統計解析

古谷知之

授業概要

* 履修者の状況に応じて変更される場合がありますが、全体としては以下のような授業構成となります。
* 講義の中でR演習を行うこともあります。

第1回　ガイダンス・単回帰分析
第2回　重回帰分析(1)
第3回　重回帰分析(2)
第4回　一般化線形回帰モデル(1)
第5回　一般化線形回帰モデル(2)
第6回　一般化線形回帰モデル(3)
第7回　一般化線形回帰モデル(4)
第8回　一般化線形回帰モデル(5)
第9回　一般化線形回帰モデル(6)
第10回　一般化線形混合モデル
第11回　状態空間モデル
第12回　R演習(1)
第13回　R演習(2)
第14回　R演習(3)

統計モデルの種類

	主な推定方法	データ分布	回帰係数
線形回帰モデル （単回帰・重回帰など）	最小二乗法	正規分布	一変数に一つ
一般化線形モデル	最尤推定法	正規分布以外の分布も可能	一変数に一つ
一般化線形混合モデル	最尤推定法	正規分布以外の分布も可能	変数の個体差に応じて推定可能
階層ベイズモデル	ベイズ推定	正規分布以外の分布も可能	変数の個体差に応じて推定可能

本授業で扱う統計モデル

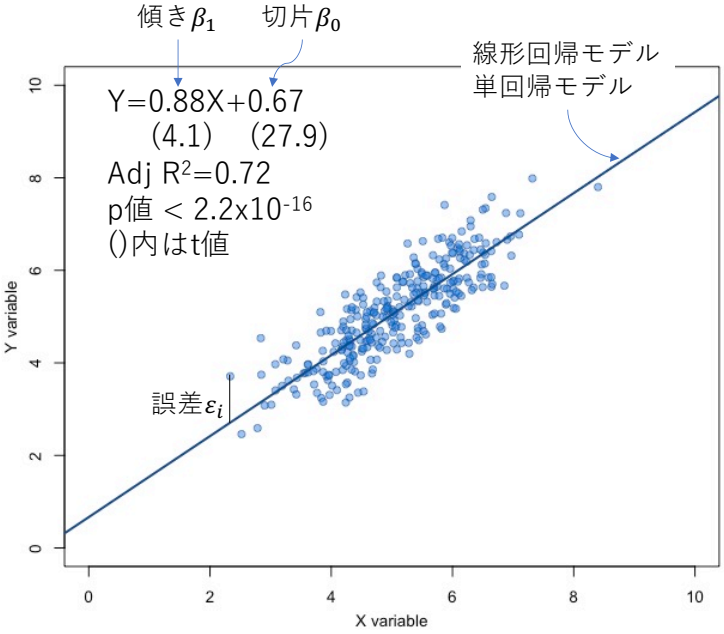
- 線形回帰モデル
  - 単回帰モデル、重回帰モデル
- 一般化線形回帰モデル
  - 離散：ポアソン回帰モデル、二項反応モデル（ロジスティック回帰モデル、プロビット回帰モデル、補対数対数モデル）、負の二項分布モデル、ゼロ過剰ポアソン回帰モデル、ゼロ過剰負の二項分布モデル
  - 連続：ガンマ回帰モデル、ベータ回帰モデル、指数-ガウス回帰モデル
  - スパース：Lasso回帰モデル、Ridge回帰モデル
- 一般化線形混合モデル
  - マルチレベルモデル
- 状態空間モデル

授業内容

- 重回帰分析
- 重回帰モデルの統計量（回帰係数の求め方）
- 最小二乗法と最尤推定法
- 回帰係数の統計的検定
- 寄与率と相関係数
- 自由度修正済みR^2 、AIC、BIC
- 残差の性質と残差解析（外れ値の検出）
- F分布とχ^2分布
- 回帰分析の分散分析
- データの標準化
- 多重共線性

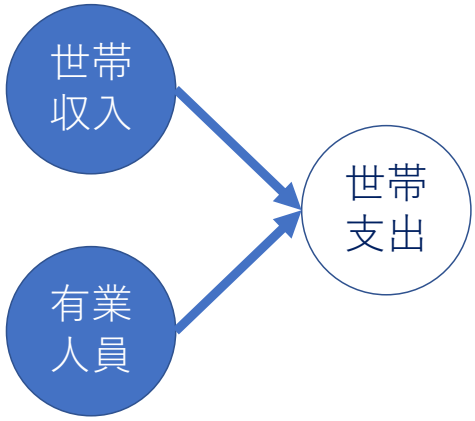
単回帰分析

- ２つの変数x, yについて、
  $$y = \beta_0 + \beta_1 x$$
  のような線形の式形状で表される統計モデルを単回帰モデルという（左図の青直線のこと）
- ここで、単回帰モデルに用いられるβ_0とβ_1は回帰係数という（β_0は切片、β_1は傾きを意味する）
- 従属変数、独立変数ともに正規分布するデータを扱う



変数を増やしてみよう

- 世帯の支出が増える原因は、所得が増えるだけではない
- 他の原因を追加できないか？
- 例えば、同じ世帯で働いている人の人数が増える場合も、支出が増える可能性がある
  - 共働き、同居する子供が働く等



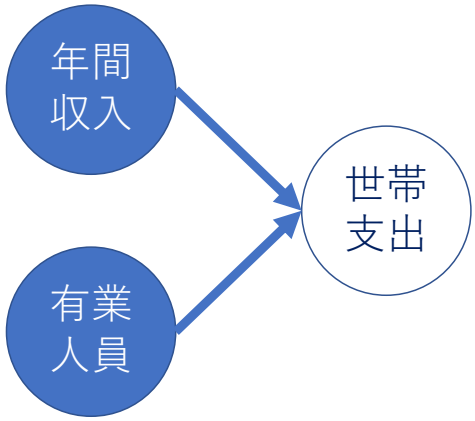
重回帰分析

- 複数の独立変数（説明変数）を用いて回帰分析を行うことを、重回帰分析という

- 独立変数をそれぞれ年間収入x_1、有業人員x_2と表す
- このとき、重回帰モデル式は

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

と表すことができる



重回帰分析

- 従属変数yとk個の独立変数$x_1, x_2, \ldots, x_k$に対する標本数がn個の重回帰モデルは以下のように記述できる$(i = 1, \ldots, n)$。

