Contents

1 Loglinear Models for Contingency Tables

As introduced before the loglinear model is a GLM with link function $g(\mu) = log(\mu)$, a Poisson random component and the linear predictor as systematic component.
We used the loglinear model for modeling count data.
In this section we will apply this model to count data in contingency tables, here the response variable is the count and the predictors are the categorical variables defining the contingency table. This model is used to investigate the interaction structure between the set of categorical variables of interest. In contrast the logistic regression model was used to model a categorical response variable in dependency on predictor variables (which of course could be categorical).

1.1 Loglinear models for 2-way tables

1.1.1 Independence Model

Consider two categorical variables with I and J categories, respectively, then $\{\pi_{ij}\}$ is the joint distribution of the two categorical variables, and if the two variables are independent then

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \quad i = 1, \ldots, I, j = 1, \ldots, J$$

The joint distribution are the parameters of a multinomial distribution, but when considering the expected cell count $\{\mu_{ij}\} = \{n\pi_{ij}\}$, we can model these using a loglinear model, where the two categorical variables are independent if

$$\mu_{ij} = n\pi_{i+}\pi_{+j}, \quad i = 1, \ldots, I, j = 1, \ldots, J$$

taking the logarithm to both sides gives

$$log(\mu_{ij}) = log(n) + log(\pi_{i+}) + log(\pi_{+j}), \quad i = 1, \ldots, I, j = 1, \ldots, J$$

Therefore the log of the expected cell counts are linear in effects by the categories of the two categorical variables.
Write

$$log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y, \quad i = 1, \ldots, I, j = 1, \ldots, J$$

This model is called the *loglinear model of independence*, because it only fits if the two variables X and Y are independent.

- λ is the intercept (independent from X and Y and needs to be included in the model to ensure model that $\sum_{ij} \mu_{ij} = n$ ($\lambda = \log(n)$.
- λ_i^X is the (main or marginal) effect of category i of variable X on the log of the expected cell count. The larger λ_i^X the larger the expected frequency for row i.
- λ_j^Y is the (main or marginal) effect of category j of variable Y on the log of the expected cell count. The larger λ_j^Y the larger the expected frequency for column j.

To test if two variables are independent can now be based on the model fit statistics for this model. The two statistics to be used are Pearson's χ^2 or the likelihood ratio statistic, which match the χ^2 test statistics for the test of independence for two-way tables.

Example 1
Movie rating of Alien versus Wall-e.

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	4823.100	4	.000
Pearson Chi-Square	4692.375	4	.000

a. Model: Poisson

b. Design: Constant + movie_n + rating

At significance level of 5% the data provide sufficient evidence that the rating is not independent from the movie (Wall-e versus Alien) (Pearson $\chi^2 = 4692.4, df = 4, P < 0.001$, or likelihood ratio= $4823.1, df = 4, P < 0.001$).
We reject the model and find it not to be a good fit for the population.

Example 2
Many nursing students take STAT 151 online because they struggle to fit the course into their otherwise busy studyplan. Does taking the course f2f or online have an effect on passing STAT 151? Of a random sample of 160 nursing students

- 90 took STAT 151 online, of which 20 failed
- 60 took STAT 151 f2f, of which 11 failed

A test for independence results in $\chi^2(1) = 1.07$ and P-value = 0.30. Indicating that at significance level of 5% there is insufficient evidence that the rate of success in STAT 151 depends on the mode of delivery.
So the loglinear model of independence should provide a good fit for the data. The four model equations are:

$$\begin{aligned} log(\mu_{f2f,no}) &= \lambda + \lambda^{Mode}_{f2f} + \lambda^{Pass}_{no}, \\ log(\mu_{f2f,yes}) &= \lambda + \lambda^{Mode}_{f2f} + \lambda^{Pass}_{yes}, \\ log(\mu_{online,no}) &= \lambda + \lambda^{Mode}_{online} + \lambda^{Pass}_{no}, \\ log(\mu_{online,yes}) &= \lambda + \lambda^{Mode}_{online} + \lambda^{Pass}_{yes}, \end{aligned}$$

