

Rと一般化線形モデル

1. 一般化線形モデル

分散分析、線形回帰分析は線形モデルであり、残差が正規分布に従う仮定に基づいている。しかし、データが常に正規分布に従うと言う保証はない。また、非線形の現象については線形になるような変換を施し、線形モデルで問題を解決することもできる。しかし、このような方法ではモデルを不自然な尺度で歪んだ解釈を行ってしまう危険性が伴う。

一般化線形モデル (Generalized Linear Model) は、正規分布を含んだ分布族 (family) にデータを対応させ、非

線形の現象を線形モデルの場合と同じく簡単に扱え、かつ不自然な尺度で解釈しないように工夫したデータ解析方法である。また、一般化線形モデルは、被説明変数(反応変数、応答変数とも呼ぶ)が2値データ、例えば「男」と「女」、「死」と「生存」、「はい」と「いいえ」のようなデータのモデルも含んでいる。

通常線形モデルは次の式で表される。

$$y = X\beta + e$$

X は説明変数の行列である。一般化線形モデルでは、 $X\beta$ という線形結合から、 $g(\mu) = X\beta$ のような変換を行った拡張である。この μ は被説明変数の平均で、 g をリンク関数と呼ぶ。

R では、パッケージ `stats` に一般化線形モデルの関数 `glm` が用意されている。関数 `glm` で対応できる主な分布を表 1 に示す。

関数 `glm` の最も簡単な書式を次に示す。

`glm(formula, family, data)`

引数 `formula` は、関数 `lm` と同じくモデルの式を、引数 `family` には表 1 の分布名を指定する。デフォルトには `gaussian` が指定されている。

一般化線形モデル関数 `glm` の使用法について例を用いて説明する。R に `airquality` というデータがある。

データ `airquality` は 1973 年 5 月から 9 月までのニューヨークの大気状態を 6 つの変数で観

表 1 関数 `glm` で使用可能な主な分布

分布族(family)	リンク関数 $g(\mu)$	y_i の範囲
正規(gaussian)	μ	$(-\infty, +\infty)$
二項(binomial)	$\log(\mu/(1-\mu))$	$\frac{0, 1, 2, \dots, n_i}{n_i}$
ポアソン(poisson)	$\log(\mu)$	$0, 1, 2, \dots$
ガンマ(Gamma)	$1/\mu$	$(0, +\infty)$
逆正規(Inverse.gaussian)	$1/\mu^2$	$(0, +\infty)$

測した 154 の観測値である。データの中の変数を次に示す。

- [,1] Ozone オゾンの量 (ppb)
- [,2] Solar.R 日射量 (lang)
- [,3] Wind 風力 (mph)
- [,4] Temp 温度 (華氏 F)
- [,5] Month 月 1~12
- [,6] Day 月のうちの日 1~31

ここでは、日射量、風力、温度の値でオゾンの量を説明できるかどうかと言う、オゾンの量を被説明変数とした重回帰モデルを考えることにする。第 5 列の月(Month)、6 列の日(Day)のデータは必要ではないので、次のように新たなデータセットを作成する。

```
>data(airquality)
>airq2<-airquality[,1:4]
>airq2
  Ozone Solar.R Wind Temp
1     41     190  7.4  67
2     36     118  8.0  72
<後略>
```

回帰分析を行う前に、まず 4 変数の対散布図で変数の相互関係を考察してみよう。対散布図関数 `pairs` に引数 `panel=panel.smooth` を用いると散布図の点の傾向を示す曲線が描かれる。

```
>pairs(airq2,panel=panel.smooth,lwd=2)
```

