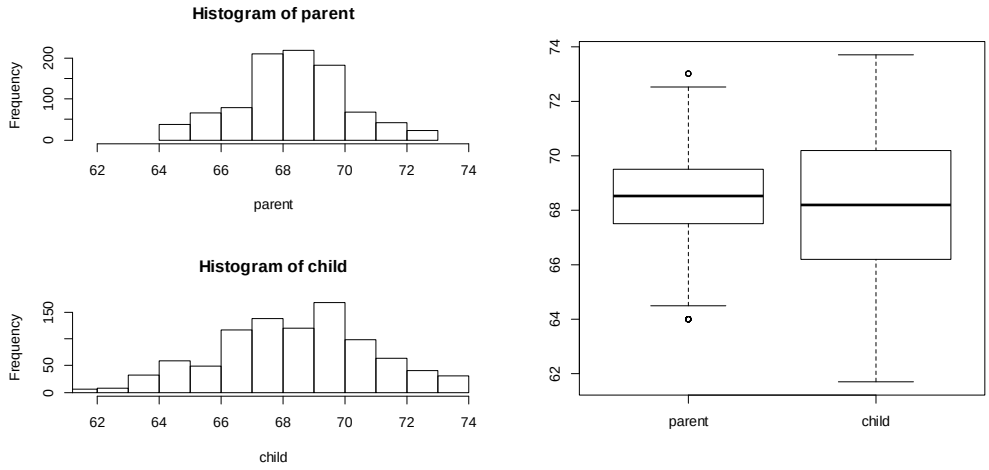
一方、二変量データを数量的に要約するには、1）個々の変数に対する通常の数値的要約と、2）二変量の関係の強さを要約する「共分散」「相関係数」の概念があります。

1）個々の変数に対する数量的要約：Galton の親子の身長

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	IQR
parent	64	67.50	68.5	68.31	69.5	73	1.79	2
child	61.7	66.20	68.2	68.09	70.2	73.7	2.52	4

parent, child 各変数の要約統計量を見ると、平均、中央値がほぼ等しく分布の中心は同じである一方、標準偏差 (SD)、四分位点間距離 (IQR) を見ると、child の散らばりの方が大きいことがわかります。このことは、以下のヒストグラム、ボックスプロットを用いた視覚的要約でも確認することが出来ます。

図 1.2



2）共分散、相関係数

二変量のデータが存在し互いに影響し合っているとき、二つの変数がいかに影響し合っているかを定量的に理解することを考えます。このとき、よく用いられる統計量に以下の**共分散**と**相関係数**があります。

定義：今、$(x_1, y_1), \cdots, (x_n, y_x)$が与えられたとする。このとき x と y の間の**共分散**を以下で定義する。

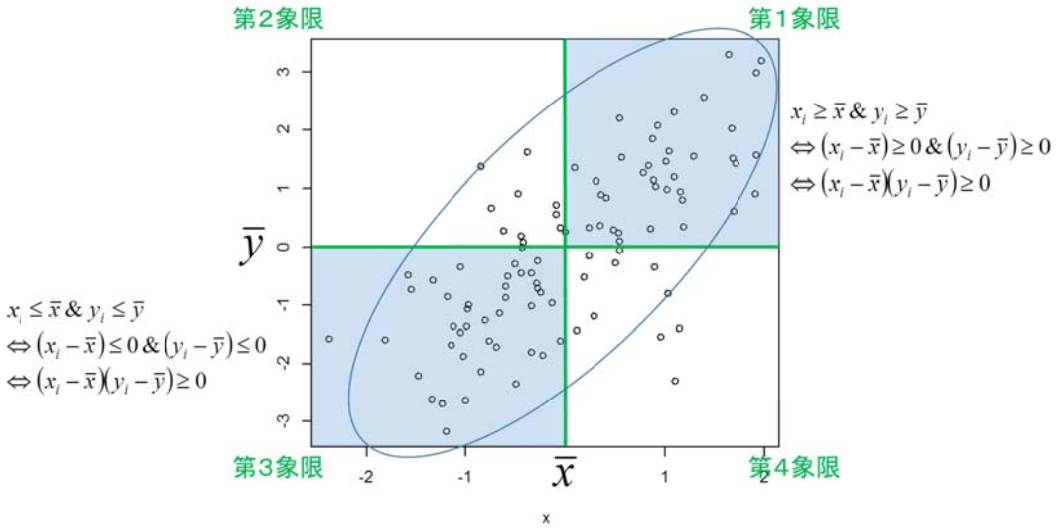
$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

共分散の性質：
x と y の間に正の(負の)相関があるとき、$\mathrm{Cov}(X, Y)$はそれぞれ正(負)になる。

従って、共分散を考えるときは共分散の符号が問題で有り、共分散の値そのものはあまり問題になりません。共分散の値は x と y の単位に依存しますが、単位は二つの変数の相関にあまり関係がありません。

二つの変数の間に正の相関があるとき，共分散が正の値をとる直感的な説明は次のようなものです。

図 1.3



x と y の間に正の相関があるときデータ領域を $\bar{x}, \bar{y}$ で分割すると、データの大部分は第 1，3 象限に存在します。第 1，3 象限のいずれでも $(x_1 - \bar{x})(y_i - \bar{y}) \ge 0$ となるので、共分散 $(1/n)\sum_{i=1}^{n}(x_1 - \bar{x})(y_i - \bar{y})$ も正になります。