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \\
y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{aligned}
$$

重回帰分析

- 次のようなベクトルと行列を用いて、

$$
\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}
$$

- 次式のように簡略化できる

$$
\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}
$$

重回帰分析で検討すべきこと

- 回帰係数の値：偏回帰係数
- 説明変数の重要性：標準化偏回帰係数
- 回帰係数の統計的有意性：t検定
- 回帰係数の信頼度：信頼区間
- 変数の選択：自由度修正済み決定係数、AIC、BIC
- 予測への適用可能性：重相関係数、決定係数（寄与率）
- 外れ値の検出：残差解析
- モデル全体の統計的有意性：F検定
- 従属変数間の相関：多重共線性

重回帰分析

- 誤差項$\boldsymbol{\varepsilon} = \boldsymbol{y} - X\boldsymbol{\beta}$の二乗和$Q$は、

$$Q = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T X^T X\boldsymbol{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$$

- 最小二乗法より、

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2X^T\boldsymbol{y} + 2X^T X\boldsymbol{\beta} = 0$$

- ここから以下の正規方程式を得る

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}$$

- 両辺に左から$(X^T X)^{-1}$をかけると、回帰係数の推定量$\widehat{\boldsymbol{\beta}}$を得る

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T\boldsymbol{y}$$

線形回帰分析を行う上での仮定（前提）

- 線形回帰分析では、独立変数と従属変数がともに正規分布に従うことを前提としている

- 独立変数行列Xが平均μ、分散Σの正規分布に従う$X \sim N(\mu, \Sigma)$とき、$X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\beta}\Sigma\boldsymbol{\beta}^T)$となる

- さらに誤差項$\boldsymbol{\varepsilon}$が平均0、分散σ^2の正規分布に従う$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$と仮定している。
- このことから従属変数$\boldsymbol{y}$は平均$X\boldsymbol{\beta}$、分散$\sigma^2 I$の正規分布に従う

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$$

重回帰モデルの統計量

- 回帰係数$\boldsymbol{\beta}$・誤差項$\boldsymbol{\varepsilon}$・従属変数$\boldsymbol{y}$の確率分布から、偏回帰係数$\hat{\boldsymbol{\beta}}$・予測値$\hat{\boldsymbol{y}}$・予測誤差$\boldsymbol{e}$の確率分布は以下のようになる

- 偏回帰係数

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

- 予測値

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \boldsymbol{y} = H\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 H)$$

$$H = X(X^T X)^{-1} X^T$$ Hはハット行列

- 予測誤差

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (I - H)\boldsymbol{y} \sim N\big(0, \sigma^2 (I - H)\big)$$

重回帰モデルの統計量

- 偏回帰係数は正規分布に従う

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

- この性質を標準化すると、以下のようになる

$$\frac{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\sigma^2 (X^T X)^{-1}}} \sim N(0, 1)$$

線形回帰モデルの尤度関数

- データX、未知パラメータ$\boldsymbol{\beta}$、誤差項の分散σ^2が与えられた条件下で、被説明変数$\boldsymbol{y}$が得られる条件付き確率を尤度関数という
- サンプルiの説明変数$x_i = (1, x_{i1}, \ldots, x_{ik})$、被説明変数$y_i$とするとき、尤度関数は以下のような正規分布となる

$$p(y_i | x_i; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_i - x_i \boldsymbol{\beta})^2}{2\sigma^2} \right]$$

尤度関数

- 尤度関数の平均と分散はそれぞれ以下のとおりとなる

- 平均

$$E(y_i|x_i;\boldsymbol{\beta},\sigma^2) = x_i\boldsymbol{\beta}$$

- 分散

$$V(y_i|x_i;\boldsymbol{\beta},\sigma^2) = \sigma^2$$

尤度関数

- 全てのサンプルiについての尤度関数は

$$p(y|X;\boldsymbol{\beta},\sigma^2) = \prod_{i=1}^{n} p(y_i|x_i;\boldsymbol{\beta},\sigma^2)$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

対数尤度関数

- 尤度関数の自然対数をとると

$$\ln[p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2)] = \ln\left[\prod_{i=1}^{n} p(y_i|x_i;\boldsymbol{\beta},\sigma^2)\right]$$

$$= \ln\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2}{2\sigma^2}\right]\right]$$

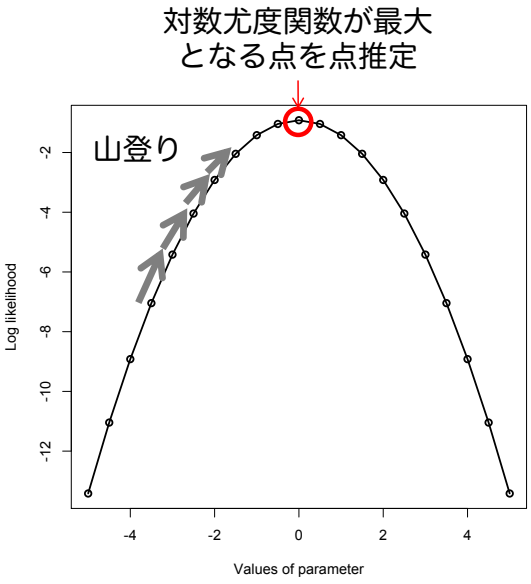
$$= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2$$

最尤法と最小二乗法

- 最尤法では(対数)尤度関数を最大化することで未知パラメータを得る⇒対数尤度関数は上に凸となる関数

- 対数尤度関数を最大化することは、次式を最小化することと同じ

$$\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2$$



最小二乗法による解

- 次式を最小化することにより得られる未知パラメータはそれぞれ以下のようになる

$$\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2$$

- 最小二乗解

$$\widehat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\boldsymbol{\beta}}\right)^2$$ 自由度: $n-(k+1)$

尤度関数(全データ)

- データX、未知パラメータ$\boldsymbol{\beta}$、分散σ^2が与えられた条件下で、被説明変数$\boldsymbol{y}$が得られる条件付き確率、すなわち尤度関数は、以下のような正規分布となる

$$p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{(y-X\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

尤度関数(全データ)

- 尤度関数の平均と分散はそれぞれ以下のとおりとなる

- 平均

$$E(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = X\boldsymbol{\beta}$$

- 分散

$$V(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = \sigma^2$$

尤度関数(全データ)

