クラスター分析

データの標準化（測定単位に意味がなければデータを標準化しておくことがすすめられる）
(新)データフレーム名 <- scale(データフレーム名)
　特に指定しなければ，平均=0，標準偏差=1に標準化する．

距離行列の作成
距離行列名 <- dist((新)データフレーム名, method="距離定義名")
　method="euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski"

階層的クラスター分析
出力名 <- hclust(距離行列名^2, method="方法名")
　method = "ward", "centroid" の場合

出力名 <- hclust(距離行列名, method="方法名")
　method = "single", "complete", "average", "mcquitty", "median" の場合

デンドログラムの表示
plot(出力名)

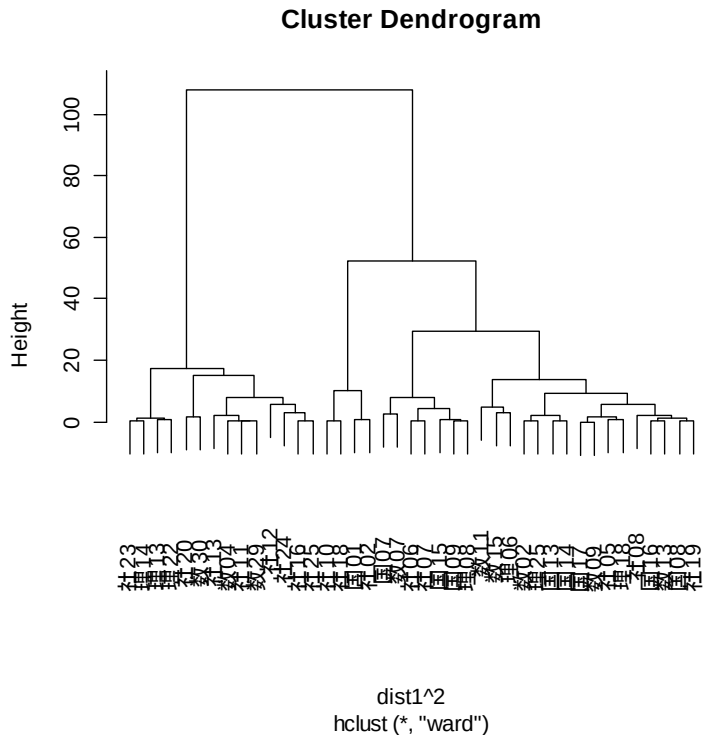
あるクラスタ数におけるクラスタリング状況の出力
cutree(出力名)

非階層的クラスター分析
kmeans(データフレーム名,クラスター数)

	A	B	C	D	E
1	item	p.na	b.na	p.correct	b.correct
2	国01	0.028	0.395	0.462	0.337
3	国07	0.019	0.377	0.944	0.251
4	国08	0.143	0.479	0.509	0.305
5	国09	0.09	0.525	0.767	0.301
6	国13	0.206	0.585	0.717	0.275
7	国14	0.237	0.565	0.846	0.283
8	国15	0.062	0.534	0.805	0.256
9	国16	0.135	0.553	0.544	0.377
10	国17	0.147	0.428	0.701	0.293
11	社02	0.054	0.303	0.454	0.344
12	社05	0.241	0.49	0.658	0.292
13	社06	0.067	0.469	0.554	0.252
14	社07	0.045	0.478	0.611	0.21
15	社08	0.081	0.498	0.384	0.299
16	社10	0.028	0.143	0.475	0.436
17	社11	0.32	0.588	0.469	0.293
18	社12	0.393	0.576	0.549	0.103
19	社13	0.326	0.596	0.632	0.215
20	社16	0.271	0.582	0.416	0.221
21	社18	0.022	0.151	0.49	0.481

```
> setwd("i:¥¥Rdocuments¥¥scripts¥¥")
> d1 <- read.table("クラスター分析_データ.csv", header=TRUE, sep=",")
> head(d1)
                item  p.na  b.na p.correct b.correct
1 国01              0.028 0.395     0.462     0.337
2 国07              0.019 0.377     0.944     0.251
3 国08              0.143 0.479     0.509     0.305
4 国09              0.090 0.525     0.767     0.301
5 国13              0.206 0.585     0.717     0.275
6 国14              0.237 0.565     0.846     0.283
>

> # データの標準化
> d2 <- d1[,c("p.na","b.na","p.correct","b.correct")]
> d2 <- scale(d2)
>

> # 距離行列の作成
> rownames(d2) <-d1$item
> dist1 <- dist(d2, method="euclidean")
>
```

```
> # 階層的クラスター分析
> hclust1 <- hclust(dist1^2, method="ward")
> plot(hclust1)
>

> #あるクラスター数における分類状況
> cutree(hclust1, k=8)

国01 国07 国08 国09
   1    2    3    2
国13 国14 国15 国16
   3    3    2    3
国17 社02 社05 社06
   3    1    3    2
社07 社08 社10 社11
   2    3    4    5
社12 社13 社16 社18
   5    5    5    4
社19 社20 社23 社24
   3    6    7    5
社25 数02 数04 数07
   5    3    5    2
数09 数11 数13 数15
   3    8    3    8
数29 数30 理06 理08
   5    6    8    2
理13 理14 理18 理22
   7    7    3    7
理25
   3
>
```



```
> # 非階層的クラスター分析
> kmeans1 <- kmeans(d2,8)
> kmeans1

K-means clustering with 8 clusters of sizes 9, 3, 7, 6, 3, 5, 6, 2

Cluster means:
        p.na        b.na  p.correct   b.correct
1 -0.9457740 -0.39880299  0.9944966 -0.53699218
2 -1.2824776 -2.71477252 -0.3252307  1.31723347
3 -0.5552738 -0.05554498 -0.2603222  0.48178486
4  0.8706532  0.82874967 -0.4454055 -1.26543797
5  0.5436963 -0.21624893  0.4929454  1.46768035
6  0.2512959  0.52501920  1.0062397  0.09822798
7  0.5343927  0.71921724 -1.0429868  0.22745799
8  2.4642359  0.42910431 -1.8661015 -0.57878298

Clustering vector:
国01 国07 国08 国09
   3    1    3    1
国13 国14 国15 国16
   6    6    1    3
国17 社02 社05 社06
   1    2    6    1
社07 社08 社10 社11
   1    3    2    7
社12 社13 社16 社18
   4    4    4    2
社19 社20 社23 社24
   3    8    7    4
社25 数02 数04 数07
   4    6    4    1
数09 数11 数13 数15
   1    5    3    5
数29 数30 理06 理08
   7    8    3    1
理13 理14 理18 理22
   7    7    5    7
理25
   6

Within cluster sum of squares by cluster:
[1] 9.6458967 2.6410792 5.3717044 7.4818942 3.4843068 2.3167858 4.0996303
[8] 0.8415862

Available components:
[1] "cluster"  "centers"  "withinss" "size"
```

階層的な方法と非階層的な方法で計算方法が異なるので，
同じクラス多数でも，必ずしも分類状況は一致しない

主成分分析

```
library(psych)
オブジェクト名 <- principal(データ名, nfactors=因子数, rotate="回転方法")
print(オブジェクト名, sort=TRUE)
```

スクリープロット

```
VSS.scree(オブジェクト名)
```

あらかじめpsychパッケージをインストールしておく必要がある.

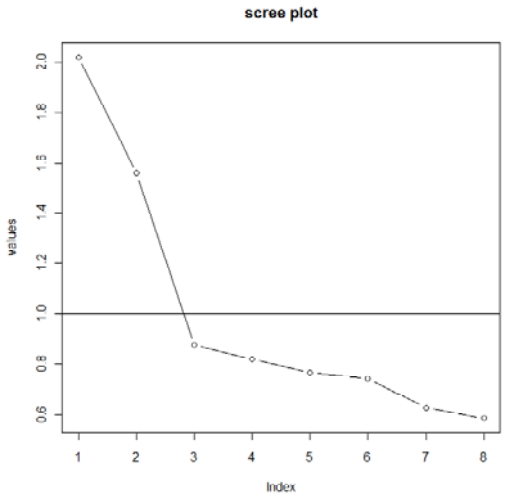
データには，素データ，相関係数行列，共分散行列のいずれかを指定する.