定義：$(x_1, y_1), \cdots, (x_n, y_x)$ が与えられたとき、x と y の間の**相関係数**を以下で定義する。

$$\mathrm{Corr}(x, y) = r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

相関係数の性質：

- 相関係数は、x と y の**線形関係**の強さを測る尺度。相関係数の絶対値が 1 に近いほど、x と y の関係は直線に近い。
- $-1 \le r \le 1$
- $r = +(-)1$：正（負）の完全な相関。線形関係。
- 相関係数は x と y の単位に依存しない。

例、共分散、相関係数：Galton の親子の身長

Galton の親子の身長のデータの場合、共分散=2.064、相関係数＝0.459 となります。共分散、相関係数が正の値ですから、親と子の身長の間には正の相関があります。また、相関係数の絶対値は 0 と 1 の間くらいであり、二つの変数の間の直線関係は強くもなく弱くもない、といったところです。

1．3　回帰分析

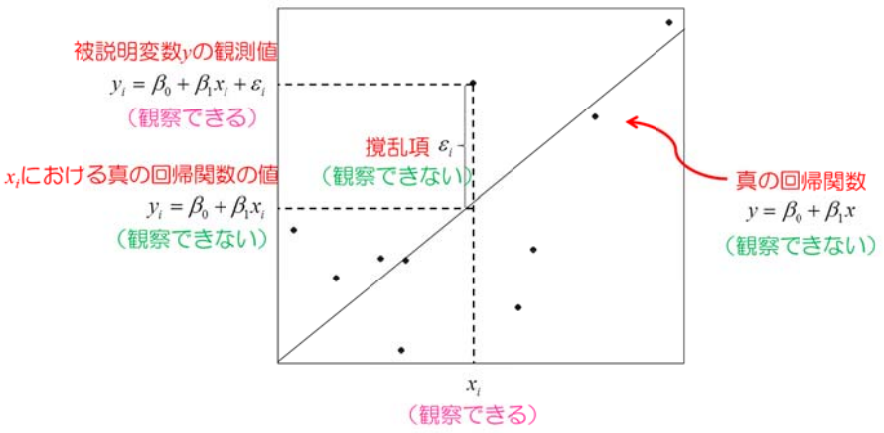
いま、二つの変数 x と y の関係が線形（直線）で近似できるとします。そのとき、x と y の関係を以下の**回帰式（regression equation）**でモデル化します。

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n$$

ただし、(左辺の) 説明される変数 y を，**従属変数、被説明変数、response variable,**（右辺の）説明する変数 x を**独立変数、説明変数、independent variable** と呼びます。また、未知のパラメター β_0, β_1 を**回帰係数、regression coefficient**、ε を**誤差項、撹乱項、error term** と呼びます。

このうち、$y = \beta_0 + \beta_1 x$ を真の回帰関数とよび、x と y の間に想定した関数関係を示しています。実際のデータにはランダムな誤差 ε が含まれているので，実際に観測される被説明変数 y の値は「真の回帰関数＋誤差」になっています。

図 1.4



回帰モデルの仮定：

- **線形性（linearity）**：被説明変数 y と説明変数 x の関係は直線で近似できる。
- **独立性（independence）**：サンプル $(x_1, y_1), \cdots, (x_n, y_x)$ は互いに独立である。すなわち，あるサンプルの値が他のサンプルの値に影響を与えない。
- **正規性（normality）**：撹乱項 ε は期待値 0，分散 σ^2 の正規分布に従う。正規分布については、以下に説明する。 $\varepsilon \sim N(0, \sigma^2)$
- **等分散性（homoscedasticity）**：撹乱項 ε の分散は（従って、被説明変数 y の分散も） σ^2 で一定である。 $V(\varepsilon) = E(\varepsilon^2) = \sigma^2$.

以上の仮定に従って、未知の回帰係数と真の回帰関数を推定するわけですが、その前に、上の回帰モデルの仮定に出てきた「正規分布」という確率分布について解説しておきます。

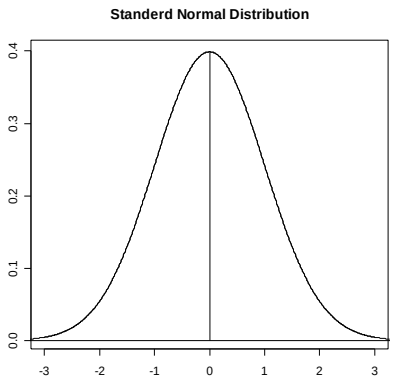
1．4　正規分布

正規分布（normal distribution）は、最も代表的な連続型の（実数値をとる）確率分布といえます。正規分布は自然界の様々な局面で登場しますが、特に計測値に含まれるランダムな測定誤差を表すのに用いられています。

正規分布の確率密度関数は、以下のように与えられます。（「確率密度関数」という概念については、数理統計学の教科書を参照してください。）

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-(x-\mu)^2 / 2\sigma^2\right\}, -\infty < x < \infty.$$

図 1.5：標準正規分布の確率密度関数



正規分布の特徴：

i. $E(X)=\mu, V(X)=\sigma^2$.　正規分布は期待値（平均）＝μ、分散＝σ^2。

ii. 正規分布の分布型は、μ と σ^2 で特徴付けられる。すなわち、μ あるいは σ^2 が異なれば、異なる種類の正規分布になる。

iii. 特に、期待値=0，分散=1 の正規分布を、**標準正規分布（standard normal distribution）**と呼ぶ。

iv. **釣り鐘型（bell-shaped）**で、**左右対称**な分布。

正規分布の極めて有用な性質：

前に述べたとおり、正規分布は様々な場面における測定誤差をモデル化するのに有用です。さらに，以下に述べるとおり様々な確率的現象が正規分布で近似される、という性質を持つため，正規分布は極めて重要な確率分布となっています。

