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

Handouts: Introduction to the Log-linear model [a]

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

[a] Basic reading list: [1, 5, 2, 3, 6, in the Section References]

1 General words

Assume classifier variables $X, Y, Z, ...$ specifying an $I \times J \times K \times ...$ contingency table $(n_{ijk...})$.

Consider the above table as a large vector, where the quantities are vectorized as

$$\begin{aligned}\boldsymbol{n} =& (n_{111...}, ..., n_{IJK...})\\ \boldsymbol{\mu} =& (\mu_{111...}, ..., \mu_{IJK...})\\ \boldsymbol{\pi} =& (\pi_{111...}, ..., \pi_{IJK...})\end{aligned}$$

As an informal definition: the log linear model is the statistical model those mean is parametrised in a form $\mu_{ijk} = \exp(u_{ijk})$ where u_{ijk} is a linear combination of terms (suitably chosen). The definition can be equivalently stated by using the probabilities π_{ijk}, however we use μ_{ijk} because it is common to more sampling schemes and hence the notation more 'universal'.

The Poisson log linear model assumes that $(n_{ijk...})$ are generated from a Poisson sampling scheme as $n_{ijk} \sim \text{Poi}(\mu_{ijk})$, and describes the relation

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} \quad \Longleftrightarrow \quad \boldsymbol{\mu} = \exp(\boldsymbol{X}\boldsymbol{\beta}) \tag{1}$$

where $\boldsymbol{\beta}$ are unknown coefficients, and $\boldsymbol{X}$ is a design matrix determining dependencies among the classifier variables in the contingency table.

The Multinomial log linear model[1] assumes that $(n_{ijk...})$ are generated from a Multinomial sampling scheme as $n \sim \text{Mu}(n_{++}, \boldsymbol{\pi})$, and describes the relation

$$\text{logit}(\boldsymbol{\pi}) = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} \quad \Longleftrightarrow \quad \boldsymbol{\pi} = \frac{\exp(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})}{1^T \exp(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})} \tag{2}$$

where $\tilde{\boldsymbol{\beta}}$ are unknown coefficients, and $\tilde{\boldsymbol{X}}$ is a design matrix determining dependencies among the classifier variables in the contingency table.

The Product-Multinomial log linear model is omitted from this handout.

[1] Multinomial log linear model is discussed in the Likelihood methods exercise sheet.

Remark. The Multinomial log linear model (2) implies the Poisson log linear models (1) as

$$\begin{aligned}\log(\boldsymbol{\mu}) = \log(n_{++}\boldsymbol{\pi}) &= \log(n_{++}) + \log(\boldsymbol{\pi}) = \\ &\overset{(2)}{=} \log(n_{++}) + \log(\exp(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})) - \log(1^T \exp(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})) \\ &= \underbrace{\log(n_{++}) - \log(1^T \exp(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}))}_{=\beta_0} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{\beta}\end{aligned}$$

where $\boldsymbol{X} = [\mathbf{1}, \tilde{\boldsymbol{X}}]$, and $\boldsymbol{\beta} = [\beta_0, \tilde{\boldsymbol{\beta}}]^T$. So the Poisson log linear model has one additional coefficient parameter β_0 depending on the total sample size. In Multinomial log linear model any intercept term like β_0 is simplified in (2), which is reasonable since n_{++} is a known/predetermined quantity in Multinomial sampling.

In this Handout, we use the Poisson log linear model (referred to it as the log linear model for simplicity) for the presentation, however the extension to the Multinomial log linear model is straightforward.

Questions to address:

1. What is the type of dependence among the classifier variables of the contingency table? =>> Model selection =>> Learn the structure of $\boldsymbol{X}$
2. What is the effect of levels of classifier variables for a given model? =>> Parameter inference =>> Learn the values of $\boldsymbol{\beta}$'s.

2 Dependency models in $I \times J$ contingency tables

Consider:

- a $I \times J$ contingency table $(n_{i,j})$
- $\pi_{i,j}$ as the (i,j)-cell probabilities, assumed to be unknown
- $\mu_{i,j} = \mathrm{E}(N_{i,j}) = n_{+,+}\pi_{i,j}$ as the expected (i,j)-cell frequency, assumed to be unknown
- the Poisson log linear model[2], referred to it as 'the log-linear model'.

The Log-linear models are parametrised with respect to the expected (i,j)-cell frequency $\mu_{i,j}$ and based on the dependency structures assumed.

[2]extension to the Multinomial is straightforward, but omitted

2.1 Model $[X, Y]$: X,Y are independent

This type of dependency pattern is expressed as

$$\pi_{i,j} = \pi_{i,+}\pi_{+,j}\,; \qquad \forall(i,j) \tag{3}$$

implying a multiplicative form for the expected (i,j)-cell frequency as $\mu_{i,j} = n_{++}\pi_{i,j}$; namely

$$\log(\mu_{i,j}) = \log(n_{++}) + \log(\pi_{i,+}) + \log(\pi_{+,j})\,; \qquad \forall(i,j). \tag{4}$$

The Log-linear model parametrise the expected (i,j)-cell frequency $\mu_{i,j}$ as

$$\log(\mu_{i,j}) = \lambda + \lambda_i^X + \lambda_j^Y\,; \qquad \forall(i,j) \tag{5}$$

where λ, λ_i^X, and λ_j^Y are unknown terms/parameters based on the sample size, the probability of i-th level of X, and the probability of j-th level of Y. By exponentiation, we observe that (5) can represent the multiplicative form of $\mu_{i,j}$ implied 3 and model $[X, Y]$.

Interpretation:

In (5), λ_i^X represents the effect of the i-th level of classification variable X, and λ_j^Y represents the effect of the j-th level of classification variable Y.

If λ's were known all the quantities describing contingency tables can be recovered in a meaningful way. E.g.,

$$\begin{aligned}\frac{\pi_{j|i}}{\pi_{j'|i}} &= \frac{\mu_{j|i}}{\mu_{j'|i}} = \exp(\lambda_j^Y - \lambda_{j'}^Y)\\ \theta_{ij}^{\mathrm{L}} &= \frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i+1,j}\mu_{i,j+1}} = \frac{\exp(\lambda_j^Y - \lambda_{j+1}^Y)}{\exp(\lambda_j^Y - \lambda_{j+1}^Y)} = 1\end{aligned}$$

etc...

Remark 1. The Log linear model in matrix form with $I = 2$, $J = 2$ is

$$\underbrace{\begin{bmatrix}\log(\mu_{11})\\ \log(\mu_{12})\\ \log(\mu_{21})\\ \log(\mu_{22})\end{bmatrix}}_{\substack{=\log(\boldsymbol{\mu})\\ \text{response}}} = \underbrace{\begin{bmatrix}1 & 1 & 0 & 1 & 0\\ 1 & 1 & 0 & 0 & 1\\ 1 & 0 & 1 & 1 & 0\\ 1 & 0 & 1 & 0 & 1\end{bmatrix}}_{\substack{=\boldsymbol{X}\\ \text{design matrix}}} \underbrace{\begin{bmatrix}\lambda\\ \lambda_1^X\\ \lambda_2^X\\ \lambda_1^Y\\ \lambda_2^Y\end{bmatrix}}_{\substack{=\boldsymbol{\beta}\\ \text{effects}}} \tag{6}$$

Suppose you want to learn $\boldsymbol{\lambda}$'s, given that $\mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{n_{++}}$; do you see any problem???

