Topics in statistics III/IV (MATH3361/4071) Michaelmas term, 2020-2021

Exercises: Log linear models

Lecturer & Author: Georgios Karagiannis georgios.karagiannis@durham.ac.uk

---

Exercise 1. Assume model (X, Y, Z), under Poisson sampling scheme. Consider corner points constraints.

1. Find the Likelihood equations.
2. Express the Log linear coefficients with respect to the expected counts
3. Find the MLEs of the Log linear coefficients

Solution.

1. I consider as a reference levels the I, J, K ones.

   The likelihood function is

$$\begin{aligned}\ell(\lambda) =& n\lambda + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z \\ & - \sum_{i,j,k} \underbrace{\exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z)}_{=\mu_{ijk}(\lambda)}\end{aligned}$$

   In order to find the likelihood equations, I need to maximize the likelihood, under any possible constraints, such as the corner points in this case. I can do this by using the Lagrange multipliers method

   The Lagrange multipliers functin is

$$\begin{aligned}\mathcal{L}(\lambda, \theta) =& n\lambda + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z \\ & - \sum_{i,j,k} \underbrace{\exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z)}_{=\mu_{ijk}(\lambda)} \\ & \underbrace{-\theta^X(\lambda_I^X) - \theta^Y(\lambda_J^Y) - \theta^Z(\lambda_K^Z)}_{=> \text{ identifiability constraints}}\end{aligned}$$

   Then, to maximize the likelihood under the corner points constraints I do:

$$0 = \nabla_{\underset{\sim}{\lambda},\underset{\sim}{\theta}} \mathcal{L}(\underset{\sim}{\lambda}, \underset{\sim}{\theta})|_{(\underset{\sim}{\lambda},\underset{\sim}{\theta})=(\hat{\underset{\sim}{\lambda}},\hat{\underset{\sim}{\theta}})} \implies$$

$$
\begin{aligned}
0 =& \frac{\mathrm{d}}{\mathrm{d}\lambda}\mathcal{L}(\underset{\sim}{\lambda},\underset{\sim}{\theta})|_{(\underset{\sim}{\lambda},\underset{\sim}{\theta})=(\hat{\underset{\sim}{\lambda}},\hat{\underset{\sim}{\theta}})} \Longrightarrow n = \mu_{+++}(\hat{\lambda});\\
0 =& \frac{\mathrm{d}}{\mathrm{d}\lambda_i^X}\mathcal{L}(\underset{\sim}{\lambda},\underset{\sim}{\theta})|_{(\underset{\sim}{\lambda},\underset{\sim}{\theta})=(\hat{\underset{\sim}{\lambda}},\hat{\underset{\sim}{\theta}})} \Longrightarrow n_{i++} = \mu_{i++}(\hat{\lambda}) + \hat{\theta}^X;\\
0 =& \frac{\mathrm{d}}{\mathrm{d}\lambda_j^Y}\mathcal{L}(\underset{\sim}{\lambda},\underset{\sim}{\theta})|_{(\underset{\sim}{\lambda},\underset{\sim}{\theta})=(\hat{\underset{\sim}{\lambda}},\hat{\underset{\sim}{\theta}})} \Longrightarrow n_{+j+} = \mu_{+j+}(\hat{\lambda}) + \hat{\theta}^Y;\\
0 =& \frac{\mathrm{d}}{\mathrm{d}\lambda_k^Z}\mathcal{L}(\underset{\sim}{\lambda},\underset{\sim}{\theta})|_{(\underset{\sim}{\lambda},\underset{\sim}{\theta})=(\hat{\underset{\sim}{\lambda}},\hat{\underset{\sim}{\theta}})} \Longrightarrow n_{++k} = \mu_{++k}(\hat{\lambda}) + \hat{\theta}^Z;
\end{aligned}
$$

$$
\begin{aligned}
0 = \frac{\mathrm{d}}{\mathrm{d}\theta^X}\mathcal{L}(\underset{\sim}{\lambda},\underset{\sim}{\theta})|_{(\underset{\sim}{\lambda},\underset{\sim}{\theta})=(\hat{\underset{\sim}{\lambda}},\hat{\underset{\sim}{\theta}})} \Longrightarrow 0 = \lambda_I^X; \quad 0 = \frac{\mathrm{d}}{\mathrm{d}\theta^Y}\mathcal{L}(\underset{\sim}{\lambda},\underset{\sim}{\theta})|_{(\underset{\sim}{\lambda},\underset{\sim}{\theta})=(\hat{\underset{\sim}{\lambda}},\hat{\underset{\sim}{\theta}})} \Longrightarrow 0 = \lambda_J^Y\\
0 = \frac{\mathrm{d}}{\mathrm{d}\theta^Z}\mathcal{L}(\underset{\sim}{\lambda},\underset{\sim}{\theta})|_{(\underset{\sim}{\lambda},\underset{\sim}{\theta})=(\hat{\underset{\sim}{\lambda}},\hat{\underset{\sim}{\theta}})} \Longrightarrow 0 = \lambda_K^Z;
\end{aligned}
$$

From the above, I get

$$
\begin{aligned}
0 =& \hat{\theta}^X = \hat{\theta}^Y = \hat{\theta}^Z;\\
n =& \mu_{+++}(\hat{\lambda}); && (1)\\
n_{i++} =& \mu_{i++}(\hat{\lambda}); && (2)\\
n_{+j+} =& \mu_{+j+}(\hat{\lambda}); && (3)\\
n_{++k} =& \mu_{++k}(\hat{\lambda}) && (4)\\
0 =& \lambda_I^X = \lambda_J^Y = \lambda_K^Z && (5)
\end{aligned}
$$

but I only need (4), and (5). This is because(4), can reproduce (1), (3), and (3).

2. The log linear model equation for model (X, Y, Z) is:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \tag{6}$$

Then under the aforesaid corner point constraints, and by using (6), I notice that

$$
\begin{aligned}
\log(\mu_{IJK}) =& \lambda + \cancelto{O}{\lambda_I^X} + \cancelto{0}{\lambda_J^Y} + \cancelto{0}{\lambda_K^Z}\\
\log(\mu_{IJk}) =& \lambda + \cancelto{0}{\lambda_I^X} + \cancelto{0}{\lambda_J^Y} + \lambda_k^Z\\
&\text{etc.....}
\end{aligned}
$$

So, I get

$$\begin{aligned}
\lambda &= \log(\mu_{IJK}) \\
\lambda_k^Z &= \log(\mu_{IJk}) - \lambda = \log(\frac{\mu_{IJk}}{\mu_{IJK}}) \\
\lambda_j^Y &= \log(\mu_{IjK}) - \lambda = \log(\frac{\mu_{IjK}}{\mu_{IJK}}) \\
\lambda_i^X &= \log(\mu_{iJK}) - \lambda = \log(\frac{\mu_{iJK}}{\mu_{IJK}})
\end{aligned}$$

3. Given model (X, Y, Z) , it is

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \implies \mu_{ijk} = \frac{\mu_{i++}\mu_{+j+}\mu_{++k}}{n^2}$$

So the MLE is

$$\hat{\mu}_{ijk} = \frac{n_{i++}n_{+j+}n_{++k}}{n^2}$$

By replacing the parameters with the MLES, the MLE of the log-linear model coefficients are

$$\begin{aligned}
\hat{\lambda} &= \log(\frac{n_{I++}n_{+J+}n_{++K}}{n^2}) \\
\hat{\lambda}_k^Z &= \log(\frac{n_{++k}}{n_{++K}}) \\
\hat{\lambda}_j^Y &= \log(\frac{n_{+j+}}{n_{+J+}}) \\
\hat{\lambda}_i^X &= \log(\frac{n_{i++}}{n_{I++}})
\end{aligned}$$

The next exercise is from Homework 2

Exercise 2. The 1988 General Social Survey compiled by the National Opinion Research Center asked: "Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms. Table 4 summarizes opinions about health care costs (Y) and the information program (X), classified also by the respondent's gender (Z).

Gender (Z)	Information Opinion (X)	Health Opinion (Y)	
		Support	Oppose
Male	Support	76	160
	Oppose	6	25
Female	Support	114	181
	Oppose	11	48

Table 1: [Source: 1988 General Social Survey, National Opinion Research Center.]

Questions: Regarding the questions below, any inference based on hypothesis tests should be performed at sig. level 5%.

1. By using appropriate statistical tools, compare the following associations of the classification variables

   - [Y , X , Z are independent] vs. [X and Y are conditionally independent on Z]

   Justify the way you addressed the statistical problem.

   {Mark 50%}

2. By using appropriate statistical tools, compare the following associations of the classification variables

   - [X and Z are jointly independent from Y] vs. [X and Y are jointly independent from Z]

   Justify the way you addressed the statistical problem.

   {Mark 50%}

{Mark 100%}

Solution.

Exercise 3. Consider a $I \times J \times K$ table n_{ijk} generated from a multinomial sampling scheme; i.e.

$$(n_{111}, ..., n_{IJK}) \sim \text{Mult}(n, \underset{\sim}{\pi})$$

where $\underset{\sim}{\pi} = (\pi_{111}, ..., \pi_{IJK})$, and $n = \sum_{ijk} n_{ijk}$. Show that the rejection area of the Goodness of fit test for model M_0 based on the deviance/likelihood ratio statistic is

$$\text{RA} = \{\underbrace{2 \sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}})}_{=G^2} > q\}$$

for some value q, where $\mu_{ijk} = n\pi_{ijk}$, and $\hat{\mu}_{ijk}$ are the MLE of μ_{ijk} under modelM_0.