確率変数の標準化（standardization）：

いま、X を期待値（平均）＝μ、分散＝σ^2 である確率変数とする。
このとき、$Z=(X-\mu)/\sigma$ とすると、

$$E(Z)=\frac{E(X)-\mu}{\sigma}=\frac{\mu-\mu}{\sigma}=0,$$

$$V(X)=\frac{1}{\sigma^2}V(X)=\frac{1}{\sigma^2}\times\sigma^2=1.$$

すなわち、任意の確率変数 X に対して、$Z=(X-\mu)/\sigma$ は必ず期待値=0、分散=1 を持つ。この、$Z=(X-\mu)/\sigma$ なる変換を確率変数 X の**標準化（standardization）**と呼ぶ。

正規確率変数の標準化：

上で述べた確率変数の標準化は任意の確率変数に対して成り立つが、特に、X が期待値＝μ、分散＝σ^2 の正規分布に従うとき（このことを $X\sim N(\mu,\sigma^2)$ と表す）、正規確率変数 X の標準化 $Z=(X-\mu)/\sigma$ は期待値=0，分散=1 の標準正規分布 $N(0,1)$ に従います。

上の正規分布の特徴（ii）で述べたとおり、正規分布は μ と σ^2 が異なれば別の正規分布になりますが、標準化により全ての正規分布は $N(0,1)$ に帰着します。

中心極限定理（Central Limit Theorem, CLT）：

中心極限定理は，任意の確率分布から得られたサンプルの標本平均の分布は，サンプル数が大きくなるとき正規分布で近似できる，という重要な定理です。

定理：中心極限定理

$X_1,\ldots,X_n$ を独立かつ同一の分布に従う確率変数とする。ただし、

$E(X)=\mu, V(X)=\sigma^2$ とする。このとき、標本平均の分布は正規分布に収束する。

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \to N(\mu,\sigma^2/n), n\to\infty$$

中心極限定理では、元のデータ $X_1,\ldots,X_n$ がどのような確率分布に従うか仮定されていないことに注意してください。すなわち、どのような確率的現象から出発しても、十分多くのサンプルを集めれば、標本平均の性質は正規分布という特定の確率分布で解析できることを示しています。

中心極限定理と正規確率変数の標準化：
さらに、上に述べた標準化と中心極限定理を組み合わせれば、以下の結果を導くことができます。

$X_1,\ldots,X_n$ を独立かつ同一の分布に従う確率変数とする。$E(X)=\mu, V(X)=\sigma^2$ とすると、$E(\overline{X})=\mu, V(\overline{X})=\sigma^2/n$ となる。このとき標準化した標本平均の分布は標準正規分布に収束する。

$$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Longrightarrow N(0,1), n\to\infty$$

この結果はきわめて強力であって、元データがいかなる確率分布に従おうとも、サンプル数が十分大きければすべての議論は標準正規分布というただ一つの分布に帰着してしまうことを意味します。

1．5　回帰係数の推定

与えられたデータ $(x_1,y_1),\cdots,(x_n,y_x)$ に対して、未知の回帰係数 β_0,β_1 を推定し、回帰式 $y_i=\beta_0+\beta_1x_i+\varepsilon_i$ を当てはめることを考えます。説明変数 $x=x_i$ が与えられた時、推定された（説明された）回帰式の値は $\beta_0+\beta_1x_i$ となります。このと

き i 番目のサンプルとして観察された被説明変数の値 y_i と回帰式の値の差は、誤差項 $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ になります。つまり、誤差項とは被説明変数の値の中で回帰によって説明できなかった部分に当たります。誤差項そのものは正負いずれの値もとりますから、被説明変数 y_i と真の回帰モデルの「乖離」を ε_i^2 で定義することにします。このとき**残差二乗和** $\sum_{i=1}^{n} \varepsilon_i^2$ は、データ全体における被説明変数の値 y_i の変動のうち、回帰によって説明できなかった変動の総和になりますから、この $\sum_{i=1}^{n} \varepsilon_i^2$ が最小になるように回帰式を推定することを考えます。

最小二乗推定量 (Ordinary Least Squares Estimator, OLSE)：

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

$$\Rightarrow \begin{cases} \dfrac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = 0 \\ \dfrac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2 = 0 \end{cases} \Rightarrow \begin{cases} n\beta_0 + \beta_1 \left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} y_i \\ \beta_0 \left(\sum_{i=1}^{n} x_i\right) + \beta_1 \left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Rightarrow \begin{cases} \hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

推定された回帰直線： $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

最小二乗推定量の性質：

- **不偏性 (unbiasedness)：** $E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1$
- **正規性 (normality)：**

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma_{\hat{\beta}_0}^2\right), \hat{\beta}_1 \sim N\left(\beta_1, \sigma_{\hat{\beta}_1}^2\right)$$

ただし、$\sigma_{\hat{\beta}_0}^2 = \dfrac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}, \sigma_{\hat{\beta}_1}^2 = \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}$

- **回帰係数の信頼区間 (Confidence Interval, CI)**：

  回帰係数 β_0, β_1 の信頼区間は，以下のように与えられます。

  β_0 の信頼区間：　$\left(\hat{\beta}_0 - s_{\hat{\beta}_0} t_{\alpha/2,n-2}, \hat{\beta}_0 + s_{\hat{\beta}_0} t_{\alpha/2,n-2}\right)$

  β_1 の信頼区間：　$\left(\hat{\beta}_1 - s_{\hat{\beta}_1} t_{\alpha/2,n-2}, \hat{\beta}_1 + s_{\hat{\beta}_1} t_{\alpha/2,n-2}\right)$