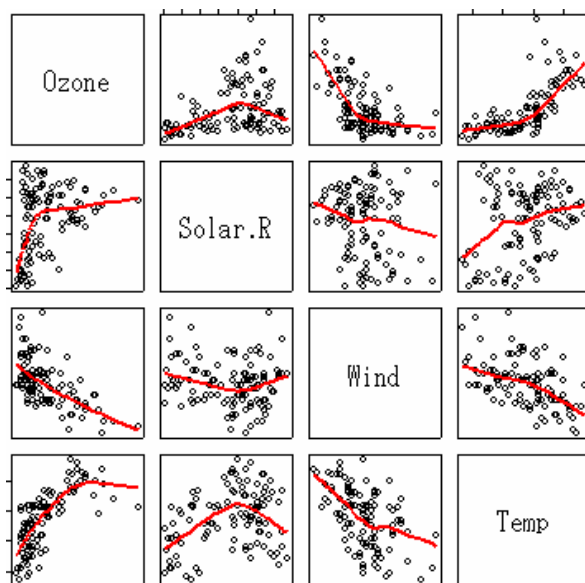


図 2 airquality の対散布図

対散布図から、日射量(Solar.R)、温度(Temp)の値が大きくなるに伴い Ozone の値が大きく、風力(Wind)の値が大きいかほどオゾン量が小さくなる相関関係および逆相関関係があることが分かる。

そこで、Ozone を被説明変数とし、残りの 3 変数を説明変数とした重回帰分析を行うことにする。

```
>airq2.lm<-lm(Ozone~.,data=airq2)
```

次に残差の Q-Q プロットを図 1 に示す。

```
> qqnorm(resid(airq2.lm))  
> qqline(resid(airq2.lm))
```

図 1 で分かるように残差が正規分布に十分良く当てはまっているとは言いがたい。

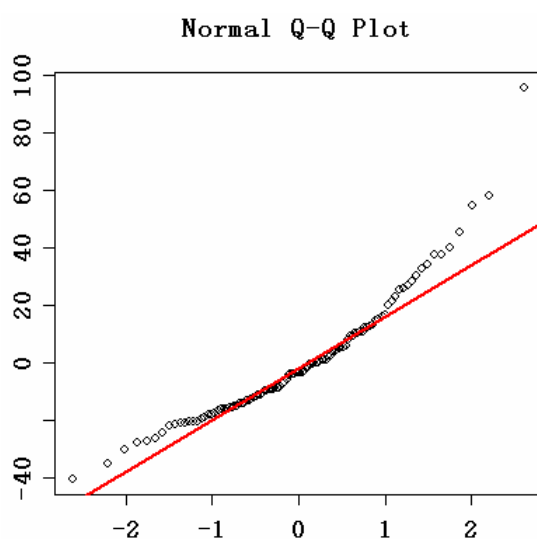


図 1. 残差の Q-Q プロット

そこで、関数 lm による重回帰モデルと一般化線形モデルによる重回帰モデルとの当てはめの良さについて比較してみることにする。モデルの当てはめの良さに関する評価は、AIC を用いることにする。

まず関数 lm による重回帰モデルの AIC を次に示す。

```
> AIC(airq2.lm1)
```

```
[1] 998.717
```

次に関数 glm の gaussian、Gamma 分布を用いた場合の AIC 値を求める。

```
>AIC(glm(Ozone~Solar.R+Wind+Temp,data=airq2,family=gaussian))
```

```
[1] 998.7171
```

```
>AIC(glm(Ozone~Solar.R+Wind+Temp,data=airq2,family=Gamma))
```

```
[1] 939.8778
```

AIC の値から分かるように、関数 lm による重回帰モデルと関数 glm の gaussian 分布を用いた結果は同じである。

Gamma 分布による AIC の値が gaussian 分布を用いた場合より小さいので、Gamma 分布によるモデルの当てはめが良いと判断される。

関数 glm の引数 family に poisson を指定した場合の回帰分析をポアソン回帰分析とも呼ぶ

2. ロジスティック回帰と一般化線形モデル

(1) ロジスティック回帰分析

次の関数をロジスティック関数と呼ぶ。

$$p = \frac{e^{\eta}}{1 + e^{\eta}}$$

ロジスティック関数がどのような形をしているかを見ることにしよう。次のコマンドで、図 1 のような横軸が -5 から 5 までの範囲内のロジスティック曲線が作成される。

```
>eta<-seq(from=-5,to=5,length=200)
```

```
>plot(eta,exp(eta)/(1+exp(eta)),type="l")
```

