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).

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

\[
\log(\mu_{ij}) = \lambda + \lambda_X^i + \lambda_Y^j, \quad i = 1, \ldots, I, j = 1, \ldots, J
\]

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

- \( \lambda \) is the overall effect, it is included in the model to ensure \( \sum_{ij} \mu_{ij} = n \).
- \( \lambda_X^i \) is the (main or marginal) effect of category \( i \) of variable \( X \) on the log of the expected cell count.
- \( \lambda_Y^j \) is the (main or marginal) effect of category \( j \) of variable \( Y \) on the log of the expected cell count.

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 \( \chi^2 \) or the likelihood ratio statistic, which match the test for independence for two-way tables.

Example 1

Movie rating of Alien versus Wall-e.
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.

### 1.1 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 \( Y = j \):

\[
\log(\theta(i_1, i_2|Y=j)) = \log(P(X = i_1, Y = j)/P(X = i_2, Y = j))
\]

\[
= \log(nP(X = i_1, Y = j)/nP(X = i_2, Y = j))
\]

\[
= \log(\mu_{i_1j}/\mu_{i_2j})
\]

\[
= \log(\mu_{i_1j}) - \log(\mu_{i_2j})
\]

\[
= \lambda + \lambda^X_{i_1} + \lambda^Y_j - (\lambda + \lambda^X_{i_2} + \lambda^Y_j)
\]

\[
= \lambda^X_{i_1} - \lambda^X_{i_2}
\]

Important: Observe that the conditional log odds do not depend on \( j \), which shows that the model truly implies that \( X \) and \( Y \) are independent.

From above it also follows that for the independence model 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 one category is omitted and interpreted as a reference category. This is similar to using one less dummy variable than levels of the categorical predictor to be included with a model.
1.2 Saturated Model versus Independence Model

Many times the variables under consideration are not independent, in this case we need to include the interaction term with the model and obtain perfect fit.

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

is called the saturated model.

- The conditional log odds, \( \log(\theta(i_1, i_2|Y = j)) \), for falling into category \( X = i_1 \) versus \( X = i_2 \), when \( Y = j \) in this model is equal to \( \lambda^X_{i_1} + \lambda^{XY}_{i_1j} - \lambda^X_{i_2} - \lambda^{XY}_{i_2j} \) depends in this case on the level of \( Y \).

- \( \lambda^{XY}_{ij} \) 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 two categorical variables.

For 2 × 2 tables the odds ratio can be expressed in terms of the parameters of the saturated model

\[
\log(\theta) = \log \left( \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} \right) = \log(\mu_{11}) + \log(\mu_{22}) - \log(\mu_{12}) - \log(\mu_{21}) \\
= \lambda^{XY}_{11} + \lambda^{XY}_{22} - \lambda^{XY}_{12} - \lambda^{XY}_{21}
\]

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, rather than coding each measurement.

Example 2
Fitting the saturated model to the data on \( X = \)rating and \( Y = \)movie results in a perfect fit which reflected by all residuals being 0.
The observed are the same as the expected counts.

If you reproduce the output with SPSS, you find that for example

\[ \lambda^{XY}_{12} = -0.952, \quad \text{with} \quad e^{-0.952} = 0.39 \]

Therefore we estimate that OR(12,910) = 0.39 for Alien, over Walle.
Which means that the odds for a 1,2 rating rather than a 9,10 rating for Alien is 0.39 times the odds for Walle (61% less).
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.

### 1.3 Loglinear Models for three-way tables

We now consider three categorical variables, \( X, Y, Z \), at \( I, J, K \) levels respectively.