- 全てのデータについての尤度関数は

$$p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}\right]$$

対数尤度関数(全データ)

• 全てのデータについての尤度関数は

$$\ln[p(\boldsymbol{y}|X;\boldsymbol{\beta},\sigma^2)] = \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}\right]\right]$$

$$= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}{2\sigma^2}$$

最小二乗法による解(全データ)

• 対数尤度関数を最大化⇔最小二乗法による不偏推定量が得られる

$$\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$$

$$\hat{\sigma}^2 = \frac{1}{\nu}\left(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\right)$$

$\nu = n-(k+1)$ …自由度

↑定数項を加えた変数の数

ピアソンの相関係数

- 相関係数r=共分散S_{xy} /標準偏差の積$S_x\ S_y$
- ２つの変数$x = (x_1, \ldots, x_n)$と$y = (y_1, \ldots, y_n)$について、標本平均をそれぞれ$\bar{x}, \bar{y}$とする。このとき、
- 共分散

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

- 標準偏差

$$S_x = \frac{\sqrt{\sum_{i=1}^{n}(\bar{x} - x_i)^2}}{n - 1}$$

$$S_y = \frac{\sqrt{\sum_{i=1}^{n}(\bar{y} - y_i)^2}}{n - 1}$$

ピアソンの相関係数

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

- 共分散を標準偏差の積で割ることで-1から1の間の値に正規化した指標

- 相関係数は$-1 \le r \le 1$の値を取る。
- $r = 0$のとき無相関
- $|r| = 1$のとき完全な相関

変数の平方和

- 従属変数yの平方和をS_{yy}、独立変数の平方和をS_{xx}とあらわす
- これらは、各変数とその平均との差分の平方和で定義される

$$S_{yy} = \sum_{i=i}^{n} (y_i - \bar{y})^2$$

$$S_{xx} = \sum_{i=i}^{n} (x_i - \bar{x})^2$$

- とくに従属変数yの平方和S_{yy}はデータ変動に関する情報量を意味し、全平方和S_Tとも呼ばれる

寄与率（決定係数）と相関係数

- ここで$S_R = S_{xy}^2 / S_{xx}$とすると、残差平方和の最小値S_eは回帰平方和S_Rを用いて次式のようにあらわすことができる

$$S_e = S_{yy} - S_R = \sum_{i=i}^{n} (y_i - \hat{y}_i)^2$$

$$S_R = S_{yy} - S_e = \sum_{i=i}^{n} (\hat{y}_i - \bar{y})^2$$

- S_{yy}：従属変数yの平方和なのでデータ変動に関する情報量
- S_e：残差平方和は回帰直線では説明できない情報量
- S_R：回帰平方和は回帰直線によって説明できる情報量

寄与率（決定係数）と相関係数

- S_RをS_{yy}で基準化すると、寄与率R^2が得られる

$$R^2 = \frac{S_R}{S_{yy}} = 1 - \frac{S_e}{S_{yy}}$$

- 寄与率（決定係数）R^2は1に近いほど、データ変動をよく説明する
- 相関係数の二乗は寄与率（決定係数）となる

$$r^2 = \left(\frac{S_{xy}}{S_x S_y}\right)^2 = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_R}{S_{yy}} = R^2$$

- 残差分散の不偏推定量$\hat{\sigma}_e^2$は S_eを用いて、$\hat{\sigma}_e^2 = \frac{S_e}{n-k}$となる

重相関係数

- 観測値と予測値との相関係数を表す指標として、決定係数R^2を元にした重相関係数Rが用いられる

$$R = \sqrt{1 - \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2}} = \sqrt{1 - \frac{S_e}{S_{yy}}}$$

変数の選択

- 例えば、説明変数がx_1とx_2の２つ（$k=2$）であるとき、理論的には①説明変数がないモデル、②説明変数がx_1のみのモデル、③説明変数がx_2のみのモデル、④説明変数がx_1とx_2の２つのモデルという、４種類のモデルが考えられ、この中から最も「よい」モデルを選ぶ必要がある
- 一般的に説明変数がk個のとき、モデルにおける変数の組み合わせは2^k通りあるが、すべてを調べることは困難
- 適当な変数を組み合わせたモデルを比較し「よい」モデルかどうかを判断する方法に、自由度修正済み決定係数（$\mathrm{Adj.}\,R^2$）、AIC、BICなどがある

自由度修正済み決定係数$\mathrm{Adj.}\,R^2$

- 説明変数の数を増やすほど決定係数（寄与率）R^2は１に近づく
- 説明変数が多い場合、そうした性質を補正した自由度修正済み決定係数$\mathrm{Adj.}\,R^2$が用いられる
- 自由度調整済み決定係数は決定係数（寄与率） R^2をS_eとS_{yy}の自由度で調整したもの

$$\mathrm{Adj.}\,R^2 = 1 - \frac{S_e/(n-k-1)}{S_{yy}/(n-1)} = 1 - \frac{\dfrac{\sum_{i-1}^{n}(y_i-\hat{y}_i)^2}{n-k-1}}{\dfrac{\sum_{i-1}^{n}(y_i-\bar{y})^2}{n-1}}$$

- ここで、 nは標本数、 kは説明変数の数である

赤池の情報量基準（AIC）

- 赤池の情報量基準AICは、対数尤度関数を最大化する解が得られた場合、対数尤度$\ln\hat{L}$と変数の数kを用いて、次式で表される

$$AIC = -2\ln\hat{L} + 2k$$

- ここで、

$$\ln\hat{L} = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)}{2\sigma^2}$$

- 誤差分散σ^2の最尤推定量$\hat{\sigma}^2$は次式のようになる

$$\hat{\sigma}^2 = \frac{\left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)}{n}$$

赤池の情報量基準（AIC）

- したがって$\ln\hat{L}$は以下のように変形できる

$$\ln\hat{L} = -\frac{n}{2}\ln\left(2\pi\frac{\left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)}{n}\right) - \frac{n}{2}$$

- するとAICは次式のように表される

$$AIC = -2\ln\hat{L} + 2k$$

$$= n\left(\ln\left(2\pi\frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}{n}\right) + 1\right) + 2k$$

赤池の情報量基準（AIC）

- AICの定義式のうち、定数（$n2\pi$とn）を除いた次式を最小にするようなモデルを見つければ良い

$$n \cdot \ln\left(\frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}{n}\right) + 2k$$