Solution. Well, the likelihood ratio test has rejection area

$$\frac{\sup_{\forall \mu: \text{ under } M_0} L(\underset{\sim}{\pi} = n\underset{\sim}{\mu})}{\sup_{\forall \mu: \text{ under } M_1} L(\underset{\sim}{\pi} = n\underset{\sim}{\mu})} < q'' \Longleftrightarrow \log(\frac{\sup_{\forall \mu: \text{ under } M_0} L(\underset{\sim}{\pi} = n\underset{\sim}{\mu})}{\sup_{\forall \mu: \text{ under } M_1} L(\underset{\sim}{\pi} = n\underset{\sim}{\mu})}) < q' \tag{7}$$

is model M_0 is the model of interest and model M_1 is the saturated model.

- The likelihood is
$$L(\underset{\sim}{\pi}) \propto \prod_{\forall i,j,k} \pi_{ijk}^{n_{ijk}} \propto \prod_{\forall i,j,k} \mu_{ij,k}^{n_{ij,k}}$$
- Under the saturated model the MLE (aka maximum likelihood estimate) is $\hat{\mu}_{ijk}^{\text{satur.}} = n_{ijk}$ and under the model M_0 the MLE is $\hat{\mu}_{ijk}$ (this is assumed).
- Then from (7), I get
$$\{\log(\frac{L(\underset{\sim}{\pi} = \underset{\sim}{\hat{\mu}}/n_{+++})}{L(\underset{\sim}{\pi} = \underset{\sim}{\hat{\mu}}^{\text{satur.}}/n_{+++})}) < q'\} \Longrightarrow \{\underbrace{2 \sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}})}_{=G^2} > q\}$$

Exercise 4. Consider a $I \times J \times K$ table n_{ijk} generated from a Poisson sampling scheme; i.e.

$$n_{ijk} \sim \text{Poi}(\mu_{ijk})$$

1. Show that the rejection area of the Goodness of fit test for model M_0 based on the deviance/likelihood ratio statistic is
$$\text{RA} = \{2 \sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}}) > q\} \tag{8}$$
for some value q, where $\hat{\mu}_{ijk}$ are the MLE of μ_{ijk} under modelM_0.
2. In an exercise/example in the Handout [Handouts: The Log-linear model], you where asked to compute the rejection area of the LR statistic under the Multinational sampling scheme, and in fact the resulting rejection area was the the same as in (8). State the assumption, we secretly took in order to get (8) under the Poisson sampling scheme in order to make the likelihood ratio under the Poisson sampling scheme to look like that under the Multinational sampling scheme.

Solution.

1. Well, the likelihood ratio test has rejection area (see your Statistical Concepts 2, term 1, notes)

$$\frac{\sup_{\forall \mu:\text{ under } M_0} L(\mu)}{\sup_{\forall \mu:\text{ under } M_1} L(\underset{\sim}{\mu})} < q' \Longleftrightarrow \log(\frac{\sup_{\forall \mu:\text{ under } M_0} L(\mu)}{\sup_{\forall \mu:\text{ under } M_1} L(\underset{\sim}{\mu})}) < q'' \tag{9}$$

is model M_0 is the model of interest and model M_1 is the saturated model.

- The likelihood is

$$L(\underset{\sim}{\pi}) \propto \prod_{\forall i,j,k} \exp(\mu_{ijk}) \mu_{ijk}^{n_{ijk}}$$

- Under the saturated model the MLE (aka maximum likelihood estimate) is $\hat{\mu}_{ijk}^{\text{satur.}} = n_{ijk}$ and under the model M_0 the MLE is $\hat{\mu}_{ijk}$ (this is assumed). Then from (7), I get

$$\begin{aligned}\{\log(\frac{L(\hat{\mu})}{L(\hat{\mu}^{\text{satur.}})}) < q''\} &\Longrightarrow \{\sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}_{ijk}}{n_{ijk}}) + \sum_{\forall i,j,k} \hat{\mu}_{ijk} - \sum_{\forall i,j,k} n_{ijk} < q''\} \\ &\Longrightarrow \{2 \sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}}) - 2 \sum_{\forall i,j,k} \hat{\mu}_{ijk} + 2 \sum_{\forall i,j,k} n_{ijk} > \underbrace{-2q''}_{=q}\}\end{aligned}$$

- If $\hat{\mu}_{+++} = n_{+++}$ then I get what I need to show.

2. The two sampling schemes lead to the same hypothesis test and inference provided that $\mu_{+++} = n_{+++}$ which is a reasonable assumption in our problems. Also is is worth recalling that Multinomial Distribution can derived from Poisson by conditioning on the total number of occurrences $\mu_{+++} = n_{+++}$.

Exercise 5. Consider a $2 \times 2 \times 2$ contingency table, with classification variables X, Y, Z.

1. State the equation of the Log-linear model of the model describing the dependency type (XZ, XY)
2. Apply the two types of the non-identifiability constraints (corner points and sum-to-zero)
3. Write down the number of the free parameters, and say how you calculated them
4. Consider the corner points only;
   (a) express the log ratio of $\log(\frac{\pi_{1|j,k}}{\pi_{2|j,k}})$ as a function of the linear model coefficients (aka the λ's).
   (b) express the log coditional odds ratio $\log(\theta_{(k)}^{XY})$ as a function of the linear model coefficients (aka the λ's). and give a sort interpretation about λ_{11}^{XY} based on this.

Solution.

1. Well, this model describes the dependency type: Z and Y are conditionally independent on X, it is given by the following Log linear model

$$\log(\mu_{i,j,k}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XY}; \qquad \forall (i,j,k) \in \{1,2\}^3$$

2. For the corner points I assume that the reference levels 2, 2, 2 , they are

$$\begin{aligned} \lambda_2^X = \lambda_2^Y = \lambda_2^Z = 0 \\ \lambda_{2j}^{XY} = \lambda_{i2}^{XY} = 0, \ \forall i,j \\ \lambda_{i2}^{XZ} = \lambda_{2k}^{XZ} = 0, \ \forall i,k \end{aligned}$$

   For the Sum-to-zero they are

$$\begin{aligned} \lambda_1^X &= -\lambda_2^X; & \lambda_1^Y &= -\lambda_2^Y & \lambda_1^Z &= -\lambda_2^Z \\ \lambda_{i1}^{XZ} &= -\lambda_{i2}^{XZ}; \ (\forall i) & \lambda_{1j}^{XY} &= -\lambda_{2j}^{XY}, \ \forall j \\ \lambda_{i1}^{XZ} &= -\lambda_{i2}^{XZ}; \ (\forall i) & \lambda_{1k}^{XZ} &= -\lambda_{2k}^{XZ}; \ (\forall k) \end{aligned}$$

3. The number of free parameters are

$$d = 1 + (2-1) + (2-1) + (2-1) + (2-1)(2-1) + (2-1)(2-1) = 6$$

   because

   - from λ I have 1 free parameter,
   - from λ_i^X I have $I = 2$ parameters but because of the constraints, 1 of them (e.g., in the corner points the $(I = 2)$-th) are set to a value 0 or can be specified from the others
     - some for λ_j^Y, λ_k^Z
   - from λ_{ik}^{XZ} I have $I \times K = 2 \times 2 = 4$ parameters, but because of the constraints (e.g., $\lambda_{2j}^{XY} = \lambda_{i2}^{XY} = 0$), $I + K - 1 = 3$ of them are set to a value 0 (corner points) or can be specified from the others (sum-to-zero)
     (a) some for λ_{ij}^{XY}

4. We compute,

- For λ_1^X it is

$$\log(\frac{\pi_{1|j,k}}{\pi_{2|j,k}}) = \log(\frac{\mu_{1jk}}{\mu_{2jk}}) = \log(\mu_{1jk}) - \log(\mu_{2jk}) = \begin{cases} \lambda_1^X + \lambda_{11}^{XY} + \lambda_{11}^{XZ} & j=1,\, k=1 \\ \lambda_1^X + \lambda_{11}^{XY} & j=1,\, k=2 \\ \lambda_1^X + \lambda_{11}^{XZ} & j=2,\, k=1 \\ \lambda_1^X & j=2,\, k=2 \end{cases}$$

- For λ_{11}^{XY} it is

$$\forall k \;\; \log(\theta_{(k)}^{XY}) = \log(\frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}) = \lambda_{11}^{XY}$$

  Here λ_{11}^{XY} is equal to the log conditional ratio of X, Y, given any level of Z.

PJC exercise (modified to Poisson sampling)

Exercise 6. Consider log-linear modelling of four-dimensional contingency tables, i.e. $d = 4$. Suppose that the table has R rows, C columns and S slices and H hyper-slices (values of l). Consider that the observations are collected based on a Multinomial sampling scheme.

- Write down the full model, for $\log(\pi_{ijkl}) = \log(\frac{\mu_{ijkl}}{n_{++++}}) = ...$

  Hint: In multinational sampling, the intercept coefficient in the linear predictor function is set to zero.

- How many parameters (degrees of freedom) are associated with each term in the full model?

- To what value must they all sum?

- How many different possible log-linear models for $d = 4$ are there if we don't require that models are hierarchical?

Solution.

$$\begin{aligned} \log(\pi_{ijkl}) = \log(\frac{\mu_{ijkl}}{n_{++++}}) = \eta_{ijkl} = \cancelto{=0}{\beta} + \\ \beta_i + \beta_j + \beta_k + \beta_l + \\ \beta_{ij} + \beta_{ik} + \beta_{il} + \beta_{jk} + \beta_{jl} + \beta_{kl} + \\ \beta_{ijk} + \beta_{ijl} + \beta_{ikl} + \beta_{jkl} + \\ \beta_{ijkl} \end{aligned} \tag{10}$$

where the sum over any single index, keeping any other indexes fixed, of any "β"-term yields 0.

