

Rとカテゴリカルデータのモデリング(3)

同志社大学文化情報学部教授

金 明哲 (Jin Mingzhe)

■中国生まれ。総合研究大学院大学数物研究科統計科学専攻博士後期課程修了。博士(学術)。1995年札幌学院大学社会情報学部、助教授、教授を経て、2005年4月より現職。E-mail: mjin@mail.doshisha.ac.jp



1. 分割表のモデリング

カテゴリカルデータを分割表にまとめ、そのセルの度数をモデリングする方法として、分割表の対数線形モデル (Log linear model) という方法がある。説明を簡単にするために、まず表1に示す2元分割表を考えよう。

表1 2元分割表の一般形式と周辺度数

	b_1	b_2	...	b_j	...	b_c	計
a_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	n_{1+}
a_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	n_{2+}
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots		
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	n_{i+}
\vdots	\vdots	\vdots	\ddots	\vdots	\ddots		
a_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	n_{r+}
計	n_{+1}	n_{+2}	...	n_{+j}	...	n_{+c}	n_{++}

2元分割表のセルの度数 n_{ij} の期待度数

$E_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$ を対数変換すると、

$$\log(E_{ij}) = -\log(n_{++}) + \log(n_{i+}) + \log(n_{+j})$$

となる。式の中の項目 $-\log(n_{++})$ は総度数の

効果、 $\log(n_{i+})$ は第 i 行の周辺度数の効果、 $\log(n_{+j})$ は第 j 列の周辺度数の効果である。分割表が独立である場合は、この式で観測データを表現することができる。しかし、観測データが独立ではない場合は、行と列の相互効果を加えて表現することが必要である。グッドマン (Goodman) は1970年代に、次に示すような式で2元分割表の観測度数をモデリングする方法を提案した。

$$\log(n_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

ただし、 $\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0$ を条件とする。式の中の λ は総度数の効果、 λ_i^A は第 i 行の周辺度数の効果、 λ_j^B は第 j 列の周辺度数の効果、 λ_{ij}^{AB} は第 i 行と第 j 列の交互作用効果を表すパラメータである。このモデルを2元分割表の飽和モデル (Saturated model) と呼ぶ。飽和モデルのパラメータの数は分割表のセルの数に等しい。交互作用効果のパラメータを持たないモデルを独立モデルと呼ぶ。

モデルの中のパラメータは観測データの対数値 $\log(n_{ij}) = v_{ij}$ を用いて、次のように求めることができる。

$$\lambda = \bar{v} = \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c v_{ij}$$

$$\lambda_i^A = \bar{v}_{i+} - \bar{v}$$

$$\lambda_j^B = \bar{v}_{\cdot j} - \bar{v}$$

$$\lambda_{ij}^{AB} = v_{ij} - \bar{v}_{i+} - \bar{v}_{\cdot j} + \bar{v}$$

$$\bar{v}_{\cdot j} = \frac{1}{r} \sum_{i=1}^r v_{ij}, \quad \bar{v}_{i+} = \frac{1}{c} \sum_{j=1}^c v_{ij}$$

パラメータを求めるプロセスを、表2に示すデータを用いて説明する。表2の対数変換の結果を表3に示す。

表2 喫煙と肺癌の分割表

	肺癌あり	肺癌なし
喫煙者	182	156
非喫煙者	72	98

表3 表2の対数変換の分割表

	肺癌あり	肺癌なし	平均
喫煙者	5.204	5.050	5.127
非喫煙者	4.277	4.585	4.431
平均	4.740	4.817	4.779

2×2の分割表におけるモデルの制約条件は、

$$\begin{cases} \lambda_1^A = -\lambda_2^A \\ \lambda_1^B = -\lambda_2^B \\ \lambda_{11}^{AB} = \lambda_{22}^{AB} = -\lambda_{12}^{AB} = -\lambda_{21}^{AB} \end{cases}$$