回転方法（rotate）
直交回転："none", "varimax", "quartimax"
斜交回転："promax", "oblimin", "simplimax", "cluster"

```
> setwd("i:¥¥Rdocuments¥¥scripts¥¥")
> d1 <- read.table("因子分析_データ.csv", header=TRUE, sep=",")
> head(d1)

  x1 x2 x3 x4 x5 x6 x7 x8
1  3  2  4  4  5  2  3  4
2  3  2  3  3  3  2  3  2
3  1  3  1  3  2  2  2  2
4  3  4  1  3  4  3  4  5
5  1  3  3  3  4  4  3  4
6  2  1  3  4  2  2  4  3
>
>
> # 記述統計量
> dtmp <- d1
> ntmp <- nrow(dtmp)
> mtmp <- colMeans(dtmp)
> stmp <- apply(dtmp, 2, sd)
> ctmp <- cor(dtmp)
> ktmp <- round(data.frame(ntmp, mtmp, stmp, ctmp),2)
> colnames(ktmp) <- c("N","Mean","SD",colnames(ctmp))
> ktmp
     N Mean   SD   x1   x2   x3   x4   x5   x6   x7   x8
x1 346 2.97 1.01 1.00 0.30 0.38 0.27 0.09 0.06 0.08 0.09
x2 346 3.00 1.02 0.30 1.00 0.24 0.17 0.04 0.06 0.01 0.06
x3 346 3.03 1.02 0.38 0.24 1.00 0.19 0.08 0.05 0.03 0.06
x4 346 3.02 1.00 0.27 0.17 0.19 1.00 0.07 0.02 0.05 0.03
x5 346 3.00 1.03 0.09 0.04 0.08 0.07 1.00 0.24 0.32 0.36
x6 346 3.04 1.03 0.06 0.06 0.05 0.02 0.24 1.00 0.19 0.20
x7 346 3.01 1.02 0.08 0.01 0.03 0.05 0.32 0.19 1.00 0.25
x8 346 3.01 1.02 0.09 0.06 0.06 0.03 0.36 0.20 0.25 1.00
>
>
```

	A	B	C	D	E	F	G	H
1	x1	x2	x3	x4	x5	x6	x7	x8
2	3	2	4	4	5	2	3	4
3	3	2	3	3	3	2	3	2
4	1	3	1	3	2	2	2	2
5	3	4	1	3	4	3	4	5
6	1	3	3	3	4	4	3	4
7	2	1	3	4	2	2	4	3
8	2	3	2	3	1	3	2	3
9	5	3	4	5	3	4	2	3
10	3	3	3	4	4	3	3	2
11	4	4	4	3	3	3	3	4
12	4	4	3	2	1	2	3	3
13	3	3	2	3	1	3	1	3
14	2	3	2	3	3	2	2	1
15	3	3	2	5	4	5	4	4
16	3	3	4	2	3	3	3	5
17	4	3	2	4	4	4	3	1
18	3	3	4	4	5	3	5	4
19	3	3	4	3	3	4	1	5
20	5	4	5	3	2	3	3	3
21	3	4	3	3	4	4	4	4

```
>  # psychパッケージの読み込み
>  library(psych)
>  # スクリープロット
>  VSS.scree(d1)

> #主成分分析の実行
> prin.1 <- principal(d1, nfactors=2, rotate="promax")
> print(prin.1, sort=TRUE)

Principal Components Analysis
Call: principal(r = d1, nfactors = 2, rotate = "promax")
Standardized loadings based upon correlation matrix
     PC2   PC1   h2   u2
x1  0.76  0.03 0.59 0.41
x3  0.70  0.00 0.49 0.51
```



```
x2  0.63 -0.02 0.40 0.60
x4  0.56 -0.01 0.31 0.69
x5  0.02  0.75 0.56 0.44
x8  0.01  0.69 0.48 0.52
x7 -0.03  0.66 0.43 0.57
x6  0.02  0.56 0.31 0.69

                 PC2  PC1
SS loadings    1.79 1.79
Proportion Var 0.22 0.22
Cumulative Var 0.22 0.45

 With factor correlations of
     PC2  PC1
PC2 1.00 0.12
PC1 0.12 1.00



Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  28  and the objective function was  0.74
The degrees of freedom for the model are 13  and the objective function was  0.34
The number of observations was  346  with Chi Square =  116.84  with prob <  8.5e-19

Fit based upon off diagonal values = 0.56>
```

```
> # 因子分析の結果との比較
```

```
> # 因子分析の実行
>  # psychパッケージの読み込み
>  library(psych)
> fac.1 <- fa(d1, nfactors=2, rotate="promax", smc=TRUE, fm="wls")
> print(fac.1, sort=TRUE)

Factor Analysis using method =  wls
Call: fa(r = d1, nfactors = 2, rotate = "promax", fm = "wls", smc = TRUE)
Standardized loadings based upon correlation matrix

    WLS1  WLS2   h2   u2
x1  0.70  0.00 0.50 0.50
x3  0.54 -0.01 0.29 0.81
x2  0.44 -0.01 0.19 0.71
x4  0.37  0.00 0.13 0.87
x5  0.00  0.66 0.44 0.56
x8  0.01  0.53 0.29 0.86
x7 -0.01  0.48 0.23 0.77
x6  0.02  0.37 0.14 0.71

# WLS1,WLS2,h2はソート順
# u2は項目順になっている


                WLS1 WLS2
SS loadings    1.12 1.09
Proportion Var 0.14 0.14
Cumulative Var 0.14 0.28


 With factor correlations of
     WLS1 WLS2
WLS1 1.00 0.21
WLS2 0.21 1.00


Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  28  and the objective function was  0.74 with Chi
Square of  253.75
The degrees of freedom for the model are 13  and the objective function was  0.01
```

```
The root mean square of the residuals is  0.01
The df corrected root mean square of the residuals is  0.02
The number of observations was  346  with Chi Square =  2.26  with prob <  1

Tucker Lewis Index of factoring reliability =  1.103
RMSEA index =  0  and the 90 % confidence intervals are  0 0.018
BIC =  -73.75
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                WLS1 WLS2
Correlation of scores with factors              0.80 0.79
Multiple R square of scores with factors        0.65 0.63
Minimum correlation of possible factor scores  0.29 0.25
>
```

主成分回帰分析・PLS回帰分析

重回帰分析において，説明変数の数が多いときに，無相関ないくつかの主成分（またはそのようなもの）に説明変数を集約して重回帰分析を行ったとしたときの，もとの説明変数の影響度を推定する方法．

説明変数が非常に多きときに，重要な説明変数を見つけるのに適した分析法．

説明変数を集約するとき，説明変数の主成分を用いる場合を「主成分回帰分析」，基準変数との共分散を最大にするような成分を合成する場合を「PLS回帰分析」と言う．

主成分回帰分析が，説明変数内だけで，より少数の変数に情報をまとめるのに対し，PLS回帰分析では，集約された変数は基準変数と関連が強いのが望ましいという制約を課している点が異なる．

主成分回帰分析は，主成分得点がどう基準変数に影響するかに関心がある「主成分得点を用いた回帰分析」とは，解釈するものが異なる．

主成分回帰分析

```
library(pls)
オブジェクト名 <- pcr(基準変数名 ~ 変数1，変数2，…，data=データ名)
summary(オブジェクト名)
オブジェクト名$coefficient
```

あらかじめplsパッケージをインストールしておく必要がある．
主成分数の上限を設定するときは，ncomp=最大主成分数 をオプションに指定する．

PLS回帰分析

```
library(pls)
オブジェクト名 <- plsr(基準変数名 ~ 変数1，変数2，…，data=データ名)
summary(オブジェクト名)
オブジェクト名$coefficient
```