- AICはどの値が良いというわけではなく、 AICがより小さい値を取るモデルが良いとされる

ベイズ情報量規準（BIC）

- ベイズ情報量規準は次式で与えられる

$$BIC = -2\ln\hat{L} + 2k \cdot \ln(n)$$

- 誤差項が正規分布に従う重回帰モデルに対しては、次式のようになる

$$BIC = n\left(\ln\left(2\pi\frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}{n}\right) + 1\right) + 2k \cdot \ln(n)$$

- AICと同様に、BICがより小さいモデルが当てはまりの良いモデルといえる

残差の性質

- 推定された回帰モデルの残差$e_i = y_i - \hat{y}_i$（実績値と予測値との差分）は、回帰分析で説明がつかない部分を意味する
- 残差e_iは互いに独立に（各i毎に）平均0、分散σ^2の正規分布に従う。$e_i \sim N(0, \sigma^2)$
- 残差平方和S_eは次のように求められる

$$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \cdots - \beta_k x_{ik}\right)^2$$

- これの自由度は$n - (k+1) - 1 = n - k$なので、回帰の残差分散s_e^2は次式のようになる

$$s_e^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - k}$$

残差の性質

- 単回帰分析の場合、残差$e_i = y_i - \hat{y}_i$は以下の 2 つの性質を持つ

- 残差の和は0になる

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

- 残差と従属変数とは直交する

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) x_i = 0$$

残差の性質

- 誤差εが正規分布に従う仮定$\varepsilon \sim N(0, \sigma^2 I)$の下で、残差$e$も正規分布に従う

$$e \sim N\left(0, \sigma^2 (I - X(X^T X)^{-1} X^T)\right)$$

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- ここで、 $H = X(X^T X)^{-1} X^T$はハット行列と呼ばれる

レバレッジ（梃子比）

梃子＝てこ

- ハット行列Hの第i対角要素をレバレッジ（梃子比） h_{ii}という
- 梃子比は以下の性質を持つ

$$\frac{1}{n} < h_{ii} < 1$$

- 従属変数の個数をk個とすると、以下の性質が成り立つ

$$\sum_{i=1}^{n} h_{ii} = k + 1$$

レバレッジと標準化残差

- レバレッジ（梃子比）を用いて、残差の分散とスチューデント化残差を以下のように得ることができる
- 残差e_iの分散

$$Var(e_i) = \sigma^2(1 - h_{ii})$$

- 残差を標準偏差の推定量で基準化（標準化）したものを標準化残差e_i'という

$$e_i' = \frac{e_i}{\hat{\sigma}}$$

スチューデント化残差

- 標準化残差e_i'と梃子比h_{ii}を用いて、スチューデント化残差を以下のように定義できる

$$\frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

- ここで、残差分散に関する不偏推定量s_e^2は次式から得られる

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}$$

- スチューデント化残差は近似的に$N(0,1)$に従う

梃子比が与える影響

- $Var(e_i) = \sigma^2(1 - h_{ii})$であることから、梃子比$h_{ii}$が 1 に近づけば残差$e_i$は本来の分散$\sigma^2$よりも小さくなる
- このとき、予測式が他のデータよりもi番目のデータに近づくことになり、 このデータがモデルに与える影響が大きいといえる
- $\hat{\boldsymbol{y}} = H\boldsymbol{y}$であることから、次式が成り立つ

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i \cdots + h_{in}y_n$$

- 梃子比h_{ii}が大きくなると、 i番目のデータについてy_iが$\hat{y}_i$に与える影響が大きくなる

回帰係数のt検定

- 偏回帰係数は正規分布する

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

- 各標本についても同様に正規分布する

$$\hat{\beta}_k = \sim N\left(\beta_k, \sigma^2 (X^T X)^{-1}_{kk}\right)$$

- この性質を標準化すると

$$\frac{\hat{\beta}_k - \beta_k}{\sigma^2 (X^T X)^{-1}_{kk}} \sim N(0, 1)$$

回帰係数のt検定

- 未知パラメータである分散σ^2を不偏推定量$\hat{\sigma}^2$で置き換えると

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}^2 (X^T X)^{-1}_{kk}} \sim t(n-k-1)$$

- ここから「偏回帰係数が統計的に有意である」ということをしめすためには、帰無仮説$H_0: \hat{\beta}_k = 0$に対するt検定を行えば良い

回帰モデルの統計的有意性の検定

- 回帰モデルが統計的に有意であることを示すには、「全ての回帰係数が統計的に有意でない」という帰無仮説を棄却すれば良い

- つまり、$H_0: \hat{\boldsymbol{\beta}} = 0$つまり$H_0: \hat{\beta}_1 = \cdots = \hat{\beta}_k = 0$を棄却できればよい

- この帰無仮説のもとでは、回帰平方和S_Rと残差分散s_e^2がともに誤差分散σ^2の不偏推定量になる
- 回帰平方和S_Rは次式により定義され、自由度kのχ^2分布に従う

$$S_R = \sum_{i=i}^{n} (\hat{y}_i - \bar{y})^2$$

回帰モデルの統計的有意性の検定

- $\hat{y}_i$は正規分布に従う$\hat{y}_i \sim N(X_i\boldsymbol{\beta}, \sigma^2 H_{ii})$
- $\bar{\hat{y}}_i = \bar{y}$であることから、 $\hat{y}_i$を標準化した値$\frac{\hat{y}_i - \bar{y}}{\sigma\sqrt{H_{ii}}}$は平均0、分散1の標準正規分布に従う

$$\frac{\hat{y}_i - \bar{y}}{\sigma\sqrt{H_{ii}}} \sim N(0,1)$$

- $\left(\frac{\hat{y}_i - \bar{y}}{\sigma\sqrt{H_{ii}}}\right)^2$は自由度$k$の$\chi^2$分布に従う
- $k+1$個の回帰係数が与えられるとわかることから、その自由度は$(k+1)-1=k$となる