である。表1のセル $n_{11} = 182$ を表現するモデルのパラメータは、

$$\lambda = 4.779$$

$$\lambda_1^A = \bar{v}_{i+} - \bar{v} = 5.127 - 4.779 = 0.348$$

$$\lambda_1^B = \bar{v}_{\cdot j} - \bar{v} = 4.740 - 4.779 = -0.039$$

$$\lambda_{11}^{AB} = 5.204 - 5.127 - 4.740 + 4.779 = 0.116$$

となり、モデルによるセル n_{11} の推測値 \hat{n}_{11} は、

$$\log(\hat{n}_{11}) = 4.779 + 0.348 - 0.039 + 0.116 = 5.204$$

$$(\hat{n}_{11} = e^{5.204} \cong 182)$$

となる。同様のモデルに基づいて求めた結果を表4に示す。

表4 表2の対数線形モデルの推測値

	肺癌あり	肺癌なし
喫煙者	182	156
非喫煙者	72	98

このモデルによる推測値は、観測値と完全に一致する。

2. 関数loglinによるモデリング

Rには対数線形モデルのパラメータの推測や検定統計量を求める関数loglinがある。関数loglinの書式を次に示す。

loglin(table, margin, fit = FALSE, param = FALSE, print = TRUE...)

引数tableにはデータの分割表を、引数marginではリスト形式で分割表の範囲を、引数fitでは分割表の推測値を返すか否かを、引数paramでは推測したパラメータを返すか否かを論理値 (FALSE, TRUE) で指定する。これ以外にも不完備表 (Incomplete tables、表の中に論理的にゼロとなるセルを含む分割表) に対応する引数startがある。

表2のデータを用いた例を次に示す。まず次のように分割表を作成する。

```
>tab1<-matrix(c(182,72,156,98),nc=2)
>rownames(tab1)<-c("喫煙者","非喫煙者")
>colnames(tab1)<-c("肺癌あり","肺癌なし")
>(tab1<-as.table(tab1))
      肺癌あり 肺癌なし
喫煙者      182      156
非喫煙者      72      98
```

データtab1を用いた関数loglinの使用例を次に示す。コマンドの中のlist(c(1,2))は2元分割表の表側と表頭を表す。

```
>log.m1<-loglin(tab1,margin=list(c(1,2)),
param=T,fit=T)
```

返す項目は関数summaryを用いて確認できる。

```
>summary(log.m1)
```

	Length	Class	Mode
lrt	1	-none-	numeric
pearson	1	-none-	numeric
df	1	-none-	numeric
margin	2	-none-	list
fit	4	table	numeric
param	4	-none-	list

\$lrtには尤度比の検定統計量、\$pearsonにはピアソンのカイ2乗検定統計量、\$dfにはモデルの当てはめの自由度、\$marginにはモデルの当てはめのために指定した値が記録されている。\$fitはモデルによって当てはめた結果、\$paramには推測されたモデルのパラメータの値が記録されている。

```
>log.m1$fit
```

	肺癌あり	肺癌なし
喫煙者	182	156
非喫煙者	72	98

返された結果から分かるように、推測値は用いた観測値と全く同じである。

```
>log.m1$param
```

```
$(Intercept)`
[1] 4.778874
$`1`
  喫煙者 非喫煙者
0.3480573 -0.3480573
$`2`
  肺癌あり 肺癌なし
-0.03853767 0.03853767
$`1.2`
  喫煙者 肺癌あり 肺癌なし
非喫煙者 -0.115613 -0.115613
0.115613 0.115613
```

パラメータ λ , λ_i^A , λ_j^B , λ_{ij}^{AB} の推測値は上から順番にそれぞれ項目 '\$(Intercept)`, '\$1`, '\$2`, '\$1.2' に返されている。