Parameter estimates (from SPSS) are:

$$\hat{\lambda} = 2.518,\ \hat{\lambda}^{Mode}_{online} = .405,\ \hat{\lambda}^{Mode}_{f2f} = 0,\ \hat{\lambda}^{Pass}_{yes} = 1.345,\ \hat{\lambda}^{Pass}_{no} = 0$$

Comments:

- For one category of row and column variable the parameter estimate is set to 0. This is similar to modelling categorical variables through dummy variables, one category is chosen as the reference level and is assigned a parameter value of 0.
- Since more students took the course online: $\hat{\lambda}^{Mode}_{online} > \hat{\lambda}^{Mode}_{f2f}$, and since more students passed than faile: $\hat{\lambda}^{Pass}_{yes} > \hat{\lambda}^{Pass}_{no}$,

Interpretation of the parameters in the loglinear model of independence

To understand the interpretation of the parameters calculate the conditional log odds, $\log(\theta(i_1, i_2|Y = j))$ for falling into category $X = i_1$ versus $X = i_2$, when $Y = j$. (Since we assume that X and Y are independent this odds ratio should be the same for all j)

		j		
i_1		$\pi_{i_1 j}$		
i_2		$\pi_{i_2 j}$		

$$\begin{aligned}
\log(\theta(i_1, i_2|Y = j)) &= \log\left(\frac{\pi_{i_1 j}}{\pi_{i_2 j}}\right) = \log\left(\frac{n\pi_{i_1 j}}{n\pi_{i_2 j}}\right) = \log\left(\frac{\mu_{i_1 j}}{\mu_{i_2 j}}\right) \\
&= \log(\mu_{i_1 j}) - \log(\mu_{i_2 j}) \\
&= \lambda + \lambda^X_{i_1} + \lambda^Y_j - (\lambda + \lambda^X_{i_2} + \lambda^Y_j) \\
&= \lambda^X_{i_1} - \lambda^X_{i_2}
\end{aligned}$$

Important conclusions for the loglinear model of independence:

- it has been confirmed that the conditional log odds for levels of X do not depend on j (which we should have expected when X and Y are independent)
- the odds for falling into category i_1 versus i_2 equals $e^{\lambda^X_{i_1} - \lambda^X_{i_2}}$ for all j, which means that in this model
$$\lambda^X_{i_1} - \lambda^X_{i_2} = log(P(X = i_1)/P(X = i_2))$$
the logarithm of the relative risk of category i_1 versus category i_2 in variable X,
or $e^{\lambda^X_{i_1} - \lambda^X_{i_2}}$ are the odds to fall within category i_1, given the individual falls either into category i_1 or i_2.

Comment: One of the parameters $\lambda^X_1, \ldots, \lambda^X_I$ is redundant (same for Y), because total expected cell count adds to n. Usually the parameter value for one category is set to zero, which turns that category into the reference category. This is similar to using one less dummy variable than levels of the categorical predictor to be included with a model.
The purpose of this discussion was to illustrate that when we assume that the model of independence describes the population we can confirm that the properties we know about independent variables hold.

1.1.2 Saturated Model

Many times the variables under consideration are not independent (e.g. movie example), in which case the appropriate model includes the interaction.

$$log(\mu_{ij}) = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^{XY}_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots, J$$

is called the saturated model for a two-way contingency table.

- The conditional log odds for falling into category $X = i_1$ versus $X = i_2$, when $Y = j$ is for the saturated model
$$\log(\theta(i_1, i_2|Y = j)) = \lambda^X_{i_1} + \lambda^{XY}_{i_1 j} - \lambda^X_{i_2} - \lambda^{XY}_{i_2 j}$$
and depends for this model on the level of Y, reflecting, that X and Y are not independent anymore. The level of Y affects the probabilities for the different levels of X (i.e. the likelihood of each rating level depends on the movie that is being evaluated).