The saturated model for a three-way table includes three way interaction terms.

\[
\log(\mu_{ijk}) = \lambda_i + \lambda_j + \lambda_k + \lambda_{ij} + \lambda_{ik} + \lambda_{jk} + \lambda_{ijk}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K
\]

Again the saturated model has the same number of parameters as observed cell counts and always will fit the data perfectly.
Example 3
Movie - rating - sex: Male

<table>
<thead>
<tr>
<th>Movie</th>
<th>1-2</th>
<th>3-4</th>
<th>5-6</th>
<th>7-8</th>
<th>9-10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alien</td>
<td>1502</td>
<td>1404</td>
<td>7090</td>
<td>49158</td>
<td>77097</td>
<td>136251</td>
</tr>
<tr>
<td>Wall-e</td>
<td>5654</td>
<td>2261</td>
<td>8199</td>
<td>43116</td>
<td>94079</td>
<td>153309</td>
</tr>
<tr>
<td>Total</td>
<td>7156</td>
<td>3665</td>
<td>15289</td>
<td>92274</td>
<td>171176</td>
<td>289560</td>
</tr>
</tbody>
</table>

Female

<table>
<thead>
<tr>
<th>Movie</th>
<th>1-2</th>
<th>3-4</th>
<th>5-6</th>
<th>7-8</th>
<th>9-10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alien</td>
<td>430</td>
<td>356</td>
<td>1313</td>
<td>5075</td>
<td>6777</td>
<td>13951</td>
</tr>
<tr>
<td>Wall-e</td>
<td>1104</td>
<td>441</td>
<td>1274</td>
<td>5998</td>
<td>19106</td>
<td>27923</td>
</tr>
<tr>
<td>Total</td>
<td>1534</td>
<td>797</td>
<td>2587</td>
<td>11073</td>
<td>25883</td>
<td>41874</td>
</tr>
</tbody>
</table>

Fitting the saturated model will give that all residuals are zero again. The table shows the first part of the estimates for this model.

<table>
<thead>
<tr>
<th>Parameter Estimates b,c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>[movie_n = 1.00]</td>
</tr>
<tr>
<td>[movie_n = 2.00]</td>
</tr>
<tr>
<td>[rating = 1.2]</td>
</tr>
<tr>
<td>[rating = 5.5]</td>
</tr>
<tr>
<td>[rating = 7.8]</td>
</tr>
<tr>
<td>[rating = 9.0]</td>
</tr>
<tr>
<td>[sex_n = 1.00]</td>
</tr>
<tr>
<td>[sex_n = 2.00]</td>
</tr>
<tr>
<td>[movie_n = 1.00] * [rating = 1.2]</td>
</tr>
<tr>
<td>[movie_n = 1.00] * [rating = 3.4]</td>
</tr>
<tr>
<td>[movie_n = 1.00] * [rating = 5.5]</td>
</tr>
<tr>
<td>[movie_n = 1.00] * [rating = 7.8]</td>
</tr>
<tr>
<td>[movie_n = 1.00] * [rating]</td>
</tr>
</tbody>
</table>

Not all of the 20 ( 2 × 5 × 2 = 20 ) model parameters are shown in this table.
Omission of any of the terms in the model will constitute a more parsimonious model.

**As a rule:** If a higher order term is included with the model all lower order terms for the variables involved should be included with the model. Such models are called hierarchical.

**The different Models**

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 omits the three-way interaction term, therefore implying that the association between each pair of two variables is the same for all levels of the third variable, which is called a homogeneous relationships.

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.

**Example 4**

If the relationship between rating of movies, sex of the rater, and the choice of movie is homogeneous it means that (for example) the effect of movie on the rating is the same for both sexes. To test if the relationship is homogeneous test the model fit of the following model

\[
\log(\mu_{ijk}) = \lambda + \lambda_i^{Sex} + \lambda_j^{Rate} + \lambda_k^{Movie} + \lambda_{ij}^{RateMovie} + \lambda_{ik}^{SexMovie} + \lambda_{ijk}^{SexRate},
\]

\[i = 1, 2, j = 1, \ldots, 5, k = 1, 2\]

For the Loglinear model the deviance (or Pearson’s \(\chi^2\)) compare the proposed model with the saturated model. The “loglinear model” user interface in SPSS calls the deviance likelihood ratio statistic.