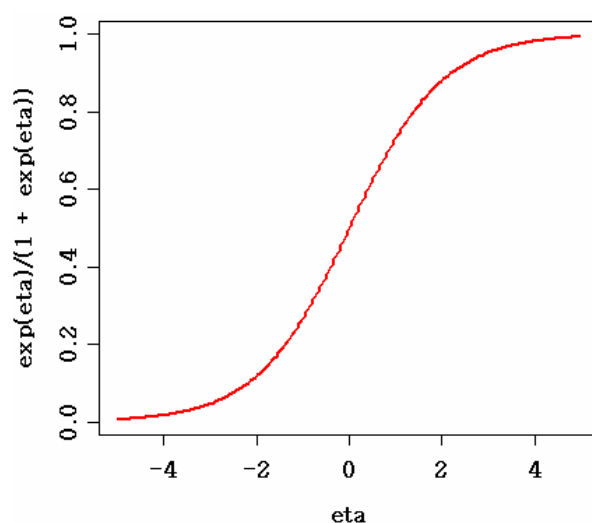


図 1. ロジスティック曲線

この S 字型曲線をロジスティック曲線と呼ぶ。ロジスティック関数は二項分布と深く関係している。例えば、ある病気にかかった場合、その死亡率を p とすると、その生存率は $1-p$ となる。このような、あることが起る確率と起らない確率の比 $\frac{p}{1-p}$ (オッズと呼ぶ) の対数変換をロジット (logit) 変換と呼ぶ。

$$\eta = g(p) = \log\left(\frac{p}{1-p}\right)$$

ロジスティック関数は、オッズのロジット変換の逆関数である。上記の式が表 1 の二項分布のリンク関数と同じであることを注意して欲しい。

ロジスティック関数は経済データ解析に用いるのに都合がよい。例えば、携帯電話やインターネットの普及率を考えた場合、普及率が大きくなり、飽和状態に近づくとその伸び率は小さくなり、普及率が 100% (確率 1) を超えることはあり得ない。このようなデータについて線形回帰分析を行うと、両側に行くほど予測値と実測値との乖離が大きくなる。そこで、このようなデータについてはロジスティック回帰分析が多く用いられている。R では一般化線形モデル関数 `glm` の二項分布を用いてロジスティック回帰分析を行うことができる。

ここで表 2 に示す日本のカラーテレビの普及率の例を用いて説明する。

表 2 カラーテレビの普及率

年度	普及率	年度	普及率	年度	普及率
1966	0.003	1973	0.758	1980	0.982
1967	0.016	1974	0.859	1981	0.985
1968	0.054	1975	0.903	1982	0.989
1969	0.139	1976	0.937	1983	0.988
1970	0.263	1977	0.954	1984	0.992
1971	0.423	1978	0.978		
1972	0.611	1979	0.978		

出处: 回帰分析の基礎、早川毅著、朝倉出版
(経済企画庁調査局、消費者動向調査による)

```
>年度<-c(1966:1984)
>普及率<-c(0.003, 0.016, 0.054, 0.139, 0.263, 0.423, 0.611, 0.758, 0.859, 0.903, 0.937, 0.954, 0.978, 0.978, 0.982, 0.985, 0.989, 0.988, 0.992)
> tv<-glm(普及率~年度, family=binomial)
```

関数 `glm` による当てはめ値は `fitted` で返すことができる。図 2 にカラーテレビの普及率の実測値と関数 `glm` を用いたロジスティック回帰モデルの予測値の折れ線プロットを示す。

```
>plot(年度, 普及率, type="l")
>lines(年度, fitted(tv), lty=2, col="red", lwd=2)
>legend(1975, 0.5, c("実測値", "予測値"), col=1:2, lty=1:2)
```

関数 `predict` でリンク関数空間上の予測値を返すことができる。ただし、次のように引数 `type` を指定すると `fitted` と同じ結果が返される。

```
>predict(tv, type="response")
```

