Categorical Data Analysis
Lecture 13

Graduate School of Advanced Science and Engineering
Rei Monden

2025.1.23

Loglinear models for contingency tables

Loglinear models for contingency tables

Loglinear models = GLMs for count data.
Predictors can be used to model counts.

Today we focus *exclusively* on modeling cell counts in contingency tables that cross-classify categorical variables.

Loglinear models allow studying the associations between the cell counts and the categorical variables that create the cell counts.

Loglinear models for counts in contingency tables

Let's go through a few loglinear models, namely:

1. Independence model for two-way contingency tables.
2. Saturated model for two-way contingency tables.

We will not study loglinear models for n-way contingency tables for $n \geq 3$ in detail.

But we will learn to fit and compare such models using R.

Loglinear models for counts in contingency tables

In general, *any* loglinear model for counts in contingency tables is a GLM with:

- Link function: $\log$.
- Random component: Poisson distribution (per cell count).

Independence model

Independence model

An *observed* $r \times c$ contingency table

	Y categories			
X categories	$Y = 1$	$Y = 2$	$\cdots$	$Y = c$
$X = 1$	n_{11}	n_{12}	$\cdots$	n_{1c}
$X = 2$	n_{21}	n_{22}	$\cdots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$X = r$	n_{r1}	n_{r2}	$\cdots$	n_{rc}

Recall from Chapter 3 that categorical variables X and Y are said to be statistically independent when all joint cell probabilities equal the product of the corresponding row and column marginal probabilities:

$$\underbrace{P(X = i, Y = j)}_{\pi_{ij}} = \underbrace{P(X = i)}_{\pi_{i+}} \underbrace{P(Y = j)}_{\pi_{+j}},$$

for all $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

Notice that $\sum_{i,j} \pi_{ij} = 1$.

Independence model

$$\boxed{\pi_{ij} = \pi_{i+}\pi_{+j}}$$

To find the expected cell counts (μ_{ij}) under the independence model, we simply multiply the joint probabilities $\{\pi_{ij}\}$ by the total sample size n:

$$\sum_{i,j} \pi_{ij} = 1 \Longrightarrow$$

$$\sum_{i,j} n\pi_{ij} = n \Longrightarrow$$

$$\sum_{i,j} \underbrace{n\pi_{i+}\pi_{+j}}_{\mu_{ij}} = n.$$

Independence model

Expected cell probabilities under the independence model

	Y categories				
X categories	$Y=1$	$Y=2$	$\cdots$	$Y=c$	Total
$X=1$	π_{11}	π_{12}	$\cdots$	π_{1c}	π_{1+}
$X=2$	π_{21}	π_{22}	$\cdots$	π_{2c}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X=r$	π_{r1}	π_{r2}	$\cdots$	π_{rc}	π_{r+}
Total	π_{+1}	π_{+2}	$\cdots$	π_{+c}	1

$\pi_{ij} = \pi_{i+}\pi_{+j}$.

Expected cell counts under the independence model

	Y categories				
X categories	$Y=1$	$Y=2$	$\cdots$	$Y=c$	Total
$X=1$	μ_{11}	μ_{12}	$\cdots$	μ_{1c}	
$X=2$	μ_{21}	μ_{22}	$\cdots$	μ_{2c}	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
$X=r$	μ_{r1}	μ_{r2}	$\cdots$	μ_{rc}	
Total					n

$\mu_{ij} = n\pi_{i+}\pi_{+j}$.

Independence model

$$\boxed{\mu_{ij} = n\pi_{i+}\pi_{+j}}$$

Taking logs we have

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y,$$

where:

- $\lambda = \log(n)$ is an intercept based on the sample size;
- $\lambda_i^X = \log(\pi_{i+})$ is the ith row effect;
- $\lambda_j^Y = \log(\pi_{+j})$ is the jth column effect.

This is the so-called loglinear model of independence.

Independence model

$$\boxed{\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y}$$

$\mathcal{H}_0$: *X and Y are independent.*

is equivalent to

$\mathcal{H}_0$: *The loglinear model of independence is true.*

But we already learned how to test the first $\mathcal{H}_0$ above:

Using the likelihood-ratio chi-squared test of independence (Chapter 3)!

Therefore, the chi-squared significance test of independence is essentially a goodness-of-fit test for the loglinear model of independence:

The likelihood-ratio chi-squared significance test of independence will reject the independence hypothesis if and only if the loglinear model does not fit the data well.