ただし、$s_{\hat{\beta}_j}$ ：$\hat{\beta}_j$ の標準誤差、$t_{\alpha/2,n-2}$ ：自由度 n-2 の t 分布の上側 $\alpha/2$ 点。

- **回帰係数に関する仮説検定**

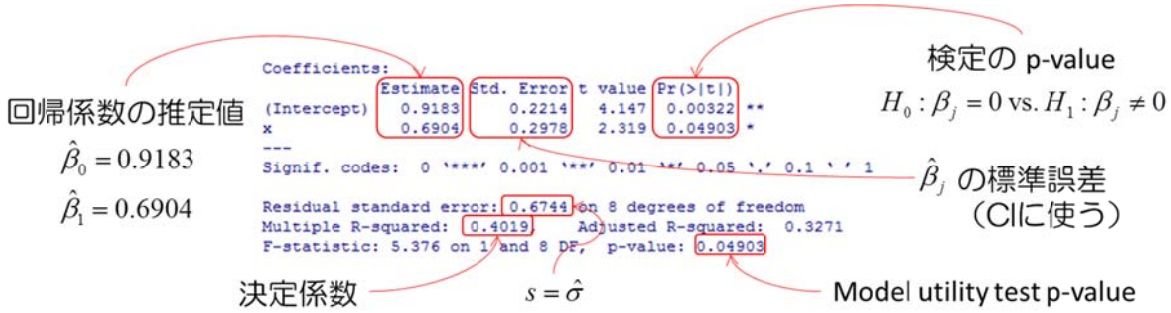
  $H_0 : \beta_j = \beta_{j0}$ vs. $H_1 : \beta_j \neq \beta_{j0}$

  検定統計量：$t = \dfrac{\hat{\beta}_j - \beta_{j0}}{s_{\hat{\beta}_j}} \sim t_{n-2}$ under H_0

ただし、β_{j0} は帰無仮説 H_0 の元で仮定される定数で、通常 $H_0 : \beta_j = 0$ が検定されることが多い。

下に示すのは、回帰分析を行う統計解析ソフトの典型的な出力例になります。
（使用したソフトは R version 3.0.1）
　回帰分析の解析結果を検討するときは，以下の点に注意します。

- 回帰係数の推定値
- 回帰係数の有意性検定の p 値
- 決定係数（被説明変数の変動のうち回帰によって説明された変動の割合）
- Model utility test（回帰モデル全体の有意性検定。後でもう一度触れます）の p 値
- 撹乱項の標準誤差



1．6　多変量回帰分析

前項までは，説明変数が一つのいわゆる単純回帰 (simple regression) について解説してきました。本項以降では，複数の説明変数を持つ多変量回帰分析 (multivariate regression analysis) について検討します。

多変量回帰の場合、モデルを記述するのには行列表示を用いた方が便利です。まず、行列を使って回帰モデルを定式化します。

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \varepsilon_i \sim N\left(0, \sigma^2, i = 1, \ldots, n\right)$$

行列表示を使うと、上の多変量回帰モデルは以下のように簡潔に表せます。

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

$$\varepsilon \sim N\left(\mathbf{0}, \sigma^2 I_n\right), I_n : n\text{-dimensional identity matrix.}$$

前項の単純回帰の場合と同様、残差二乗和 $\sum_{i=1}^{n} \varepsilon_i^2$ を最小化すると回帰係数 $\beta_0, \beta_1, \ldots, \beta_k$ の最小二乗推定量 (Least Squares Estimators, LSE) は、以下のように得られます。

$$\min_{\beta_0, \beta_1, \ldots, \beta_k} \sum_{i=1}^{n} \left\{ y_i - \left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}\right) \right\}^2$$

$$\Rightarrow \hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

最小二乗推定量の性質：

- **不偏性 (unbiasedness)：** $E\left(\hat{\beta}\right) = \beta$
- **正規性 (normality)：** $\hat{\beta} \sim N\left(\beta, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$

- **回帰係数の信頼区間 (Confidence Interval, CI)**：

  β_j の信頼区間：　$\left(\hat{\beta}_j - s_{\hat{\beta}_j} t_{\alpha/2,n-(k+1)}, \hat{\beta}_j + s_{\hat{\beta}_j} t_{\alpha/2,n-(k+1)}\right)$

ただし、$s_{\hat{\beta}_j} = \left(\left[s^2 (X'X)^{-1}\right]_{jj}\right)^{1/2}$ ： $\hat{\beta}_j$ の標準誤差、$t_{\alpha/2,n-(k+1)}$ ：自由度 n-(k+1)の t 分布の上側 $\alpha/2$ 点。

$$s^2 = \hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-(k+1)}, E(s^2) = \sigma^2$$

- **回帰係数に関する仮説検定**

  $H_0 : \beta_j = \beta_{j0}$ vs. $H_1 : \beta_j \neq \beta_{j0}$

  検定統計量： $t = \dfrac{\hat{\beta}_j - \beta_{j0}}{s_{\hat{\beta}_j}} \sim t_{n-(k+1)}$ under H_0

ただし、β_{j0} は帰無仮説 H_0 の元で仮定される定数で、通常 $H_0 : \beta_j = 0$ が検定されることが多い。

決定係数：

上で述べた回帰係数の有意性検定は，個々の係数が有意か（主として 0 に等しいか否か）を検定するものでした。しかし、回帰モデル全体の有意性を議論するためには，別の概念が必要になります。そのために、まず以下を定義します。