関数loglinは多元分割表のデータを扱うことも可能である。パッケージdatasetの中に髪の色（4種類）と目の色（4種類）を性別に分けた3元分割表データHairEyeColorがある。

```
>class(HairEyeColor)
```

```
[1] "table"
>dim(HairEyeColor)
[1] 4 4 2
>HairEyeColor
```

```
, , Sex = Male
```

Hair	Eye			
	Brown	Blue	Hazel	Green
Black	32	11	10	3
Brown	38	50	25	15
Red	10	10	7	7
Blond	3	30	5	8

```
, , Sex = Female
```

Hair	Eye			
	Brown	Blue	Hazel	Green
Black	36	9	5	2
Brown	81	34	29	14
Red	16	7	7	7
Blond	4	64	5	8

この3元分割表のカテゴリカル変数である髪の色、目の色、性別をそれぞれ $A = \text{Hair}$, $B = \text{Eye}$, $C = \text{Sex}$ にした飽和モデルの一般式を次に示す。

$$\log(n_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

ただし、パラメータの制約条件は、各カテゴリカル変数のパラメータの合計がゼロであることである。

$$\sum_i \lambda_i^A = \dots = \sum_k \lambda_k^C = \sum_i \lambda_{ij}^{AB} = \dots = \sum_k \lambda_{jk}^{BC} = \sum_k \lambda_{ijk}^{ABC} = 0$$

データHairEyeColorの3つのカテゴリカル変数の飽和モデルの推測例を次に示す。コマンド中の引数list(c(1,2), c(1,3), c(2,3))は、3つのカテゴリカル変数を組み合わせるモデルの書式である。

```
>fm <- loglin(HairEyeColor,list(c(1,2),
c(1,3),c(2,3)),fit=T,param=T)
```

比較のために、モデルにより推測された分割表の値を整数に丸めて次に示す。

```
>round(fm$fit)
```

```
, , Sex = Male
```

```
      Eye
Hair  Brown Blue Hazel Green
Black 31 13 9 3
Brown 41 44 27 16
Red 10 9 7 8
Blond 2 35 3 6
```

```
, , Sex = Female
```

```
      Eye
Hair  Brown Blue Hazel Green
Black 37 7 6 2
Brown 78 40 27 13
Red 16 8 7 6
Blond 5 59 7 10
```

推測値の残差を次に示す。

```
>HairEyeColor-round(fm$fit)
```

```
, , Sex = Male
```

```
      Eye
Hair  Brown Blue Hazel Green
Black 1 -2 1 0
Brown -3 6 -2 -1
Red 0 1 0 -1
Blond 1 -5 2 2
```

```
, , Sex = Female
```

```
      Eye
Hair  Brown Blue Hazel Green
Black -1 2 -1 0
Brown 3 -6 2 1
Red 0 -1 0 1
Blond -1 5 -2 -2
```

3. 関数loglmによるモデリング

パッケージMASSには分割表を対数線形モデルで当てはめる関数loglmがある。関数loglmの書式は、線形回帰関数lmや一般線形回帰関数glmと似ている。

関数loglmでは、分割表(表2)のデータを表5のような度数分布表の形式にして用いる。

表5 表2の度数分布表

喫煙歴	肺癌歴	度数
あり	あり	182
あり	なし	156
なし	あり	72
なし	なし	98

```
>tab2<-data.frame(
喫煙=c(rep("あり",2),rep("なし",2)),
肺癌=c(rep(c("あり","なし"),2)),
度数=c(182,156,72,98))
```

```
>tab2
```

```
  喫煙  肺癌  度数
1 あり あり 182
2 あり なし 156
3 なし あり 72
4 なし なし 98
```

作成したデータセットtab2を飽和モデルに当てはめる例を次に示す。

```
>library(MASS)
```

```
>tab.m2<-loglm(度数~喫煙+肺癌+喫煙:肺癌,data=tab2)
```

```
>tab.m2
```

```
Call:
```

```
loglm(formula = 度数 ~ 喫煙 + 肺癌 + 喫煙:肺癌, data = tab2)
```

```
Statistics:
```

```
                X^2 df P(> X^2)
Likelihood Ratio  0  0 1
Pearson           0  0 1
```