Independence model

$$\boxed{\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y}$$

Categorical predictor X has r levels, so it requires $(r-1)$ indicator variables to be coded.

As such, the set of the r row effects, $\{\lambda_1^X, \lambda_2^X, \ldots, \lambda_r^X\}$ is not identified (it has one parameter too many).

To constrain the model, R defaults to setting $\lambda_1^X = 0$.

Similarly, R by default sets $\lambda_1^Y = 0$.

Thus, the loglinear model of independence has

$$\underbrace{1}_{\text{intercept}} + \underbrace{(r-1)}_{X} + \underbrace{(c-1)}_{Y} = r + c - 1$$

parameters (i.e., model df's).

Independence model

$$\boxed{\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y}$$

Thus, assume that $\lambda_1^X = \lambda_1^Y = 0$.

To interpret the row effect λ_i^X, focus on rows 1 (constrained, acting as the reference) and i of the contingency table, *at any column* j:

	Y categories		
X categories	$\cdots$	$Y = j$	$\cdots$
$X = 1$	$\cdots$	μ_{1j}	$\cdots$
$X = i$	$\cdots$	μ_{ij}	$\cdots$

We have that

$$\begin{aligned}
\text{logit}[P(X = i)] &= \log\left[\frac{P(X = i)}{P(X = 1)}\right] = \log\left(\frac{\mu_{ij}}{\mu_{1j}}\right) \\
&= \log(\mu_{ij}) - \log(\mu_{1j}) \\
&= \left(\lambda + \lambda_i^X + \lambda_j^Y\right) - \left(\lambda + \lambda_j^Y\right) \\
&= \lambda_i^X.
\end{aligned}$$

Independence model

$$\text{logit}[P(X = i)] = \lambda_i^X \quad \Longrightarrow \quad \text{odds} = \exp(\lambda_i^X)$$

Conclusion:
$\exp(\lambda_i^X) =$ odds of response in row i over row 1, *at any column* j (since X and Y are independent)!

In general,

$$\exp\left(\lambda_{i_1}^X - \lambda_{i_2}^X\right)$$

is the odds of response in row i_1 over row i_2, *at any column* j.

Similar reasoning applies to interpreting λ_j^Y.

Independence model

	Heaven		
Happy	no	yes	Total
not	32	190	222
pretty	113	611	724
very	51	326	377
Total	196	1127	1323

- $X =$ happiness (not too happy, pretty happy, very happy)
- $Y =$ belief in heaven (no, yes).

This is a 3×2 contingency table.

Independence model

```
# Import data frame from file:
heaven.df <- read.table("HappyHeaven.dat", header = TRUE)
heaven.df
```

Output:

```
   happy heaven count
1    not     no    32
2    not    yes   190
3 pretty     no   113
4 pretty    yes   611
5   very     no    51
6   very    yes   326
```

Independence model

Are X and Y independent?

First let's use the likelihood-ratio chi-squared test of independence:

```
heaven.chisq <- chisq.test(
  matrix(heaven.df$count, ncol=2, byrow = TRUE),
  correct = FALSE
)

G2      <- with(heaven.chisq,
                2 * sum(observed * log(observed / expected))
                )
p.value <- 1 - pchisq(G2, (3-1) * (2-1))
c(G2, p.value)
```

Output:

```
[1] 0.8911136 0.6404675
```

Conclusion: $\chi^2(2) = .89$, $p = .64$.
At $\alpha = 5\%$ significance level, we cannot reject the null hypothesis that X and Y are independent.

Independence model

Are X and Y independent?

Now let's fit the loglinear model of independence:

```
heaven.fit <- glm(count  ~ happy + heaven,
                  family = poisson(link = "log"),
                  data   = heaven.df)

summary(heaven.fit)
```

Output:

```
-----------------------------------------------------
Residual deviance:    0.89111  on 2  degrees of freedom
-----------------------------------------------------
```

Recall that $D = .891111 \sim \chi^2(6-4)$ and this test compares our model to the saturated model (i.e., the model that fits the data perfectly).

Conclusion: $\chi^2(2) = .89$, $p = .64$.
Exactly the same result as from the likelihood-ratio chi-squared test of independence.

Independence model

Although the loglinear model of independence and the likelihood-ratio chi-squared test of independence lead to the same *goodness-of-fit* result, the loglinear model is much richer!