- SST (Total Sum of Squares): $\sum_{i=1}^{n} (y_i - \bar{y})^2$　被説明変数 y の、データ全体における変動（全変動）を示します。SST を(n-1)で割ると y の分散になりますね。
- SSE (Error Sum of Squares): $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$　$\hat{y}_i$ は回帰モデルによって推定された y の値ですから、$(y_i - \hat{y}_i)$ は y の中で回帰によって説明されなかった部分でこれを**残差 (residuals)** といいます。SSE は y の回帰では説明されなかった変動を表します。
- SSR (Regression Sum of Squares): $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$　$\hat{y}_i$ は回帰によって説明された（推定された） y の値です。一方もし回帰モデルが存在せず、説明変数 x

の情報なしに y の値を推定するとすれば、それは y の平均値 $\bar{y}$ で推定するしかありません。すなわち $(\hat{y}_i - \bar{y})$ は、回帰モデルを適用することで初めて説明できた y の変化を示しており、SSR は y の回帰によって説明された変動を表しています。

このとき、以下の定理が成立します。

定理：SST =SSR + SSE
（証明略）

以上の概念を用いて、回帰モデル全体のパフォーマンスを評価する尺度として、以下のものを定義します。

定義：決定係数 **(coefficient of determination)** $R^2 = SSR/SST$

すなわち決定係数とは、被説明変数 y の変動のうち、「回帰によって説明された変動の割合」を示しています。上の定理から、$0 \le R^2 \le 1$ であり、R^2 が 1 に近いほど回帰は有効である、といえます。

Model Utility Test: さらに、決定係数 R^2 を使って、以下の仮説を検定することができます。

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{vs. not } H_0$$

この検定を Model utility test といいます。まず、帰無仮説の意味を考えてみます。もし帰無仮説 H_0 が真であるとすると、回帰式は $y = \beta_0 + \varepsilon$ となり、説明変数 x は被説明変数 y を説明するのに、全く役に立たないということになります。もし対立仮説 H_1 が正しければ、いずれかの説明変数がなにがしかの説明力を持つということになります。
この Model utility test の検定統計量は、以下で与えられることが知られています。

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{SSR/k}{SSE/[n-(k+1)]} \sim F_{k,n-(k+1)} \text{ under } H_0$$

1．7　回帰診断 (Regression diagnostics)

これまでの解析で、多変量の回帰モデルを推定し、その結果を解釈するところまで来ました。しかし、ここで解析を終わらせるわけにはいきません。なぜなら、仮にいま手元にあるデータに回帰分析を適用したとして、そのデータにおいて回帰モデルに必要な前提条件が満たされているとは限らないからです。

ここで、回帰モデルの仮定を再掲すると以下の通りです。

回帰モデルの仮定：

- **線形性（linearity）**：被説明変数 y と説明変数 x の関係は直線で近似できる。
- **独立性（independence）**：サンプル $(x_1, y_1), \cdots, (x_n, y_x)$ は互いに独立である。すなわち，あるサンプルの値が他のサンプルの値に影響を与えない。
- **正規性（normality）**：撹乱項 ε は期待値 0，分散 σ^2 の正規分布に従う。正規分布については、以下に説明する。 $\varepsilon \sim N(0, \sigma^2)$
- **等分散性（homoscedasticity）**：撹乱項 ε の分散は（従って、被説明変数 y の分散も） σ^2 で一定である。 $V(\varepsilon) = E(\varepsilon^2) = \sigma^2.$

この回帰モデルの仮定がすべて満たされていない限り、推定や仮設検定の結果は（仮に計算できたとしても）まったくナンセンスなものとなります。

この回帰モデルを成り立たせている前提条件を事後的に確認することを、**回帰診断 (regression diagnostics)** といいます。

また、多重回帰モデルに特有の問題として、もし説明変数の間に線形関係があるならば、パラメターの推定が不安定になる、という**多重共線性 (multicollinearity)** という現象が知られています。これは、数学的には回帰式 $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ において、もし説明変数が一次従属の関係にあると $\mathbf{X}'\mathbf{X}$ が「特異」行列になり、最小二乗推定量 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ の中の $(\mathbf{X}'\mathbf{X})^{-1}$ が定義できなくなるのが原因です。（$\mathbf{X}'\mathbf{X}$ が完全に特異でなくても、$(\mathbf{X}'\mathbf{X})^{-1}$ の計算が非常に不安定なんる）直感的には、複数の説明変数が比例関係にあるとき、それらの変数が本質的に同じ情報を持ち冗長であるのが、多重共線性であるといえます。この多重共線性を発見することも、回帰診断の目的の一つになります。

上に示した回帰モデルの仮定は、1）x, y にかかわるもの、2）撹乱項にかかわるもの、の二つに分けられます。

線形性の仮定の確認と多重共線性の有無の確認は 1）にかかわる問題ですが、これは各変数間の散布図を用いるのが適当です。

線形性の仮定：被説明変数と説明変数の間に、非線形な関係が存在しないことを確認する。（x, y の間に相関がないように見える場合も、モデルに含めて結構です。相関がなければ、「有意ではない」という結果が出るだけです。）

多重共線性：説明変数相互の間に、線形関係が存在しないことを確認する。

一方、独立性、正規性、等分散性の仮定の確認は、2）にかかわります。しかし撹乱項そのものはデータから観察することはできませんので、それに代わるものが必要になります。

定義：残差 **(residuals)** $e_i = y_i - \hat{y}_i$．すなわち残差とは、被説明変数 y と推定された回帰式の値の差になります。

定義：残差プロット **(residual vs. fitted value plot)**　縦軸に残差 e_i、横軸に推定された回帰式の値 $\hat{y}_i$ をとった図。撹乱項と被説明変数の関係を示すものとして、独立性、等分散性の仮定の確認に用いられる。