χ^2分布

- 確率変数X_iが平均μ、分散σ^2の正規分布に従う$X_i \sim N(\mu, \sigma^2)$とき、その標準化された値$Z_i = \frac{X_i - \mu}{\sigma}$は標準正規分布に従う

$$Z_i \sim N(0,1)$$

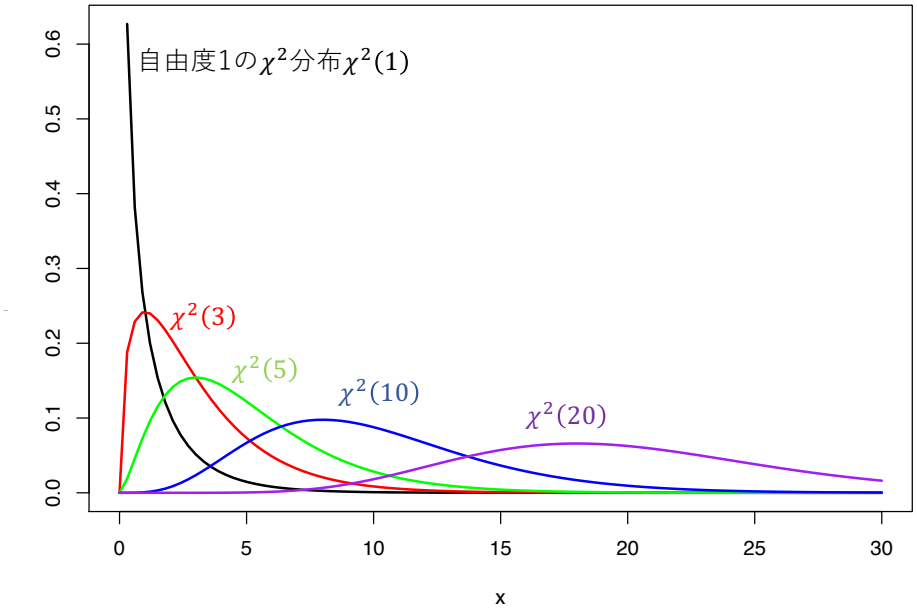
- このときZ_i^2は自由度1のχ^2分布$\chi^2(1)$に従うことが知られている

$$Z_i^2 = \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1)$$

- $W = \sum_{i=1}^{k} Z_i^2$は自由度kのχ^2分布$\chi^2(k)$に従う

$$W = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$$

χ^2分布



χ^2分布

- 自由度kのχ^2分布$\chi^2(k)$の確率密度分布$f(x;k)$は$0 \leq x$の範囲で次式で表すことができる($x < 0$のときは0)

$$f(x;k) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} e^{-\frac{x}{2}} x^{\frac{k}{2}-1}$$

- ここで$\Gamma(\alpha)$はガンマ関数
- 自由度kのχ^2分布の確率密度関数について、期待値と分散はそれぞれkおよび$2k$となる

回帰平方和と残差分散のF値

- 回帰平方和S_Rを自由度kで割った値と残差分散s_e^2の比率をとり、これをF値と呼ぶ

$$F = \frac{S_R/k}{s_e^2}$$

- F値は次式のようになる

$$F = \frac{S_R/k}{s_e^2} = \frac{\sum_{i=i}^{n}(\hat{y}_i - \bar{y})^2}{k} \Big/ \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$

- これは、自由度kのχ^2分布と自由度$n-k-1$のχ^2分布との比率

F分布とχ^2分布

- W_1が自由度k_1のχ^2分布$\chi^2(k_1)$に従い、 W_2が自由度k_2のχ^2分布$\chi^2(k_2)$に従い、 W_1とW_2が互いに独立であるとする

- このとき、 W_1とW_2を各々の自由度で割った値の比率$\frac{W_1/k_1}{W_2/k_2}$をF値といい、自由度k_1, k_2のF分布$F(k_1, k_2)$に従う

$$F = \frac{W_1/k_1}{W_2/k_2} \sim F(k_1, k_2)$$

F分布

- 自由度k_1, k_2のF分布$F(k_1, k_2)$の確率密度関数$f(x; k_1, k_2)$は、次式で表すことができる

$$f(x; k_1, k_2) = \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right) x^{\frac{k_1 - 2}{2}}}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)\left(1 + \frac{k_1}{k_2}x\right)^{\frac{k_1+k_2}{2}}} \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}$$

F分布とt分布

- 標準正規分布$N(0,1)$に従うZと、自由度kのχ^2分布$\chi^2(k)$に従うWがあるとき、次式から得られるt値は自由度kのt分布に従う

$$t = \frac{Z}{\sqrt{W/k}}$$

- 両辺を2乗すると

$$t^2 = \frac{Z^2}{W/k}$$

- Z^2は自由度1のχ^2分布$\chi^2(1)$に従うことから、 t^2は自由度$1, k$のF分布$F(1, k)$に従う

回帰分析の分散分析

- 従属変数とその平均値との差の平方和を全平方和S_Tといい、自由度$n-1$のχ^2分布に従う（全平方和S_Tは前出のS_{yy}と同じ）

$$S_T = \sum_{i=i}^{n} (y_i - \bar{y})^2$$

- 残差平方和S_eは以下のように表され、自由度$n-k-1$のχ^2分布に従う

$$S_e = \sum_{i=i}^{n} (y_i - \hat{y}_i)^2$$

回帰分析の分散分析

- このとき全平方和S_Tは残差平方和S_eと回帰平方和S_Rとに分解される

$$S_T = S_e + S_R$$

- すなわち、次式のように表すことができる。

$$\sum_{i=i}^{n} (y_i - \bar{y})^2 = \sum_{i=i}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=i}^{n} (\hat{y}_i - \bar{y})^2$$

自由度　　$n-1$　　$n-k-1$　　k

回帰分布の分散分析

- 回帰平方和S_Rを自由度kで割った値と残差分散s_e^2の比率F値は、自由度$k, n-k-1$のF分布$F(k, n-k-1)$に従う

$$F = \frac{S_R/k}{s_e^2} = \frac{\sum_{i=i}^{n}(\hat{y}_i - \bar{y})^2}{k} \Big/ \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$

- 棄却すべき帰無仮説は$H_0: \hat{\beta}_1 = \cdots = \hat{\beta}_k = 0$である
- 有意水準5%でこの帰無仮説を棄却する場合、自由度$k, n-k-1$のF値の95%値をF_αとすると
- $F(k, n-k-1) > F_\alpha$なら帰無仮説を棄却できる
- $F(k, n-k-1) \leq F_\alpha$なら帰無仮説を棄却できない