  The calculations are similar to the one for the independence model, only use the model equation for the saturated model.

- λ^{XY}_{ij} is the interaction effect of category i and category j on the expected cell count. It measures the association between X and Y, and the degree of deviation from independence for the level of the categorical variables.

- for a 2×2 contingency table the log odds ratio equals

$$\begin{aligned}
\log(\theta) &= \log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) \\
&= \log(\mu_{11}) + \log(\mu_{22}) - \log(\mu_{12}) - \log(\mu_{21}) \\
&= (\lambda + \lambda^X_1 + \lambda^Y_1 + \lambda^{XY}_{11}) + (\lambda + \lambda^X_2 + \lambda^Y_2 + \lambda^{XY}_{22}) \\
&\quad -(\lambda + \lambda^X_1 + \lambda^Y_2 + \lambda^{XY}_{12}) - (\lambda + \lambda^X_2 + \lambda^Y_1 + \lambda^{XY}_{21}) \\
&= \lambda^{XY}_{11} + \lambda^{XY}_{22} - \lambda^{XY}_{12} - \lambda^{XY}_{21}
\end{aligned}$$

  The odds ratio only depends on the interaction terms. But if the interaction terms are all zero (independent X and Y) then the log(odds ratio) is 0, therefore the odds ratio 1, indicating independence.

Similar to ANOVA model in $I \times J$ tables we only need $(I-1) \times (J-1)$ interaction terms. (Interaction between each pair of dummies for X and Y).

Comment: The number of parameters in the saturated model is equal to the number of cells, therefore this model gives a perfect fit for any sample. Usually this is not the goal because we try to find out if a smaller set of parameters provides a more parsimonious model which describes the population already well, rather than "coding" each measurement.

Example 3
Fitting the saturated model to the data on X = rating and Y = movie results in a perfect fit which is reflected by all residuals being 0.

Cell Counts and Residuals[a,b]

rating	movie_n	Observed Count	Observed %	Expected Count	Expected %	Residual	Standardized Residual	Adjusted Residual	Deviance
12	1.00	1932.500	.6%	1932.500	.6%	.000	.000	.000	.000
	2.00	6758.500	2.0%	6758.500	2.0%	.000	.000	.000	.000
34	1.00	1760.500	.5%	1760.500	.5%	.000	.000	.000	.000
	2.00	2702.500	.8%	2702.500	.8%	.000	.000	.000	.000
56	1.00	8403.500	2.5%	8403.500	2.5%	.000	.000	.	.000
	2.00	9473.500	2.9%	9473.500	2.9%	.000	.000	.000	.000
78	1.00	54233.500	16.4%	54233.500	16.4%	.000	.000	.000	.000
	2.00	49114.500	14.8%	49114.500	14.8%	.000	.000	.000	.000
910	1.00	83874.500	25.3%	83874.500	25.3%	.000	.000	.	.000
	2.00	113185.5	34.1%	113185.5	34.1%	.000	.000	.000	.000

a. Model: Poisson

b. Design: Constant + rating + movie_n + rating * movie_n

The observed are the same as the expected counts.

When reproducing the output with SPSS, one finds that for example

$$\hat{\lambda}^{XY}_{12,Alien} = -0.952, \quad \text{with} \quad e^{-0.952} = 0.39,$$

$$\hat{\lambda}^{XY}_{12,Wall-e} = 0,$$

$$\hat{\lambda}^{XY}_{910,Alien} = 0, \text{ and}$$

$$\hat{\lambda}^{XY}_{910,Wall-e} = 0$$

Therefore, we estimate that the odds ratio for 1,2 versus 9,10 and Alien versus Wall-e is $OR = e^{-0.952+0-0-0} = 0.39$.
The odds for a rating of 1,2 versus a rating of 9,10 is 61% lower for Alien than for Wall-e. (61%=100% -39%)
This is the same result we get, if we use the numbers from the contingency table to directly calculate the OR. This is not surprising since we are using the saturated model, which reproduces the data perfectly.