Term	Parameters
β_i	$R-1$
β_j	$C-1$
β_k	$S-1$
β_l	$H-1$
β_{ij}	$(R-1)(C-1)$
β_{ik}	$(R-1)(S-1)$
β_{il}	$(R-1)(H-1)$
β_{jk}	$(C-1)(S-1)$
β_{jl}	$(C-1)(H-1)$
β_{kl}	$(S-1)(H-1)$
β_{ijk}	$(R-1)(C-1)(S-1)$
β_{ijl}	$(R-1)(C-1)(H-1)$
β_{ikl}	$(R-1)(S-1)(H-1)$
β_{jkl}	$(C-1)(S-1)(H-1)$
β_{ijkl}	$(R-1)(C-1)(S-1)(H-1)$

from which the total is clearly

$$[1+(R-1)][1+(C-1)][1+(S-1)][1+(H-1)]-1 = RCSH-1$$

They must sum to this because our full model is an arbitrary probability distribution for a multinomial with $R \times C \times S \times H$ categories.

If we don't require that the model is hierarchical then a model can consist of any subset of terms from (1) except that the unindexed β must always be present to ensure that all the probabilities sum to 1. There are 15 other terms (rows in the table above) and so the answer is $2^{15} = 32768$ models.

Exercise 7. The following 2×3 contingency table shows data reported by Fox et al. (1993). In this question, you will use it to work through the Iterative Proportional Fitting process for the independence model. Note that IPF is not necessary in this model as we have closed-form maximum likelihood estimates but it is being used as a simple example.

Anti-emetic response data after 2 days:

	3 Level of response		
	1 None	1 Partial	1 Complete
Control	12	3	7
Treatment	3	7	12

Find the values of $\hat{p}_i$ $(i = 1, 2)$ and $\hat{p}_j$ $(j = 1, 2, 3)$ for this table.

Starting from the "approximation" $\hat{p}_{ij}^{(0)} = 1$ for all i and j, carry out one full iteration of the IPF algorithm.

Why is there no point in carrying out further iterations?

Solution.

1. Here $B = \{i, j\}$ (or $\{1, 2\}$ if you prefer numerical representation). Consequently, the maximal terms are i and j: both are in $\mathbb{B}$; all terms are "contained" in one or other of i and j; and neither is contained in the other.

   IPF consists of alternating over maximal terms, making the marginal probability table corresponding to a maximal term match the corresponding empirical table. So, for maximal term i, the empirical probabilities are just the proportion in each row of data and we have to make the approximation have the same marginal row probabilities.

   (a) For i, the empirical probabilities are $(12 + 3 + 7)/44 = 0.5$ and $(3 + 7 + 12)/44 = 0.5$. For j, they are $15/44 = 0.341$, $10/44 = 0.227$ and $19/44 = 0.432$ (using three decimal places in this written solution).

   (b) We start from the table

   $$\begin{matrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{matrix}$$

   To adjust the i probabilities, we replace $\hat{p}_{ij}^{(}0)$ by $(\hat{p}_i/\hat{p}_i^{(0)})\hat{p}_{ij}^{(0)}$. In other words we rescale each row to have the correct sum.

   The row sums $\hat{p}_i^{(0)}$ of the starting approximation are both 3. So we multiply every entry in the first row by $0.5/3 = 0.167$ and every entry in the second row by $0.5/3 = 0.167$ which produces the table

   $$\begin{matrix} 0.167 & 0.167 & 0.167 \\ 0.167 & 0.167 & 0.167 \end{matrix}$$

   Now to adjust the j probabilities, we need to rescale each column. The current column sums are all 0.333 and so we need to multiply every entry in the first column by $0.341/0.333 = 1.024$, every entry in the second column by $0.227/0.333 = 0.682$ and every entry in the third column by $0.432/0.333 = 1.297$. The result is

   $$\begin{matrix} 0.171 & 0.114 & 0.217 \\ 0.171 & 0.114 & 0.217 \end{matrix}$$

We have now completed one iteration of IPF.

The results would actually add up exactly to 1 if I hadn't rounded to three decimal places in my calculations.

(c) We know for the independence model that the exact answer is simply $\hat{p}_{ij} = \hat{p}_i \hat{p}_j$ which you will find gives the same table.

Why? Using the notation from the final lecture:

- Initially $\phi_{ij}^{(1,0)} = 1$
- After the first step in the algorithm, the entries are $\phi_{ij}^{(1,1)} = \hat{p}_i / C$ so that row i sums to $\hat{p}_i$ and column j sums to $1/C$.
- After the second step in the algorithm, the entries are $\phi_{ij}^{(1,1)} = \hat{p}_j / (1/C) \phi_{ij}^{(1,1)} = \hat{p}_i \hat{p}_j$ which is precisely the independence model maximum likelihood estimate.

Comment: In fact it turns out that one iteration of IPF is enough for the independence model in any dimension. This generalises a bit further to any hierarchical model where the maximal terms are non-overlapping, i.e. no variable appears in more than one maximal term. Such a model consists of a number of independent components where each component is a "full model" in some variables.

Exercise 8. The R output at the end of the question shows the result of forwards and backwards step-wise model selection using BIC for a 4-dimensional contingency table starting from the independence table. The column labelled AIC is in fact the BIC value in the form where lower values are preferred to higher values.

1. At the third stage of the procedure (begins with `Step:  AIC=1486.01`), explain why the particular model terms were considered for addition and deletion.

2. What models were considered along the way that were nearly as good as the model selected (say with BIC values at most 4 higher than the final model).

3. Interpret the final model considering (in turn) each of hs, phs and fol as a response variable of interest.

```
> step(loglm(f~., minn38), scope=list(lower=~., upper=~.^4), k=log(sum(minn38$f)))
Start:  AIC=3836.09
f ~ hs + phs + fol + sex

          Df    AIC
+ phs:fol 18 2543.0
```

```
+ hs:phs   6 2779.1
+ hs:sex   2 3433.9
+ phs:sex  3 3448.1
+ hs:fol  12 3792.6
+ fol:sex  6 3819.7
<none>       3836.1

Step:  AIC=2542.99
f ~ hs + phs + fol + sex + phs:fol

          Df    AIC
+ hs:phs   6 1486.0
+ hs:sex   2 2140.8
+ phs:sex  3 2155.0
+ hs:fol  12 2499.5
+ fol:sex  6 2526.6
<none>       2543.0
- phs:fol 18 3836.1

Step:  AIC=1486.01
f ~ hs + phs + fol + sex + phs:fol + hs:phs

          Df    AIC
+ hs:sex   2 1083.8
+ phs:sex  3 1098.0
+ fol:sex  6 1469.6
<none>       1486.0
+ hs:fol  12 1569.0
- hs:phs   6 2543.0
- phs:fol 18 2779.1

Step:  AIC=1083.83
f ~ hs + phs + fol + sex + phs:fol + hs:phs + hs:sex

          Df     AIC
+ phs:sex  3  718.53
+ fol:sex  6 1043.04
```

```
<none>       1083.83
+ hs:fol  12 1166.83
- hs:sex   2 1486.01
- hs:phs   6 2140.80
- phs:fol 18 2376.92

Step:  AIC=718.53
f ~ hs + phs + fol + sex + phs:fol + hs:phs + hs:sex + phs:sex

             Df     AIC
+ fol:sex     6  714.66
<none>           718.53
+ hs:phs:sex  6  767.92
+ hs:fol     12  801.54
- phs:sex     3 1083.83
- hs:sex      2 1098.04
- hs:phs      6 1752.83
- phs:fol    18 2011.63

Step:  AIC=714.66
f ~ hs + phs + fol + sex + phs:fol + hs:phs + hs:sex + phs:sex +
    fol:sex

              Df     AIC
<none>           714.66
- fol:sex      6  718.53
+ hs:phs:sex   6  764.04
+ hs:fol      12  793.14
+ phs:fol:sex 18  826.08
- phs:sex      3 1043.04
- hs:sex       2 1094.17
- hs:phs       6 1748.96
- phs:fol     18 1995.25
Call:
loglm(formula = f ~ hs + phs + fol + sex + phs:fol + hs:phs +
    hs:sex + phs:sex + fol:sex, data = minn38, evaluate = FALSE)
```

```
Statistics:
                      X^2  df     P(> X^2)
Likelihood Ratio 256.1798 120 6.849299e-12
Pearson          262.6046 120 1.165845e-12

\begin{sol*}
~
\begin{enumerate}
\item At this stage we have the model
\begin{center}
hs + phs + fol + sex + phs:fol + hs:phs
\par\end{center}

which is hierarchical. We can only add or delete terms which make
the new model hierarchical.

So we can't delete hs, phs or fol but could delete sex, phs:fol or
hs:phs. However, we always keep at least the independence model terms
and so we can't delete sex.

We can't add phs:fol:hs:sex as the 3-way interactions are all missing
and we can't add any 3-way interaction as each one needs all three
corresponding 2-way interactions to be present. So we can only add
all the missing 2-way interactions.
\item Only one: the model obtained by removing fol:sex which has a BIC of
718.53.
\item hs has interactions with phs and sex and therefore depends on them. Similarly
phs depends on fol, hs and sex; fol depends on phs and sex.

In each case, there are no higher-order interactions involved and
so the odds-ratios for the dependence of a reponse on one of the covariates
are not affected by any of the other covariates.
\end{enumerate}
\end{sol*}
```

The next exercise was addressed in the Last Lecture

Exercise 9. In 1968, 715 blue collar workers, selected from Danish Industry, were asked a number of questions concerning their job satisfaction. Some of these questions were summarized in a measure of job satisfaction. Based on similar questions the job satisfaction of the supervisors were measured. Also included in the investigation was an external evaluation of the quality of management for each factory. Table 2 shows the 715 workers distributed on the three variables

Y: Own job satisfaction,

X: Supervisors job satisfaction,

Z: Quality of management.