For example, let's look at the parameter estimates:

```
summary(heaven.fit)
```

Output:

```
-----------------------------------------------------
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.49313    0.09408   37.13  < 2e-16 ***
happypretty  1.18211    0.07672   15.41  < 2e-16 ***
happyvery    0.52957    0.08460    6.26 3.86e-10 ***
heavenyes    1.74920    0.07739   22.60  < 2e-16 ***
-----------------------------------------------------
```

Independence model

```
-----------------------------------------------------
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.49313    0.09408   37.13  < 2e-16 ***
happypretty  1.18211    0.07672   15.41  < 2e-16 ***
happyvery    0.52957    0.08460    6.26 3.86e-10 ***
heavenyes    1.74920    0.07739   22.60  < 2e-16 ***
-----------------------------------------------------
```

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y \quad \Longrightarrow \quad \boxed{\mu_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y)}$$

Here are the predicted cell counts under the fitted loglinear model of independence:

	Heaven		
Happy	no	yes	Total
not	$e^{3.493} = 32.89$	$e^{3.493+1.749} = 189.11$	222
pretty	$e^{3.493+1.182} = 107.26$	$e^{3.493+1.182+1.749} = 616.74$	724
very	$e^{3.493+0.530} = 55.85$	$e^{3.493+0.530+1.749} = 321.15$	377
Total	196	1127	1323

Independence model

Compare:

Observed counts

	Heaven		
Happy	no	yes	Total
not	32	190	222
pretty	113	611	724
very	51	326	377
Total	196	1127	1323

Predicted counts (loglin. model of ind.)

	Heaven		
Happy	no	yes	Total
not	32.89	189.11	222
pretty	107.26	616.74	724
very	55.85	321.15	377
Total	196	1127	1323

The observed and predicted counts look quite similar.

This is the reason why the likelihood-ratio chi-squared test of independence failed to reject the null hypothesis of independence!

Independence model

```
--------------------
Coefficients:
            Estimate
(Intercept)  3.49313
happypretty  1.18211
happyvery    0.52957
heavenyes    1.74920
--------------------
```

Observed counts

	Heaven		
Happy	no	yes	Total
not	32	190	222
pretty	113	611	724
very	51	326	377
Total	196	1127	1323

$$\mu_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y)$$

Interpretation:

- Intercept: $\exp(\widehat{\lambda}) = \exp(3.493) = 32.89$.
  Estimated $(\text{happy}_{\text{not}}, \text{heaven}_{\text{no}})$ cell size.

Independence model

```
--------------------
Coefficients:
            Estimate
(Intercept)  3.49313
happypretty  1.18211
happyvery    0.52957
heavenyes    1.74920
--------------------
```

Observed counts

	Heaven		
Happy	no	yes	Total
not	32	190	222
pretty	113	611	724
very	51	326	377
Total	196	1127	1323

$$\mu_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y)$$

Interpretation:

- Row-effect happy$_{\text{pretty}}$: $\exp(\hat{\lambda}_2^X) = \exp(1.182) = 3.26$.
  Estimated odds of response in row 'happy=pretty' over row 'happy=not', *at any column*.

The *observed* values are actually

$$\frac{113}{32} = 3.53 \text{ (column 1)}, \qquad \frac{611}{190} = 3.22 \text{ (column 2)}.$$

Independence model

```
--------------------
Coefficients:
             Estimate
(Intercept)   3.49313
happypretty   1.18211
happyvery     0.52957
heavenyes     1.74920
--------------------
```

Observed counts

	Heaven		
Happy	no	yes	Total
not	32	190	222
pretty	113	611	724
very	51	326	377
Total	196	1127	1323

$$\mu_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y)$$

Interpretation:

- Row-effect happy$_{\text{very}}$: $\exp(\hat{\lambda}_3^X) = \exp(0.530) = 1.70$.
  Estimated odds of response in row 'happy=very' over row 'happy=not', *at any column*.

The *observed* values are actually

$$\frac{51}{32} = 1.59 \text{ (column 1)}, \qquad \frac{326}{190} = 1.72 \text{ (column 2)}.$$

Independence model

```
--------------------
Coefficients:
            Estimate
(Intercept)  3.49313
happypretty  1.18211
happyvery    0.52957
heavenyes    1.74920
--------------------
```

Observed counts

	Heaven		
Happy	no	yes	Total
not	32	190	222
pretty	113	611	724
very	51	326	377
Total	196	1127	1323

$$\mu_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y)$$

Interpretation:

- Column-effect heaven$_{\text{yes}}$: $\exp(\widehat{\lambda}_2^Y) = \exp(1.749) = 5.75$.
  Estimated odds of response in column 'heaven=yes' over column 'heaven=no', *at any row*.