1.2 Loglinear Models of 3-way Tables

We now consider three categorical variables, X, Y, Z, at I, J, K levels respectively.
Recall that the purpose of studying loglinear models for contingency tables is to study the association pattern of the variables creating the table.
In this section we will present different models implying different degrees of association and independence. In statistics we are then using sample data to choose between these models the most parsimonious one which is a good description of the population. The dependency structure implied by the chosen model then informs us how the variables studied relate to each other. To understand the dependency structure for each model the implications for the conditional odds ratios are presented,

The **saturated model** for a three-way table includes the intercept, main effects, two-way interaction and three way interaction terms:

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$$

As for two-way tables, the saturated model has the same number of parameters as observed cell counts and always will fit the data perfectly.
The odds ratios describing the relationship between two of the variables can be different for the different levels of the third variable.

Example 4
Movie - rating - gender:
Male

	Rating					
Movie	1-2	3-4	5-6	7-8	9-10	Total
Alien	1502	1404	7090	49158	77097	136251
Wall-é	5654	2261	8199	43116	94079	153309
Total	7156	3665	15289	92274	171176	289560

Female

	Rating					
Movie	1-2	3-4	5-6	7-8	9-10	Total
Alien	430	356	1313	5075	6777	13951
Wall-é	1104	441	1274	5998	19106	27923
Total	1534	797	2587	11073	25883	41874

Fitting the saturated model will result in estimates perfectly match the data and all residuals are zero again.
The table shows the top of the following table provides some estimates for this model:

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	11.452	.003	3512.569	.000	11.446	11.458
[movie_n = 1.00]	-.199	.005	-40.978	.000	-.209	-.190
[movie_n = 2.00]	0[a]	.	.	.	.	.
[rating = 12]	-2.812	.014	-205.348	.000	-2.839	-2.785
[rating = 34]	-3.728	.021	-175.198	.000	-3.770	-3.686
[rating = 56]	-2.440	.012	-211.909	.000	-2.463	-2.417
[rating = 78]	-.780	.006	-134.160	.000	-.792	-.769
[rating = 910]	0[a]	.	.	.	.	.
[sex_n = 1.00]	-1.594	.008	-200.891	.000	-1.610	-1.579
[sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 12]	-1.126	.029	-38.271	.000	-1.184	-1.069
[movie_n = 1.00] * [rating = 34]	-.277	.034	-8.079	.000	-.345	-.210
[movie_n = 1.00] * [rating = 56]	.054	.017	3.175	.001	.021	.087
[movie_n = 1.00] * [rating = 78]	.330	.008	40.301	.000	.314	.346
[movie_n = 1.00] * [rating	0[a]	.	.	.	.	.

Not all of the 20 ($2 \times 5 \times 2 = 20$) model parameters are shown in this table.

Omission of any of the terms from the saturated model will constitute a more parsimonious model. If a more parsimonious model"fits as well" as the saturated model, we learn that less parameters capture the dependency structure between the variables, and interpret the degree of independence implied.

As a general rule in model building: If a higher order term is included with the model all lower order terms for the variables involved should be included with the model. The eligible models are then hierarchical or nested within each other.

1.2.1 Models of Interest

1. **Saturated Model** (see above)
2. **Homogeneous Model**

   $$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$$

   only omit the three-way interaction term, therefore implying that the association between each pair of variables is the same at all levels of the third variable. This is called a homogeneous relationship.

   The homogeneous model implies that all odds ratios describing the relationship between two of the variables are the same at all levels of the third. This means, that there might be a relationship between each pair of variables,, but that the relationship is the same at all levels of the third variable, i.e. the third variable does not affect the relation ship between the other two.