関数 `glm` の要約の出力は関数 `lm` と同じく `summary` を用いる。

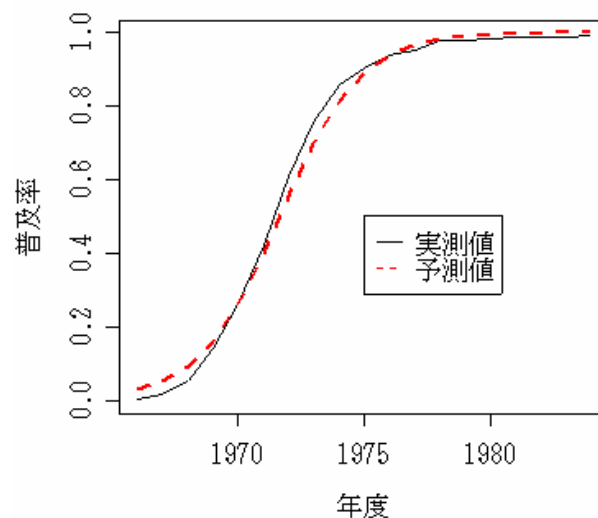


図2 実測値と予測値の折れ線プロット

(2) 2値データのロジスティック回帰分析

関数 `glm` の二項分布を用いて、2値(binary)になっている被説明変数を予測するモデルを構築することも可能である。ここでは、Rの中のデータ `ToothGrowth` を用いて説明する。

データ `ToothGrowth` は各々10匹のギニアピッグ(モルモット)の造歯細胞(歯)の成長について、ビタミンCの投与量(0.5, 1, 2mg)を異なる摂取法(オレンジジュースまたはアスコルビン酸)で、計測を行った60行×3列の実験データである。3変数のラベルを次に示す。

```
[,1] len  歯の長さ
[,2] supp 摂取法 (VC 又は 0J)
[,3] dose 投与量 (0.5, 1, 2mg)
```

通常では、このデータは「歯の長さ」が「摂取法」と「投与量」の影響を受けているかを分析するのに用いられているが、ここでは、「歯の長さ」と「投与量」を説明変数とし、どのような「摂取法」を用いたかを予測する例題の題材とする。データ `ToothGrowth` の中から5行をランダムサンプリングしたデータを次に示す。

```
>data(ToothGrowth)
> samp<-sample(60,5)
> ToothGrowth[samp,]
      len supp dose
50 27.3   OJ  1.0
53 22.4   OJ  2.0
11 16.5   VC  1.0
13 15.2   VC  1.0
32 21.5   OJ  0.5
```

関数 `glm` では、被説明変数が 2 値の場合は、引数 `family` に二項分布 `binomial` を指定する。関数 `glm` による使用例を次に示す。

```
>attach(ToothGrowth)
>Tooth.glm<-glm(supp~len+dose, family=binomial)
```

結果の要約は `summary` で返されるが、ここでは省略する。ここで興味を持っているのは予測値がどのような形式であり、実測値とどのような関係を持っているかである。次に実測値と予測値の対応のサンプルを示す。

```
>実測値<-supp[samp]
>予測値<-fitted(Tooth.glm)
> data.frame(実測値, 予測値[samp])
      実測値  予測値.samp.
50      OJ      0.1015566
53      OJ      0.7212634
11      VC      0.5292075
13      VC      0.5971142
32      OJ      0.1201745
```

関数 `glm` では、2 値のカテゴリカルデータを 1、0 のダミー変数に自動的に置き換えて計算を行う。返された、予測値は確率データであり、確率の値はダミー変数 1 (ここでは VC) に対する予測確率である。よって、得られた予測値の値が小さいとダミー変数 0 (ここでは OJ) に対応するカテゴリを予測したことになる

2 値データであるので、確率値 0.5 を境として、0.5 より大きければダミー変数 1、0.5 より小さければダミー変数 0 であると見なすこともできる。

四捨五入関数 `round` を用いることで、予測値を 0、1 で返すことができる。

```
>予測値_1<-round(予測値)
> data.frame(実測値, 予測値_1[samp])
      実測値  予測値_1.samp.
50      OJ              0
53      OJ              1
11      VC              1
13      VC              1
```

```
32    OJ          0
```

次のように関数 `teable` を用いて実測値と予測値のクロス表を作成することができる。

```
> table(supp,予測値 1)
  予測値 1
supp  0   1
  OJ 17  13
  VC  7  23
```

返された実測値と予測値のクロス表から分かるように、カテゴリ `OJ` の場合は、30 の中の 17 が正しく予測され、`VC` では 30 の中の 23 が正しく予測されている。このような回帰分析応用方法は、一種の 2 群判別分析として解釈することもできる。

関数 `attach(ToothGrowth)` を用いた場合は、`ToothGrowth` の解析が終わったら、次のように関数 `detach` を用いて、検索リストから切り離すことをお薦めする。

```
>detach(ToothGrowth)
```