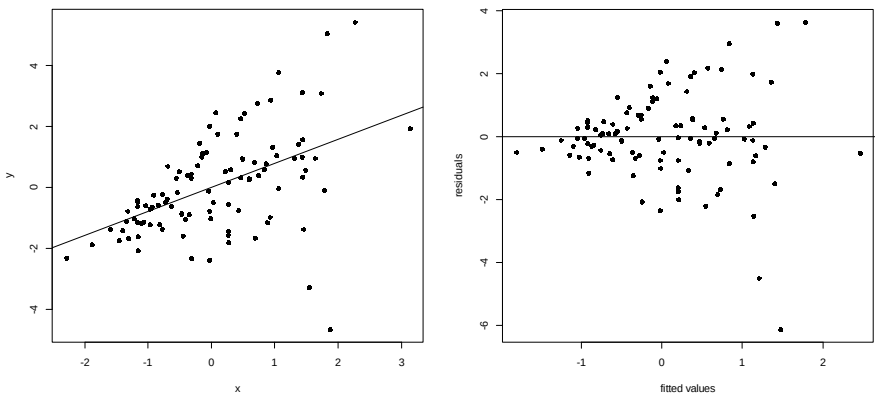
図 1.6 左は、説明変数 x の値が大きくなるにつれ被説明変数 y の分散が増大する傾向のあるデータの例と、それにあてはめられた回帰直線を示しています。一方図 1.6 右は、左図のデータから構成した残差プロットを示していますが、図の右のほうに行くにしたがって残差の分散が大きくなることがわかります。

このように、残差プロットではプロットの中で残差の範囲が変動することを見ることで、**等分散性の仮定**が満たされているかを判断できます。等分散性の仮定が満たされていれば、残差の範囲は均一になります。

また、**独立性の仮定**が満たされる場合、残差プロットには特異なパターンが現

れず、残差プロット一面に一様に点が現れることが知られています。

図 1.6



正規性の確認：
標本分布の正規性の確認は、適切なモデルを選択する上で重要なものとなります。正規性の仮定は後述する“QQ-norm plot”という図で確認しますが、QQ-norm plot を定義するため、まず次の概念を導入します。

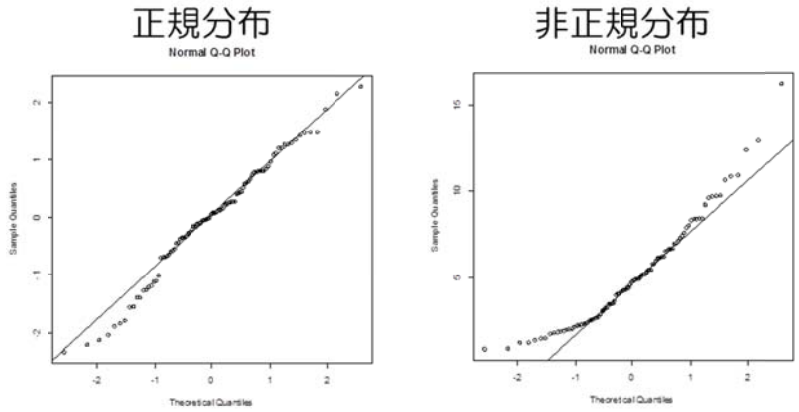
定義：n 個の標本を大きさ順に並べたとき、i 番目に小さな標本は $[100(i\text{-}.5)/n]$ **標本パーセント点 (sample percentile)** であるという。

例えば標本が、正規分布など特定の確率分布から抽出されたとします。このとき、その特定の分布の理論上の $[100(i\text{-}.5)/n]$ パーセント点はデータの $[100(i\text{-}.5)/n]$ 標本パーセント点の近くにあるはずです。したがって、正規分布の理論上のパーセント点と標本パーセント点をプロットすれば、もし元のデータが正規分布から生成されていた場合プロットが直線状に並ぶはずです。

定義：n 個の標本が得られたとき，標準正規分布の$[100(i\text{-}.5)/n]$ パーセント点と，i 番目に小さな観測値= $[100(i\text{-}.5)/n]$ 標本パーセント点のプロットを、QQ-norm plot (Normal probability plot, 正規確率プロット) という。

回帰モデルの**正規性の仮定**を確認するためには、回帰から得られた**残散の QQ-norm plot** を描き、プロットが直線状に並ぶかどうかを確認すればよい、ということになります。

図 1.7：QQ-norm plot



例：CPU データ

209 種類のコンピュータの中央演算装置（CPU）の性能と、各種の特性値をまとめたデータ。特性値の値から、CPU の性能 (perf) を予測するのが目的。

syct	cycle time in nanoseconds
mmin	minimum main memory in kilobytes
mmax	maximum main memory in kilobytes
cach	cache size in kilobytes
chmin	minimum number of channels
chmax	maximum number of channels
perf	published performance on a benchmark mix relative to an IBM 370/158-3