Non-Identifiability issues:

We observe that (5) has $I + J + 1$ parameters (to be learned), however we know that (4) has $1 + (I-1) + (J-1)$ non-redundant (or free) parameters, under the Poisson sampling scheme. This is because, in (4), n_{++} is not fixed –hence we have to learn it, and because each marginal probabilities π_{i+}, and π_{+j} has to sum up to 1 –hence we we need to learn only $(I-1)$ and $(J-1)$

of each. Also, several choices for $(\lambda, \lambda_i^X, \lambda_j^Y)$, e.g., $(\log(n_{++}), \log(\pi_{i,+}), \log(\pi_{+,j}))$, and $(\log(n_{++}) + C, \log(\pi_{i,+}) - C, \log(\pi_{+,j}))$, for any $C \in \mathbb{R}$ may satisfy (4). This creates, non-identifiability issues in the estimation of λ's, as well as problems to the interpretation of λ's, as there their estimates are not unique. To fix this, we impose practical constrains.

Identifiability Constraints:

We artificially impose constraints on λ's aiming at overcoming the non-identifiability issue above, and making the formulation (3) interpretable. Two popular choices are:

- Corner points: Set
$$\lambda_I^X = \lambda_J^Y = 0$$
where levels I, J are called reference levels for variables X, Y. Then λ_i^X and λ_j^Y account for deviations from the reference levels I, and J.

- Sum-to-zero: Set
$$\sum_{\forall i} \lambda_i^X = \sum_{\forall j} \lambda_j^Y = 0$$
Then λ_i^X and λ_j^Y account for deviations of X and Y from the overall effect λ.

For instance the default option in the function 'glm {stats}' in R is the Cornell points with reference levels the first levels of each category; e.g. $\lambda_1^X = \lambda_1^Y = 0$.

Remark 2. Different identifiability constraints lead to different interpretation for λ's (and hence different values when estimated), but to the same inference.
In a 2×2 table it is

$$\frac{\pi_{2|i}}{\pi_{1|i}} = \begin{cases} \exp(-\lambda_1^Y) & \text{, Corner points } \lambda_2^X = \lambda_2^X = 0 \\ \exp(2\lambda_1^Y) & \text{, Sum-ro-zeros } \lambda_1^X + \lambda_2^X = \lambda_1^Y + \lambda_2^Y = 0 \end{cases}$$

Number of free parameters (after imposing the constrains)

$$d \quad = 1 + (I-1) + (J-1) \quad = ...$$

2.2 Model $[XY]$: The saturated model. X,Y are not independent

The Log-linear model can be specified as

$$\log(\mu_{i,j}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}; \qquad \forall (i,j) \tag{7}$$

as $\pi_{i,j}$ do not have any pattern.

Non-Identifiability issue:

Similar problem, there are IJ cells and $1 + I + J + IJ$ λ's. Similar treatment

Identifiability Constraints:

- Corner points: Set

$$\begin{aligned}\lambda_I^X = \lambda_J^Y &= 0\\ \lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} &= 0; \qquad \forall (i,j)\end{aligned}$$

  with reference levels I, J fof X and Y.

- Sum-to-zero: Set

$$\begin{aligned}\sum_{\forall i} \lambda_i^X = \sum_{\forall j} \lambda_j^Y &= 0\\ \sum_{\forall i} \lambda_{ij}^{XY} = \sum_{\forall j} \lambda_{ij}^{XY} &= 0; \qquad \forall (i,j)\end{aligned}$$

Number of free parameters

$$d \quad = 1 + (I-1) + (J-1) + (I-1)(J-1) \quad = ...$$

Example 3. Consider 2×2 table, then

$$\log(\theta) = \begin{cases} \lambda_{11}^{XY} & \text{, Corner points } \lambda_2^X = \lambda_2^X = \lambda_{21}^{XY} = \lambda_{12}^{XY} = \lambda_{22}^{XY} = 0 \\ 2\lambda_{11}^{XY} + 2\lambda_{11}^{XY} & \text{, Sum-ro-zeros } \lambda_1^X + \lambda_2^X = \lambda_1^Y + \lambda_2^Y = 0 \\ & \quad \text{and } \lambda_{11}^{XY} + \lambda_{12}^{XY} = \lambda_{11}^{XY} + \lambda_{21}^{XY} = \lambda_{21}^{XY} + \lambda_{22}^{XY} = 0 \end{cases}$$

2.3 MLE of λ's

The MLE is discussed in Section 5. For now assume that $\hat{\mu}_{ij}$, $\hat{\pi}_{ij}$ are the MLEs of μ_{ij}, π_{ij}

2.4 Goodness of fit test (Model selection)

The pair of hypothesis test are as follows

$$\begin{cases} \mathrm{H}_0: & (X,Y) \\ \mathrm{H}_1: & (XY) \end{cases} \Longrightarrow \begin{cases} \mathrm{H}_0: & X, Y \text{ are independent} \\ \mathrm{H}_1: & X, Y \text{ are not independent} \end{cases}$$

The test can be performed based on the following two statistics:

- Pearson's statistic

$$X^2 \overset{\mathrm{H}_0}{=} \sum_{\forall i,j} \frac{(n_{ij} - \hat{\mu}_{ij}^{(0)})^2}{\hat{\mu}_{ij}^{(0)}} \xrightarrow[D]{} \chi^2_{\mathrm{df}}$$

- Likelihood ratio statistic

$$G^2 \overset{\mathrm{H}_0}{=} 2\sum_{\forall i,j} n_{ij} \log(\frac{n_{ij}}{\hat{\mu}_{ij}^{(0)}}) \underset{D}{\longrightarrow} \chi^2_{\mathrm{df}}$$

The degrees of freedom are

$$\mathrm{df} = d_1 - d_0$$

where

$$d_0 : \# \text{ of free parameters in the model under } \mathrm{H}_0$$
$$d_1 : \# \text{ of free parameters in the model under } \mathrm{H}_1$$

and

$$\hat{\mu}_{ij}^{(0)} = \frac{n_{i+}n_{+j}}{n} : \text{the MLE of } \mu_{ij} \text{ under the model in } \mathrm{H}_0$$

The rejection areas at sig. level a are:

- For Pearson's statistic

$$R(\{n_{ij}\}) = \{X^2_{\mathrm{obs}} \geq \chi^2_{\mathrm{df},1-a}\}$$

- For Likelihood ratio statistic

$$R(\{n_{ij}\}) = \{G^2_{\mathrm{obs}} \geq \chi^2_{\mathrm{df},1-a}\}$$

3 Dependency models in $I \times J \times K$ contingency tables

Consider:

- a $I \times J \times K$ contingency table $(n_{i,j,k})$
- $\pi_{i,j,k}$ as the (i,j,k)-cell probabilities
- $\mu_{i,j,k} = \mathrm{E}(N_{i,j,k}) = n_{+++}\pi_{i,j,k}$ as the expected (i,j,k)-cell frequencies
- the Poisson log linear model[3].

3.1 Model $[X,Y,Z]$: X,Y,Z are independent

Because

$$\pi_{i,j,k} = \pi_{i,+,+}\pi_{+j+}\pi_{+,+,k}; \qquad \forall(i,j,k)$$

the Log-linear model is specified as

$$\log(\mu_{i,j,k}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z; \qquad \forall(i,j,k)$$

Identifiability Constraints:

[3] extension to the Multinomial is straightforward, but omitted

- Corner points: Set
$$\lambda_I^X = \lambda_J^Y = \lambda_K^Z = 0$$
with reference levels I, J, K for variables X, Y, Z.