The *observed* values are actually

$$\frac{190}{32} = 5.94 \text{ (row 1)}, \qquad \frac{611}{113} = 5.41 \text{ (row 2)}, \qquad \frac{326}{51} = 6.39 \text{ (row 3)}.$$

Saturated model

Saturated model

The saturated model is given by

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

It is called *saturated* because it has as many parameters as cells:

$$\underbrace{1}_{\text{intercept}} + \underbrace{(r-1)}_{X} + \underbrace{(c-1)}_{Y} + \underbrace{(r-1)(c-1)}_{XY} = rc.$$

(identification constraint: $\lambda_{i1}^{XY} = 0 = \lambda_{1j}^{XY}$ for all i, j).

This means that it fits the cell counts *perfectly*:

$$\hat{\mu}_{ij} = n_{ij}, \text{for any cell } (i,j).$$

Saturated model

$$\boxed{\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}}$$

The $\{\lambda_{ij}^{XY}\}$ parameters describe association (i.e., lack of independence) between X and Y.

These are interaction terms.

Saturated model

$$\boxed{\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}}$$

Thus, assume that $\lambda_1^X = \lambda_1^Y = \lambda_{i1}^{XY} = \lambda_{1j}^{XY} = 0$ for all i, j.

To interpret the interaction effect λ_{ij}^{XY}, focus on rows 1 and i and columns 1 and j of the contingency table:

	Y categories	
X categories	$Y = 1$	$Y = j$
$X = 1$	μ_{11}	μ_{1j}
$X = i$	μ_{i1}	μ_{ij}

Saturated model

	Y categories	
X categories	$Y = 1$	$Y = j$
$X = 1$	μ_{11}	μ_{1j}
$X = i$	μ_{i1}	μ_{ij}

We have that

$$\begin{aligned}
\log(\text{odds ratio}) &= \log\left(\frac{\mu_{ij}/\mu_{1j}}{\underbrace{\mu_{i1}/\mu_{11}}_{\text{columnwise}}}\right) = \log\left(\frac{\mu_{ij}/\mu_{i1}}{\underbrace{\mu_{1j}/\mu_{11}}_{\text{rowwise}}}\right) \\
&= \log\left(\frac{\mu_{11}\mu_{ij}}{\mu_{1j}\mu_{i1}}\right) \\
&= \log(\mu_{11}) + \log(\mu_{ij}) - \log(\mu_{1j}) - \log(\mu_{i1}) \\
&= \cdots \\
&= \lambda_{ij}^{XY}.
\end{aligned}$$

Saturated model

$$\log(\text{odds ratio}) = \lambda_{ij}^{XY} \quad \Longrightarrow \quad \text{odds ratio} = \exp(\lambda_{ij}^{XY})$$

Conclusion:

- *Columnwise:*
  $\exp(\lambda_{ij}^{XY})$ = ratio of odds of response in row i over row 1 for column j, to the odds of response in row i over row 1 for column 1.
- *Rowwise:*
  $\exp(\lambda_{ij}^{XY})$ = ratio of odds of response in column j over column 1 for row i, to the odds of response in column j over column 1 for row 1.

Saturated model

Let's fit the saturated model:

```
heaven.sat <- glm(count  ~ happy * heaven,
                  family = poisson(link = "log"),
                  data   = heaven.df)

summary(heaven.sat)
```

Output:

```
----------------------------------------------------
Residual deviance: 2.2204e-16  on 0  degrees of freedom
----------------------------------------------------
```

The deviance is of course 0 since $D = 2(L_S - L_M)$, where in this case model M is the saturated model.