Example 5
If the relationship between rating of movies, gender, and movie would be homogeneous it would mean that (for example) the relationship between rating and movie is the same for both genders. I.e. the rating of the movies might be different, but male and female raters rated them the same way.

To test if the relationship is homogeneous test the model fit of the following model

$$log(\mu_{ijk}) = \lambda + \lambda_i^{Gender} + \lambda_j^{Rate} + \lambda_k^{Movie} + \lambda_{jk}^{RateMovie} + \lambda_{ik}^{GenderMovie} + \lambda_{ij}^{GenderRate},$$

$$i = 1, 2, j = 1, \ldots, 5, k = 1, 2$$

For the Loglinear model the deviance (or Pearson's χ^2) compare the proposed model with the saturated model. The "loglinear model" user interface in SPSS calls the deviance likelihood ratio statistic.

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	1121.297	4	.000
Pearson Chi-Square	1162.251	4	.000

a. Model: Poisson

b. Design: Constant + movie_n + rating + sex_n + rating * movie_n + movie_n * sex_n + rating * sex_n

$H_0 : \lambda_{ijk}^{XYZ} = 0$ for all i, j, k in the saturated model.

At significance level of 5% the data provide sufficient evidence (Pearson $\chi^2 = 1162.3, df = 4, P < 0.001$, or deviance= $1121.3, df = 4, P < 0.001$) to reject that the threeway interaction terms are all 0, and find that the relationship between gender, rating, and movie is not homogeneous.

It is concluded that the relationship between movie and rating depends on the gender of the rater. To illustrate, we can find from the contingency table, that

$$\text{OR(rating 1 versus 5} \mid \text{male)} = \frac{1502(94079)}{5654(77097)} = 0.32, \text{ and}$$

$$\text{OR(rating 1 versus 5} \mid \text{female)} = \frac{430(19106)}{1104(6777)} = 1.09.$$

The odds to rate Alien 1-2 rather than 9-10 is for male raters 68% lower for Alien than for Wall-é, but the same odds are for female 10% higher for Alien than for Wall-é. Illustrating the non-homogeneous relationship.

3. **Model of Conditional Independence:**

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$$

For this model remove the three-way interaction and the two-way interaction term for Y and Z from the saturated model. Which means that the model implies a homogeneous relationship between X, Y, and Z, and that Y and Z are independent conditioned under X, i.e. the relationship between Y and Z vanishes when keeping the level of X constant.

The homogeneous model implied that the odds ratio for every choice of two categories from Y and from Z is the same for all levels of X, the conditional independence now means that these odds ratios are one.

Example 6
This example illustrates the process of testing for conditional independence. We already know from the previous result, that the relationship between movie, gender, and rating is not homogeneous, so it can also not be conditional independent on any of the variables.

To test if the relationship between rating and the raters gender is conditionally independent from movie, i.e. for each movie gender and rating are independent, i.e. each movie is rated the same by male and female raters, test the model fit for

$$log(\mu_{ijk}) = \lambda + \lambda_i^{Gender} + \lambda_j^{Rate} + \lambda_k^{Movie} + \lambda_{ij}^{RateMovie} + \lambda_{ik}^{GenderMovie}, \quad i = 1, 2, j = 1, \ldots, 5, k = 1, 2$$

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	1599.934	8	.000
Pearson Chi-Square	1788.937	8	.000

a. Model: Poisson

b. Design: Constant + movie_n + rating + sex_n + rating * movie_n + movie_n * sex_n

This table gives the information for testing $H_0 : \lambda_{jk}^{RateGender} = \lambda_{ijk}^{RateGenderMovie} = 0$ in the saturated model.

At significance level of 5% the data provide sufficient evidence (Pearson $\chi^2 = 1788.9, df = 8, P < 0.001$, or deviance= $1599.9, df = 8, P < 0.001$) that gender and rating are not independent for both movies, i.e. female and male raters do not rate both movies the same.

One should have been able to predict this result from the previous test. Since the relationship is not homogeneous each pair of two variables can not be conditionally independent.