回帰分布の分散分析

- 平方和と自由度、F値などを表にまとめたものを、分散分析表という

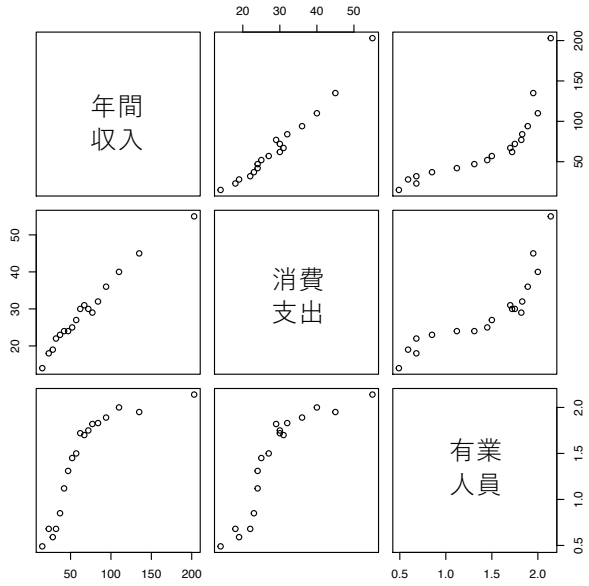
変動要因	平方和	自由度	平均平方	F値
回帰平方和	S_R	k	S_R	$F = \frac{S_R/k}{s_e^2}$
残差平方和	S_e	$n-k-1$	$s_e^2 = \frac{S_e}{n-k-1}$	
全平方和	S_T	$n-1$		

分析に用いるデータ

- 『家計調査』2017年9月、第２－６表「年間収入階級別１世帯当たり１か月間の収入と支出」
- データは政府統計ポータルサイトe-statから入手可能
- 「二人以上の世帯」
- 所得x_{1i}＝「年間収入(10万円)」、有業人員数x_{1i}＝「有業人員(人)」、支出y_i＝「消費支出(千円)」を用いる
- 「年間年収階級」の18階級を分析単位とする

散布図を描いてみる

- 有業人員と消費支出との間にも正の相関関係がありそう？
  - 直線的な関係ではない？
- 有業人員と年間収入との間にも正の相関関係がありそう？
  - これについては後で検討する



変数間の相関を計算する

- 消費支出、年間収入、有業人員数それぞれの間での相関係数を計算すると下表のようになる

	消費支出	年間収入	有業人員数
消費支出	1.000	0.985	0.860
年間収入	0.985	1.000	0.819
有業人員数	0.860	0.819	1.000

重回帰分析の分析結果

- 単回帰分析の結果は以下のようになった ()内はt値

$$\hat{y} = \underset{(10.82)}{12.04} + \underset{(13.18)}{0.187}x_1 + \underset{(2.50)}{2.98}x_2$$

自由度修正済み$R^2 = 0.977$

- $\hat{\beta}_1 = 0.19 > 0$かつ$\hat{\beta}_1$のt値=13.18と5%水準で統計的に有意
- $\hat{\beta}_2 = 2.98 > 0$かつ$\hat{\beta}_2$のt値=2.50と5%水準で統計的に有意
- 自由度修正済み$R^2 = 0.977$と1に近い

- 偏回帰係数$\hat{\beta}_1$と$\hat{\beta}_2$を比較したとき、消費支出に与える影響はどちらが大きいのか？

重回帰分析の分析結果

- 単回帰分析の結果は以下のようになった

$$\hat{y} = \underset{(10.82)}{12.04} + \underset{(13.18)}{0.187}x_1 + \underset{(2.50)}{2.98}x_2 \qquad ()\text{内は}t\text{値}$$

自由度修正済み$R^2 = 0.977$

- 偏回帰係数$\hat{\beta}_1$と$\hat{\beta}_2$を比較したとき、消費支出に与える影響はどちらが大きいのか？
- 年間収入x_1が１単位(10万円)増えると消費支出が187円(=0.187X1000円）増える
- 有業人員x_2が１単位(１人)増えると消費支出が2980円(=2.98X1000円）増える
- 単位が異なると偏回帰係数の影響を比較しづらい？

モデル適合度の比較

- モデル１： $y = \beta_0 + \beta_1 x_1$
- モデル２： $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- モデル３： $y = \beta_0 + \beta_2 x_2$
- 自由度修正済み決定係数、AIC、BICを計算し比較すると、以下のようになる

	Adj. R^2	AIC	BIC
モデル１	0.969	75.6	78.3
モデル２	0.977	71.3	74.9
モデル３	0.723	114.9	117.6

残差解析と外れ値の検出

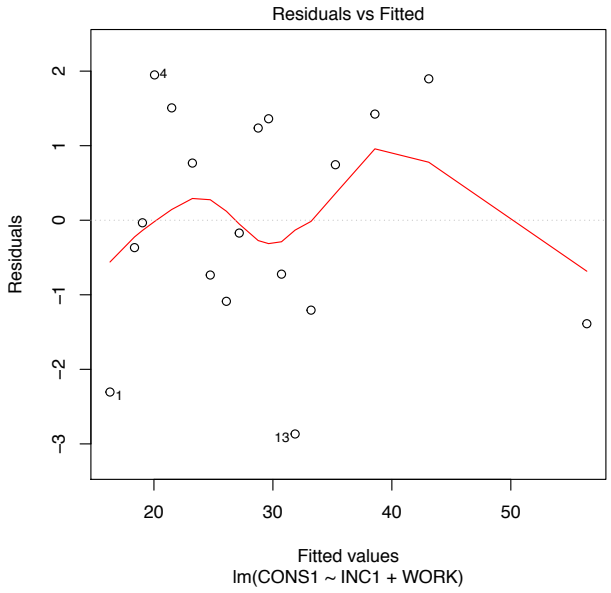
- 残差についての性質を調べることで、回帰分析用いられたデータが外れ値なのかどうかを判断する材料を提供できる。主に以下の手法が用いられる