- Sum-to-zero: Set
$$\sum_{\forall i} \lambda_i^X = \sum_{\forall j} \lambda_j^Y = \sum_{\forall k} \lambda_k^Z = 0$$

Number of free parameters

$$d \quad = 1 + (I-1) + (J-1) + (K-1) \quad = ...$$

3.2 Model $[X, YZ]$: Joint independent X from Y and Z

Because
$$\pi_{i,j,k} = \pi_{i++}\pi_{+jk}; \qquad \forall (i,j,k)$$
the Log-linear model is specified as
$$\log(\mu_{i,j,k}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}; \qquad \forall (i,j,k)$$

Identifiability Constraints:

- Corner points: Set
$$\lambda_I^X = \lambda_J^Y = \lambda_K^Z = 0$$
$$\lambda_{jK}^{YZ} = \lambda_{Jk}^{YZ} = 0, \ \forall j,k$$
with reference levels I, J, K for variables X, Y, Z.

- Sum-to-zero: Set
$$\sum_{\forall i} \lambda_i^X = \sum_{\forall j} \lambda_j^Y = \sum_{\forall k} \lambda_k^Z = 0$$
$$\sum_{\forall j} \lambda_{jk}^{YZ} = \sum_{\forall k} \lambda_{jk}^{YZ} = 0, \ \forall j,k$$

Number of free parameters

$$d \quad = 1 + (I-1) + (J-1) + (K-1) + (J-1)(K-1) \quad = ...$$

3.3 Model $[XZ, YZ]$: X and Y are conditionally independent on Z

Because

$$\pi_{i,j,k} = \frac{\pi_{i,+,k}\pi_{+,j,k}}{\pi_{+,+,k}}; \qquad \forall(i,j,k)$$

the Log-linear model is specified as

$$\log(\mu_{i,j,k}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}; \qquad \forall(i,j,k)$$

Identifiability Constraints:

- Corner points: Set

$$\begin{aligned} \lambda_I^X = \lambda_J^Y = \lambda_K^Z = 0 \\ \lambda_{jK}^{YZ} = \lambda_{Jk}^{YZ} = 0,\ \forall j,k \\ \lambda_{iK}^{XZ} = \lambda_{Ik}^{XZ} = 0,\ \forall i,k \end{aligned}$$

  with reference levels I, J, K for variables X, Y, Z.

- Sum-to-zero: Set

$$\begin{aligned} \sum_{\forall i} \lambda_i^X = \sum_{\forall j} \lambda_j^Y = \sum_{\forall k} \lambda_k^Z = 0 \\ \sum_{\forall j} \lambda_{jk}^{YZ} = \sum_{\forall k} \lambda_{jk}^{YZ} = 0,\ \forall j,k \\ \sum_{\forall i} \lambda_{ik}^{XZ} = \sum_{\forall k} \lambda_{ik}^{XZ} = 0,\ \forall i,k \end{aligned}$$

Number of free parameters

$$d \quad = 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) + (J-1)(K-1) \quad = ...$$

3.4 Model $[XY, XZ, YZ]$: The homogeneous association model

Homogeneous association model in a 3-way contingency table represents the type of independence where each of the possible three pairs of classification variables is conditionally independent on the third one at the same time.

It is easy to check that this can can happen if the (i,j,k)-cell probability is of the form

$$\pi_{i,j,k} = \psi_{ij}\phi_{jk}\omega_{ik}; \qquad \forall(i,j,k) \tag{8}$$

for some terms ψ_{ij}, ϕ_{jk}, and ω_{ik}. No closed form expression exists for ψ_{ij}, ϕ_{jk}, and ω_{ik} in terms of the marginal cell-probabilities.

Hence, the Log-linear model is specified as

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

after a log transformation of $\mu_{i,j,k} = n_{+++}\pi_{i,j,k}$ and given (8).

Identifiability Constraints:

- Corner points: Set

$$\begin{aligned}\lambda_I^X = \lambda_J^Y = \lambda_K^Z = 0\\ \lambda_{jK}^{YZ} = \lambda_{Jk}^{YZ} = 0,\ \forall j,k\\ \lambda_{iK}^{XZ} = \lambda_{Ik}^{XZ} = 0,\ \forall i,k\\ \lambda_{iJ}^{XY} = \lambda_{Ij}^{XY} = 0,\ \forall i,j\end{aligned}$$

  with reference levels I, J, K for variables X, Y, Z.

- Sum-to-zero: Set

$$\begin{aligned}\sum_{\forall i}\lambda_i^X = \sum_{\forall j}\lambda_j^Y = \sum_{\forall k}\lambda_k^Z = 0\\ \sum_{\forall j}\lambda_{jk}^{YZ} = \sum_{\forall k}\lambda_{jk}^{YZ} = 0,\ \forall j,k\\ \sum_{\forall i}\lambda_{ik}^{XZ} = \sum_{\forall k}\lambda_{ik}^{XZ} = 0,\ \forall i,k\\ \sum_{\forall i}\lambda_{ij}^{XY} = \sum_{\forall j}\lambda_{ij}^{XY} = 0,\ \forall i,j\end{aligned}$$

Number of free parameters

$d \quad = 1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1) \quad = ...$

This model is also called,

- the homogeneous association model. This is because the model represents a homogeneous association between each pair of variables at each level of the third variable.

  Explain: It is

$$\log(\theta_{ij(k)}^L) = \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i+1,j}^{XY} - \lambda_{i,j+1}^{XY}$$

  and hence independent of k.

  Check that the same applies for $\log(\theta_{i(j)k}^L)$ and $\log(\theta_{(i)jk}^L)$.

- no 3 factor model
- 3 interactions model

3.5 Model $[XYZ]$: The saturated model

Well, this is a quite boring model as it does not imply any structure or any type of independence.

The Log-linear model can be defined as

$$\log(\mu_{i,j,k}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}; \qquad \forall(i,j,k)$$

Identifiability Constraints:

- Corner points: Set

$$\begin{aligned}
\lambda_I^X = \lambda_J^Y = \lambda_K^Z &= 0 \\
\lambda_{jK}^{YZ} = \lambda_{Jk}^{YZ} &= 0,\ \forall j,k \\
\lambda_{iK}^{XZ} = \lambda_{Ik}^{XZ} &= 0,\ \forall i,k \\
\lambda_{iJ}^{XY} = \lambda_{Ij}^{XY} &= 0,\ \forall i,j \\
\lambda_{iJK}^{XYZ} = \lambda_{IjK}^{XYZ} = \lambda_{IJk}^{XYZ} &= 0,\ \forall i,j,k
\end{aligned}$$

  with reference levels I, J, K for variables X, Y, Z.