あらかじめplsパッケージをインストールしておく必要がある．
主成分数の上限を設定するときは，ncomp=最大主成分数 をオプションに指定する．

```
> setwd("i:¥¥Rdocuments¥¥scripts¥¥")
> d1 <- read.table("回帰分析データ.csv", header=TRUE, sep=",")
> head(d1)
  id stress kyoufu support utsu work result
1  1     20    2.2      17   18    0      0
2  2     23    4.8      18   21    1      0
3  3     30    5.8      12   29    1      1
4  4     25    5.2      18   29    0      1
5  5     26    2.0       8   22    1      0
6  6     21    5.0      26   19    1      0
>
>
> # 記述統計量
> dtmp <- d1[,c(-1)]
> ntmp <- nrow(dtmp)
> mtmp <- colMeans(dtmp)
> stmp <- apply(dtmp, 2, sd)
> ctmp <- cor(dtmp)
> ktmp <- round(data.frame(ntmp, mtmp, stmp, ctmp),2)
> colnames(ktmp) <- c("N","Mean","SD",colnames(ctmp))
> ktmp
          N  Mean   SD stress kyoufu support  utsu  work result
stress  245 22.94 5.25   1.00   0.40   -0.34  0.62  0.03   0.44
kyoufu  245  4.05 1.17   0.40   1.00   -0.03  0.31  0.10   0.20
support 245 18.42 4.96  -0.34  -0.03    1.00 -0.51 -0.03  -0.39
utsu    245 20.29 6.49   0.62   0.31   -0.51  1.00  0.02   0.76
work    245  0.50 0.50   0.03   0.10   -0.03  0.02  1.00  -0.08
result  245  0.26 0.44   0.44   0.20   -0.39  0.76 -0.08   1.00
>
```

1	id	stress	kyoufu	support	utsu	work	result
2	1	20	2.2	17	18	0	0
3	2	23	4.8	18	21	1	0
4	3	30	5.8	12	29	1	1
5	4	25	5.2	18	29	0	1
6	5	26	2	8	22	1	0
7	6	21	5	26	19	1	0
8	7	14	2.2	24	12	0	0
9	8	22	4.4	17	19	1	0
10	9	26	4.2	11	27	0	1
11	10	26	4.2	18	18	1	0
12	11	21	2	27	18	0	0
13	12	24	4.8	19	29	1	1
14	13	24	3.4	23	19	1	0
15	14	23	2.2	10	24	1	0
16	15	23	4.4	24	20	0	0
17	16	35	7.2	12	31	0	1
18	17	25	3.2	17	19	1	0
19	18	33	3.4	14	26	0	1
20	19	30	4.4	20	30	1	1
21	20	19	5.6	18	12	1	0

```
>
>
> #標準偏回帰係数の推定
> d2 <- as.data.frame(scale(d1[,c("stress","kyoufu","utsu","work")]))
> result.2 <- lm(utsu ~ stress + kyoufu , data=d2)
> summary(result.2)

Call:
lm(formula = utsu ~ stress + kyoufu, data = d2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74165 -0.49430 -0.00665  0.47948  2.46771

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.235e-16  5.023e-02   0.000    1.000
stress      5.880e-01  5.501e-02  10.690   <2e-16 ***
kyoufu      7.502e-02  5.501e-02   1.364    0.174
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7861 on 242 degrees of freedom
Multiple R-squared:  0.387,     Adjusted R-squared:  0.382
F-statistic:  76.4 on 2 and 242 DF,  p-value: < 2.2e-16

> confint(result.2)
                  2.5 %     97.5 %
(Intercept) -0.09893404 0.09893404
stress       0.47968584 0.69640299
kyoufu      -0.03334233 0.18337482
>
>




> # 主成分回帰分析
> library(pls)
> res3 <- pcr(utsu ~ stress + kyoufu + support, data=d2)
> summary(res3)
Data:   X dimension: 245 3
        Y dimension: 245 1
Fit method: svdpc
Number of components considered: 3
TRAINING: % variance explained
      1 comps  2 comps  3 comps      # その主成分までを用いて何%の分散が説明されるか
X       51.32    83.78   100.00
utsu    45.50    48.92    49.43


> res3$coefficient
, , 1 comps

                utsu          # 第１主成分のみを採用したときの，もとの変数にかかる係数
stress    0.3795492
kyoufu    0.2963544
support -0.2522451

, , 2 comps

                utsu          # 第２主成分まで採用したときの，もとの変数にかかる係数
stress    0.3778659
kyoufu    0.1761000
support -0.3960610

, , 3 comps

                utsu          # 第３主成分まで採用したときの，もとの変数にかかる係数
stress    0.4509472           # もともと３変数なので，重回帰分析の結果と一致する
kyoufu    0.1209941
support -0.3508386
```

```
>
>
> # PLS 回帰分析
> library(pls)
> res4 <- plsr(utsu ~ stress + kyoufu + support, data=d2)
> summary(res4)
Data:   X dimension: 245 3
        Y dimension: 245 1
Fit method: kernelpls
Number of components considered: 3
TRAINING: % variance explained
      1 comps  2 comps  3 comps          # その成分までを用いて何%の分散が説明されるか
X       50.74    81.91   100.00
utsu    48.75    49.38    49.43
> res4$coefficient
, , 1 comps

                utsu            # 第１成分のみを採用したときの，もとの変数にかかる係数
stress    0.4093243
kyoufu    0.2068089
support -0.3352484

, , 2 comps

                utsu            # 第２成分まで採用したときの，もとの変数にかかる係数
stress    0.4283727
kyoufu    0.1324056
support -0.3703227

, , 3 comps

                utsu            # 第３成分まで採用したときの，もとの変数にかかる係数
stress    0.4509472             # もともと３変数なので，重回帰分析の結果と一致する
kyoufu    0.1209941
support -0.3508386

>
>
>
```

ロジスティック回帰分析 ― 素データの場合

オブジェクト名 <- glm(基準変数 ~ 説明変数, family=binomial, データフレーム名)
summary(オブジェクト名)

glm()でモデルを指定し，結果を「オブジェクト名」に保存する．その内容をsummary()で表示する．

```
> setwd("i:¥¥Rdocuments¥¥scripts¥¥")
> d1 <- read.table("ロジスティック回帰分析_データ.csv", header=TRUE, sep=",")
> head(d1)
  kyoufu support stress result
1     29      16     28      0
2     27      19     17      1
3     27      16     21      1
4     26      23     22      0
5     26      22     23      1
6     26      20     25      1
>
```

	A	B	C	D
1	kyoufu	support	stress	result
2	29	16	28	0
3	27	19	17	1
4	27	16	21	1
5	26	23	22	0
6	26	22	23	1
7	26	20	25	1
8	26	20	21	0
9	26	19	25	0
10	26	16	26	0
11	25	32	24	1
12	25	15	21	0
13	25	13	29	0
14	25	12	26	0
15	24	21	20	1
16	24	11	28	0
17	23	31	26	1
18	23	29	20	1
19	23	28	23	1
20	23	21	24	0
21	23	14	29	0

```
> # 記述統計量
> dtmp <- d1
> ntmp <- nrow(dtmp)
> mtmp <- colMeans(dtmp)
> stmp <- apply(dtmp, 2, sd)
> ctmp <- cor(dtmp)
> ktmp <- round(data.frame(ntmp, mtmp, stmp, ctmp),2)
> colnames(ktmp) <- c("N","Mean","SD",colnames(ctmp))
> ktmp
          N  Mean   SD kyoufu support stress result
kyoufu  245 15.22 5.37   1.00   -0.10   0.43  -0.21
support 245 20.32 6.08  -0.10    1.00  -0.32   0.48
stress  245 18.52 5.12   0.43   -0.32   1.00  -0.39
result  245  0.55 0.50  -0.21    0.48  -0.39   1.00

>
> # 説明変数の標準化
> d2 <- as.data.frame(scale(d1[,c("kyoufu","support","stress")]))
> d2 <- data.frame(d2, d1$result)
> colnames(d2) <- c("kyoufu","support","stress","result")


>  #共分散行列の確認
>  cov(d2)
            kyoufu    support     stress     result
kyoufu   1.0000000 -0.1034273  0.4255222 -0.1032508
support -0.1034273  1.0000000 -0.3231937  0.2385116
stress   0.4255222 -0.3231937  1.0000000 -0.1966482
result  -0.1032508  0.2385116 -0.1966482  0.2484108


> #標準偏回帰係数の推定
> result.2 <- glm(result ~ kyoufu + support + stress, family=binomial, d2)
> summary(result.2)

Call:
glm(formula = result ~ kyoufu + support + stress, family = binomial,
    data = d2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5189  -0.8689   0.3166   0.7980   2.0672

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3065     0.1568   1.954 0.050708 .
kyoufu       -0.1587     0.1715  -0.926 0.354625
support       1.1552     0.1954   5.912 3.38e-09 ***
stress       -0.7613     0.1980  -3.845 0.000121 ***
```