- 残差プロット
- 残差の正規Q-Qプロット
- S-Lプロット
- 梃子（てこ）比とクックの距離

残差プロット

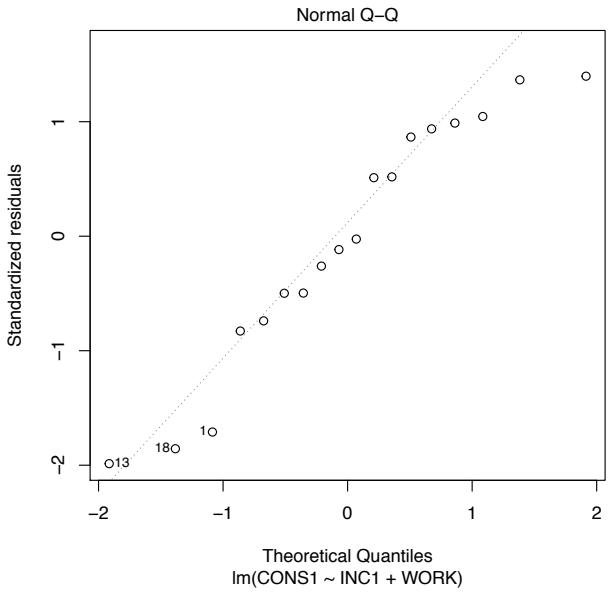
- 予測値$\hat{y}_i$を横軸、残差e_iを縦軸に描いた散布図
- 縦軸の絶対値が$2\hat{\sigma}$を超える残差が多く見られるようなら、データを採用するという仮定を疑うべき→そのデータは外れ値である可能性がある

- 右の結果の場合、 $i = 1, 4, 13$ が外れ値の可能性がある



残差の正規QQプロット

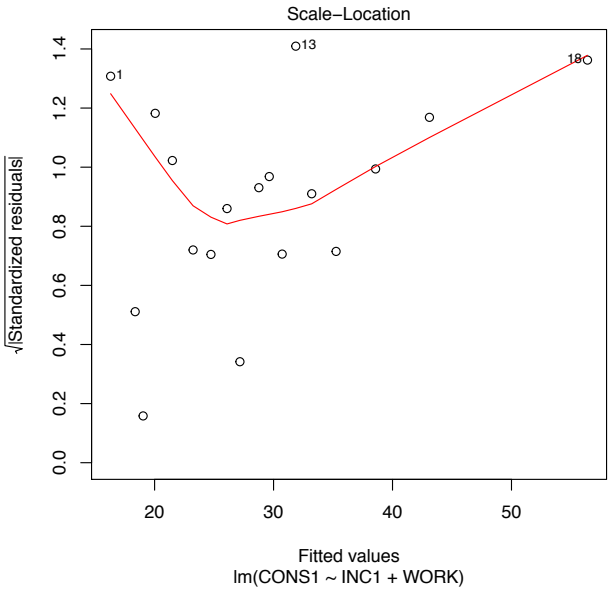
- スチューデント化残差に対する正規Q-Qプロット
- Q-Qプロットではデータ正規分布に従うとき45°線上にデータが乗ってくる性質がある。残差が正規分布に近ければ45°線上にプロットされる
- 縦軸で絶対値が 2 を超える残差が多い場合は、外れ値の可能性がある
- 右の結果の場合、$i = 1, 13, 18$が外れ値の可能性がある



S-Lプロット

- スチューデント化残差の絶対値の平方根を予測値に対して描いた散布図
- 縦軸が$\sqrt{2}$を超えるようなら、データが外れ値であることへの注意が必要

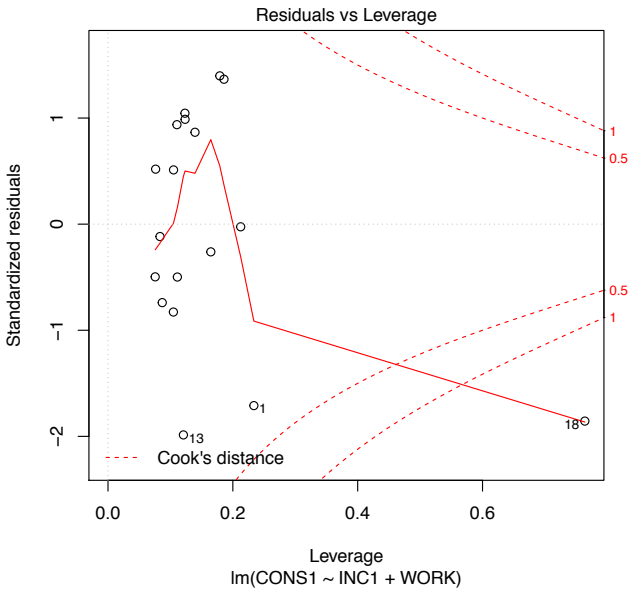
- 右の結果の場合、$i = 1, 13, 18$が外れ値の可能性がある



梃子比とクックの距離

- 梃子比h_{ii}が$2(k+1)/n$より小さければ注意が必要
- 実践的にはクックの距離D_iが $0.2 < D_i \leq 0.5$なら「要注意」、 $0.5 < D_i$なら当該データを「解析から除去」するのが望ましい

- 右の結果の場合、 $i = 18$を除去したほうがよく、 $i = 1, 13$は要注意データだと判断できる



クックの距離

- １つのデータを除去して推計される予測値と、全データを用いて推計される予測値との差の平方和を、誤差分散の推定値で割ったもの

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p \cdot s_e^2} = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{(k+1) \cdot s_e^2}$$

- ここで、 $\hat{y}_j$は全データを用いて得られる予測値、 $\hat{y}_{j(i)}$はデータiを除去して得られる予測値、 pは説明変数の数（重回帰分析のとき$p = k + 1$）、 s_e^2は残差分散の推定値である。