Quality of management (Z)	Supervisors job satisfaction(X)	Own Job satisfaction(Y)	
		Low	High
Bad	Low	103	87
	High	32	42
Good	Low	59	109
	High	78	205

Table 2: Own job satisfaction, supervisors job satisfaction and the quality of management for 715 blue collar workers in Denmark in 1968. (Source: Petersen (1968), table M/7)

1. Is the own job satisfaction and the supervisors satisfaction conditionally independent of the quality management?
2. Test the hypothesis that the Quality of management, Supervisors job satisfaction, and Own Job satisfaction are mutually independent, against the hypothesis that the own job satisfaction and the supervisors satisfaction conditionally independent of the quality management
3. [Y and X are conditionally independent on Z] vs. [X and Y jointly independent on Z and Y]

Any inference based on hypothesis tests should be performed at sig. level 5%.

Solution.

1. The question can be addressed as a goodness of fit test, if we test the consider the model (XZ, YZ) representing that 'the own job satisfaction and the supervisors satisfaction conditionally independent of the quality management', against the model (XYZ) representing 'the saturated model' or otherwise 'no-independence'.

   So I ll perform likelihood ratio hypothesis test

$$\begin{cases} \mathrm{H}_0: & (XZ, YZ) \\ \mathrm{H}_1: & (XYZ) \end{cases}$$

Here, I will perform the Pearson's χ^2 test, however, I could also do the Likelihood ratio (or Deviance) test.

Then the hypothesis test statistic is

$$X^2 \overset{\text{H}_0}{=} \sum_{\forall i,j,k} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}} \underset{D}{\rightarrow} \chi^2_{\text{df}}$$

where

$$\hat{\mu}_{ijk} = \frac{n_{ij+} n_{+jk}}{n_{++k}}$$

Let

$$X^2_{i,j,k} = \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$$

Then I do the calculations, and I get:

i	j	k	$n_{i,j,k}$	$\hat{\mu}_{i,j,k}$	$X^2_{i,j,k}$
1	1	1	103	97.15	0.35
2	1	1	32	37.84	0.90
1	2	1	87	92.84	0.36
2	2	1	42	36.15	0.94
1	1	2	59	51.03	1.24
2	1	2	78	85.96	0.73
1	2	2	109	116.96	0.54
2	2	2	205	197.03	0.32

- So $X^2_{obs} = 5.41$
- The degrees of freedom are df=$(I-1)(J-1) + (I-1)(J-1)(K-1) = 2$ as the two models differ only on terms $\lambda^{XY}_{ij}, \lambda^{XYZ}_{ijk}$.
- The critical value is $\chi^2_{2,0.95} = 5.991465$.
- I cannot reject the null hypothesis at sig level 5% because $X^2_{obs} < \chi^2_{2,0.95}$. Therefore, I can base my inference on model (XZ, YZ) with caution.

2. This is a comparison between Nested models.

   In particular, it can be performed as a hypothesis test where I compare model of independence (X, Y, Z) representing that 'Quality of management, Supervisors job satisfaction, and Own Job satisfaction are mutually independent', against the model (XZ, YZ) representing 'the own job satisfaction and the supervisors satisfaction conditionally independent of the quality management'.

So I ll perform likelihood ratio hypothesis test

$$\begin{cases} \mathrm{H}_0: & (X, Y, Z) \\ \mathrm{H}_1: & (XZ, YZ) \end{cases}$$

the Likelihood ratio test statistic is

$$G^2(H_0, H_1) \overset{\mathrm{H}_0}{=} 2 \sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}^{(1)}_{ijk}}{\hat{\mu}^{(0)}_{ijk}}) \underset{D}{\to} \chi^2_{\mathrm{df}}$$

where

$$\hat{\mu}^{(0)}_{ijk} = \frac{n_{i++} n_{+j+} n_{++k}}{n^2}$$

and

$$\hat{\mu}^{(1)}_{ijk} = \frac{n_{i+k} n_{+jk}}{n_{++k}}$$

Let

$$G_{i,j,k} = 2 n_{ijk} \log(\frac{\hat{\mu}^{(1)}_{ijk}}{\hat{\mu}^{(0)}_{ijk}})$$

Then I do the calculations, and I get:

i	j	k	$n_{i,j,k}$	$\hat{\mu}^{(0)}_{i,j,k}$	$\hat{\mu}^{(1)}_{i,j,k}$	$G_{i,j,k}$
1	1	1	103	50.28	97.15	135.67
2	1	1	32	50.14	37.84	-18.01
1	2	1	87	81.89	92.84	21.81
2	2	1	42	81.67	36.15	68.43
1	1	2	59	85.90	51.03	-61.44
2	1	2	78	85.66	85.96	0.54
1	2	2	109	139.91	116.96	-39.04
2	2	2	205	139.51	197.03	141.51

- So $G^2_{obs} = 112.60$
- The degrees of freedom are df=$(I-1)(K-1) + (I-1)(J-1) = 2$ as the two models differ only on terms $\lambda^{XZ}_{ik}, \lambda^{XY}_{ij}$.
- The critical value is $\chi^2_{2,0.95} = 5.991465$.
- I reject the null hypothesis at sig level 5% because $G^2_{obs} > \chi^2_{2,0.95}$.

3.

This is a comparison between non-nested models.

In particular, the models under comparison are the $M_0 : (XZ, YZ)$ representing 'Y is joint independent on X and Z' and the $M_1 : (YX, Z)$ representing 'Z is marginally independent on X and Y'.

I will use one of the information criteria, because I compare non-nested models. In particular I ll use both of AIC and BIC, which are defined as follows

$$\begin{aligned}\text{AIC}(M) = &-2\sum_{ijk}\log f_M(n_{ijk}|\hat{\mu}^M_{ijk}) + 2p_M \\ &- 2\sum_{ijk}\log(\text{Poi}(n_{ijk}|\hat{\mu}^M_{ijk})) + 2p_M\end{aligned}$$

and

$$\begin{aligned}\text{BIC}(M) = &-2\sum_{ijk}\log f_M(n_{ijk}|\hat{\mu}^M_{ijk}) + \log(n_{+++})p_M \\ &- 2\sum_{ijk}\log(\text{Poi}(n_{ijk}|\hat{\mu}^M_{ijk})) + \log(n_{+++})p_M\end{aligned}$$

where

$$\text{Poi}(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}1(y \in \mathbb{N})$$

is the probability function of the Poisson distribution with mean μ.

- For model $M_0 : (XZ, YZ)$, the MLE of the $\mu_{i,j,k}$ are

$$\hat{\mu}^{(0)}_{ijk} = \frac{n_{i+k}n_{+jk}}{n_{++k}}$$

then after calculations

i	j	k	$n_{i,j,k}$	$\hat{\mu}^{(0)}_{i,j,k}$	$\log f_0(n_{ijk}\|\hat{\mu}^{(0)}_{ijk})$	$n_{ijk}\log(\hat{\mu}^{(0)}_{ijk})$
1	1	1	103	97.15	-3.40	471.36
2	1	1	32	37.84	-3.13	116.26
1	2	1	87	92.84	-3.34	394.18
2	2	1	42	36.15	-3.23	150.69
1	1	2	59	51.03	-3.55	232.01
2	1	2	78	85.96	-3.47	347.40
1	2	2	109	116.96	-3.54	519.04
2	2	2	205	197.03	-3.73	1083.09

- The number of free parameters is

$$p_0 = 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) + (I-1)(K-1) = 6$$

- So

$$\begin{aligned}
\text{AIC}(0) &= -2\sum_{ijk} n_{ijk}\log(\hat{\mu}^{(0)}_{ijk}) + 2\sum_{ijk}\hat{\mu}^{(0)}_{ijk} \quad (= n_{+++}) \\
&\quad -2\sum_{ijk}\log(n_{ijk}!) + 2\times 5 \\
&\propto -2\sum_{ijk} n_{ijk}\log(\hat{\mu}^{(0)}_{ijk}) + 2n_{+++} + 2\times 5 \\
&= -5186.15
\end{aligned}$$

and

$$\begin{aligned}
\text{BIC}(0) &= -2\sum_{ijk} n_{ijk}\log(\hat{\mu}^{(0)}_{ijk}) + \sum_{ijk}\hat{\mu}^{(0)}_{ijk} \quad (= n_{+++}) \\
&\quad +\sum_{ijk}\log(n_{ijk}!) + \log(n_{+++})\times 5 \\
&\propto -2\sum_{ijk} n_{ijk}\log(\hat{\mu}^{(0)}_{ijk}) + n_{+++} + \log(n_{+++})\times 5 \\
&= -5873.717
\end{aligned}$$

- For model $M_1 : (YX, Z)$ the MLE of the $\mu_{i,j,k}$ are

$$\hat{\mu}^{(1)}_{ijk} = \frac{n_{ij+}n_{++k}}{n_{+++}}$$

Then after calculations

i	j	k	$n_{i,j,k}$	$\hat{\mu}^{(1)}_{i,j,k}$	$\log f_1(x_{ijk}\|\hat{\mu}^{(1)}_{ijk})$	$n_{ijk}\log(\hat{\mu}^{(1)}_{ijk})$
1	1	1	103	59.81	-16.02	421.40
2	1	1	32	40.61	-3.64	118.53
1	2	1	87	72.36	-4.54	372.51
2	2	1	42	91.20	-19.42	189.54
1	1	2	59	102.18	-13.73	272.98
2	1	2	78	69.38	-3.61	330.69
1	2	2	109	123.63	-4.16	525.08
2	2	2	205	155.80	-10.64	1034.95