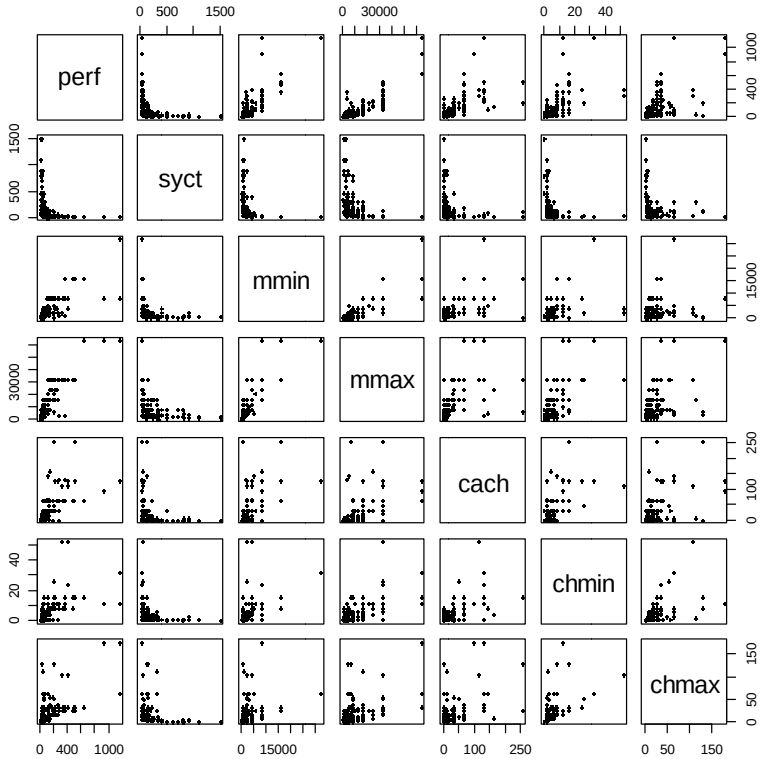
P. Ein-Dor and J. Feldmesser (1987) Attributes of the performance of central processing units: a relative performance prediction model. *Comm. ACM.* **30**, 308–317.
Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth edition. Springer.

まず、各変数の散布図を示すと図 1.8 のようになります。

CPU データの散布図を観察すると、以下のような問題点があるのがわかります。

- y と x の間に，非線形な関係が存在しないか？
  - perf と syct の間に，明らかな非線形関係がある。
- 誤差項の分散は一定か？
  - mmin 等に，明らかな分差増大傾向がある。
- x 同士の間に，線形な関係が存在しないか？
  - 例えば，mmin と mmax の間に明らかに線形関係がある。

図 1.8



したがって、予備的な視覚的要約の段階でも、線形回帰モデルを当てはめるのは不適切であることがわかります。しかしともかく、回帰モデルを当てはめると、結果は次のようになります。