<table>
<thead>
<tr>
<th>Goodness-of-Fit Tests (^{ab})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio 1121.297</td>
</tr>
<tr>
<td>Pearson Chi-Square 1162.251</td>
</tr>
</tbody>
</table>

\(^{a}\) Model: Poisson

\(^{b}\) Design: Constant + movie_n + rating + sex_n +
rate_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 three-way interaction terms are all 0, and find that the relationship between Sex, Rating, and movie is not homogeneous.

It is concluded that the relationship between movie and rating depends on the sex of the rater. To illustrate, we can find from the contingency table, that
OR(rating 1 versus 5 | male) = \frac{1502(94079)}{5654(77097)} = 0.32, and

OR(rating 1 versus 5 | 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 Walle, but the same odds are for female 10% higher for Alien than for Walle. 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 it implies a homogeneous relationship between X, Y, and Z, but also that Y and Z are independent conditioned under X. (Which means that the relationship between Y and Z vanishes when keeping the level of X constant.)

In other words the odds ratio for every choice of two categories from Y and from Z is one for all levels of X.

Example 5

This example only illustrates the process. We already know from the results above, that a more parsimonious than the saturated model will not fit well, since the relationship between the three variables is not homogeneous.

To test if the rating of movies is independent from the raters sex is conditionally independent from movie, i.e. for each movie sex and rating are independent, test the model fit for

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

| Goodness-of-Fit Tests\(^a\)\(^b\) |
|-------------------|-----|-----|
| Likelihood Ratio  | 1599.934 | 8  | .000 |
| Pearson Chi-Square | 1788.937 | 8  | .000 |

This table gives the information for testing \( H_0 : \lambda_{jk}^{RateSex} = \lambda_{ijk}^{RateSexMovie} = 0 \) 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 Sex and Rating are not independent for both movies.

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.
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 the odds ratios are all 1.

**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 \(Y\).

### 1.4 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 \(X=\text{Rating} \quad Y=\text{Movie} \quad Z=\text{Sex}\). 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

\[
\log(\text{odds}) = \log(\hat{\mu}_{4Wf}/\hat{\mu}_{5Wm}) = \log \hat{\mu}_{4Wf} + \log \hat{\mu}_{5Wm} - \log \hat{\mu}_{5Wf} - \log \hat{\mu}_{4Wm}
\]

\[
= \lambda + \lambda_4^X + \lambda_5^Y + \lambda_5^Z + \lambda_{4f}^{XY} + \lambda_{4m}^{XZ} + \lambda_{4w}^{YZ} + \lambda_{4Wf}^{XYZ} + \lambda_5^X + \lambda_5^Y + \lambda_5^Z + \lambda_{5f}^{XY} + \lambda_{5m}^{XZ} + \lambda_{5w}^{YZ} + \lambda_{5Wf}^{XYZ} - \lambda_{5f}^{XY} - \lambda_{5m}^{XZ} - \lambda_{5w}^{YZ} - \lambda_{5Wf}^{XYZ}
\]

\[
= \lambda_{4f}^{XZ} + \lambda_{5m}^{XZ} - \lambda_{5f}^{XZ} - \lambda_{4m}^{XZ} + \lambda_{4Wf}^{XYZ} + \lambda_{5Wm}^{XYZ} - \lambda_{5Wf}^{XYZ} - \lambda_{4Wm}^{XYZ}
\]

This is estimated to be

\[
\log(\hat{\text{odds}}) = \log(\hat{\mu}_{4Wf}/\hat{\mu}_{5Wm}) = \lambda_{4f}^{XZ} + \lambda_{5m}^{XZ} - \lambda_{5f}^{XZ} - \lambda_{4m}^{XZ} + \lambda_{4Wf}^{XYZ} + \lambda_{5Wm}^{XYZ} - \lambda_{5Wf}^{XYZ} - \lambda_{4Wm}^{XYZ}
\]
1.5 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:

<table>
<thead>
<tr>
<th>Supervisor’s Satisfaction</th>
<th>Bad Management</th>
<th>Good Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>103 87 190</td>
<td>Low 59 109 168</td>
</tr>
<tr>
<td>High</td>
<td>32 42 74</td>
<td>High 78 205 283</td>
</tr>
<tr>
<td></td>
<td>135 129 264</td>
<td></td>
</tr>
</tbody>
</table>

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

\[ \hat{\theta}_{bad} = \frac{(103 \times 42)}{(32 \times 87)} = 1.554 \]

\[ \hat{\theta}_{good} = \frac{(59 \times 205)}{(78 \times 109)} = 1.423 \]

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

\[ \theta_{bad} : (.904; 2.670) \]

\[ \theta_{good} : (.944; 2.144) \]

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: \( X = \text{Management}, Y = \text{Supervisor}, Z = \text{Worker} \)

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:

\[ \log(\mu_{ijk}) = \lambda_i + \lambda^X_j + \lambda^Y_k, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K \]

Actual and expected cell counts for the different models

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

Fitted Odds ratios for the different models:

<table>
<thead>
<tr>
<th>Model</th>
<th>S and W (cond. on M)</th>
<th>M and W (cond. on S)</th>
<th>M and S (cond. on W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>saturated - bad, low</td>
<td>1.55</td>
<td>2.19</td>
<td>4.26</td>
</tr>
<tr>
<td>saturated - good, high</td>
<td>1.42</td>
<td>2.00</td>
<td>3.90</td>
</tr>
<tr>
<td>homogeneous</td>
<td>1.47</td>
<td>2.11</td>
<td>4.04</td>
</tr>
<tr>
<td>cond. indep.(super-worker)</td>
<td>1</td>
<td>2.40</td>
<td>4.32</td>
</tr>
<tr>
<td>independence</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
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.

**Model Fit**

<table>
<thead>
<tr>
<th>Model</th>
<th>likelihood ratio = deviance</th>
<th>df</th>
<th>P</th>
<th>Pearson</th>
<th>df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>saturated</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>homogeneous</td>
<td>0.065</td>
<td>1</td>
<td>.799</td>
<td>0.065</td>
<td>1</td>
<td>0.799</td>
</tr>
<tr>
<td>cond. indep.(super-worker)</td>
<td>5.387</td>
<td>2</td>
<td>0.068</td>
<td>5.410</td>
<td>2</td>
<td>0.067</td>
</tr>
<tr>
<td>independence</td>
<td>117.997</td>
<td>4</td>
<td>&lt; 0.001</td>
<td>128.086</td>
<td>4</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Model fit for the conditional independent model is acceptable (the model is not rejected at 5% significance).

The residuals for the conditional independence model of workers and supervisor satisfaction are

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

Do the data provide sufficient evidence that the interaction term $\lambda_{ij}^{SuperWorker} \neq 0$?

Which is asking if there is an association between the supervisor and worker satisfaction when correcting for the style of management (good versus bad).

To find an answer compare the likelihood ratio statistic for the homogeneous model and the model of conditional independence of worker and supervisor satisfaction.

1. $H_0 : \lambda_{ij}^{SuperWorker} = 0$ versus $H_a : \lambda_{ij}^{Super Worker} \neq 0$, $\alpha = 0.05$

2. $dofs$

3. $\chi^2 = 5.386 - 0.065 = 5.321$, $df = 2 - 1 = 1$

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.

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 $H_0 : \lambda_{ij}^{\text{SuperWorker}} = 0$ was significant, one might want to go with the simpler model, especially after reviewing the residuals again which indicate an appropriate fit for the simpler 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>homogeneous</td>
<td>-13.94</td>
</tr>
<tr>
<td>cond. indep.(super-worker)</td>
<td>-6.61</td>
</tr>
</tbody>
</table>

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