If the model would have been found to be homogeneous, one should test for conditional independence of each pair of variables, (gender, rating), (gender, movie), and (rating, movie).

4. **The Independence Model**

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$$

implies that the variables X, Y and Z are all independent. Therefore all odds ratios imaginable are 1.

1.2.2 Process: Loglinear Regression Analysis of three way tables:

1. test if the relationship is homogeneous,
   if not stop, otherwise continue with step 2

2. test for conditional independence for each pair of variables,
   if more than one pair is conditionally independent, fit model removing the two-way interaction terms for all those pairs and test model fit. If all pairs are conditionally independent continue with step 3.

3. test for independence of all three variables.

Instead of always comparing with the saturated model it is meaningful to test for conditional independence in an homogeneous model, i.e. to test $H_0 : \lambda_{ij}^{XY} = 0$. For this test one would use as test statistic the difference in the log likelihood ratio (deviance) for the homogeneous model and the model implying conditional independence for X and Y.

1.2.3 Estimating odds ratios

Assume the saturated model holds true:

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$$

with X=Rating Y=Movie Z=Gender. The odds ratio for rating a movie 7,8 (code=4) rather than 9,10 (code=5) comparing female with male raters for Movie=Wall-e

$$\begin{aligned}
log(odds) &= log(\tfrac{\mu_{4Wf}\mu_{5Wm}}{\mu_{5Wf}\mu_{4Wm}}) \\
&= \log \mu_{4Wf} + \log \mu_{5Wm} - \log \mu_{5Wf} - \log \mu_{4Wm} \\
&= \quad \lambda + \lambda_4^X + \lambda_W^Y + \lambda_f^Z + \lambda_{4W}^{XY} + \lambda_{4f}^{XZ} + \lambda_{Wf}^{YZ} + \lambda_{4Wf}^{XYZ} \\
&\quad +\lambda + \lambda_5^X + \lambda_W^Y + \lambda_m^Z + \lambda_{5W}^{XY} + \lambda_{5m}^{XZ} + \lambda_{Wm}^{YZ} + \lambda_{5Wm}^{XYZ} \\
&\quad -\lambda - \lambda_5^X - \lambda_W^Y - \lambda_f^Z - \lambda_{5W}^{XY} - \lambda_{5f}^{XZ} - \lambda_{Wf}^{YZ} - \lambda_{5Wf}^{XYZ} \\
&\quad -\lambda - \lambda_4^X - \lambda_W^Y - \lambda_m^Z - \lambda_{4W}^{XY} - \lambda_{4m}^{XZ} - \lambda_{Wm}^{YZ} - \lambda_{4Wm}^{XYZ} \\
&= \lambda_{4f}^{XZ} + \lambda_{5m}^{XZ} - \lambda_{5f}^{XZ} - \lambda_{4m}^{XZ} + \lambda_{4Wf}^{XYZ} + \lambda_{5Wm}^{XYZ} - \lambda_{5Wf}^{XYZ} - \lambda_{4Wm}^{XYZ}
\end{aligned}$$

This is estimated to be

$$\begin{aligned}
log(\widehat{odds}) &= \log \widehat{\mu}_{4Wf} + \log \widehat{\mu}_{5Wm} - \log \widehat{\mu}_{5Wf} - \log \widehat{\mu}_{4Wm} \\
&= \hat{\lambda}_{4f}^{XZ} + \hat{\lambda}_{5m}^{XZ} - \hat{\lambda}_{5f}^{XZ} - \hat{\lambda}_{4m}^{XZ} + \hat{\lambda}_{4Wf}^{XYZ} + \hat{\lambda}_{5Wm}^{XYZ} - \hat{\lambda}_{5Wf}^{XYZ} - \hat{\lambda}_{4Wm}^{XYZ}
\end{aligned}$$