このように関数loglmは、関数loglinと同じく尤度比とピアソンカイ2乗統計量を返す。次のように関数fittedとresidualsで当てはめ値と残差を返すことができる。

```
>fitted(tab.m2)
```

```
Re-fitting to get fitted values
```

```
      肺癌
喫煙 あり  なし
あり  182  156
なし  72   98
```

```
>residuals(tab.m2)
```

```
Re-fitting to get frequencies and fitted values
```

```
      肺癌
喫煙 あり  なし
あり  0   0
なし  0   0
```

この結果は、関数loglinを用いた場合と同じであることが確認できる。

Rでは関数updateを用いて、作成したモデルの更新を行うことができる。モデルtab.m2の交互作用のパラメータを除いたモデルの書式を例として次に示す。コマンドの中の「~」の右に追加、あるいは除くパラメータを指定する。追加する場合は+、除く場合は-を付ける。

```
>(tab.m3<-update(tab.m2,`~喫煙:肺癌`))
Call:
loglm(formula = 度数 ~ 喫煙 + 肺癌, data = tab2)
```

```
Statistics:
                X^2 df      P(> X^2)
Likelihood Ratio 5.994097 1 0.01435383
Pearson          5.976471 1 0.01449799
```

```
> fitted(tab.m3)
Re-fitting to get fitted values
```

喫煙	肺癌あり	肺癌なし
あり	169	169
なし	85	85

```
>residuals(tab.m3)
Re-fitting to get frequencies and fitted values
```

喫煙	肺癌あり	肺癌なし
あり	0.9875737	-1.013250
なし	-1.4484958	1.376219

4. モデルの選択

分割表（あるいは度数分布表）のモデリングを行う際、飽和モデルを用いると当てはめはよいが、用いるパラメータの数が多いため問題である。データをモデリングする際に重要なのは、少ないパラメータで簡潔なモデルを構築することである。

同一の分割表の対数線形モデルは唯一ではない。例えば、2元分割表のモデルとしては、次に示すようなモデルが考えられる。

$$\log(n_{ij}) = \lambda$$

$$\log(n_{ij}) = \lambda + \lambda_i^A$$

$$\log(n_{ij}) = \lambda + \lambda_j^B$$

$$\log(n_{ij}) = \lambda + \lambda_i^A + \lambda_j^B$$

$$\log(n_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

⋮

これらのモデルの中で、どのモデルを用いるべきであるかが1つの問題となる。

通常、モデルの評価にはAICやBICなどの情報量規準がよく用いられている。これらは、返された尤度比とカイ2乗統計量を用いて簡単に計算することができる。関数loglmが返

す尤度比 (G^2 : Likelihood Ratio) を次の式の「 -2 (対数尤度)」に代入するとAICが計算できる。

$$AIC = -2(\text{対数尤度}) + 2(\text{パラメータの数})$$

2×2の分割表の飽和モデルのパラメータの数は4 (松田, p.86) であるため、表2 (あるいは表5) の飽和モデルのAICは $0 + 2 \times 4 = 8$ になり、交互作用効果を除いたモデル (独立モデル) のAICは $5.994 + 2 \times 3 = 11.994$ になる。

AICを求める式から分かるように、AICは分割表の総度数の影響を無視している。しかし、分割表のカイ2乗値および尤度比は総度数の影響を受ける。各セルの比率は全く同じであっても総度数の増加に伴い分割表のカイ2乗値、尤度比が増大する。例えば、tab2の総度数は508であり、独立モデルの尤度比は5.994097である。各セルの比率が同じであり、総度数が2,000の場合の独立モデルの尤度比は23.8992になる。

```
>tab3<-tab2
>tab3[,3]<-round(2000*tab3[,3]/sum(tab3[,3]),0)
>loglm(度数~喫煙+肺癌,data=tab3)
```

```
Call:
loglm(formula = 度数 ~ 喫煙 + 肺癌, data = tab3)
```

```
Statistics:
                X^2 df      P(> X^2)
Likelihood Ratio 23.89922 1 1.014759e-06
Pearson          23.82270 1 1.053009e-06
```