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.590e+01  8.045e+00  -6.948 4.99e-11 ***
syct         4.886e-02  1.752e-02   2.789  0.00579 **
mmin         1.529e-02  1.827e-03   8.371 9.42e-15 ***
mmax         5.571e-03  6.418e-04   8.680 1.33e-15 ***
cach         6.412e-01  1.396e-01   4.594 7.64e-06 ***
chmin       -2.701e-01  8.557e-01  -0.316  0.75263
chmax        1.483e+00  2.201e-01   6.738 1.64e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.99 on 202 degrees of freedom
Multiple R-squared:  0.8649,    Adjusted R-squared:  0.8609
F-statistic: 215.5 on 6 and 202 DF,  p-value: < 2.2e-16
```

元データの回帰分析の結果、次のようなことがわかります。

- 説明変数の有意性検定は，chmin を除き，ほとんどが強く有意.
- 決定係数：R2=0.8649　被説明変数の変動の 86.49%が説明できた.
- Model utility test: p-value < 2.2×10-16

すなわち、回帰分析の出力結果を見る限り、この回帰分析は成功しているとしか言いようがない、ということになります。しかし、もともと図 1.8 の散布図による視覚的なデータの要約の結果、線形性の仮定、等分散性の仮定には大きな疑問があり、また、多重共線性の存在も予想されました。そこで、モデルの仮定が満たされているか確認するため、残差のプロットによる**回帰診断**を行いました。

図 1.9

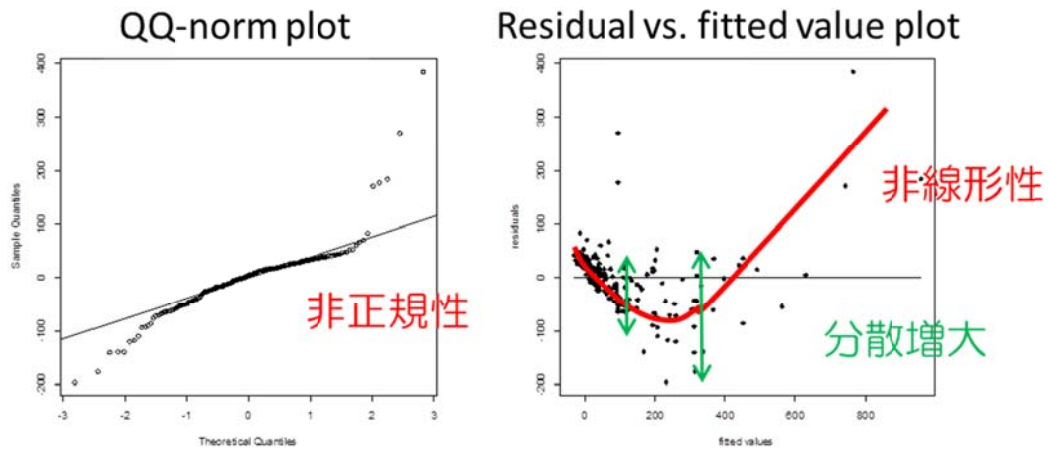


図 1.9 左は､元データに対する回帰分析による残差の QQ-norm plot になります。図から明らかなとおり、QQ-norm plot の点は直線からかけ離れた逆 S 字型 (inverse S-shape) を描いており、**残差の非正規性**を示しています。
一方図 1.9 左は、残差プロットを示しています。まず図からは残差の分散が増大する傾向がみられ、**等分散性の仮定が破れ**ていることがわかります。また、プロットは明らかに非線形な傾向を示しており、線形性の仮定も満たされていないことがわかります。このように、回帰診断を行うことで元のデータでは回帰分析の仮定が満たされておらず、回帰分析を適用することは不適切であることがわかります。
最後に、このように元データにおいて回帰モデルの仮定が破綻していることは、計算の出力からは決してわからず回帰診断によってはじめてわかる、ということを強調しておきます。**（回帰診断の重要性）**

1．8　変数変換

前項でみたとおり、回帰モデルの仮定が満たされていないときは、直接回帰モデルを適用することはできません。（モデルを適用すること自体はできるかもしれませんが、解析結果の解釈は不能で、仮設検定その他の推測も理論的に正当化できません。）

線形回帰モデルの仮定（線形性，正規性，等分散性）が満たされないとき、変数に何らかの変換を施すことで、モデルを改善できる場合があります。

例えば、撹乱項の分散が説明変数の値とともに大きくなる場合、**対数変換 (logarithmic transformation)、冪変換 (power transformation)** によって、モデルの仮定が満たされるようにモデルを修正できる場合があります。

被説明変数の予測値を得るには、まず変換された被説明変数に対して回帰モデルを当てはめ、そのあとで元のモデルに逆変換します。（例えば、対数変換→指数変換）もっともよい返還を選ぶため、いくつかの返還を試してみる必要があります。

ただしこれら対数変換、冪変換などは、その変換を選択した根拠が恣意的なものとならざるを得ず、また、その変換によってモデルが改善したことを理論的に示すことも困難です。このようなとき、対数変換，冪変換を組み合わせた **Box-Cox 変換**により、**分散の安定化と正規性の改善**を同時に達成できる場合があります。

定義：Box-Cox 変換

$$y^{(\lambda)} = \begin{cases} \left(y^{\lambda} - 1\right) / \lambda : \lambda \neq 0 \\ \log(\lambda): \quad \lambda = 0 \end{cases}$$

Box, George E. P.; Cox, D. R. (1964). "An analysis of transformations". *Journal of the Royal Statistical Society*, Series B 26 (2): 211–252.

Box-Cox 変換は、パラメーターλによって特徴付けられる。パラメーターλは、モデルの適合度を最適化するように、ソフトウエアにより自動的に選択される。
例えば、統計解析ソフト R などは、Box-Cox 変換を実装している。

参考文献：Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth edition. Springer.

例：CPU データの Box－Cox 変換

Venables and Ripley (2002), §6.8 と R version 3.0.1, MASS パッケージを用い、CPU データに対して Box-Cox 変換を行いました。

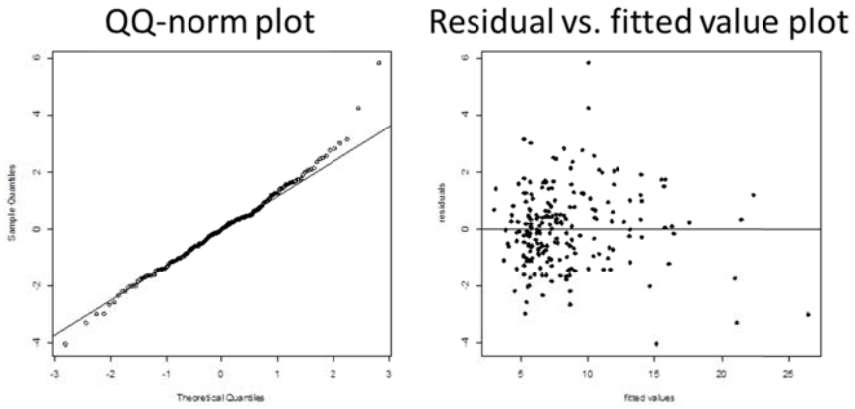
その結果、Box-Cox 変換の最適なλは、λ= 0.2929 となりました。返還後のデータに対する回帰分析の結果は、以下の通りです。

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.214e+00  1.843e-01  28.284  < 2e-16 ***
syct        -1.681e-03  4.014e-04  -4.187 4.21e-05 ***
mmin         1.868e-04  4.186e-05   4.463 1.34e-05 ***
mmax         1.607e-04  1.471e-05  10.924  < 2e-16 ***
cach         2.792e-02  3.198e-03   8.731 9.56e-16 ***
chmin        2.774e-02  1.961e-02   1.415    0.159
chmax        8.330e-03  5.042e-03   1.652    0.100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.375 on 202 degrees of freedom
Multiple R-squared:  0.8821,    Adjusted R-squared:  0.8786
F-statistic: 251.8 on 6 and 202 DF,  p-value: < 2.2e-16
```

- chmin，chmax は有意ではない.
- 決定係数：R2=0.8821　被説明変数の変動の 88.21%が説明できた。
- Model utility test: p-value < 2.2×10-16

回帰診断のための残差プロットは、以下の通りです。QQ-norm plot はほぼ直線上にプロットされ、正規性が向上したことがわかります。また、残差プロットの範囲は均一で、分散が安定化され全体として回帰モデルの仮定が満たされたことがわかります。



Take Home Message

1. 回帰分析

2. 共分散と相関係数

3. 線形回帰モデル
- 回帰係数の推定．最小二乗推定量の性質

4. 回帰診断：回帰モデルの仮定の確認
- 散布図：線形性の確認
- QQ-norm プロット：残差の正規性の確認
- 残差プロット：等分散性，独立性の確認

5. Box-Cox 変換：分散の安定化と正規性の向上

以上