Recap: Interaction terms in the log-linear model
- Main (binary) effects from JMP = $2 \times \hat{\lambda}_i^X$ => odds (sign tells you direction of effect)
- Interaction (binary) effects from JMP = $4 \times \hat{\lambda}_{ij}^{XY}$ => odds ratio
- $H_0$: $\hat{\lambda}_{ij} = 0$ for all $i, j$ (no association between $X$ and $Y$)
  - $df =$ number of nonredundant parameters $(I - 1)(J - 1)$
  - equivalent to goodness of fit test from the independence model
- Remember – risky to interpret the “main effects” if the model includes interactions.

Example 1: These data give the frequencies of infant vitality in a group of German women who were either pregnant for the first time or had experienced complications in earlier pregnancies. The data (Wermuth, 1976; McCullagh and Nelder, 1989; Zelterman, 2002) are cross-classified by four binary variables: gestational age, maternal age, smoking habit, and whether or not the infant lives.

<table>
<thead>
<tr>
<th>Gestational age (days)</th>
<th>Mother’s age (years)</th>
<th>Cigarettes smoked/day</th>
<th>Perinatal mortality</th>
<th>Live births</th>
</tr>
</thead>
<tbody>
<tr>
<td>197-260</td>
<td>&lt; 30</td>
<td>≤ 5</td>
<td>50</td>
<td>315</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6+</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>30+</td>
<td>≤ 5</td>
<td>41</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6+</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>261+</td>
<td>&lt; 30</td>
<td>≤ 5</td>
<td>24</td>
<td>4012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6+</td>
<td>6</td>
<td>459</td>
</tr>
<tr>
<td></td>
<td>30+</td>
<td>≤ 5</td>
<td>14</td>
<td>1494</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6+</td>
<td>1</td>
<td>124</td>
</tr>
</tbody>
</table>

(a) How could you use logistic regression with these data? What information would you not be able to gain from logistic regression?

(b) Suppose we looked at just smoking and maternal age for the live births. Find the conditional odds ratios at the two levels of gestational age:

Gestational age 197-260:

\[
\begin{array}{ccc}
& <30 & 30+ \\
<5 & 5 & 315 \\
6+ & 40 & 11 \\
\end{array}
\]

$OR = 0.589$

Gestational age 261+:

\[
\begin{array}{ccc}
& <30 & 30+ \\
<5 & 5 & 199 \\
6+ & 49 & 124 \\
\end{array}
\]

$OR = 0.725$

(b) What is the two-way table if we collapse across the gestational age?

<table>
<thead>
<tr>
<th></th>
<th>&lt;30</th>
<th>30+</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤5</td>
<td>4327</td>
<td>1641</td>
</tr>
<tr>
<td>6+</td>
<td>699</td>
<td>135</td>
</tr>
</tbody>
</table>

Find the marginal odds ratio of smoking and age.

$OR = 0.7134$
Recall:
- The marginal odds ratio ignores the third factor, whereas the conditional odds ratio controls for it.
- Homogenous XY association implies that the odds ratio between X and Y is the same for every
  value of Z. (This is also symmetric: homogenous XY association implies homogenous XZ
  association and homogenous YX association.)
- Conditional independence says that common odds ratio = 1, does not imply marginal independence.

(c) If maternal age and smoking have homogenous association, roughly what value would you estimate

for this common odds ratio?

\[ \text{something between, probably close to} \, .7 \]

(d) Open the InfantVitality.jmp data file. How many rows are there? Fit a loglinear model with “mutual

independence of the four factors” (just the main effects). How parameters are in this model? Is this

model a good fit for the data?

\[ \text{model: GMCV} \quad 5 \text{ parameters} \]

(e) Fit a loglinear model including all two-way interactions. (Shortcut is to highlight the 4 variables and

then choose Macros > Factorial to Degree, which is set to 2.) Does (GM, GC, GV, MC, MV, CV) appear

to be a reasonable model for these data? What does this imply?

\[ \text{yes} \Rightarrow \text{don’t need + three way interactions} \]

(f) Use the pull-down menu to select Save Columns > Predicted values. Look at the column in the data

sheet. Are the predicted counts similar to the observed counts? How does the residual plot look?

\[ \text{much better fit} \]

(g) Now include all possible three-way interactions (change degree to 3). How many parameters are

there? How would you interpret a three-way interaction? Are any of them statistically significant?

What do you suggest next? What does this tell you? And then? How would you interpret your final

model?

\[ \text{gestational age, maternal age, vitality all exhibit pairwise interactions} \]

\[ \text{cigarette smoking is conditionally independent of both vitality & gest. age given the maternal age } \]

(h) Using these estimated counts (from e), create the predicted two way-way tables for the different

gestational ages for the live births. Then find the conditional odds ratios.

<table>
<thead>
<tr>
<th>Gestational age</th>
<th>Cigarette use</th>
<th>Mom ≤ 30</th>
<th>Mom 30+</th>
</tr>
</thead>
<tbody>
<tr>
<td>197-260</td>
<td>&lt; 5</td>
<td>51.3</td>
<td>38.69</td>
</tr>
<tr>
<td></td>
<td>6+</td>
<td>9.166</td>
<td>4.85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gestational age</th>
<th>Cigarette use</th>
<th>Mom ≤ 30</th>
<th>Mom 30+</th>
</tr>
</thead>
<tbody>
<tr>
<td>261+</td>
<td>&lt; 5</td>
<td>24.33</td>
<td>14.69</td>
</tr>
<tr>
<td></td>
<td>6+</td>
<td>4.21</td>
<td>1.78</td>
</tr>
</tbody>
</table>

\[ \text{OR} = .70 \]

\[ \text{OR} = .70 \]