Saturated model

Parameter estimates:

```
summary(heaven.sat)
```

Output:

```
-----------------------------------------------------
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.46574    0.17678  19.605  < 2e-16 ***
happypretty            1.26165    0.20025   6.300 2.97e-10 ***
happyvery              0.46609    0.22552   2.067   0.0388 *
heavenyes              1.78129    0.19108   9.322  < 2e-16 ***
happypretty:heavenyes -0.09358    0.21679  -0.432   0.6660
happyvery:heavenyes    0.07378    0.24329   0.303   0.7617
-----------------------------------------------------
```

Saturated model

```
-----------------------------------------------------
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.46574    0.17678  19.605  < 2e-16 ***
happypretty            1.26165    0.20025   6.300 2.97e-10 ***
happyvery              0.46609    0.22552   2.067   0.0388 *
heavenyes              1.78129    0.19108   9.322  < 2e-16 ***
happypretty:heavenyes -0.09358    0.21679  -0.432   0.6660
happyvery:heavenyes    0.07378    0.24329   0.303   0.7617
-----------------------------------------------------
```

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad \Longrightarrow \quad \boxed{\mu_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})}$$

It is easy to see that, for any cell ij, $\widehat{\mu}_{ij} = n_{ij}$.

For example, $n_{22} = 611$ (happy=pretty, heaven=yes) and

$$e^{3.467+1.262+1.781-0.094} \equiv 611.$$

Saturated model

Of course, the saturated model on its own is useless:

- It has as many parameters as cells
- It captures all signal *and all noise* in the data.

But, the saturated model is useful in relative terms, to compare to any model of interest.

In fact, this is what we did when we earlier tested the deviance of the loglinear model of independence.

Saturated model

Compare:

```
summary(heaven.fit)
```

Output:

```
-----------------------------------------------------
Residual deviance:    0.89111  on 2  degrees of freedom
-----------------------------------------------------
```

```
anova(heaven.fit, heaven.sat, test = "LRT")
```

Output:

```
-----------------------------------------------------
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
2         0    0.00000  2  0.89111   0.6405
-----------------------------------------------------
```

Statistical inference for loglinear models

Statistical inference for loglinear models

We have already done this through goodness-of-fit testing based on deviances.

Let's go through it once more, now with a new data set.

Although now we will also fit loglinear models with more than two categorical factors, we will not interpret their parameters like before.

Instead, we will only do model comparison.

Statistical inference for loglinear models

Alcohol use	Cigarette use	Marijuana use: yes	Marijuana use: no
yes	yes	911	538
	no	44	456
no	yes	3	43
	no	2	279

Survey among high school students.

This is a $2 \times 2 \times 2$ contingency table.

Statistical inference for loglinear models

```
# Import data frame from file:
marij.df <- read.table("Substance.dat", header = TRUE)
marij.df
```

Output:

```
  alcohol cigarettes marijuana count
1     yes        yes       yes   911
2     yes        yes        no   538
3     yes         no       yes    44
4     yes         no        no   456
5      no        yes       yes     3
6      no        yes        no    43
7      no         no       yes     2
8      no         no        no   279
```

Statistical inference for loglinear models

For example, does a model including the three factors fair significantly worse than the saturated model (which includes all three factors plus their interactions)?

```
marij.fit <- glm(count  ~ alcohol + cigarettes + marijuana,
                 family = poisson(link = "log"),
                 data   = marij.df)

summary(marij.fit)
```

Output:

```
Residual deviance: 1286.0  on 4  degrees of freedom
```

Conclusion: $\chi^2(4) = 1286$, $p < .001$.
This model fits significantly worse than the saturated model.

We need to include more effects to improve it!

Statistical inference for loglinear models

Equivalent way of running the same model comparison:

```
marij.sat <- glm(count  ~ alcohol * cigarettes * marijuana,
                 family = poisson(link = "log"),
                 data   = marij.df)

anova(marij.fit, marij.sat, test = "LRT")
```

Output:

```
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
2         0          0  4     1286 < 2.2e-16 ***
```

Statistical inference for loglinear models

Let's extend the model by adding all two-way interactions:

```
marij.fit2 <- glm(count  ~ alcohol + cigarettes + marijuana +
                           alcohol:cigarettes + alcohol:marijuana +
                           cigarettes:marijuana,
                  family = poisson(link = "log"),
                  data   = marij.df)

summary(marij.fit2)
```

Output:

```
Residual deviance:    0.37399  on 1  degrees of freedom
```

Conclusion: $\chi^2(1) = 0.37$, $p = .54$.
This model does not fit significantly worse than the saturated model.

We therefore keep it as it is (very slightly...) simpler than the saturated model.