- Sum-to-zero: Set

$$\begin{aligned}
\sum_{\forall i} \lambda_i^X = \sum_{\forall j} \lambda_j^Y = \sum_{\forall k} \lambda_k^Z &= 0 \\
\sum_{\forall j} \lambda_{jk}^{YZ} = \sum_{\forall k} \lambda_{jk}^{YZ} &= 0,\ \forall j,k \\
\sum_{\forall i} \lambda_{ik}^{XZ} = \sum_{\forall k} \lambda_{ik}^{XZ} &= 0,\ \forall i,k \\
\sum_{\forall i} \lambda_{ij}^{XY} = \sum_{\forall j} \lambda_{ij}^{XY} &= 0,\ \forall i,j \\
\sum_{\forall i} \lambda_{ijk}^{XYZ} = \sum_{\forall j} \lambda_{ijk}^{XYZ} = \sum_{\forall k} \lambda_{ijk}^{XYZ} &= 0,\ \forall i,j,k
\end{aligned}$$

Number of free parameters

$$\begin{aligned}
d =& 1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1) \\
&+ (I-1)(J-1)(K-1) \\
=& IJK
\end{aligned}$$

3.6 Hierarchical models

We consider only hierarchical models. It means that if an interaction term of two variables is in the model, then all the marginal terms (lower lever interations) should be included in the model. This facilitates inference and interpretation of λ's.

This is a consequence of the 'Principle of marginality', which says that it is wrong to

- test, estimate, or interpret main effects of explanatory variables where the variables interact, or
- model interaction effects but delete main effects that are marginal to them.

This is because non-hierarchical models lack applicability as they ignore the dependence of a variable's effect upon another variable's value.

- An example of hierarchical model is

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

because it contains both $\lambda + \lambda_i^X + \lambda_j^Y$ given that it contains λ_{ij}^{XY}

- An example of a NON hierarchical model is

$$\log(\mu_{i,j,k}) = \lambda + \lambda_i^X + \lambda_k^Z + \lambda_{ij}^{XY}; \qquad \forall(i,j,k)$$

because it does not contain λ_j^Y while it contains λ_{ij}^{XY}

4 Dependency models in multi-way tables

Well... in multi-way tables, the notation gets really crazy. It is straightforward how to define the types of dependencies in the in multi-way tables, based on what we have discussed.

E.g., in a $I \times J \times K \times Q$ table with classifiers X, Y, Z, D; the homogeneous association model is $[XY, XZ, YZ, XQ, XQ, ZQ]$

$$\log(\mu_{i,j,k,q}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_q^Q + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{iq}^{XQ} + \lambda_{jq}^{YQ} + \lambda_{kq}^{ZQ}; \qquad \forall(i,j,k,q)$$

, etc...

5 Maximum Likelihood Estimation

We focus on the case of the $I \times J \times K$ contingency tables $(n_{i,j,k})$, and the Poisson log linear model; i.e. $(n_{i,j,k})$ are generated by a Poisson sampling scheme. We consider both identifiability constraints Corner points and sum-to-zero.

The following procedure can be further extended to more general cases; e,g, 100-way Tables....

5.1 Minimal sufficient statistics

Consider $I \times J \times K$ contingency table $(n_{i,j,k})$, and a Poisson sampling scheme $N_{i,j,k} \sim \text{Poi}(\mu_{i,j,k})$. So

$$f(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{\mu}) = \prod_{i,j,k} \frac{\exp(-\mu_{i,j,k})\mu_{i,j,k}^{n_{i,j,k}}}{n_{i,j,k}!}$$

where $\boldsymbol{N} = (N_{111}, ..., N_{IJK})$ and $\boldsymbol{\mu} = (\mu_{111}, ..., \mu_{IJK})$. The saturated Log-linear model is

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{i,j}^{X,Y} + \lambda_{i,k}^{XZ} + \lambda_{j,k}^{YZ} + \lambda_{i,j,k}^{XYZ}$$

- Simpler models result by letting the corresponding λ terms as zero. E.g., model (XY, Z) results by setting $\lambda_{i,k}^{XZ} = \lambda_{j,k}^{YZ} = \lambda_{i,j,k}^{XYZ} = 0$.

The log-likelihood $\ell(\boldsymbol{\mu}) = \log(L(\boldsymbol{\mu}))$ is

$$
\begin{aligned}
\ell(\boldsymbol{\lambda}) =& \sum_{i,j,k} n_{ijk} \log(\mu_{ijk}) - \sum_{i,j,k} \mu_{ijk} \qquad (9)\\
=& n_{+++}\lambda + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z \\
&+ \sum_i \sum_j n_{ij+}\lambda_{ij}^{XY} + \sum_i \sum_k n_{i+k}\lambda_{i,k}^{XZ} + \sum_j \sum_k n_{+jk}\lambda_{jk}^{YZ} \\
&+ \sum_i \sum_j \sum_k n_{ijk}\lambda_{ijk}^{XYZ} \\
&- \sum_{i,j,k} \underbrace{\exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{i,j}^{X,Y} + \lambda_{i,k}^{XZ} + \lambda_{j,k}^{YZ} + \lambda_{i,j,k}^{XYZ})}_{=\mu_{ijk}(\boldsymbol{\lambda})} - \sum_{i,j,k} \log(n_{ijk}!)
\end{aligned}
$$

- Since the Poisson distribution is member of the exponential family of distributions, coefficients of the parameters are sufficient statistics.
- For model (XYZ) the sufficient statistics are obviously all the data $\{n_{ijk}\}$. This is because all the coefficients of $\boldsymbol{\lambda}$'s can be computed by $\{n_{ijk}\}$.
- For model (XY, Z) the sufficient statistics are only $\{n_{ij+}, n_{++k}\}$. This is because, after setting $\lambda_{i,k}^{XZ} = \lambda_{j,k}^{YZ} = \lambda_{i,j,k}^{XYZ} = 0$, the remaining coefficients of $\boldsymbol{\lambda}$'s can be computed by $\{n_{ij+}, n_{++k}\}$.
- The sufficient statistics are summarized in Tables 1 and 2.

Model	Feature	
(X, Y, Z)	Sufficient statistics	$n_{i++}, n_{+j+}, n_{++k}$
	Likelihood equations	$\hat{\mu}_{i++} = n_{i++}, \hat{\mu}_{+j+} = n_{+j+}, \hat{\mu}_{++k} = n_{++k}$
	Probability form	$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$
	Fitted Value	$\hat{\mu}_{ijk} = \frac{n_{i++}n_{+j+}n_{++k}}{n^2}$
(XY, Z)	Sufficient statistics	$\{n_{ij+}, n_{++k}\}$
	Likelihood equations	$\hat{\mu}_{ij+} = n_{ij+}, \hat{\mu}_{++k} = n_{++k}$
	Probability form	$\pi_{ijk} = \pi_{ij++}\pi_{++k}$
	Fitted Value	$\hat{\mu}_{ijk} = \frac{n_{ij+}n_{++k}}{n}$
(XY, YZ)	Sufficient statistics	n_{ij+}, n_{+jk}
	Likelihood equations	$\hat{\mu}_{ij+} = n_{ij+}, \hat{\mu}_{+jk} = n_{+jk}$
	Probability form	$\pi_{ijk} = \frac{\pi_{ij+}\pi_{+jk}}{\pi_{+j+}}$
	Fitted Value	$\hat{\mu}_{ijk} = \frac{n_{ij+}n_{+jk}}{n_{+j+}}$
(XY, XZ, YZ)	Sufficient statistics	$\{n_{ij+}, n_{i+k}, n_{+jk}\}$
	Likelihood equations	$\hat{\mu}_{ij+} = n_{ij+}, \hat{\mu}_{+jk} = n_{+jk}, \hat{\mu}_{i+k} = n_{i+k}$
	Probability form	Not available
	Fitted Value	Newton Method
(XYZ)	Sufficient statistics	$\{n_{ijk}\}$
	Likelihood equations	$\hat{\mu}_{ijk} = n_{ijk}$
	Probability form	No structure
	Fitted Value	$\hat{\mu}_{ijk} = n_{ijk}$