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.09  on 244  degrees of freedom
Residual deviance: 249.58  on 241  degrees of freedom
AIC: 257.58

Number of Fisher Scoring iterations: 5


> #（標準）偏回帰係数の信頼区間
> confint(result.2)
Waiting for profiling to be done...
                   2.5 %     97.5 %
(Intercept)  0.002021354  0.6187654
kyoufu      -0.498291957  0.1768211
support      0.789789998  1.5587577
stress      -1.164559209 -0.3850987


> #ステップワイズ分析
>  #MASSパッケージの読み込み
>  library(MASS)
> result.3 <- stepAIC(result.2)
Start:  AIC=257.58
result ~ kyoufu + support + stress

          Df Deviance    AIC
- kyoufu   1   250.44 256.44
<none>         249.58 257.58
- stress   1   266.10 272.10
- support  1   295.24 301.24

Step:  AIC=256.44
result ~ support + stress

          Df Deviance    AIC
<none>         250.44 256.44
- stress   1   273.83 277.83
- support  1   295.74 299.74

> summary(result.3)
Call:
glm(formula = result ~ support + stress, family = binomial, data = d2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4713  -0.8605   0.3192   0.8171   2.1162

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3070     0.1565   1.961   0.0498 *
support       1.1502     0.1954   5.886 3.95e-09 ***
stress       -0.8292     0.1850  -4.482 7.39e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.09  on 244  degrees of freedom
Residual deviance: 250.44  on 242  degrees of freedom
AIC: 256.44

Number of Fisher Scoring iterations: 5
```

ロジスティック回帰分析 ― 集計データの場合

```
オブジェクト名 <- glm(割合変数 ~ 独立変数群, data=データフレーム名, weights=ウエイト変数,
                 family=binomial(link="リンク関数名"))
```
または
```
オブジェクト名 <- glm(二値変数 ~ 独立変数群, data=データフレーム名,
                 family=binomial(link="リンク関数名"))

summary(オブジェクト名)
confint(オブジェクト名)
```

データが集計されている場合には，独立変数の各水準における事象発生割合を基準変数とし，ウエイト変数（独立変数の各水準のサイズ）を指定する．
データが集計されていない（0,1のまま）場合には，二値変数を基準変数に用いる．

ランダム成分は二項分布とする．

リンク関数		
	identity：恒等リンク	線形回帰
	logit： ロジット関数	ロジスティック回帰
	probit： プロビット関数	プロビット回帰

```
> # Data from Norton P. G. & Dunn, E. V. (1985). Br. Med. J., 291, 630-632.
> # Agresti,A. (2007). An introduction to categorical data analysis, 2nd ed. Wiley, p69.

> setwd("i:¥¥Rdocuments¥¥programs¥¥")
> d1 <- read.table("snoringData.csv", header=TRUE, sep=",")
> d1$snr <- c(0,2,4,5)
> d1$subtotal <- d1$yes + d1$no
> d1$pyes <- d1$yes / d1$subtotal
> d1
     snoring yes   no snr subtotal       pyes
1      never  24 1355   0     1379 0.01740392
2 occasional  35  603   2      638 0.05485893
3     nearly  21  192   4      213 0.09859155
4      every  30  224   5      254 0.11811024
>
>
> # GLM
> #Binomial(identity) linear regression
> result.1 <- glm(pyes ~ snr, data=d1, weights=subtotal,
+                 family=binomial(link="identity"))
> summary(result.1)

Call:
glm(formula = pyes ~ snr, family = binomial(link = "identity"),
    data = d1, weights = subtotal)

Deviance Residuals:
       1         2         3         4
 0.04478  -0.21322   0.11010   0.09798

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.017247   0.003451   4.998 5.80e-07 ***
snr         0.019778   0.002805   7.051 1.77e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.904481  on 3  degrees of freedom
Residual deviance:  0.069191  on 2  degrees of freedom
AIC: 24.322

Number of Fisher Scoring iterations: 3

> confint(result.1)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) 0.01132939 0.02483298
```

	A	B	C
1	snoring	yes	no
2	never	24	1355
3	occasional	35	603
4	nearly	21	192
5	every	30	224
6			

```
snr         0.01451898 0.02550672
>
> #Binomial(logit) logistic regression
> result.2 <- glm(pyes ~ snr, data=d1, weights=subtotal,
+                 family=binomial(link="logit"))
> summary(result.2)

Call:
glm(formula = pyes ~ snr, family = binomial(link = "logit"),
    data = d1, weights = subtotal)

Deviance Residuals:
      1       2       3       4
-0.8346  1.2521  0.2758 -0.6845

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.86625    0.16621 -23.261  < 2e-16 ***
snr          0.39734    0.05001   7.945 1.94e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.9045  on 3  degrees of freedom
Residual deviance:  2.8089  on 2  degrees of freedom
AIC: 27.061

Number of Fisher Scoring iterations: 4

> confint(result.2)
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) -4.2072190 -3.5544117
snr          0.2999362  0.4963887
>
>
> #Binomial(probit) probit regression
> result.3 <- glm(pyes ~ snr, data=d1, weights=subtotal,
+                 family=binomial(link="probit"))
> summary(result.3)

Call:
glm(formula = pyes ~ snr, family = binomial(link = "probit"),
    data = d1, weights = subtotal)

Deviance Residuals:
      1       2       3       4
-0.6188  1.0388  0.1684 -0.6175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.06055    0.07017 -29.367  < 2e-16 ***
snr          0.18777    0.02348   7.997 1.28e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.9045  on 3  degrees of freedom
Residual deviance:  1.8716  on 2  degrees of freedom
AIC: 26.124

Number of Fisher Scoring iterations: 4

> confint(result.3)
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) -2.2026985 -1.9262643
snr          0.1416397  0.2343393
>
```

多項ロジスティック回帰分析

```
library(VGAM)
オブジェクト名 <- vglm(基準変数 ~ 説明変数, family=multinomial, data=d1)
summary(オブジェクト名)
```

あらかじめVGAMパッケージをインストールしておく必要がある.
vglm は glm を含む, より一般化された分析モデルである.