1.2.4 Example for finding the appropriate loglinear model

Example from Carolyn J. Anderson, Department of Educational Psychology, Illinois. (Could not locate the original article.)
With this data the relationship between Worker's and Supervisor's job satisfaction in companies with good and bad management shall be compared
The data:

Bad Management

Worker's Satisfaction

		Low	High	
Supervisor's	Low	103	87	190
Satisfaction	High	32	42	74
		135	129	264

Good Management

Worker's Satisfaction

	Low	High	
Low	59	109	168
High	78	205	283
	137	314	451

To describe the association between the job satisfaction of workers and supervisors compute the conditional odds ratios for bad and good management:

$\hat{\theta}_{bad} = (103 \times 42)/(32 \times 87) = 1.554$

$\hat{\theta}_{good} = (59 \times 205)/(78 \times 109) = 1.423$

with 95% confidence intervals (from SPSS (descriptives>crosstab, check risk in statistics)

θ_{bad} : (.904; 2.670)

θ_{good}: (.944; 2.144)

The estimates are similar and the confidence intervals overlap a lot, which probably means that it is reasonable to assume that the odds ratios for bad and good management are not different, the relationship between the three variables seems to be homogeneous.
But also 1 is included in the confidence intervals, therefore it seems that the conditional odds ratios could be one, which could mean that supervisor and worker are conditionally independent, and the interaction for worker and supervisor will not be necessary in the model. We have conditional independence between supervisor and worker (conditioned on management).
The models: (X = Management, Y = Supervisor, Z= Worker)

1. Saturated:

   $log(\mu_{ijk}) = \lambda + \lambda_i^M + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$

2. Homogeneous:

   $log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$

3. Conditional independence of Supervisor and Worker (on management):

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$$

4. Independent:

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$$

Actual and expected cell counts for the different models

Manage	Superv.	Worker	observed	saturated	homog.	cond. indep. sprvsr-wrkr	independ.
bad	low	low	103.00	103.00	102.26	97.16	50.29
bad	low	high	87.00	87.00	87.74	92.84	81.90
bad	high	low	32.00	32.00	32.74	37.84	50.15
bad	high	high	42.00	42.00	41.26	36.16	81.67
good	low	low	59.00	59.00	59.74	51.03	85.90
good	low	high	109.00	109.00	108.26	116.97	139.91
good	high	low	78.00	78.00	77.26	85.97	85.66
good	high	high	205.00	205.00	205.74	197.03	139.52

Fitted Odds ratios for the different models:

Model	S and W (cond. on M)	M and W (cond. on S)	M and S (cond. on W)
saturated - bad, low	1.55	2.19	4.26
saturated - good, high	1.42	2.00	3.90
homogeneous	1.47	2.11	4.04
cond. indep.(super-worker)	1	2.40	4.32
independence	1	1	1

To assess the model fit for the four models we will look at measures of model fit (Pearson, deviance), residuals, and compare the models by testing which parameters are zero.
To make sure that a more parsimonious model has not been missed, all possible models and their fit are listed in the table below:

Model Fit

Model	likelihood ratio	*df*	P	Pearson	*df*	P
saturated	0	0	1	0	0	1
homogeneous	0.065	1	.799	0.065	1	0.799
cond. indep.(management)	5.387	2	0.068	5.410	2	0.067
cond. indep.(supervisor)	71.902	2	$< .001$	70.877	2	$< .001$
cond. indep.(worker)	19.711	2	$< .001$	19.884	2	$< .001$
cond. indep.(management, super)	35.597	3	$< .001$	35.719	3	$< .001$
cond. indep.(management, worker)	87.787	3	$< .001$	85.023	3	$< .001$
cond. indep.(super, worker)	102.111	3	$< .001$	99.094	3	$< .001$
independence	117.997	4	$< .001$	128.086	4	$< .001$

Model fit for the conditional independent (conditioned on management) model is acceptable. It is the most parsimonious model which is not rejected at 5% significance. It is also obvious that every other

more parsimonious model has much larger G^2 and X^2 values, indicating much larger discrepancy between data and model.
The residuals for the conditional independence model of workers and supervisor satisfaction are