Table 1: Sufficient statistics, Likelihood equations, Probability forms, and Fitted Values for $I \times J \times K$ contingency tables $(n_{i,j,k})$ under Poisson sampling scheme

Model	Feature	
(X, Y)	Sufficient statistics	n_{i+}, n_{+j}
	Likelihood equations	$\hat{\mu}_{i+} = n_{i+}, \hat{\mu}_{+j} = n_{+j}$
	Probability form	$\pi_{ij} = \pi_{i+}\pi_{+j}$
	Fitted Value	$\hat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n}$
(XY)	Sufficient statistics	$\{n_{ij}\}$
	Likelihood equations	$\hat{\mu}_{ij} = n_{ij}$
	Probability form	No structure
	Fitted Value	$\hat{\mu}_{ij} = n_{ij}$

Table 2: Sufficient statistics, Likelihood equations, Probability forms, and Fitted Values for $I \times J$ contingency tables $(n_{i,j})$ under Poisson sampling scheme

5.2 Calculate the Likelihood equations

Under a specific model (type of dependence) and under the identifiability constraints, MLE $\hat{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda}$'s can be found by solving:

$$\begin{aligned} &\text{maximize } \ell(\boldsymbol{\lambda}) \\ &\text{subject to } g_1(\boldsymbol{\lambda}) = ... = g_C(\boldsymbol{\lambda}) = 0 \end{aligned}$$

where $\ell(\boldsymbol{\lambda})$ is the log likelihood in (9), and $g_1(\boldsymbol{\lambda}) = 0$,...,$g_C(\boldsymbol{\lambda}) = 0$ are the identifiability constraints considered (e.g., corner points, or sum-to-zero).

The likelihood equations can be produced by using the method of Lagrange multipliers, as:

$$0 = \nabla_{\boldsymbol{\lambda},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\theta})|_{(\boldsymbol{\lambda},\boldsymbol{\theta})=(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}})} \tag{10}$$

where

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \ell(\boldsymbol{\lambda}) - \sum_{c=1}^{C} \theta_c g_c(\boldsymbol{\lambda}),$$

is the Lagrange function, and $\theta_1, ..., \theta_C$ are real values called Lagrange multipliers.

Example 4. Assume model (XY, Z), under Poisson sampling scheme. Consider sum-to-zero constraints. Find the Likelihood equations.

Solution. It is

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

So the Lagrange function is

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

Hence

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

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

$$
0 = \frac{\mathrm{d}}{\mathrm{d}\theta^X}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta})|_{(\boldsymbol{\lambda},\boldsymbol{\theta})=(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}})} \Longrightarrow 0 = \sum_i \lambda_i^X; \quad 0 = \frac{\mathrm{d}}{\mathrm{d}\theta^Y}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta})|_{(\boldsymbol{\lambda},\boldsymbol{\theta})=(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}})} \Longrightarrow 0 = \sum_j \lambda_j^Y
$$

$$
0 = \frac{\mathrm{d}}{\mathrm{d}\theta^Z}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta})|_{(\boldsymbol{\lambda},\boldsymbol{\theta})=(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}})} \Longrightarrow 0 = \sum_k \lambda_k^Z; \quad 0 = \frac{\mathrm{d}}{\mathrm{d}\theta_{i*}^{XY}}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta})|_{(\boldsymbol{\lambda},\boldsymbol{\theta})=(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}})} \Longrightarrow 0 = \sum_j \lambda_{ij}^{XY}
$$

$$
0 = \frac{\mathrm{d}}{\mathrm{d}\theta_{*j}^{XY}}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta})|_{(\boldsymbol{\lambda},\boldsymbol{\theta})=(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}})} \Longrightarrow 0 = \sum_i \lambda_{ij}^{XY}
$$

From this I get

$$
\begin{aligned}
0 =& \hat{\theta}^X = \hat{\theta}^Y = \hat{\theta}^Z = \hat{\theta}_{i*}^{XY} + \hat{\theta}_{*j}^{XY};\\
n =& \mu_{+++}(\hat{\boldsymbol{\lambda}}); && (11)\\
n_{i++} =& \mu_{i++}(\hat{\boldsymbol{\lambda}}); && (12)\\
n_{+j+} =& \mu_{+j+}(\hat{\boldsymbol{\lambda}}); && (13)\\
n_{ij+} =& \mu_{ij+}(\hat{\boldsymbol{\lambda}}); && (14)\\
n_{++k} =& \mu_{++k}(\hat{\boldsymbol{\lambda}}) && (15)\\
0 =& \sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_k \lambda_k^Z = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} && (16)
\end{aligned}
$$

but I only need (14), (15), and (16). This is because (14), (15), can reproduce (11), (13), and (13).

The matrix form An alternative manner to derive the likelihood equations is to vectorize the quantities in a Linear system. The resulting system can be easily coded as an R routine

- Let, $\boldsymbol{n} = (n_{111}, ..., n_{IJK})$ and $\boldsymbol{\mu} = (\mu_{111}, ..., \mu_{IJK})^T$ denote the vectorized observed and expected counts of an $I \times J \times K$ table with IJK cells.
- The Log linear model equation is re-shaped as

$$
\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} \tag{17}
$$

where $\boldsymbol{\beta}$ is the vector that contains the $\boldsymbol{\lambda}$'s, $\boldsymbol{\mu} = (\mu_{111}, ..., \mu_{IJK})^T$ is the vector that contains the $\boldsymbol{\mu}$'s, and $\boldsymbol{X}$ is the design matrix that links $\boldsymbol{\beta}$ and $\log(\boldsymbol{\mu})$ such that the identification constraints are satisfied. (See Examples 5, and 6)

Example 5. (Cont. Example 4) if we have

- a $2 \times 2 \times 2$ table, with (n_{ijk})
- a Log-linear model (XY, Z), and hence

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

- Corner points indentifiability constraints by using as reference levels the level 2 of each classification variables X, Y, Z; namely $\lambda_2^X = \lambda_2^Y = \lambda_2^Z = \lambda_{i2}^{XY} = \lambda_{2j}^{XY} = 0$

then the matrix form is

$$\underbrace{\begin{bmatrix} \log(\mu_{111}) \\ \log(\mu_{112}) \\ \log(\mu_{121}) \\ \log(\mu_{122}) \\ \log(\mu_{211}) \\ \log(\mu_{212}) \\ \log(\mu_{221}) \\ \log(\mu_{222}) \end{bmatrix}}_{=\log(\boldsymbol{\mu})} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}}_{=\boldsymbol{X}} \underbrace{\begin{bmatrix} \lambda \\ \lambda_1^X \\ \lambda_1^Y \\ \lambda_1^Z \\ \lambda_{11}^{XY} \end{bmatrix}}_{=\boldsymbol{\beta}}$$

Example 6. (Cont. Example 4) if we have

- a $2 \times 2 \times 2$ table, with (n_{ijk})
- a Log-linear model (XY, Z), and hence