```
> setwd("i:¥¥Rdocuments¥¥scripts¥¥")
> d1 <- read.table("多項ロジスティック回帰分析_データ.csv", header=TRUE, sep=",")
> head(d1)

  kyoufu support stress result
1     29      16     28      1
2     27      19     17      2
3     27      16     21      2
4     26      23     22      2
5     26      22     23      3
6     26      20     25      2
>
```

	A	B	C	D
1	kyoufu	support	stress	result
2	29	16	28	1
3	27	19	17	2
4	27	16	21	2
5	26	23	22	2
6	26	22	23	3
7	26	20	25	2
8	26	20	21	1
9	26	19	25	1
10	26	16	26	1
11	25	32	24	3
12	25	15	21	1
13	25	13	29	1
14	25	12	26	1
15	24	21	20	2
16	24	11	28	1
17	23	31	26	3
18	23	29	20	3
19	23	28	23	2
20	23	21	24	1
21	23	14	29	1

```
> # 記述統計量
> dtmp <- d1
> ntmp <- nrow(dtmp)
> mtmp <- colMeans(dtmp)
> stmp <- apply(dtmp, 2, sd)
> ctmp <- cor(dtmp)
> ktmp <- round(data.frame(ntmp, mtmp, stmp, ctmp),2)
> colnames(ktmp) <- c("N","Mean","SD",colnames(ctmp))
> ktmp
          N  Mean   SD kyoufu support stress result
kyoufu  245 15.22 5.37   1.00   -0.10   0.43  -0.21
support 245 20.32 6.08  -0.10    1.00  -0.32   0.48
stress  245 18.52 5.12   0.43   -0.32   1.00  -0.39
result  245  0.55 0.50  -0.21    0.48  -0.39   1.00

>
> #偏回帰係数の推定と分析結果の表示
>  # VGAMパッケージの読み込み
> library(VGAM)
> result.1 <- vglm(result ~ kyoufu + support + stress,
+                  family=multinomial, data=d1)
> summary(result.1)

Call:
vglm(formula = result ~ kyoufu + support + stress,
    family = multinomial, data = d1)

Pearson Residuals:
                      Min       1Q   Median      3Q    Max
log(mu[,1]/mu[,3]) -2.7306 -0.53116 -0.19290 0.57759 7.7165
log(mu[,2]/mu[,3]) -2.3776 -0.62207 -0.38850 1.07377 2.0425

Coefficients:
                  Value Std. Error  t value
(Intercept):1  0.6413635   1.204523  0.53246
(Intercept):2  1.8121174   1.021766  1.77352
kyoufu:1       0.0522839   0.040671  1.28554
kyoufu:2      -0.0095065   0.035416 -0.26842
support:1     -0.2430842   0.040026 -6.07318
support:2     -0.1317942   0.032755 -4.02369
stress:1       0.1897292   0.048025  3.95061
stress:2       0.0804411   0.039218  2.05114
```

```
Number of linear predictors:  2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Dispersion Parameter for multinomial family:   1

Residual Deviance: 440.6041 on 482 degrees of freedom

Log-likelihood: -220.3020 on 482 degrees of freedom

Number of Iterations: 5
>

> #標準偏回帰係数の推定
> d2 <- as.data.frame(scale(d1[,c("kyoufu","support","stress")]))
> d2 <- data.frame(d2,d1$result)
> colnames(d2) <- c("kyoufu","support","stress","result")
> head(d2)
    kyoufu     support     stress result
1 2.568375 -0.70988932  1.8521250      1
2 2.195706 -0.21672425 -0.2975205      2
3 2.195706 -0.70988932  0.4841688      2
4 2.009371  0.44082919  0.6795911      2
5 2.009371  0.27644083  0.8750134      3
6 2.009371 -0.05233589  1.2658581      2


>  #共分散行列の確認
>  cov(d2)
            kyoufu    support     stress     result
kyoufu    1.0000000 -0.1034273  0.4255222 -0.1850753
support -0.1034273  1.0000000 -0.3231937  0.3927501
stress    0.4255222 -0.3231937  1.0000000 -0.3303219
result   -0.1850753  0.3927501 -0.3303219  0.6432921

> result.2 <- vglm(result ~ kyoufu + support + stress,
+                  family=multinomial, data=d2)

> summary(result.2)

Call:
vglm(formula = result ~ kyoufu + support + stress, family = multinomial,
    data = d2)

Pearson Residuals:
                      Min       1Q   Median      3Q    Max
log(mu[,1]/mu[,3]) -2.7306 -0.53116 -0.19290 0.57759 7.7165
log(mu[,2]/mu[,3]) -2.3776 -0.62207 -0.38850 1.07377 2.0425

Coefficients:
                Value Std. Error   t value
(Intercept):1  0.012106    0.21474  0.056374
(Intercept):2  0.479586    0.18733  2.560111
kyoufu:1       0.280591    0.21827  1.285542
kyoufu:2      -0.051018    0.19007 -0.268423
support:1     -1.478719    0.24348 -6.073177
support:2     -0.801725    0.19925 -4.023692
stress:1       0.970867    0.24575  3.950605
stress:2       0.411627    0.20068  2.051143

Number of linear predictors:  2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Dispersion Parameter for multinomial family:   1

Residual Deviance: 440.6041 on 482 degrees of freedom

Log-likelihood: -220.3020 on 482 degrees of freedom

Number of Iterations: 5
>
```

対数線形モデル ― glm関数を使う方法

データ形式（一覧データ（フラットなクロス表））
テーブル名 <- ftable(データフレーム名[,c("変数名1","変数名2",…)], row.vars=c("変数名1","変数名2",…))

各カテゴリ変数の水準の組み合わせに対する度数の一覧表「sex=F, method=1, grade=A」の度数 = 7　など

sex	method	grade	Freq
F	1	A	7
F	1	B	11
F	1	C	2
F	2	A	10
F	2	B	4
F	2	C	4
M	1	A	8
M	1	B	8
M	1	C	7
M	2	A	13
M	2	B	7
M	2	C	16

モデルの指定法
glmオブジェクト名 <- glm(度数変数 ~ 説明変数1 + 説明変数2 + …, family=poisson, データフレーム名)

交互作用の指定法
説明変数名1 : 説明変数名2　　# : 区切り … 交互作用のみ
説明変数名1 * 説明変数名2　　# * 区切り … 主効果も含んだ交互作用

パラメタ推定値の表示
summary(glmオブジェクト名)

ステップワイズ分析
library(MASS)
AICオブジェクト名 <- stepAIC(glmオブジェクト名)
summary(AICオブジェクト名)

MASSパッケージは最初からインストールされている

要因の主効果，交互作用の検討
library(car)
Anova(glmオブジェクト名)

あらかじめ car パッケージをインストールしておく必要がある.
Anova は，anova とは異なる関数であることに注意.

残差分析
オブジェクト名 <- glm(度数変数~見たい効果以下の効果変数名の指定, family=poisson,データフレーム名)

モデルによる予測値
xtabs(fitted.values(オブジェクト名) ~ 変数名の指定)

残差・標準化残差（ピアソン）
xtabs(residuals(オブジェクト名, type="pearson") ~ 変数名の指定)
xtabs(rstandard(オブジェクト名, type="pearson") ~ 変数名の指定)

デビアンス残差を得るには，type="deviance" とするか，typeの指定を省略する.

```
> setwd("d:¥¥")
> d1 <- read.table("対数線形モデル_データ.csv", header=TRUE, sep=",")
> head(d1)
  id sex method grade
1  1   F      1     B
2  2   M      1     B
3  3   F      2     A
4  4   M      2     C
5  5   M      2     C
6  6   M      3     A
>
```

```
> #クロス表
> (t1 <- table(d1[,c("method","grade","sex")], dnn=list("method","grade","sex")))
, , sex = F

      grade
method  A  B  C
     1  7 11  2
     2 10  4  4
     3  3  3  2

, , sex = M

      grade
method  A  B  C
     1  8  8  7
     2 13  7 16
     3  6  5  6
```

```
> #一覧データ
> ft1 <- ftable(d1[,c("sex","method","grade")], row.vars=c("sex","method","grade"))
> (d2 <- as.data.frame(ft1))

   sex method grade Freq
1    F      1     A    7
2    M      1     A    8
3    F      2     A   10
4    M      2     A   13
5    F      3     A    3
6    M      3     A    6
7    F      1     B   11
8    M      1     B    8
9    F      2     B    4
10   M      2     B    7
11   F      3     B    3
12   M      3     B    5
13   F      1     C    2
14   M      1     C    7
15   F      2     C    4
16   M      2     C   16
17   F      3     C    2
18   M      3     C    6
>
```

	A	B	C	D
1	id	sex	method	grade
2	1	F	1	B
3	2	M	1	B
4	3	F	2	A
5	4	M	2	C
6	5	M	2	C
7	6	M	3	A
8	7	M	2	C
9	8	M	2	A
10	9	M	2	A
11	10	M	2	A
12	11	F	2	A
13	12	F	2	B
14	13	M	2	B
15	14	M	1	A
16	15	M	2	C
17	16	M	3	A
18	17	M	1	B
19	18	M	3	C
20	19	F	2	A
21	20	M	2	C