- The number of free parameters is

$$p_1 = 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) = 5$$

  - So

$$\begin{aligned}
\mathrm{AIC}(1) &= -2\sum_{ijk} n_{ijk}\log(\hat{\mu}^{(1)}_{ijk}) + 2\sum_{ijk}\hat{\mu}^{(1)}_{ijk} \quad (\hat{\mu}^{(1)}_{ijk} \to n_{+++}) \\
&\quad -2\sum_{ijk}\log(n_{ijk}!) + 2\times 5 \\
&\propto -2\sum_{ijk} n_{ijk}\log(\hat{\mu}^{(1)}_{ijk}) + 2n_{+++} + 2\times 5 \\
&= -5091.42
\end{aligned}$$

    and

$$\begin{aligned}
\mathrm{BIC}(1) &= -2\sum_{ijk} n_{ijk}\log(\hat{\mu}^{(1)}_{ijk}) + \sum_{ijk}\hat{\mu}^{(1)}_{ijk} \quad (\hat{\mu}^{(1)}_{ijk} \to n_{+++}) \\
&\quad +\sum_{ijk}\log(n_{ijk}!) + \log(n_{+++})\times 5 \\
&\propto -2\sum_{ijk} n_{ijk}\log(\hat{\mu}^{(1)}_{ijk}) + n_{+++} + \log(n_{+++})\times 5 \\
&= -5783.565
\end{aligned}$$

- Result: Both the AIC and the BIC indicate that the model $M_0 : (YX, YZ)$ is better supported by the data and hence it is preferable.

PJC exercise

Exercise 10. In log-linear modelling of contingency tables, we represent the logarithms of the cell probabilities $p_{i_1\ldots i_p}$ linearly in terms of tables involving subsets of the indices; we constrain those tables by requiring that the sum over any index in one of those tables yields zero.

For example, when $p = 2$, the full model is

$$\log p_{ij} = \eta_{ij} = \lambda + \lambda^{(R)}_i + \lambda^{(C)}_j + \lambda^{(RC)}_{ij} \tag{11}$$

and the constraints are $\sum_i \lambda^{(R)}_i = 0$, $\sum_j \lambda^{(C)}_j = 0$, $\sum_i \lambda^{(RC)}_{ij} = 0$ for each j and $\sum_j \lambda^{(RC)}_{ij} = 0$ for each i.

1. Show that, for any table of cell probabilities p_{ij}, we can find λ, one-dimensional tables $\lambda^{(R)}$, $\lambda^{(C)}$ and two-dimensional table $\lambda^{(RC)}$ satisifying the constraints and so that (11) holds for all i, j.

2. Show also that the resulting values of the various different kinds of "λ" are unique.

3. Without doing detailed calculations, describe how the argument generalises to $p = 3$ and beyond.

I have used the superscripts (R), (C) and (RC) in the interests of clarity but you may omit them provided you are careful about your use of letters!

Solution.

Q18. Write η_{ij} for $\log p_{ij}$

From (1),

$$\sum_j \eta_{ij} = C\lambda + C\lambda_i + 0 + 0$$

$$= C(\lambda + \lambda_i)$$

from the constraints on λ_j and λ_{ij}

$$\Rightarrow \sum_{ij} \eta_{ij} = RC\lambda + 0 \quad \text{since } \sum_i \lambda_i = 0$$

$$\Rightarrow \lambda = \frac{1}{RC} \sum_{ij} \eta_{ij}$$

$$\Rightarrow \lambda_i = \left[\frac{1}{C} \sum_j \eta_{ij}\right] - \lambda$$

Similarly to above, we have $\sum_i \eta_{ij} = R(\lambda + \lambda_j)$

$$\Rightarrow \lambda_j = \left[\frac{1}{R} \sum_i \eta_{ij}\right] - \lambda$$

and finally from (1) we have

$$\lambda_{ij} = \eta_{ij} - \lambda - \lambda_i - \lambda_j$$

We have just seen how to (uniquely) construct λ, λ_i, λ_j, λ_{ij} from ρ_{ij}

For $p = 3$ and beyond, this generalises to

- Find unsubscripted λ by ~~summing~~ averaging η with respect to all indices.

- Find single subscript λ tables by ~~summing~~ averaging η with respect to all but one index and subtract ~~appropriate multiples~~ unsubscripted λ

- Find two subscript λ tables by ~~summing~~ averaging η with respect to all but two indices and subtract ~~appropriate multiples~~ the relevant single subscripted λ values and the unsubscripted λ

  ⋮

- Find final λ table by subtracting all other λ from η

PJC exercise

Exercise 11. In the lectures, maximum likelihood estimates were derived using Lagrange multipliers. We considered hierarchical log-linear models for $d = 2$ (independence and full/saturated versions).

Some key general features of the process turn out to be that:

- one of the constraints is always that the sum of the cell probabilities is 1 and we always find that the Lagrange multiplier for that constraint is the total number of data n;
- all other Lagrange multipliers turn out to be zero;
- for any log-linear model coefficient "β" (in the Lectures denoted as λ's), the estimates of the table of probabilities having the same indices are simply the corresponding data proportions, i.e. for the $d = 2$ independence model, $\hat{p}_i = y_i/n$ and $\hat{p}_j = y_j/n$ and in the full model $\hat{p}_{ij} = y_{ij}/n$;
- the method used to deduce that $\hat{p}_i = y_i/n$ and $\hat{p}_j = y_j/n$ for $d = 2$ generalises to all single-index marginals
- for $d = 2$, the equations for the saturated model are the same as those for the independence model except that an additional equation is introduced involving β_{ij} and the solution of the additional equation for β_{ij} is obtained by (i) summing the equation with respect to i and exploiting the solution to the equation for β_i to show that the Lagrange multipliers, associated with the constraints that $\sum_i \beta_{ij} = 0$ for each j, all take the same value; (ii) summing with respect to j in order to show that the Lagrange multipliers for the row-sum constraints on β_{ij} all take the same value; (iii) summing with respect to both i and j to show that the sum of the row-sum and column-sum Lagrange multipliers is 0 for each i and j. This generalises when $d > 2$ to all pair-wise interactions present in the model and also generalizes to more complex terms in the model.

Now consider $d = 3$ and the hierarchical model

$$\log p_{ijk} = \eta_{ijk} = \beta + \beta_i + \beta_j + \beta_k + \beta_{ij} + \beta_{ik} + \beta_{jk}$$

Recall that the data here are y_{ijk} and the model is for cell probabilities p_{ijk} and that we use notation involving fewer indices (or alternatively "dots" as subscripts) to mean that we have summed over the indices which have been dropped. For example $y_k = y_{..k} = \sum_{i,j} y_{ijk}$ and $p_{ik} = \sum_j p_{ijk}$.

1. Write down the log-likelihood to be maximised and the constraints to be imposed.

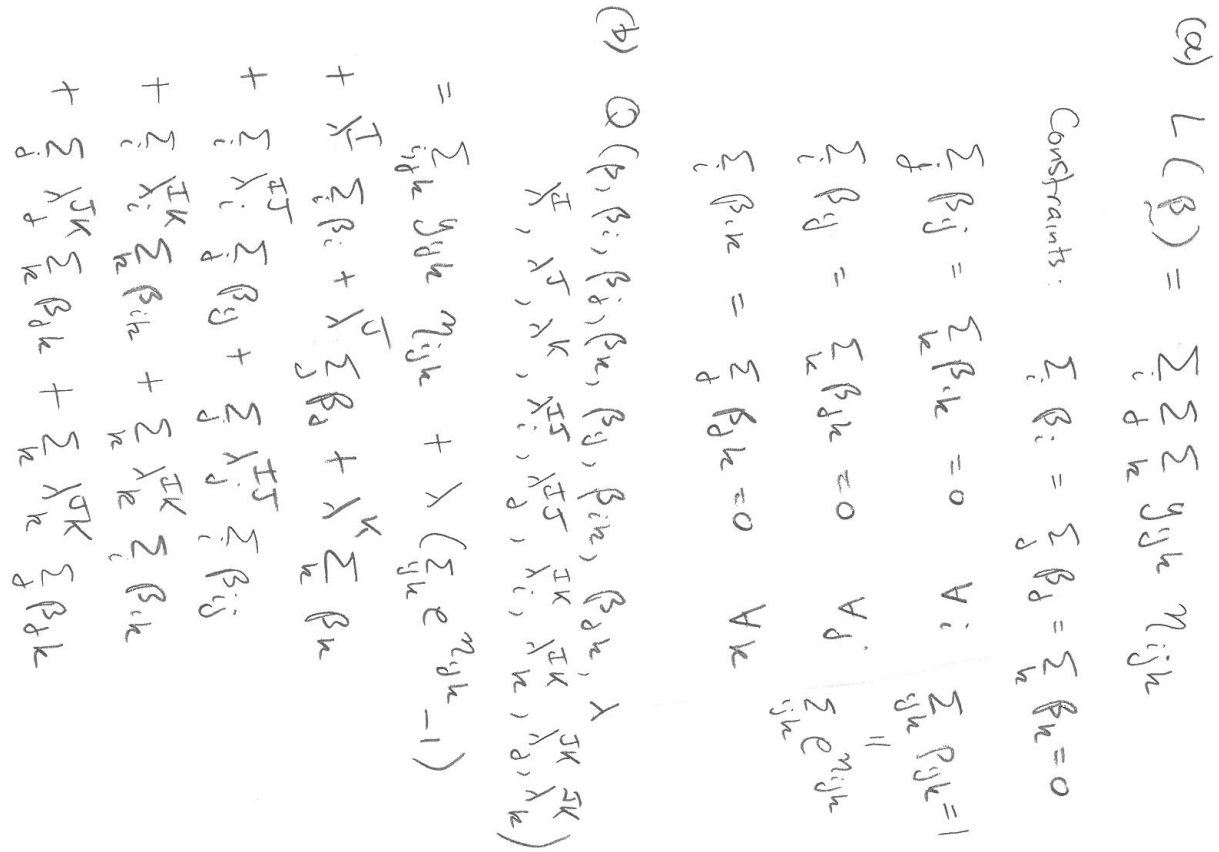
2. Hence write down the expression involving Lagrange multipliers for which a turning point must be found in order to obtain the maximum likelihood estimates.