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

- Sum-to-zero indentifiability constraints, namely $\lambda_1^X + \lambda_2^X = 0$, $\lambda_1^Y + \lambda_2^Y = 0$, $\lambda_1^Z + \lambda_2^Z = 0$, $\lambda_{11}^{XY} + \lambda_{12}^{XY} = 0$, $\lambda_{11}^{XY} + \lambda_{21}^{XY} = 0$, and $\lambda_{21}^{XY} + \lambda_{22}^{XY} = 0$

then the matrix form is

$$\underbrace{\begin{bmatrix}\log(\mu_{111})\\ \log(\mu_{112})\\ \log(\mu_{121})\\ \log(\mu_{122})\\ \log(\mu_{211})\\ \log(\mu_{212})\\ \log(\mu_{221})\\ \log(\mu_{222})\end{bmatrix}}_{=\log(\boldsymbol{\mu})} = \underbrace{\begin{bmatrix}1 & 1 & 1 & 1 & 1\\ 1 & 1 & 1 & -1 & 1\\ 1 & 1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1 & -1\\ 1 & -1 & 1 & 1 & -1\\ 1 & -1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1 & 1\\ 1 & -1 & -1 & -1 & 1\end{bmatrix}}_{=\boldsymbol{X}} \underbrace{\begin{bmatrix}\lambda\\ \lambda_1^X\\ \lambda_1^Y\\ \lambda_1^Z\\ \lambda_1^{XY}\end{bmatrix}}_{=\boldsymbol{\beta}}$$

- Hence, from (17) I have,

$$\log(\mu_i) = \sum_j X_{i,j}\beta_j, \quad \forall i = 1, ..., N$$

- The log-likelihood of the Poisson sampling scheme is

$$\begin{aligned}\ell(\boldsymbol{\mu}) &= \sum_i n_i \log(\mu_i) - \sum_i \mu_i - \underbrace{\sum_i \log(n_i!)}_{=\text{const.}}\\ &= \sum_i n_i \log(\mu_i) - \sum_i \mu_i - \text{const.}\\ &= \sum_i n_i (\sum_j X_{i,j}\beta_j) - \sum_i \exp(\sum_j X_{i,j}\beta_j) - \text{const.}\end{aligned}$$

- To derive the likelihood equations, we compute the derivative and set it equal to zero.

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}\beta_j}\ell(\boldsymbol{\mu})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\\ &= \sum_i n_i X_{i,j} - \sum_i \exp(\sum_k X_{i,k}\hat{\beta}_k) X_{i,j}\\ &= \sum_i n_i X_{i,j} - \sum_i \hat{\mu}_i X_{i,j}\end{aligned}$$

Namely, for each j

$$\sum_i n_i X_{i,j} = \sum_i \hat{\mu}_i X_{i,j}$$

where $\hat{\mu}_i(\hat{\boldsymbol{\beta}}) = \exp(\sum_j X_{i,j}\hat{\beta}_j)$. Obviously, in the matrix world, it becomes:

$$\boldsymbol{X}^T\boldsymbol{n} = \boldsymbol{X}^T\hat{\boldsymbol{\mu}}(\boldsymbol{\beta}) \tag{18}$$

Example 7. (Cont. Example 4)

Examples 5, and 6, produce the same equations:

$$n_{ij+} = \mu_{ij+}(\hat{\boldsymbol{\lambda}})$$
$$n_{++k} = \mu_{++k}(\hat{\boldsymbol{\lambda}})$$

for all i, j, k. You can double check for the $2 \times 2 \times 2$ table.

5.3 Find the fitted values

- It has been proven [6], in general that, the likelihood equations (10) log-linear models (aka the form $\mu_{ijk} = \exp(...\text{linear}...)$) (for instance ((18))) match minimal sufficient statistics to their expected values.
- A summary of the likelihood equations for different models is given in Tables 1 and 2.
- Moreover, it has been proven [6] that a unique set of fitted values both satisfy the model and the data in the minimal sufficient statistics.
- MLE of functions of parameters are the same functions of the ML estimates of those parameters.
- Therefore, the fitted values for $\hat{\mu}_{ijk}$ can be found by representing them as functions of the expected marginal counts from which the sufficient statistics are available.

Example 8. (Cont. Example 4)

Find the fitted values for μ_{ijk}

Solution. Because of model (XY, Z) it is

$$\mu_{ijk} = n_{+++}\pi_{ijk} = n_{+++}\pi_{ij+}\pi_{++k} = \frac{\mu_{ij+}\mu_{++k}}{n_{+++}}$$

So

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

- VERY IMPORTANT: Regarding the Log linear model (XY, YZ, XZ), we cannot find a tractable expression linking μ with the sufficient statistics $\{n_{i++}, n_{+j+}, n_{++k}\}$; (Recall Equation 8). Therefore, the system of non-linear equations

$$\boldsymbol{X}^T\boldsymbol{n} = \boldsymbol{X}^T\hat{\boldsymbol{\mu}}(\boldsymbol{\beta})$$

can be solved by using numerical solvers, such as Newton-Raphson (to be discussed in the Computer Practical).

5.4 Link λ's with the expected counts μ_{ijk} and their marginals.

In order to find the MLE's for the $\boldsymbol{\lambda}$'s, express $\boldsymbol{\lambda}$'s as functions of the expected counts μ_{ijk} and their marginals by using

- The log-linear model
$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{i,j}^{X,Y} + \lambda_{i,k}^{XZ} + \lambda_{j,k}^{YZ} + \lambda_{i,j,k}^{XYZ}$$
or a simpler equation associated to the models representing the underline dependencies
- The indentifiability constraints
  - Corner points, or sum-to-zero
- Then the MLE $\hat{\boldsymbol{\lambda}}$ is found by using the corresponding solutions of the likelihood equations

Example 9. (Cont. Example 4) if we have

- a $2 \times 2 \times 2$ table, with (n_{ijk})
- a Log-linear model (XY, Z), and hence
$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$$
- Corner points indentifiability constraints with reference level at $(2,2,2)$, namely $\lambda_2^X = \lambda_2^Y = \lambda_2^Z = \lambda_{i2}^{XY} = \lambda_{2j}^{XY} = 0$
- Because
$$\underbrace{\begin{bmatrix}\log(\mu_{111})\\ \log(\mu_{112})\\ \log(\mu_{121})\\ \log(\mu_{122})\\ \log(\mu_{211})\\ \log(\mu_{212})\\ \log(\mu_{221})\\ \log(\mu_{222})\end{bmatrix}}_{=\log(\boldsymbol{\mu})} = \underbrace{\begin{bmatrix}1&1&1&1&1\\ 1&1&1&0&1\\ 1&1&0&1&0\\ 1&1&0&0&0\\ 1&0&1&1&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&0\end{bmatrix}}_{=\boldsymbol{X}} \underbrace{\begin{bmatrix}\lambda\\ \lambda_1^X\\ \lambda_1^Y\\ \lambda_1^Z\\ \lambda_{11}^{XY}\end{bmatrix}}_{=\boldsymbol{\beta}}$$

it is (and notice the interpretation of $\lambda's$)

$$\begin{aligned}
\lambda &= \log(\mu_{222})\\
\lambda_1^Z &= \log(\mu_{221}) - \lambda = \log(\mu_{221}) - \log(\mu_{222})\\
\lambda_1^Y &= \log(\mu_{212}) - \lambda = \log(\mu_{212}) - \log(\mu_{222})\\
\lambda_1^X &= \log(\mu_{221}) - \lambda = \log(\mu_{221}) - \log(\mu_{222})\\
\lambda_1^{XY} &= \log(\mu_{112}) - \lambda - \lambda_1^X - \lambda_1^Y - \lambda_1^Z = \log(\mu_{112}) - \log(\mu_{122}) - \log(\mu_{212}) + \log(\mu_{222})
\end{aligned}$$

So

$$\begin{aligned}
\hat{\lambda} &= \log(\hat{\mu}_{222}) \\
\hat{\lambda}_1^Z &= \log(\hat{\mu}_{221}) - \log(\hat{\mu}_{222}) \\
\hat{\lambda}_1^Y &= \log(\hat{\mu}_{212}) - \log(\hat{\mu}_{222}) \\
\hat{\lambda}_1^X &= \log(\hat{\mu}_{221}) - \log(\hat{\mu}_{222}) \\
\hat{\lambda}_1^{XY} &= \log(\hat{\mu}_{112}) - \log(\hat{\mu}_{122}) - \log(\hat{\mu}_{212}) + \log(\hat{\mu}_{222})
\end{aligned}$$

where

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

Example 10. Consider a $2 \times 2 \times 2$ table (n_{ijk}) generated from Poisson sampling scheme. Assume a model (X, Y, Z). Consider sum-to-zero identifiability constraints. Find the MLE's for λ's.