3. 分散分析と一般化線形モデル

通常の分散分析は線形モデルである。例えば、一元配置の分散分析モデルは次のように被説明変数は平均と誤差の線形式で表し、残差 ε_{ij} は正規分布に従うと仮定している。

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

R では回帰モデルの結果から分散分析の結果を返すことができる。ここでは、異なる 6 種類の農薬を散布し、昆虫への薬剤噴霧の効果調べた農業実験データ `InsectSprays` を用いることにする。

データ `InsectSprays` は 2 変数、72 個の観測値を持つデータフレームである。次にその 2 変数のラベルを示す。

```
[,1]count  昆虫の数
[,2]spray  噴霧剤の種類(A,B,C,D,E,F)
```

```
> data(InsectSprays)
> InsectSprays[1,]
```

```
count  spray
1     10    A
```

6 種類噴霧剤ごとの箱ひげ図を図 3 に示す。図 3 から、噴霧剤の種類によって殺虫効果が明らかに異なることが読み取られる。

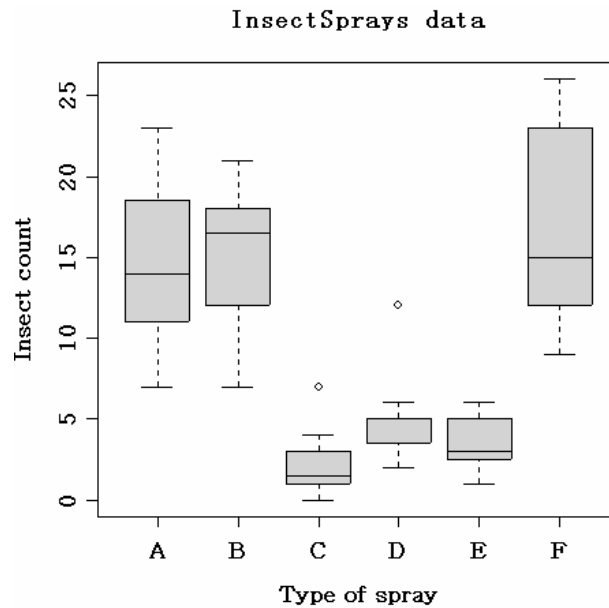


図3 InsectSprays の箱ひげ図

次に関数 `lm,glm` の結果に関数 `aov`、`anova` を用いた分散分析の例を示す。結果から3種類の結果は基本的に同じであることが分かる。

```
>attach(InsectSprays)
>summary(aov(count~spray))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
spray	5	2668.83	533.77	34.702	< 2.2e-16 ***
Residuals	66	1015.17	15.38		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
>anova(lm(count~spray))
```

Analysis of Variance Table

Response: count

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
spray	5	2668.83	533.77	34.702	< 2.2e-16 ***
Residuals	66	1015.17	15.38		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
>anova(glm(count~spray,family=gaussian),test="F")
```

Analysis of Deviance Table

Model: gaussian, link: identity

Response: count

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			71	3684.0		
spray	5	2668.8	66	1015.2	34.702	< 2.2e-16 ***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

次に関数 `glm` の正規分布を用いた場合とポアソン分布を用いた場合の当てはめの良さについて比較してみる。当てはめの良さの判断基準は AIC を用いる。

```
>AIC(glm(count~spray,family=gaussian))
[1] 408.8494
>AIC(glm(count~spray,family=poisson))
[1] 376.5892
```

AIC の値からポアソン分布を用いた場合の当てはめが、正規分布を用いた場合より良いと判断される。次にポアソン分布による分散分析の結果を示す。

```
>anova(glm(count~spray,family=poisson),test="F")
Analysis of Deviance Table
Model: poisson, link: log
Response: count
Terms added sequentially (first to last)
      Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                    71     409.04
spray  5     310.71      66     98.33 62.142 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
>detach(InsectSprays)
```

このデータではいずれの方法でも $\text{Pr}(>F)$ が非常に小さいので、「 $< 2.2e-16$ 」 (2.2×10^{-16} より小さい値) が返されている。しかし、 F 値はポアソン分布の場合は 62.142 で正規分布の 34.702 より大きく、同じではないことが分かる。