3. Show that the methods from $d = 2$ can be used hierarchically to derive the following equations satisfied by the maximum likelihood estimates: $\hat{p}_{ij} = y_{ij}/n$, $\hat{p}_{ik} = y_{ik}/n$ and $\hat{p}_{jk} = y_{jk}/n$. Note that you will also show $\hat{p}_i = y_i/n$, $\hat{p}_j = y_j/n$ and $\hat{p}_k = y_k/n$ but these are redundant as they follow from the equations for $\hat{p}_{ij}$ etc.

4. [This is too advanced...] Show that knowing all p_{ij}, p_{ik} and p_{jk} uniquely determines the values of all the "βs". In other words, we have uniquely determined the m.l.e. in the previous part of the question.

For the full model, show how the argument from $d = 2$ generalises to use the solution of the equations for β_{ij}, β_{ik} and β_{jk} to help to show that $\hat{p}_{ijk} = y_{ijk}/n$.

Solution. I should like to thank Professor Steffen Lauritzen (Oxford) for his online lecture material which enabled me to find a solution to part(d).



(c) $\frac{\partial n_{ijk}}{\partial \beta} = 1 \quad \forall\, i,j,k$

$$\Rightarrow \frac{\partial Q}{\partial \beta} = \sum_{i,j,k} y_{ijk} + \lambda \sum_{ijk} e^{n_{ijk}}$$

$$= n + \lambda \cdot 1$$

↑ from constraint

$$= 0 \text{ at m.l.e.}$$

$$\Rightarrow \lambda = -n$$

$$\frac{\partial n_{ijk}}{\partial \beta_{ii'}} = g_{ii'}$$

$$\Rightarrow \frac{\partial Q}{\partial \beta_{ii'}} = \sum_{jk} y_{ii'jk} + \lambda \sum_{jk} e^{n_{ii'jk}} + \lambda_I$$

$$= y_{ii'} + \lambda\, p_{ii'} + \lambda_I$$

$$= 0 \text{ at m.l.e.}$$

$$\Rightarrow y_{i\cdot} - n\, p_{i\cdot} + \lambda_I = 0 \quad \forall\, i$$

Sum over i: $n - n \cdot 1 + R\lambda_I = 0$

$$\Rightarrow \lambda_I = 0, \quad p_{i\cdot} = g_{i\cdot}/n \quad \forall\, i$$

Similarly: $\hat{\lambda}^J = 0 \quad \hat{p}_j = y_j/n \quad \forall j$

$\hat{\lambda}^K = 0 \quad \hat{p}_k = y_k/n \quad \forall k$

$$\frac{\partial \eta_{ijk}}{\partial \beta_{i'j'}} = \delta_{ii'} \delta_{jj'}$$

$$\Rightarrow \frac{\partial Q}{\partial \beta_{i'j'}} = \sum_k y_{i'j'k} + \lambda \sum_k e^{\eta_{i'j'k}} + \lambda^{IJ}_{i'} + \lambda^{IJ}_{j'}$$

$$= y_{i'j'} + \lambda \beta_{i'j'} + \lambda^{IJ}_{i'} + \lambda^{IJ}_{j'}$$

$$= 0 \text{ at m.l.e.}$$

$$\Rightarrow y_{ij} - n\hat{p}_{ij} + \hat{\lambda}^{IJ}_i + \hat{\lambda}^{IJ}_j = 0 \quad (*) \quad \forall i, j$$

Summing over i:

$$\underbrace{y_j - n\hat{p}_j}_{0} + \sum_i \hat{\lambda}^{IJ}_i + R\hat{\lambda}^{IJ}_j = 0 \quad \forall j$$

$\Rightarrow \hat{\lambda}^{IJ}_j$ does not depend on j

Similarly $\hat{\gamma}_i^{IT}$ does not depend on i

So (*) becomes

$$\gamma_{ij} - n\hat{p}_{ij} + \phi = 0 \quad \forall i,j$$

for some ϕ

Sum over i and j

$$n - n \cdot 1 + RC\phi = 0$$

$$\Rightarrow \phi = 0 \Rightarrow \hat{p}_{ij} = \gamma_{ij} / n \; \forall i,j$$

Similarly we find that

$$\hat{p}_{ik} = \gamma_{ik}/n \quad \forall i,k$$

$$\hat{p}_{jk} = \gamma_{jk}/n \quad \forall j,k$$

(d) Suppose that $\tilde{p}_{ijk}$ satisfies the log-linear model, i.e.

$$\tilde{\gamma}_{ijk} = \tilde{\beta} + \tilde{\beta}_i^{(1)} + \tilde{\beta}_j^{(2)} + \tilde{\beta}_k^{(3)} + \tilde{\beta}_{ij}^{(12)} + \tilde{\beta}_{ik}^{(13)} + \tilde{\beta}_{jk}^{(23)}$$

and

the m.l.e equations does (cf), namely

$$\tilde{p}_{ij} = \hat{p}_{ij}(=y_{ij}/n) \quad \tilde{p}_{ik} = \hat{p}_{ik} \quad \tilde{p}_{jk} = \hat{p}_{jk}$$

Now for any p_{ijk} satisfying the log-linear model

$$L(\tilde{p}) = \sum_{ijk} y_{ijk} \eta_{ijk}$$

$$= \beta \sum_{ijk} y_{ijk} + \sum_i \beta_i \sum_{jk} y_{ijk}$$

$$+ \dots + \sum_{jk} \beta_{jk} \sum_i y_{ijk}$$

$$= n\beta + \sum_i \beta_i y_{i\cdot\cdot} + \sum_j \beta_j y_{\cdot j\cdot} + \sum_k \beta_k y_{\cdot\cdot k}$$

$$+ \sum_{ij} \beta_{ij} y_{ij\cdot} + \sum_{ik} \beta_{ik} y_{i\cdot k} + \sum_{jk} \beta_{jk} y_{\cdot jk}$$

$$= n\{\beta + \sum_i \beta_i \tilde{p}_{i\cdot\cdot} + \dots + \sum_{jk} \beta_{jk} \tilde{p}_{\cdot jk}\}$$

$$= n\{\beta \sum_{ijk} \tilde{p}_{ijk} + \sum_i \beta_i \sum_{jk} \tilde{p}_{ijk} + \dots + \sum_{jk} \beta_{jk} \sum_i \tilde{p}_{ijk}\}$$

$$= n \sum_{ijk} \eta_{ijk} \tilde{p}_{ijk}$$

But by Jensen (see proof of Theorem 3.1 on likelihood theory handout),

$$\sum_{ijk} \eta_{ijk} \tilde{P}_{ijk} < \sum \tilde{\eta}_{ijk} \tilde{P}_{ijk}$$

unless $P \equiv \tilde{P}$.

Therefore $\tilde{P}_{ijk}$ maximises the likelihood

Moreover, suppose that $P^{(1)}_{ijk}$ and $P^{(2)}_{ijk}$ are not identical and that both satisfy the model and maximise L.

Put $$P_{ijk} = C\sqrt{P^{(1)}_{ijk} P^{(2)}_{ijk}}$$

where C is chosen so that $\sum_{ijk} P_{ijk} = 1$

Then η_{ijk} satisfies the model.

Moreover, $C > 1$

since $$\sum_{ijk} \sqrt{P^{(1)}_{ijk}} \sqrt{P^{(2)}_{ijk}} < \sqrt{\sum_{ijk} P^{(1)}_{ijk} \sum_{ijk} P^{(2)}_{ijk}}$$

$= 1$ (dot product less than product of norms of vectors unless vectors parallel)

But $L(\tilde{p})$

$= \sum_{ijk} y_{ijk} \left[\log C + \frac{1}{2} m^{(1)}_{ijk} + \frac{1}{2} m^{(2)}_{ijk} \right]$

$= \underbrace{n \log C + \frac{1}{2} L(\tilde{p}^{(1)}) + \frac{1}{2} L(\tilde{p}^{(2)})}_{>0}$

which would mean that $p^{(1)}_{ijk}$ would have higher likelihood than $p^{(2)}_{ijk}$.

Contradiction $\Rightarrow \tilde{p}_{ijk}$ is unique.

For the final part of the question, we add β_{ijk} to m_{ijk} in the following transformation to Q

$\sum_{ijk} \lambda^{IJK}_{ijk} \sum_{\alpha} \beta_{\alpha ijk} + \sum_{ijk} \lambda^{IJK}_{ijk} \sum_{\alpha} \beta_{\alpha ijk}$

$$\frac{\partial Q}{\partial \hat{P}_{i'j'k'}} = g_{i'j'k'} + \lambda e^{\eta_{i'j'k'}} + \lambda \sum_{i,j,k} \hat{\lambda}^{IJK}_{ij'k'} + \lambda \sum_{i,j,k} \hat{\lambda}^{IJK}_{i'jk'} + \lambda \sum_{i,j,k} \hat{\lambda}^{IJK}_{i'j'k}$$

$$\Rightarrow g_{ijk} - n\hat{P}_{ijk} + \sum_{i} \hat{\lambda}^{IJK}_{ijk} + \sum_{j} \hat{\lambda}^{IJK}_{ijk} + \sum_{k} \hat{\lambda}^{IJK}_{ijk} = 0 \quad (**)$$

Sum over i:

$$g_{jk} - n\hat{P}_{jk} + \sum_{i,j} \hat{\lambda}^{IJK}_{ij} + \sum_{i,k} \hat{\lambda}^{IJK}_{ik} + \sum_{j,k} \hat{\lambda}^{IJK}_{jk} = 0$$

which shows that $\sum_{j,k} \hat{\lambda}^{IJK}_{jk}$ is the sum of something which depends only on I and something else which depends only on k.