Statistical inference for loglinear models

Equivalent way of running the same model comparison:

```
anova(marij.fit2, marij.sat, test = "LRT")
```

Output:

```
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
2         0    0.00000  1  0.37399   0.5408
```

Statistical inference for loglinear models

We can also look at the statistical significance of each effect in a model via likelihood-ratio tests.

For example:

```
library(car)
Anova(marij.fit2)
```

Output:

```
                    LR Chisq Df Pr(>Chisq)
alcohol              1281.71  1  < 2.2e-16 ***
cigarettes            227.81  1  < 2.2e-16 ***
marijuana              55.91  1  7.575e-14 ***
alcohol:cigarettes    187.38  1  < 2.2e-16 ***
alcohol:marijuana      91.64  1  < 2.2e-16 ***
cigarettes:marijuana  497.00  1  < 2.2e-16 ***
```

Conclusion: All effects are statistically significant.

Statistical inference for loglinear models

Statistical significance is not enough.
After a model is selected, it remains to see how well it actually fits the data.

We can look at the cell Pearson standardized residuals to help us.

In general

$$\text{std. residual} = \frac{\text{obs. count} - \text{exp. count}}{SE}.$$

Standardized residuals are approximately $\mathcal{N}(0, 1)$ distributed when the model fits well and the observed cell counts are not too small.

Statistical inference for loglinear models

```
cbind(
  Res.fit  = rstandard(marij.fit,  type = "pearson"),
  Res.fit2 = rstandard(marij.fit2, type = "pearson")
  )
```

Output:

```
     Res.fit   Res.fit2
1  30.514739  0.6333249
2 -16.103300 -0.6333249
3 -21.176588 -0.6333249
4   5.936302  0.6333249
5 -11.436320 -0.6333250
6  -9.807251  0.6333249
7  -7.427416  0.6333250
8  31.193633 -0.6333249
```

Clearly, the model with all first- and second-order effects fits the data much better.

Statistical inference for loglinear models

Instead of reporting only the estimated parameters, it is good practice to also report associated CIs.

```
marij.fit2CI <- confint(marij.fit2)

cbind("95% LB" = marij.fit2CI[, 1],
          Est  = coef(marij.fit2),
      "95% UB" = marij.fit2CI[, 2]
        )
```

Output:

```
                          95% LB    Est 95% UB
(Intercept)                5.514  5.633  5.748
alcoholyes                 0.340  0.488  0.637
cigarettesyes             -2.218 -1.887 -1.579
marijuanayes              -6.377 -5.309 -4.477
alcoholyes:cigarettesyes   1.723  2.055  2.407
alcoholyes:marijuanayes    2.176  2.986  4.037
cigarettesyes:marijuanayes 2.537  2.848  3.181
```

Statistical inference for loglinear models

Another interesting quantity to report is the so-called dissimilarity index:

$$D_{\text{ind}} = \sum_{ij} \frac{|n_{ij} - \hat{\mu}_{ij}|}{2n} = \sum_{ij} \frac{|p_{ij} - \hat{\pi}_{ij}|}{2}.$$

D_{i} captures the difference between the *observed* and the *fitted* cell counts (or probabilities):

> D_i = *proportion of sample observations that must move to different cells for model fit to become perfect.*

Unlike significance testing, D_{ind} is not affected by the sample size.

The smaller D_{i}, the better.

Statistical inference for loglinear models

```
# D_ind for the main effects only model:
sum(abs(marij.df$count - fitted(marij.fit)))/(2*sum(marij.df$count))

# D_ind for the main plus interaction effects model:
sum(abs(marij.df$count - fitted(marij.fit2)))/(2*sum(marij.df$count))
```

Output:

```
[1] 0.2875384

[1] 0.00108406
```

Conclusion: The model with both main and two-way interaction effects fits better.

Exercise 13-1

The `DeathPenalty.dat` file contains data from an article that studied effects of racial characteristics on whether subjects convicted of homicide receive the death penalty. The 274 subjects were the defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987. Let $D =$ defendant's race, $V =$ victim's race, and $P =$ death penalty verdict.

a. Fit the loglinear model with all possible main and two-way interaction effects. Do provide all the R code used.
b. Report the estimated conditional odds ratio between D and P at each level of V. Interpret.
c. Test the goodness of fit of this model. Interpret.
d. Compute the estimated count for the cell (white defendant, black victim, no death penalty verdict) and compare this to the observed count in the same condition.