```
> #飽和モデル
> result.full <- glm(Freq~ method + grade + sex
+                  + sex:method + sex:grade + method:grade
+                  + sex:method:grade, family=poisson, d2)

> summary(result.full)

Call:
glm(formula = Freq ~ method + grade + sex + sex:method + sex:grade +
    method:grade + sex:method:grade, family = poisson, data = d2)
```

```
Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.945910   0.377964   5.148 2.63e-07 ***
method2              0.356675   0.492805   0.724   0.4692
method3             -0.847298   0.690066  -1.228   0.2195
gradeB               0.451985   0.483494   0.935   0.3499
gradeC              -1.252763   0.801784  -1.562   0.1182
sexM                 0.133531   0.517549   0.258   0.7964
method2:sexM         0.128833   0.666918   0.193   0.8468
method3:sexM         0.559616   0.876275   0.639   0.5231
gradeB:sexM         -0.451985   0.695533  -0.650   0.5158
gradeC:sexM          1.119232   0.954313   1.173   0.2409
method2:gradeB      -1.368276   0.764046  -1.791   0.0733 .
method3:gradeB      -0.451985   0.948911  -0.476   0.6338
method2:gradeC       0.336472   0.996422   0.338   0.7356
method3:gradeC       0.847298   1.214985   0.697   0.4856
method2:gradeB:sexM  0.749237   1.026424   0.730   0.4654
method3:gradeB:sexM  0.269664   1.231706   0.219   0.8267
method2:gradeC:sexM  0.004699   1.183274   0.004   0.9968
method3:gradeC:sexM -0.713766   1.441312  -0.495   0.6204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  3.5563e+01  on 17  degrees of freedom
Residual deviance: -2.2013e-25  on  0  degrees of freedom
AIC: 101.28

Number of Fisher Scoring iterations: 3



> # 要因の主効果，交互作用の検討
> library(car)
> Anova(result.full)

Analysis of Deviance Table (Type II tests)

Response: Freq
                LR Chisq Df Pr(>Chisq)
method           11.0975  2   0.003892 **
grade             1.4586  2   0.482244
sex               7.4533  1   0.006332 **
method:sex        1.2316  2   0.540205
grade:sex         5.4028  2   0.067110 .
method:grade      6.0188  4   0.197751
method:grade:sex  0.9904  4   0.911255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>


> #ステップワイズ分析
>  #MASSパッケージの読み込み
>  library(MASS)
> result.aic <- stepAIC(result.full)

Start:  AIC=101.28
Freq ~ method + grade + sex + sex:method + sex:grade + method:grade +
    sex:method:grade

                   Df Deviance     AIC
- method:grade:sex  4  0.99035  94.272
<none>              0.00000 101.282

Step:  AIC=94.27
Freq ~ method + grade + sex + method:sex + grade:sex + method:grade
```

```
               Df Deviance    AIC
- method:sex    2   2.2220 91.504
- method:grade  4   7.0091 92.291
<none>              0.9904 94.272
- grade:sex     2   6.3932 95.675

Step:  AIC=91.5
Freq ~ method + grade + sex + grade:sex + method:grade

               Df Deviance    AIC
- method:grade  4   9.1956 90.477
<none>              2.2220 91.504
- grade:sex     2   8.5797 93.861

Step:  AIC=90.48
Freq ~ method + grade + sex + grade:sex

            Df Deviance    AIC
<none>           9.1956 90.477
- grade:sex  2  15.5533 92.835
- method     2  20.2931 97.575


> summary(result.aic)

Call:
glm(formula = Freq ~ method + grade + sex + grade:sex, family = poisson,
    data = d2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5563  -0.5128   0.1097   0.3377   1.6722

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9529     0.2551   7.656 1.91e-14 ***
method2       0.2278     0.2044   1.114   0.2651
method3      -0.5423     0.2515  -2.156   0.0311 *
gradeB       -0.1054     0.3249  -0.324   0.7457
gradeC       -0.9163     0.4183  -2.190   0.0285 *
sexM          0.3001     0.2950   1.017   0.3090
gradeB:sexM  -0.1947     0.4389  -0.444   0.6572
gradeC:sexM   0.9877     0.4965   1.989   0.0467 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 35.5627  on 17  degrees of freedom
Residual deviance:  9.1956  on 10  degrees of freedom
AIC: 90.477

Number of Fisher Scoring iterations: 4


> Anova(result.aic)
Analysis of Deviance Table (Type II tests)

Response: Freq
          LR Chisq Df Pr(>Chisq)
method     11.0975  2   0.003892 **
grade       1.4586  2   0.482244
sex         7.4533  1   0.006332 **
grade:sex   6.3577  2   0.041634 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

```
> # 主効果モデル
> result.1 <- glm(Freq ~ method + grade + sex,
+                 family=poisson, d2)
> summary(result.1)

Call:
glm(formula = Freq ~ method + grade + sex, family = poisson,
    data = d2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.49529  -0.69830  -0.05981   0.23877   2.28630

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8319     0.2233   8.203 2.35e-16 ***
method2       0.2278     0.2044   1.114  0.26507
method3      -0.5423     0.2515  -2.156  0.03106 *
gradeB       -0.2126     0.2182  -0.974  0.32988
gradeC       -0.2392     0.2198  -1.088  0.27637
sexM          0.5021     0.1868   2.688  0.00719 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 35.563  on 17  degrees of freedom
Residual deviance: 15.553  on 12  degrees of freedom
AIC: 92.835

Number of Fisher Scoring iterations: 4



> # sex;grade の交互作用を見るための残差分析
>  #主効果とsex:method, method:gradeの交互作用までを仮定したモデル

> result.2 <- glm(Freq~ sex + method + grade + sex:method + method:grade,family=poisson, d2)
> summary(result.2)

Call:
glm(formula = Freq ~ sex + method + grade + sex:method + method:grade,
    family = poisson, data = d2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.19065  -0.32089   0.00029   0.25977   0.93241

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.9426     0.3056   6.356 2.07e-10 ***
sexM            0.1398     0.3057   0.457   0.6476
method2         0.0943     0.4170   0.226   0.8211
method3        -0.8848     0.5381  -1.644   0.1001
gradeB          0.2364     0.3454   0.684   0.4937
gradeC         -0.5108     0.4216  -1.212   0.2257
sexM:method2    0.5534     0.4205   1.316   0.1882
sexM:method3    0.6140     0.5266   1.166   0.2436
method2:gradeB -0.9740     0.5037  -1.934   0.0531 .
method3:gradeB -0.3542     0.5962  -0.594   0.5525
method2:gradeC  0.3711     0.5208   0.712   0.4762
method3:gradeC  0.3930     0.6433   0.611   0.5412
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 35.5627  on 17  degrees of freedom
Residual deviance:  6.3932  on  6  degrees of freedom
AIC: 95.675

Number of Fisher Scoring iterations: 4
```

```
> # 予測値
> xtabs(fitted.values(result.2) ~ sex + grade + method , data=d2)
, , method = 1

   grade
sex         A         B         C
  F  6.976744  8.837209  4.186047
  M  8.023256 10.162791  4.813953

, , method = 2

   grade
sex         A         B         C
  F  7.666667  3.666667  6.666667
  M 15.333333  7.333333 13.333333

, , method = 3

   grade
sex         A         B         C
  F  2.880000  2.560000  2.560000
  M  6.120000  5.440000  5.440000

>
> # 残差（ピアソン）
> xtabs(residuals(result.2, type="pearson") ~ sex + grade + method, data=d2)
, , method = 1

   grade
sex            A            B            C
  F  0.008804509  0.727540078 -1.068457799
  M -0.008210247 -0.678434621  0.996342033

, , method = 2

   grade
sex            A            B            C
  F  0.842700972  0.174077656 -1.032795559
  M -0.595879572 -0.123091491  0.730296743

, , method = 3

   grade
sex            A            B            C
  F  0.070710678  0.275000000 -0.350000000
  M -0.048507125 -0.188648444  0.240098019

>
> # 標準化残差（ピアソン）
> xtabs(rstandard(result.2, type="pearson") ~ sex + grade + method, data=d2)
, , method = 1

   grade
sex           A           B           C
  F  0.01491867  1.33154217 -1.64294963
  M -0.01491865 -1.33154032  1.64294735

, , method = 2

   grade
sex           A           B           C
  F  1.36218225  0.23891939 -1.59410745
  M -1.36218123 -0.23891921  1.59410626

, , method = 3

   grade
sex           A           B           C
  F  0.10718662  0.40441176 -0.51470588
  M -0.10718662 -0.40441176  0.51470588
```