A similar argument applies to $\sum_{i,j} \hat{\lambda}^{IJK}_{ij}$

Summing over two indices in $(**)$ shows that the part of the $\hat{\lambda}^{IJK}$s which depend only on the remaining index is in fact constant and so all the $\hat{\lambda}^{IJK}$s are constant and their contribution to $(**)$ is in fact 0.

Therefore $\hat{P}_{ijk} = g_{ijk}$ in

Exercise 12. Consider log-linear modelling of three-dimensional contingency tables, i.e. $p = 3$. Suppose that the table has R rows, C columns and S slices (values of k).

- Write down the full model.
- How many parameters (degrees of freedom) are associated with each term in the full model?

  Hint: remember that summing over any index of a "λ" yields zero; therefore the number of parameters for λ_i is $R - 1$ since we can determine the value for the first row from the values for the other rows.

- Enumerate all possible hierarchical log-linear models for $p = 3$ which include all main-effect terms.

Solution.

$$\log p_{ijk} = \lambda + \lambda_i + \lambda_j + \lambda_k + \lambda_{ij} + \lambda_{ik} + \lambda_{jk} + \lambda_{ijk}$$

The constant term λ has no degrees of freedom. It exists purely to ensure that $\sum_{i,j,k} p_{ijk} = 1$ and its value is determined by the values of all the other terms in the model.

For a term involving a single variable, there is just one constraint which is that all the entries sum to zero. This means that the final entry in the table is determined by the values of the remainder and so the number of parameters is one less than the number of entries. So λ_i, λ_j and λ_k have respectively $R - 1$, $C - 1$ and $S - 1$ parameters.

For a "λ" involving two variables, we have a two-dimensional table of entries in which each row and column sums to zero. So if we put arbitrary numbers in all except the final row and column, we can make the table satisfy the constraints by filling in the final column (except for the last row) so each row except the last sums to zero and then fill in the final row so that each column sums to one; note that these numbers in the final row and column are determined by the original arbitrary selection which occupied a rectangle with one less row and one less column. Therefore λ_{ij}, λ_{ik} and λ_{jk} have respectively $(R - 1)(C - 1)$, $(R - 1)(S - 1)$ and $(C - 1)(S - 1)$ parameters.

Finally, for λ_{ijk}, do the same thing by putting arbitrary numbers in all except the final row, column and slice of the three-dimensional table and note that the remaining entries are uniquely determined to satisfy the constraints by first filling in the final row in all except the final column and slice, then filling in the final column in all except the final slices and finishing by filling in the final slice. The arbitrary starting point occupied a three-dimensional table with one less row, one less column and one less slice. Therefore the number of parameters is $(R - 1)(C - 1)(S - 1)$.

Note that $(R-1)+(C-1)+(S-1)+(R-1)(C-1)+(R-1)(S-1)+(C-1)(S-1)+(R-1)(C-1)(S-1) = RCS - 1$ is the number of freely varying parameters and this is clearly the number required to specify a discrete probability distribution having RCS distinct possible outcomes.

Exercise 13. All possible hierarchical log-linear models were fitted to a contingency table for three factors: hs (3 levels), phs (4 levels) and fol (7 levels). In total there were 7861 data in the table.

The following table summarises the results of fitting each model by maximum likelihood estimation. The "log-likelihood" column is the value of the log-likelihood at the maximum. The "deviance" column is twice the difference between the maximum of the log-likelihood for the full model and the maximum of the log-likelihood for the model specified.

Model	Deviance	Log-likelihood
Full	0	40293
hs+phs+fol+hs:phs+hs:fol+phs:fol	87	40249.5
hs+phs+fol+phs:fol+hs:fol	637	39974.5
hs+phs+fol+phs:fol+hs:phs	125	40230.5
hs+phs+fol+hs:phs+hs:fol	787	39899.5
hs+phs+fol+hs:phs	899	39843.5
hs+phs+fol+hs:fol	1412	39587.0
hs+phs+fol+phs:fol	750	39918.0
Independence (hs+phs+fol)	1524	39531.0

1. Compute the number of free parameters (degrees of freedom) for each model in the table.
2. Which model would be chosen using AIC?
3. Which model would be chosen using BIC?
4. For the model chosen in part (c), what would we have learned if were interested in treating "phs" as the response variable?
5. For the model chosen in part (c), carry out the likelihood ratio test for that model versus the Independence model.

Solution. The following table extends the one provided in the question to include the number of parameters and the AIC and BIC for each model.

Model	1 Deviance	1 Log-likelihood	p	AIC	BIC
Full	0	40293	83	-80420	-79841.52
hs+phs+fol+hs:phs+hs:fol+phs:fol	87	40249.5	47	-80405	-80077.43
hs+phs+fol+phs:fol+hs:fol	637	39974.5	41	-79867	-79581.24
hs+phs+fol+phs:fol+hs:phs	125	40230.5	35	-80391	-80147.06
hs+phs+fol+hs:phs+hs:fol	787	39899.5	29	-79741	-79538.88
hs+phs+fol+hs:phs	899	39843.5	17	-79653	-79534.52
hs+phs+fol+hs:fol	1412	39587.0	23	-79128	-78967.70
hs+phs+fol+phs:fol	750	39918.0	29	-79778	-79575.88
Independence (hs+phs+fol)	1524	39531.0	11	-79040	-78963.33

1. Numbers of parameters are obtained by adding up the number of parameters for each term appearing in the model. The number of parameters for a single term was given as a formula in the lectures (see also the answer to question **??**): find the product over variables appearing in the term of $D_i - 1$ where D_i is the number of categories for variable i.

   Here $R = 3$ (hs), $C = 4$ (phs) and $S = 7$ (fol).

   So, for example, the indepdence model has $(3-1)+(4-1)+(7-1) = 11$ parameters and the model labelled `hs+phs+fol+hs:phs+hs:fol` has $11 + 2 \times 3 + 2 \times 6 = 29$.

2. The AIC column is obtained as $-2L + 2p$ where L is the value in the "Log-likelihood" column. For example $-79741 = -2 \times 39899.5 + 2 \times 29$.

   The model with the lowest AIC is the full model and so it would be selected.

   Alternatively, one can compute AIC using the deviance as Deviance $+ 2p$. This simply add the same number on to each computed AIC value and so does not change which model is selected.

3. The BIC column is obtained as $-2 : L + p \log n$ where n is the total number of data in the table, here 7861. For example $-79538.88 = -2 \times 39899.5 + 29 \log 7861$.

   The model with the lowest BIC is `hs+phs+fol+phs:fol+hs:phs` and so it would be selected.

   As with AIC, one can also cmopute BIC based on the deviance.

4. Since the model includes both `hs:phs` and `fol:phs`, the distribution of phs depends on both hs and fol.

   However, the model does not incude `hs:fol:phs`. Consequently, the effects of hs and fol on the distribution of phs are additive on log-odds scale. Another way os saying the same thing is that odds-ratios for phs depending on hs would be unaffected by the value of fol and vice-versa.

5. The test statistic is twice the difference in the log-likelihoods which is $2(40230.5 - 39531.0) = 1399$ which should be compared to the chi-squared distribution with $35 - 11 = 24$ degrees of freedom. 1399 is an absolutely enormous value for that distribution and we reject the independence model incredibly strongly.

   Aside: Note that if we test the full model versus the model chosen by BIC, the test statistic is 125 which is still very large compared to the chi-squared distribution with 48 degrees of freedom. So on the basis of the likelihood ratio test, we should prefer the full model to the BIC model. However, many applied statisticians would still prefer to use BIC for model selection as it actually does a better job of finding the "correct" model than significance tests such as the likelihood-ratio test and we like the parsimonious approach.

Revision

Exercise 14. Consider a sample of 261 individuals, for three factors: Y (2 levels), X (3 levels), and Z (4 levels). The sampling scheme that was implemented was the Poisson sampling scheme. The deviance and the log-likelihood of a number of models are given in Table 3.

Model	Deviance	Log-likelihood	BIC	free parameters	AIC
$[X, Y, Z]$	133.37	-60.63			
$[XY, Z]$	126.62	-54.01			
$[XZ, Y]$	132.91	-54.59			
$[YZ, X]$	57.12	-44.46			
$[XY, XZ]$	126.16	-53.96			
$[XY, YZ]$	50.37	-42.95			
$[XZ, YZ]$	56.66	-44.36			
$[XY, XZ, YZ]$	49.91	-42.84			
$[XYZ]$		-42.00			

Table 3:

1. Define the Akaike Information Criterion (AIC), and Bayes Information Criterion (BIC). Fill in Table 3. Which model is selected by AIC, and BIC?

2. Intuitively discuss how AIC and BIC work to address model selection problems.

Solution.

1. It is

- Assume a set of observations $\{x_i\}_{i=1}^N$. Let M be a model of interest (i.e. expressing a type of dependency). The AIC of model M is

$$\mathrm{AIC}(M) = -2\underbrace{\sum_{i=1}^{N} \log f_M(x_i|\hat{\theta}_M)}_{=\text{log-likelihood}} + 2 \underbrace{p_M}_{=\text{num. of free parameters}} \tag{12}$$

  where $\theta_M \in \Theta_M$ are the free parameters of the model M , $p_M = \dim(\Theta_M)$, and $\hat{\theta}_M$ is the MLE of θ_M.