Solution. From the table I get

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

I express the $\boldsymbol{\lambda}$'s as functions of $\boldsymbol{\mu}$'s. My Log linear model is

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

From the Sum-to-zero constraints, I get (and notice the interpretation of $\lambda's$)

$$\begin{cases} \sum_{i,j,k} \log(\mu_{ijk}) = & IJK\lambda \\ \sum_{j,k} \log(\mu_{ijk}) = & JK\lambda + JK\lambda_i^X \\ \sum_{i,k} \log(\mu_{ijk}) = & IK\lambda + IK\lambda_j^Y \\ \sum_{i,j} \log(\mu_{ijk}) = & IJ\lambda + IJ\lambda_k^Z \end{cases} \Longrightarrow \begin{cases} \lambda = & \frac{1}{IJK}\sum_{i,j,k} \log(\mu_{ijk}) \\ \lambda_i^X = & \frac{1}{JK}\sum_{i,k} \log(\mu_{ijk}) - \lambda \\ \lambda_j^Y = & \frac{1}{IK}\sum_{i,k} \log(\mu_{ijk}) - \lambda \\ \lambda_k^Z = & \frac{1}{IJ}\sum_{i,j} \log(\mu_{ijk}) - \lambda \end{cases}$$

$$\Longrightarrow \begin{cases} \hat{\lambda} = & \frac{1}{IJK}\sum_{i,j,k} \log(\hat{\mu}_{ijk}) \\ \hat{\lambda}_i^X = & \frac{1}{JK}\sum_{i,k} \log(\hat{\mu}_{ijk}) - \hat{\lambda} \\ \hat{\lambda}_j^Y = & \frac{1}{IK}\sum_{i,k} \log(\hat{\mu}_{ijk}) - \hat{\lambda} \\ \hat{\lambda}_k^Z = & \frac{1}{IJ}\sum_{i,j} \log(\hat{\mu}_{ijk}) - \hat{\lambda} \end{cases}$$

Exercise 11. Consider a $2 \times 2 \times 2$ table (n_{ijk}) generated from Poisson sampling scheme. Assume a model (X, Y, Z). Consider sum-to-zero identifiability constraints. Find the MLE's for $\boldsymbol{\lambda}$'s.

Solution. From the table I get

$$\hat{\mu}_{ij} = \frac{n_{i++}n_{+j+}}{n_{+++}}$$

I express the $\boldsymbol{\lambda}$'s as functions of μ's. My Log linear model is

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

From the Sum-to-zero constraints, I get

$$\begin{cases} \sum_{i,j} \log(\mu_{ijk}) = & IJ\lambda \\ \sum_{j} \log(\mu_{ijk}) = & J\lambda + J\lambda_i^X \\ \sum_{i} \log(\mu_{ijk}) = & I\lambda + I\lambda_j^Y \end{cases} \Longrightarrow \begin{cases} \lambda = & \frac{1}{IJ}\sum_{i,j,k} \log(\mu_{ij}) \\ \lambda_i^X = & \frac{1}{J}\sum_{j,k} \log(\mu_{ij}) - \lambda \\ \lambda_j^Y = & \frac{1}{I}\sum_{i,k} \log(\mu_{ij}) - \lambda \end{cases}$$

$$\Longrightarrow \begin{cases} \hat{\lambda} = & \frac{1}{IJ}\sum_{i,j,k} \log(\hat{\mu}_{ij}) \\ \hat{\lambda}_i^X = & \frac{1}{J}\sum_{j} \log(\hat{\mu}_{ij}) - \hat{\lambda} \\ \hat{\lambda}_j^Y = & \frac{1}{I}\sum_{i} \log(\hat{\mu}_{ij}) - \hat{\lambda} \end{cases}$$

6 Model comparison

- The aim is to learn the model structure (namely the dependence type) supported by the data.
- Assume a Log-linear model for a 3 way table, i.e. $I \times J \times K$ table. [4]
- The tools are can be categorized in the following 3 categories.

6.1 Goodness of fit test

The pair of hypothesis test are as follows

$$\begin{cases} \mathrm{H}_0: & \text{Simpler model } M_0 \\ \mathrm{H}_1: & \text{Saturated model} \end{cases}$$

- M_0 denotes a specific type of dependency e.g., (X, Y, XZ)
- The saturated model is the one with all the terms e.g., (XYZ) in 3 way tables.

The test can be based on the following two statistics:

- Pearson's statistic

$$X^2(M_0) \overset{\mathrm{H}_0}{=} \sum_{\forall i,j,k} \frac{(n_{ijk} - \hat{\mu}_{ijk}^{(0)})^2}{\hat{\mu}_{ijk}^{(0)}} \underset{D}{\longrightarrow} \chi^2_{\mathrm{df}}$$

[4] The extension of the method to N-way tables is straight forward.

- Likelihood ratio statistic

$$G^2(M_0) \overset{\mathrm{H}_0}{=} 2 \sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}^{(0)}}) \underset{D}{\to} \chi^2_{\mathrm{df}}$$

The degrees of freedom are

$$\mathrm{df} = d_1 - d_0$$

where

$$d_0 : \# \text{ of free parameters in the model under } \mathrm{H}_0$$
$$d_1 : \# \text{ of free parameters in the model under } \mathrm{H}_1$$

and

$$\hat{\mu}_{ijk}^{(0)} : \text{the MLE of } \mu_{ijk} \text{ under the model in } \mathrm{H}_0$$

The rejection areas at sig. level a are:

- For Pearson's statistic

$$R(\{n_{ijk}\}) = \{X^2_{\mathrm{obs}} \geq \chi^2_{\mathrm{df},1-a}\}$$

- For Likelihood ratio statistic

$$R(\{n_{ijk}\}) = \{G^2_{\mathrm{obs}} \geq \chi^2_{\mathrm{df},1-a}\}$$

Remarks

- Well, it can be computed by using Taylor expansion (to be proofed in likelihood methods) that $G^2(M_0) - X^2(M_0) \approx 0$ in a specific manner. So the two statistics are asymptotically equivalent.

Example 12. Consider a $I \times J \times K$ table. Design the Goodness of fit test to test if the interdependent model is desirable.