データの標準化と標準化偏回帰係数

- 単位が異なる複数の変数を用いる場合や、単位に意味がない変数（例：５段階評価等）を用いる場合
- 偏回帰係数を比較するために、独立変数と従属変数を平均0・標準偏差1となるデータに標準化する
- 変数xに対する標準化データz_xは以下のように得られる

$$z_x = \frac{x - \bar{x}}{sd(x)}$$

- 標準化した変数を用いて回帰分析をした結果、得られた偏回帰係数を標準化偏回帰係数という

データの標準化

- 例えば消費支出y_i (千円)のデータ

14	18	19	22	23	24	24	25	27	30	31	30	29	32	36	40	45	55

- 平均29.11、標準偏差10.05より

$$\frac{y_i - 29.11}{10.05}$$

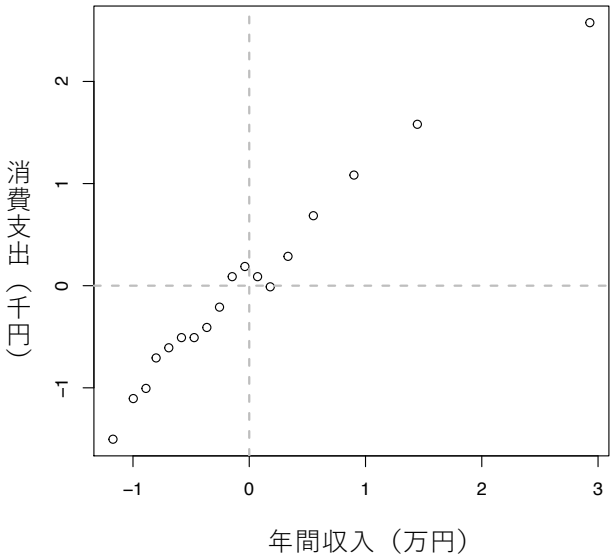
を計算すると、標準化後のデータz_yが得られる

-1.50	-1.11	-1.01	-0.71	-0.61	-0.51	-0.51	-0.41	-0.21	0.09	0.19	0.09	-0.01	0.29	0.69	1.08	1.58	2.58

- 標準化後のデータz_yは、必ず平均0・標準偏差1となる

データの標準化

- 標準化後の年間収入と消費支出の散布図をプロットすると、右図のようになる
- いずれのデータも、平均0±1の辺りに分布していることがわかる



標準化後のデータを用いた回帰分析

- 標準化後の年間収入z_{x1}、有業人員z_{x2}および消費支出z_yを用いて再度重回帰分析をすると、以下のような結果が得られる

$$\widehat{z_y} = \underset{(0.00)}{0.00} + \underset{(13.18)}{0.853} z_{x1} + \underset{(2.50)}{0.162} z_{x2}$$

()内はt値

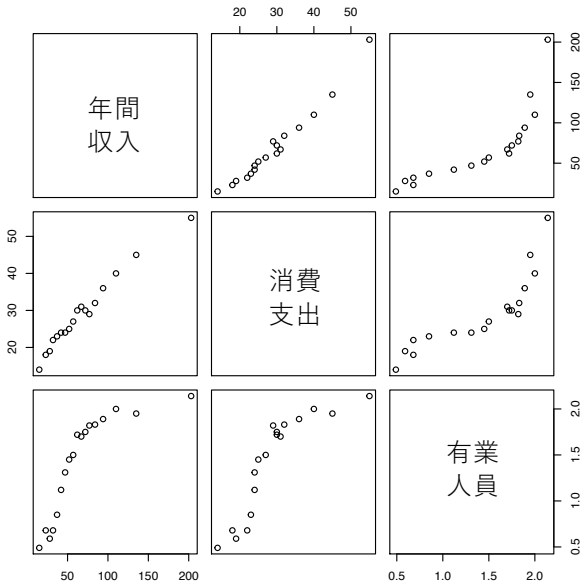
自由度修正済み$R^2 = 0.977$

- 標準化後の独立変数に対する標準化偏回帰係数はいずれも正で5%水準で統計的に有意である
- 年間収入の方が回帰係数が大きいからといって、消費支出に与える影響は有業人員より年間収入のほうが大きい訳ではない
- 1標準偏差変化した場合の変化量を示しているに過ぎず、元のデータの分散を把握しておく必要がある。

変数間の相関

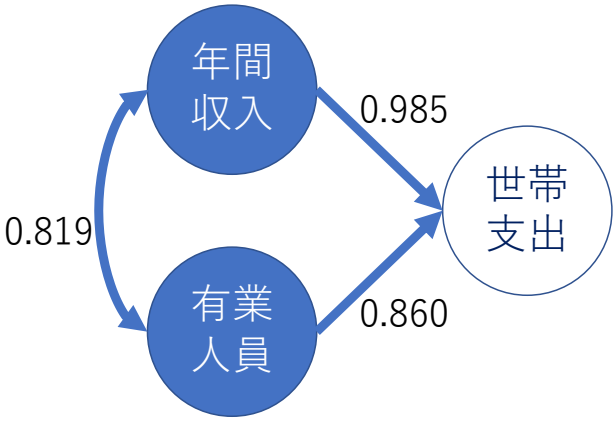
- 独立変数である年間収入と有業人員は、互いに正の相関関係にあるかもしれない
- 実際に、３変数間の相関係数を求めると以下のようになる

	消費支出	年間収入	有業人員
消費支出	1.000		
年間収入	0.985	1.000	
有業人員	0.860	0.819	1.000



変数間の相関

- （当然だが）世帯内で働いている人の数（有業人員）が増えると、世帯の年間収入は増加するだろう
- 独立変数同士が相関する場合、両方の変数を同時に独立変数として採用してよいのか？
- どちらかを採用するとしても、どちらを採用するべきか？



多重共線性

- 独立変数間に関連がある場合、以下のような状況が生じることがある

- t値が過小評価される（実際に有意でも有意でなくなる等）
- 偏回帰係数の標準誤差（分散）が大きくなる（回帰が歪む）
- 決定係数の値が大きくなる
- 偏回帰係数の符号が本来なるべき符号とは逆の符号になる

多重共線性の測定

- 分散拡大係数 (variance inflation factor: VIF) を用いて多重共線性の深刻度を測定する
- 独立変数x_1とx_2の標準偏差σ_{x1}とσ_{x2}および共分散σ_{x1x2}から相関係数rが次式より計算できる

$$r = \frac{\sigma_{x1x2}}{\sigma_{x1}\sigma_{x2}}$$

- 相関係数rを二乗した重相関係数r^2を用いて、VIFは以下のように計算される

$$VIF = \frac{1}{1 - r^2}$$

多重共線性の測定

- VIFが10以下のとき多重共線性がないと判断される（理想的には2以下）。VIFが10以上のとき、どちらかの変数を外して再度回帰分析を行う。

- 年間収入x_1と有業人員x_2の$VIF = \frac{1}{1-0.819^2} \approx 3.04$となる
- 従って、この 2 変数の間に多重共線性があるとはいえない