分割表の総度数の影響を考慮して、モデルを評価する情報量規準としてBIC (Bayesian Information Criterion) がある。分割表のBICは次の式で定義されている。

$$BIC = -2(\text{対数尤度}) + \log(n_{++}) \times (\text{パラメータの数})$$

データtab2の飽和モデルのBICは $0 + \log(508) \times 4 = 24.9219$ 、独立モデルのBICは $5.994097 + \log(508) \times 3 = 24.6855$ となる。

情報量規準を用いてモデルの選択を行う場合には、値が小さいモデルを選択する。上記のデータに対してどのモデルを用いるかに関しては、用いる情報量規準によって結果が異なる。このデータにおいては、AICを用いると飽和モデルが選択され、BICを用いると独立モデルが選択される。

対数線形モデルのAICとBICは関数 **extractAIC** で求めることができる。引数 $k=2$ のときはAIC、 $k=\log(n_{++})$ のときはBICを返す。デフォルトは $k=2$ になっている。モデル `tab.m2` を用いた例を次に示す。

```
>extractAIC(tab.m2)
[1] 4.8 #AIC
>extractAIC(tab.m2,k=log(sum(tab2[,3])))
[1] 4.00000 24.92193 #BIC
```

5. 多元度数表と対数線形モデル

パッケージ `MASS` の中に、住宅環境に対する満足度に関する調査結果を4元度数表にまとめたデータセット `housing` がある。

```
>data(housing)
>head(housing,n=3)
  Sat Infl Type Cont Freq
1 Low Low Tower Low 21
2 Medium Low Tower Low 21
3 High Low Tower Low 28
```

変数 `Sat` は現在の住宅状況に対する満足度 (高、中、低)、変数 `Infl` は不動産のマネジメントにおける影響力の認知度 (高、中、低)、変数 `Type` は賃貸住宅のタイプ (高層ビル、中庭付き、アパート、テラス)、変数 `Cont` は他の住民とのふれあいの度合い (高、低)、変数 `Freq` は度数である。変数のすべての組み合わせは $3 \times 3 \times 4 \times 2 = 72$ 通りであるので、データセットは72行5列である。

データ `housing` の分割表形式は次のコマンドで確認できる。

```
>xtabs(Freq~Sat+Infl+Type+Cont, data=
housing)
<結果は省略>
```

関数 `loglm` を用いてモデルを作成する例を次に示す。

```
>hous.log<-loglm(Freq~Sat+Infl+Type+
Cont,data=housing)
>hous.log
loglm(formula = Freq ~ Sat + Infl + Type +
Cont, data = housing, fitted = T, param = T)
```

```
Statistics:
                X^2      df  P(> X^2)
Likelihood Ratio 295.3518  63         0
Pearson          305.9267  63         0
>hous.log$param
<省略>
```

関数 `glm` では、リンク関数をポアソン (`poisson`) に指定することにより対数線形モデルに当てはめることができる。その例を次に示す。

```
>hous.glm<-glm(Freq~Sat+Infl+Type+Cont,
data=housing,family=poisson)
```

計算された尤離度は関数 `loglm` の尤度比に等しい。

```
>deviance(hous.glm)
[1] 295.3518
```

関数 `summary` は回帰係数と標準誤差などを返す。

```
>summary(hous.glm)
<結果は省略>
```

関数 `dumy.coef` は、すべてのパラメータを返す。

```
>dumy.coef(hous.glm)
<結果は省略>
```

*参考文献

松田紀之(1988): 質的情報の多変量解析: 朝倉出版 (<http://www.sci.kagoshima-u.ac.jp/~ebsa/matsuda01/index.html>).