対数線形モデル ― loglm関数を使う方法

データ形式（テーブルデータ　または　一覧データ）

テーブルデータ（クロス表）
テーブル名 <- table(データフレーム名[,c("変数名1","変数名2",…)], dnn=c("変数名1","変数名2",…))

一覧データ（フラットなクロス表）
テーブル名 <- ftable(データフレーム名[,c("変数名1","変数名2",…)], row.vars=c("変数名1","変数名2",…))

各カテゴリ変数の水準の組み合わせに対する度数の一覧表「sex=F, method=1, grade=A」の度数 = 7　など

sex	method	grade	Freq
F	1	A	7
F	1	B	11
F	1	C	2
F	2	A	10
F	2	B	4
F	2	C	4
M	1	A	8
M	1	B	8
M	1	C	7
M	2	A	13
M	2	B	7
M	2	C	16

モデルの指定法

テーブルデータを使う場合
library(MASS)
オブジェクト名 <- loglm(~ 変数1 +変数2 + 変数1:変数2 + …, テーブル名)

一覧データを使う場合
library(MASS)
オブジェクト名 <- loglm(度数変数 ~ 説明変数1 + 説明変数2 + 変数1:変数2 + …, データフレーム名)

MASSパッケージは最初からインストールされている

交互作用の指定法
説明変数名1 : 説明変数名2　　# : 区切り … 交互作用のみ
説明変数名1 * 説明変数名2　　# * 区切り … 主効果も含んだ交互作用

パラメタ推定値の表示
coef(オブジェクト名)

ステップワイズ分析
library(MASS)
AICオブジェクト名 <- stepAIC(loglmオブジェクト名)
coef(AICオブジェクト名)

尤度比検定
anova(loglmオブジェクト名1, loglmオブジェクト名2)

AICオブジェクト名も設定可能
一覧データを用いた場合は，モデルの記述に不具合が生じるようだが，計算は正しくしてくれる

残差分析
テーブルデータを使う場合
オブジェクト名 <- loglm(~変数名の指定, テーブル名)

一覧データを使う場合
オブジェクト名 <- loglm(度数変数~変数名の指定, データフレーム名)

観測値のモデル予測からのデビアンス残差の表
apply(residuals(オブジェクト名), c(配列の次元を指定), sum)

```
> setwd("d:¥¥")
> d1 <- read.table("対数線形モデル_データ.csv", header=TRUE, sep=",")
> head(d1)
  id sex method grade
1  1   F      1     B
2  2   M      1     B
3  3   F      2     A
4  4   M      2     C
5  5   M      2     C
6  6   M      3     A
>

> #クロス表
> (t1 <- table(d1[,c("method","grade","sex")], dnn=list("method","grade","sex")))
, , sex = F

      grade
method  A  B  C
     1  7 11  2
     2 10  4  4
     3  3  3  2

, , sex = M

      grade
method  A  B  C
     1  8  8  7
     2 13  7 16
     3  6  5  6

>
> #一覧データ
> ft1 <- ftable(d1[,c("sex","method","grade")], row.vars=c("sex","method","grade"))
> (d2 <- as.data.frame(ft1))
   sex method grade Freq
1    F      1     A    7
2    M      1     A    8
3    F      2     A   10
4    M      2     A   13
5    F      3     A    3
6    M      3     A    6
7    F      1     B   11
8    M      1     B    8
9    F      2     B    4
10   M      2     B    7
11   F      3     B    3
12   M      3     B    5
13   F      1     C    2
14   M      1     C    7
15   F      2     C    4
16   M      2     C   16
17   F      3     C    2
18   M      3     C    6
>
>
> #MASSパッケージの読み込み
> library(MASS)
>
>
> # 飽和モデル
> result.full <- loglm(~ method + grade + sex
+                  + sex:method + sex:grade + method:grade
+                  + sex:method:grade, t1)
> # または
> # result.full <- loglm(Freq~ method + grade + sex
> #                   + sex:method + sex:grade + method:grade
> #                   + sex:method:grade, d2)
>
> coef(result.full)
$`(Intercept)`
[1] 1.75465
```