- Assume a set of observations $\{x_i\}_{i=1}^N$. Let M be a model of interest (i.e. expressing a type of dependency). The BIC of model M is

$$\mathrm{BIC}(M) = -2\underbrace{\sum_{i=1}^{N} \log f_M(x_i|\hat{\theta}_M)}_{=\text{log-likelihood}} + \log(N)p_M \tag{13}$$

  where $\theta_M \in \Theta_M$ are the free parameters of the model M , $p_M = \dim(\Theta_M)$, $\hat{\theta}_M$ is the MLE of θ_M, and N is the number of parameters.

The number of free parameters are computed as in the lecture notes. For instance, let $I = 2$, $J = 3$, $K = 4$ be the number of the levels of the categorical variables X, Y, Z, then

- for $[X, Y, Z]$, it is

$$\begin{aligned} p_{[X,Y,Z]} &= 1 + (I-1) + (J-1) + (K-1) \\ &= 1 + (2-1) + (3-1) + (4-1) = 7 \end{aligned}$$

  because it denoted the log-linear model with equation

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

- for $[XY]$ (or equivalently denoted as $[Z, XY]$) it

$$\begin{aligned} p_{[XY]} &= 1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) \\ &= 1 + (2-1) + (3-1) + (4-1) + (2-1)(3-1) = 9 \end{aligned}$$

  because it denoted the log-linear model with equation

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

etc ...

Then by using (12), (13), and $N = 261$ we compute

Model	Deviance	Log-likelihood	BIC	free parameters	AIC
$[X, Y, Z]$	133.37	-60.63	160.21	7	135.26
$[XY]$	126.62	-54.01	158.10	9	126.02
$[XZ]$	132.91	-54.59	164.83	10	129.18
$[YZ]$	57.12	-44.46	161.26	13	114.92
$[XY, XZ]$	126.16	-53.96	174.71	12	131.93
$[XY, YZ]$	50.37	-42.95	169.37	15	115.90
$[XZ, YZ]$	56.66	-44.36	177.76	16	120.72
$[XY, XZ, YZ]$	49.91	-42.84	185.84	18	121.68
$[XYZ]$		-42.00	217.00	24	132.00

AIC suggests $[YZ]$, while BIC suggests $[XY]$ as the 'best' model.

2.

- AIC suggests as preferred model the one with the smaller AIC value. It penalizes the complexity of the model where complexity refers to the number of parameters in the model.
- BIC suggests as preferred model the one with the smaller BIC value. It penalizes the complexity of the model where complexity refers to the number of parameters in the model. It can measure the efficiency of the parameterized model in terms of predicting the data.

Revision

Exercise 15. During a particular summer, an experiment was conducted to find out the preference between two types of beverages: soda and lemonade. The data was drawn from two locations: city and rural. In each location, the gender and the choice of drinks were collected. The results are summarized in Table 4.

Location (Z)	Gender (X)	Preferred drink (Y)	
		Lemonade	Soda
City	Female	9	1
	Male	70	20
Rural	Female	30	60
	Male	2	8

Table 4: Dataset

1. Calculate the marginal contingency table of the observed counts with classifiers Gender and Preferred drink. Calculate a 95% confidence interval for the marginal odds ratio of the Gender and Drink. Interpret the result.

2. Compute and interpret the results of:

   (a) conditional odds ratio of the Gender and Drink at each Location level.

   (b) marginal odds ratio of the Gender and Location.

   (c) marginal odds ratio of the Location and Drink.

3. Compare the results above, investigate the phenomenon, and explain why this might happen.

4. Specify the log-linear model equations for the dependency types:

   (a) Gender, Drink, and Location are mutually independent

   (b) Gender and the Drink are conditionally independent given the Location

   as well as specify the number of parameters for associated log-linear models that result after setting non-identifiability constraints.

   Solution.

   (a) This is the marginal XY-contingency table, namely:

Gender (X)	Preferred drink (Y)		Total
	Lemonade	Soda	
Female	39	61	100
Male	72	28	100
total	111	99	200

   The odds ratio is

$$\hat{\theta}_{i,j}^{XY} = \frac{n_{i,j,+} n_{i+1,j+1,+}}{n_{i+1,j,+} n_{i,j+1,+}}$$

   with asymptotic distribution such as

$$\frac{\log(\hat{\theta}_{i,j}) - \log(\theta_{i,j})}{\sqrt{\frac{1}{n_{i,j}} + \frac{1}{n_{i,j+1}} + \frac{1}{n_{i+1,j}} + \frac{1}{n_{i+1,j+1}}}} \underset{D}{\longrightarrow} \mathrm{N}(0,1)$$

   So the (1-a) confidence interval is

$$(\log(\hat{\theta}_{i,j}) \pm z_{1-\frac{a}{2}} \sqrt{\frac{1}{n_{i,j}} + \frac{1}{n_{i,j+1}} + \frac{1}{n_{i+1,j}} + \frac{1}{n_{i+1,j+1}}})$$

After substitution, we get $\hat{\theta}^{XY}_{1,1} = 0.2486339$, $\log(\hat{\theta}^{XY}_{1,1}) = -1.391774$, and CI of $\log(\hat{\theta}^{XY}_{1,1})$ $(-1.9850884, -0.7984592)$, and CI of $\hat{\theta}^{XY}_{1,1}$ is $(0.1373685, 0.4500218)$. We observe that at sig. level, 5% , we reject the hypothesis that the Gender and the Drink are not associated. The odds ratio suggests that the there is a negative association between the Gender and the Preferred drink. Namely, females are less likely to drink nectar than males.

(b)

i. The MLE of the conditional odds ratio requested is

$$\hat{\theta}^{XY}_{i,j,(k)} = \frac{n_{i,j,k} n_{i+1,j+1,k}}{n_{i+1,j,k} n_{i,j+1,k}}$$

where $\hat{\theta}^{XY}_{1,1,city} = 2.571429$, and $\hat{\theta}^{XY}_{1,1,rural} = 2$. There is positive association between the Gender and the Drink, given the location. Namely, it is more likely for a female to prefer lemonade than a male, given the location.

ii. This is a marginal ZX-contingency table is

Gender (X)	Location (Z)		Total
	City	Rural	
Female	10	90	100
Male	90	10	100
total	100	100	200

The MLE of the marginal ZX odds ratio is

$$\hat{\theta}^{XZ}_{i,k} = \frac{n_{i,k} n_{i+1,k+1}}{n_{i+1,k} n_{i,k+1}} = 0.0123$$

After substitution, we get $\hat{\theta}^{XZ}_{1,1} = 0.0123$, $\log(\hat{\theta}^{XZ}_{1,1}) = -4.394449$, and CI of $\log(\hat{\theta}^{XZ}_{1,1})$ $(-5.318385, -3.470513)$, and CI of $\hat{\theta}^{XZ}_{1,1}$ is $(0.004900662, 0.031101063)$. We observe that at sig. level, 5% , we reject the hypothesis that the Gender and the Location are independent. The odds ratio indicates that the there is a negative association between the Gender and the Location , namely, female is less likely to be in a city than a male.

iii. This is a marginal YZ-contingency table

Location (Z)	Drink (Y)		Total
	Lemonade	Soda	
City	79	21	
Rural	32	68	
total			200

The MLE of the marginal ZX odds ratio is

$$\hat{\theta}^{YZ}_{j,k} = \frac{n_{j,k}n_{j+1,k+1}}{n_{j+1,k}n_{j,k+1}} = 7.994048$$

After substitution, we get $\hat{\theta}^{YZ}_{1,1} = 7.994048$, $\log(\hat{\theta}^{YZ}_{1,1}) = 2.078697$, and CI of $\log(\hat{\theta}^{YZ}_{1,1})$ (1.439878, 2.717517), and CI of $\hat{\theta}^{YZ}_{1,1}$ is (4.22018, 15.14267). We observe that at sig. level, 5% , we reject the hypothesis that the Drink and the Location are not associated. The odds ratio suggests that the there is a positive association between the Location and Drink, namely, City people are more likely to prefer Lemonade than Rural people.

(c) We have observed that, on one hand we get $\hat{\theta}^{XY}_{i,j} = 0.2486339$ marginally which implies that lemonade is less likely to be preferred if the person is a female, which on the other hand we get $\hat{\theta}^{XY}_{city} = 2.571429$, and $\hat{\theta}^{XY}_{rural} = 2$ conditionally which imply that it is more likely for a female to prefer lemonade than a male one given the location. This is the Simpson's paradox: there is inconsistency between conditional and marginal interpretation, while one does not imply the other. The confounding variable here is the Location. (i.) A female is more likely to be in a rural area than male. (ii.) Rural area people are less likely to prefer lemonade than the city ones. So since females are more likely to be in rural areas, they fail to prefer lemonade (marginally) –this can be a possible explanation.

(d)

i. Regarding the independency type "Gender, Drink, and Location are mutually independent": This is the log-linear model for the full conditional model is $[X, Y, Z]$. It has a log-linear model equation

$$\log(\mu_{ijk}) = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k$$

, and the number of free parameters is equal to

$$1 + (I-1) + (J-1) + (K-1) = 1 + 1 + 1 + 1 = 4$$

ii. Regarding the independency type: "Gender and the Drink are conditionally independent given the Location": This is the log-linear model for $[XZ, YZ]$ with a log-linear model equation

$$\log(\mu_{ijk}) = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XZ}_{ik} + \lambda^{YZ}_{jk}$$

, and the number of free parameters is equal to

$$1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) + (J-1)(K-1)$$
$$= 1 + 1 + 1 + 1 + 1 + 1 = 6$$