Exercise 13-2

The `MBTI.dat` file contains data about 4 scales of a personality test. Four two-level factors are included: Extroversion/Introversion (E/I), Sensing/iNtuitive (S/N), Thinking/Feeling (T/F) and Judging/Perceiving (J/P). There are therefore 16 personality types.

a. Fit the loglinear model by which the variables are mutually independent. Do provide all the R code used.
b. Report the results of the goodness-of-fit test.
c. Fit the loglinear model with all main and two-order interaction effects. Based on the fit, show that the estimated conditional association is strongest between the S/N and J/P scales.
d. Using the model in (c), show that there is not strong evidence of conditional association (i.e., interaction) between the E/I and T/F scales or between the E/I and J/P scales.

Exercise 13-3

Refer to the previous exercise.

a. Run the model that assumes conditional independence between E/I and T/F and between E/I and J/P but has the other pairwise associations. Do provide all the R code used.
b. Compare this to the fit of the model containing all the pairwise associations. What do you conclude?
c. Compute the 95% likelihood-ratio confidence interval for the conditional odds ratio between the S/N and J/P scales.

Exercise 13-1 (Japanese/日本語)

Moodle の Example data フォルダー内の `DeathPenalty.dat` は人種的特徴が殺人で有罪判決を受けた被告が死刑判決を受けるかどうかに及ぼす影響を調査した記事のデータが含まれています. 対象となった 274 人の被告は 1976 年から 1987 年までのフロリダ州で発生した複数の殺人事件に関連する起訴の被告です. $D =$ 被告の人種, $V =$ 被害者の人種, and $P =$ 死刑判決をそれぞれ表しています.

a. すべての主要効果と 2 次の交互作用効果を含む対数線形モデルを適用せよ. 分析に用いた R コードは全てレポートに表示すること.
b. 各レベルの （被害者の人種）における （被告の人種）と （死刑判決）の条件付きオッズ比の推定値を求め, 解釈せよ.
c. このモデルの適合度を検定し, 解釈せよ.
d. (被告: 白人、被害者: 黒人、死刑判決: なし）のセルの推定度数を計算し, 同じ条件における観測度数と比較せよ.

Exercise 13-2　(Japanese/日本語)

Moodle の Example data フォルダー内の `MBTI.dat` は性格テストの 4 つの尺度に関するデータが含まれています. 各尺度には 2 つの水準が存在し, それぞれ外向性/内向性（E/I）, 感覚/直観（S/N）, 思考/感情（T/F）, 判断/知覚（J/P）となることから, 全部で 16 種類の性格タイプに分かれます.

a. 変数が相互に独立であると仮定した対数線形モデルを適合させなさい. 分析に用いた R コードは全てレポートに表示すること.
b. 適合度検定の結果を示せ.
c. すべての主要効果および 2 次の交互作用効果を含む対数線形モデルを適用せよ. この適合度に基づいて, S/N 尺度と J/P 尺度の間の条件付き関連が最も強いことを示せ.
d. (c) のモデルを用いて, E/I 尺度と T/F 尺度、または E/I 尺度と J/P 尺度の間に強い条件付き関連（すなわち交互作用）が存在しないことを示せ.

Exercise 13-3　(Japanese/日本語)

Exercise 13-2 と同じデータを用います.

a. E/I と T/F, E/I と J/P の条件付き独立性を仮定し、他のすべてのペア間の関連を含むモデルを適用せよ. また, 分析に用いた R コードは全てレポートに表示すること.
b. a. のモデルを全てのペア間の関連を含むモデルと比較せよ. どのような結論が導き出せるか述べよ.
c. S/N 尺度と J/P 尺度の条件付きオッズ比に対する 95% 尤度比信頼区間を計算せよ.

Final Exam

- The final exam will be given on Thursday, Jauary 30th 8:45 – 10:30 at **EDU** K201.
- The exam will consist entirely of multiple-choice questions, and you will need a black pen to fill out the answer sheet. Please ensure you bring your own black pen to the exam.
- The questions will cover the material discussed in the lectures up to and including Lecture 14.

最終筆記試験

- 最終筆記試験は第 1 回の講義でアナウンスした通り、2025 年 1 月 30 日 (木)8:45 – 10:30に教 K201 で行います。
- 試験はすべて選択式の問題で構成され、回答用紙を記入するために黒のペンが必要です。必ずご自身で黒のペンを持参してください。
- テスト範囲は、第 14 回講義までに扱った内容です。