	A	B	C	D
1	id	sex	method	grade
2	1	F	1	B
3	2	M	1	B
4	3	F	2	A
5	4	M	2	C
6	5	M	2	C
7	6	M	3	A
8	7	M	2	C
9	8	M	2	A
10	9	M	2	A
11	10	M	2	A
12	11	F	2	A
13	12	F	2	B
14	13	M	2	B
15	14	M	1	A
16	15	M	2	C
17	16	M	3	A
18	17	M	1	B
19	18	M	3	C
20	19	F	2	A
21	20	M	2	C

```
$method
          1           2           3
 0.1023078   0.3051205 -0.4074284

$grade
           A             B             C
 0.209226512 -0.001717884 -0.207508628

$sex
          F           M
-0.3099278   0.3099278

$method.grade
       grade
method           A           B          C
     1 -0.05350831   0.38342865 -0.3299203
     2   0.16477037 -0.39195020  0.2271798
     3 -0.11126207   0.00852155  0.1027405

$method.sex
       sex
method           F           M
     1   0.13195435 -0.13195435
     2 -0.05811795   0.05811795
     3 -0.07383640   0.07383640

$grade.sex
     sex
grade          F          M
    A   0.1284206 -0.1284206
    B   0.1845965 -0.1845965
    C -0.3130172   0.3130172

$method.grade.sex
, , sex = F

       grade
method           A           B           C
     1 -0.01721290   0.15260380 -0.13539089
     2   0.10844296 -0.09635866 -0.01208429
     3 -0.09123005 -0.05624513  0.14747519

, , sex = M

       grade
method           A           B           C
     1   0.01721290 -0.15260380  0.13539089
     2 -0.10844296   0.09635866  0.01208429
     3   0.09123005   0.05624513 -0.14747519


> # 飽和モデルの残差
> residuals(result.full)
Re-fitting to get frequencies and fitted values
, , sex = F

       grade
method A B C
     1 0 0 0
     2 0 0 0
     3 0 0 0

, , sex = M

       grade
method A B C
     1 0 0 0
     2 0 0 0
     3 0 0 0
>
>
```

```
> # ステップワイズ分析
> (result.aic <- stepAIC(result.full))

Start:  AIC=36
~method + grade + sex + sex:method + sex:grade + method:grade +
    sex:method:grade

                   Df   AIC
- method:grade:sex  4 28.99
<none>                36.00

Step:  AIC=28.99
~method + grade + sex + method:sex + grade:sex + method:grade

               Df    AIC
- method:sex    2 26.222
- method:grade  4 27.009
<none>            28.990
- grade:sex     2 30.393

Step:  AIC=26.22
~method + grade + sex + grade:sex + method:grade

               Df    AIC
- method:grade  4 25.196
<none>            26.222
- grade:sex     2 28.580

Step:  AIC=25.2
~method + grade + sex + grade:sex

            Df    AIC
<none>         25.196
- grade:sex  2 27.553
- method     2 32.293
Call:
loglm(formula = ~method + grade + sex + grade:sex, data = t1,
    evaluate = FALSE)

Statistics:
                     X^2 df  P(> X^2)
Likelihood Ratio 9.195595 10 0.5136471
Pearson          9.224254 10 0.5109616


> coef(result.aic)

$`(Intercept)`
[1] 1.789734

$method
         1          2          3
 0.1048468  0.3326307 -0.4374775

$grade
           A            B            C
 0.208382813  0.005650259 -0.214033071

$sex
         F          M
-0.2822199  0.2822199

$grade.sex
     sex
grade          F          M
    A  0.1321676 -0.1321676
    B  0.2295396 -0.2295396
    C -0.3617072  0.3617072

>

>
```

```
> # 主効果モデル
> result.1 <- loglm(~ method + grade + sex, t1)
> # または
> #result.1 <- loglm(Freq ~ method + grade + sex, d2)
>
>

> # 尤度比検定
> anova(result.1, result.2)
LR tests for hierarchical log-linear models

Model 1:
 ~method + grade + sex
Model 2:
 Freq ~ grade + method + sex + grade:sex

           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   15.553284 12
Model 2    6.393183  6   9.160101         6        0.16477
Saturated  0.000000  0   6.393183         6        0.38062
>
>

> # sex:grade の交互作用を見るための残差分析
>  #主効果とsex:method, method:gradeの交互作用までを仮定したモデル
>
> result.2 <- loglm(~ sex + method + grade + sex:method + method:grade, t1)
> # または
> #result.2 <- loglm(Freq~ method + grade + sex + sex:method + method:grade, d2)
>
> coef(result.2)
$`(Intercept)`
[1] 1.78938

$method
         1          2          3
 0.1316042  0.3016217 -0.4332260

$grade
          A           B           C
 0.15415153 -0.05217953 -0.10197200

$sex
         F          M
-0.2644468  0.2644468

$method.grade
      grade
method           A           B           C
     1 -0.06267259  0.38004726 -0.31737467
     2  0.13830209 -0.39296578  0.25466369
     3 -0.07562951  0.01291852  0.06271099

$method.sex
      sex
method           F           M
     1  0.19456585 -0.19456585
     2 -0.08212677  0.08212677
     3 -0.11243908  0.11243908

>

> #予測値
> fitted.values(result.2)
Re-fitting to get fitted values
, , sex = F

      grade
method        A        B        C
     1 6.976744 8.837209 4.186047
     2 7.666667 3.666667 6.666667
     3 2.880000 2.560000 2.560000
```

```
, , sex = M

      grade
method         A         B         C
     1  8.023256 10.162791  4.813953
     2 15.333333  7.333333 13.333333
     3  6.120000  5.440000  5.440000

>
>
> #ピアソン残差
> residuals(result.2, type="pearson")
Re-fitting to get frequencies and fitted values
, , sex = F

      grade
method            A            B            C
     1  0.008804509  0.727540078 -1.068457799
     2  0.842700972  0.174077656 -1.032795559
     3  0.070710678  0.275000000 -0.350000000

, , sex = M

      grade
method            A            B            C
     1 -0.008210247 -0.678434621  0.996342035
     2 -0.595879572 -0.123091491  0.730296743
     3 -0.048507125 -0.188648444  0.240098019

>
>
> #デビアンス残差
> residuals(result.2)
Re-fitting to get frequencies and fitted values
, , sex = F

      grade
method            A            B            C
     1  0.008799624  0.700535218 -1.190653100
     2  0.804609625  0.171535271 -1.116569901
     3  0.070227965  0.267638153 -0.364087473

, , sex = M

      grade
method            A            B            C
     1 -0.008214218 -0.704923252  0.932409555
     2 -0.612040160 -0.124042121  0.707782794
     3 -0.048666954 -0.191281248  0.236145930

>
```

カウントデータの分析

ポアソンモデル

```
オブジェクト名 <- glm(計数変数 ~ 独立変数 + offset(log(全体数変数)), data=データフレーム名,
            family=poisson(link="log"))
```

負の二項分布モデル

```
library(MASS)
オブジェクト名 <- glm.nb(計数変数 ~ 独立変数 + offset(log(全体数変数)), data=データフレーム名,
            link=log)

summary(オブジェクト名)
confint(オブジェクト名)
```

```
> # Data from British department of transport.
> # Agresti,A. (2007). An introduction to categorical data analysis, 2nd ed. Wiley, p83.
```

```
> setwd("i:¥¥Rdocuments¥¥programs¥¥")
>
> d1 <- read.table("collisionsData.csv", header=TRUE, sep=",")
> d1$x <- d1$year - 1975
> head(d1)
  year trainKm collTT collTR  x
1 2003     518      0      3 28
2 2002     516      1      3 27
3 2001     508      0      4 26
4 2000     503      1      3 25
5 1999     505      1      2 24
6 1998     487      0      4 23
>
```

	A	B	C	D
1	year	trainKm	collTT	collTR
2	2003	518	0	3
3	2002	516	1	3
4	2001	508	0	4
5	2000	503	1	3
6	1999	505	1	2
7	1998	487	0	4
8	1997	463	1	1
9	1996	437	2	2
10	1995	423	1	2
11	1994	415	2	4
12	1993	425	0	4
13	1992	430	1	4
14	1991	439	2	6
15	1990	431	1	2
16	1989	436	4	4
17	1988	443	2	4
18	1987	397	1	6
19	1986	414	2	13
20	1985	418	0	5
21	1984	389	5	3
22	1983	401	2	7
23	1982	372	2	3
24	1981	417	2	2
25	1980	430	2	2
26	1979	426	3	3
27	1978	430	2	4
28	1977	425	1	8
29	1976	426	2	12
30	1975	436	5	2

```
> # 記述統計量
> dtmp <- d1
> ntmp <- nrow(dtmp)
> mtmp <- colMeans(dtmp)
> stmp <- apply(dtmp, 2, sd)
> ctmp <- cor(dtmp)
> ktmp <- round(data.frame(ntmp, mtmp, stmp, ctmp),2)
> colnames(ktmp) <- c("N","Mean","SD",colnames(ctmp))
> ktmp
          N    Mean    SD  year trainKm collTT collTR     x
year     29 1989.00  8.51  1.00    0.72  -0.56  -0.34  1.00
trainKm  29  440.00 38.68  0.72    1.00  -0.44  -0.26  0.72
collTT   29    1.66  1.32 -0.56   -0.44   1.00  -0.03 -0.56
collTR   29    4.21  2.81 -0.34   -0.26  -0.03   1.00 -0.34
x        29   14.00  8.51  1.00    0.72  -0.56  -0.34  1.00>
>
```

```
> # GLM
>
> #Poisson loglinear
> result.1 <- glm(collTR ~ x + offset(log(trainKm)), data=d1,
+             family=poisson(link="log"))
> summary(result.1)

Call:
glm(formula = collTR ~ x + offset(log(trainKm)), family = poisson(link = "log"),
    data = d1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0580  -0.7825  -0.0826   0.3775   3.3873

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.21142    0.15892  -26.50  < 2e-16 ***
x           -0.03292    0.01076   -3.06  0.00222 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 47.376  on 28  degrees of freedom
Residual deviance: 37.853  on 27  degrees of freedom
AIC: 133.52

Number of Fisher Scoring iterations: 5

> confint(result.1)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -4.5342219 -3.9105409
x           -0.0542105 -0.0119652
>
>

> # Negative binomial loglinear
> library(MASS)
> result.2 <- glm.nb(collTR ~ x + offset(log(trainKm)), data=d1,
+             link=log)
> summary(result.2)

Call:
glm.nb(formula = collTR ~ x + offset(log(trainKm)), data = d1,
    link = log, init.theta = 10.11828724)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.72370  -0.65461  -0.05868   0.32984   2.64065

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.19999    0.19584 -21.446  < 2e-16 ***
x           -0.03367    0.01288  -2.615  0.00893 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(10.1183) family taken to be 1)

    Null deviance: 32.045  on 28  degrees of freedom
Residual deviance: 25.264  on 27  degrees of freedom
AIC: 132.69

Number of Fisher Scoring iterations: 1


              Theta:  10.12
          Std. Err.:  8.00

 2 x log-likelihood:  -126.69

> confint(result.2)
Waiting for profiling to be done...
                 2.5 %       97.5 %
(Intercept) -4.593077 -3.815438383
x           -0.059499 -0.008290411
>
```