Cell Counts and Residuals[a,b]

manage_n	superv_n	worker_n	Observed Count	Observed %	Expected Count	Expected %	Residual	Standardized Residual	Adjusted Residual	Deviance
.00	.00	.00	59	8.3%	51.033	7.1%	7.967	1.115	1.687	1.088
		1.00	109	15.2%	116.967	16.4%	-7.967	-.737	-1.687	-.745
	1.00	.00	78	10.9%	85.967	12.0%	-7.967	-.859	-1.687	-.873
		1.00	205	28.7%	197.033	27.6%	7.967	.568	1.684	.564
1.00	.00	.00	103	14.4%	97.159	13.6%	5.841	.593	1.601	.587
		1.00	87	12.2%	92.841	13.0%	-5.841	-.606	-1.601	-.613
	1.00	.00	32	4.5%	37.841	5.3%	-5.841	-.950	-1.600	-.976
		1.00	42	5.9%	36.159	5.1%	5.841	.971	1.600	.947

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n

The residuals do not cause any concern about the model fit.

Do the data provide sufficient evidence that in the homogeneous model the interaction term $\lambda_{ij}^{SuperWorker} \neq 0$?
Which is asking if there is an association between the supervisor and worker satisfaction for both levels of management. (From the homogeneous relationship we already know that it would be the same).
To find an answer compare the likelihood ratio statistic for the homogeneous model with the one of the model of conditional independence of worker and supervisor satisfaction.

1. $H_0 : \lambda_{ij}^{SuperWorker} = 0$ versus $H_a : \lambda_{ij}^{SuperWorker} \neq 0$, $\alpha = 0.05$
2. *dots*
3. $\chi_0^2 = 5.386 - 0.065 = 5.321, df = 2 - 1 = 1$
4. $0.01 < P - value < 0.025$ (table 7)
5. At significance level of 5% the data provide sufficient evidence that the interaction term is different from 0, and should remain in the model.

Interestingly the result is in contradiction to the previous result, that the model of conditional independence of worker and supervisor satisfaction is not rejected (p=0.068), when comparing with the saturated model.

In summary we found two "reasonable" models, the homogeneous and the conditional independent model for workers and supervisors.
Even though the test for $H_0 : \lambda_{ij}^{SuperWorker} = 0$ was significant, one might want to go with the simpler more parsimonious model, especially after reviewing the residuals again which indicate an appropriate fit for the more parsimonious model.
Also the sample size was quite big $n = 751$, which will make all tests quite sensitive and indicate statistically significant results, without pointing towards practical relevance.

The decision, which model to choose, also depends on the purpose of the model, the more complex model will provide better predictions, where the simpler model provides easier interpretations. Another criteria for choosing between the two models is Akaike's information criterion (AIC), which gives a measure of fit while correcting for the complexity of the model.

Model	AIC
homogeneous	-13.94
cond. indep.(super-worker)	-6.61

The smaller the AIC the better the model, therefore the AIC favours the homogeneous model.

Example 7

Classical example from Fienberg (1977, p. 101)

The table classifies 4991 Wisconsin male high school seniors according to socio-economic status (low, lower middle, upper middle, and high), the degree of parental encouragement they receive (low and high) and whether or not they have plans to attend college (no, yes).

Socio-economic Status, Parental Encouragement and Educational Aspirations of High School Seniors

Social	Parental	College Plans		
Stratum	Encouragement	No	Yes	Total
Lower	Low	749	357	784
	High	233	133	366
Lower Middle	Low	627	38	665
	High	330	303	633
Upper Middle	Low	420	377	457
	High	374	467	841
Higher	Low	153	26	179
	High	266	800	1066
	Total	3152	1938	4991

In the analysis of these data view all three variables as responses, and study the extent to which they are associated. In this process we test various hypotheses of complete and partial independence.