- This means, I compare (X, Y, Z) vs the Saturated (XYZ)
- Hypothesis test

$$\begin{cases} \mathrm{H}_0 : & (X, Y, Z) \\ \mathrm{H}_1 : & (XYZ) \end{cases}$$
$$\Leftrightarrow \begin{cases} \mathrm{H}_0 : & \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \\ \mathrm{H}_1 : & \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \end{cases}$$

- The GoF test above can be used with the following settings:

- MLE under the model in H_0 is

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

- For the degrees of freedom

$$\begin{aligned} d_0 &: 1 + (I-1) + (J-1) + (K-1) \\ d_1 &: IJK \end{aligned}$$

so

$$\mathrm{df} = d_1 - d_0 = IJK - I - J - K + 2$$

Example 13. Consider a $I \times J \times K$ table. Design the Goodness of fit test to test if the homogeneous association model (3 factor model) is significant

- This means, I compare (XY, XZ, YZ) vs the Saturated (XYZ)
- Hypothesis test

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

$$\Leftrightarrow \begin{cases} \mathrm{H}_0: & \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \\ \mathrm{H}_1: & \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \end{cases}$$

- The GoF test above can be used with the following settings:
- MLE under the model in H_0 is

$$\hat{\mu}_{i,j,k} = (???) \Rightarrow \text{Newton-Raphson method ...}$$

actually it is intractable and has to be computed numerically. This will be discussed in the Computer Practical 1.

- The degree of freedom is

$$\begin{aligned} \mathrm{df} &= d_1 - d_0 \\ &IJK - [IJ + IK + JK - I - J - K + 1] \\ &= (I-1)(J-1)(K-1) \end{aligned}$$

as the two models differ only on the term λ_{ijk}^{XYZ} which under the identifiability constraints adds $(I-1)(J-1)(K-1)$ parameters.

6.2 Compare nested models

Definition 14. Model M_0 is nested in model M_1, iff M_1 can be reduced to the simpler M_0 by assigning specific fixed values to the parameters of M_1. Then it is denoted as

$$M_0 \subseteq M_1$$

Example 15. Model (Y, XZ) is nested in model (XY, XZ) where

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

because the former results from the later by setting $\lambda_{ik}^{XZ} = 0$, $\forall i, k$.

The hypothesis test: The pair of hypothesis test are as follows

$$\begin{cases} \mathrm{H}_0 : & \text{model (aka dependency type) } M_0 \\ \mathrm{H}_1 : & \text{model (aka dependency type) } M_1 \end{cases}$$

where M_0 is nested in M_1.

Consider the Likelihood ratio statistic

$$G^2(M_0|M_1) \overset{\mathrm{H}_0}{=} 2 \sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}_{ijk}^{(1)}}{\hat{\mu}_{ijk}^{(0)}}) \underset{D}{\longrightarrow} \chi^2_{\mathrm{df}} \tag{19}$$

with MLEs,

$$\hat{\mu}_{ijk}^{(0)} : \text{the MLE of } \mu_{ijk}^{(0)} \text{ under the model in } M_0$$
$$\hat{\mu}_{ijk}^{(1)} : \text{the MLE of } \mu_{ijk}^{(1)} \text{ under the model in } M_1$$

and degrees of freedom

$$\mathrm{df} = d_1 - d_0$$

where

$$d_0 : \# \text{ of free parameters in the model } M_0$$
$$d_1 : \# \text{ of free parameters in the model } M_1$$

The corresponding rejection areas at sig. level a is:

$$R(\{n_{ijk}\}) = \{G^2(M_0|M_1) \geq \chi^2_{\mathrm{df},1-a}\}$$

Remarks

1. $G^2(M_0|M_1)$ has the memorizing notation interpreted as a metric measuring the adequacy of M_0 given that M_1 hold as reliable.
2. If M_1 corresponds to the saturated model, then the above nested test of hypotheses is the Goodness of Fit test with deviance/likelihood ratio statistic in Section 6.1 ; aka $G^2(M_0|M_1 = \text{saturated}) \equiv G^2(M_0)$.
3. If we have models M_0, M_1, and M_2 such that $M_0 \subseteq M_1 \subseteq M_2$ and

$$G^2(M_0|M_2) = 2\sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}^{(2)}_{ijk}}{\hat{\mu}^{(0)}_{ijk}}) \xrightarrow[D]{\text{H}_0} \chi^2_{\text{df}_0}$$

$$G^2(M_1|M_2) = 2\sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}^{(1)}_{ijk}}{\hat{\mu}^{(0)}_{ijk}}) \xrightarrow[D]{\text{H}_0} \chi^2_{\text{df}_1}$$

then

$$G^2(M_0|M_1) = G^2(M_0|M_2) - G^2(M_1|M_2) \xrightarrow[D]{(?)} \chi^2_{\text{df}_1-\text{df}_0} \tag{20}$$

the rational is because

$$\begin{aligned} G^2(M_0|M_1) &\overset{\text{H}_0}{=} 2\sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}^{(1)}_{ijk}}{\hat{\mu}^{(0)}_{ijk}}) \\ &= \underbrace{2\sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}^{(2)}_{ijk}}{\hat{\mu}^{(0)}_{ijk}})}_{G^2(M_0|M_2)\xrightarrow[D]{}\chi^2_{\text{df}_0}} - \underbrace{2\sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}^{(2)}_{ijk}}{\hat{\mu}^{(1)}_{ijk}})}_{G^2(M_1|M_2)\xrightarrow[D]{}\chi^2_{\text{df}_1}} \\ &\xrightarrow[D]{(?)} \chi^2_{\text{df}_0-\text{df}_1} \end{aligned}$$

how the asymptotic distribution $\chi^2_{\text{df}_1-\text{df}_0}$ results will become clear later on.

4. The derivation of the asymptotic distribution of the LR statistic (19) will be discussed later. However, given that we know the Goodness of fit test in Section 6.1 (with likelihood statistic), we can speculate that

$$\begin{aligned} G^2(M_0|M_1) &\overset{\text{H}_0}{=} 2\sum_{\forall i,j,k} n_{ijk} \log(\frac{\hat{\mu}^{(1)}_{ijk}}{\hat{\mu}^{(0)}_{ijk}}) \\ &= \underbrace{2\sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}^{(0)}_{ijk}})}_{G^2(M_0)\xrightarrow[D]{}\chi^2_{\text{df}_0}} - \underbrace{2\sum_{\forall i,j,k} n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}^{(1)}_{ijk}})}_{G^2(M_1)\xrightarrow[D]{}\chi^2_{\text{df}_1}} \\ &\xrightarrow[D]{(?)} \chi^2_{\text{df}_0-\text{df}_1} \equiv \chi^2_{d_1-d_0} \end{aligned}$$

where $\text{df}_0 = d_{\text{saturated}} - d_0$ and $\text{df}_1 = d_{\text{saturated}} - d_1$. We will be able to justify the step (?) later on.

Example 16. (Cont. Example 15)

- The MLE estimates are

$$\hat{\mu}_{ijk}^{(0)} = \frac{n_{ij+}n_{++k}}{n}$$
$$\hat{\mu}_{ijk}^{(1)} = \frac{n_{ij+}n_{i+k}}{n_{i++}}$$

- The degrees of freedom are

$$\text{df} = d_1 - d_0 = (I-1)(K-1)$$

since they differ only in the term λ_{ik}^{XZ}.

Example 17. Consider a $I \times J \times K$ table n_{ijk} generated from a multinomial sampling scheme; i.e.

$$(n_{111}, ..., n_{IJK}) \sim \text{Mult}(n_{+++}, \boldsymbol{\pi})$$

where $\boldsymbol{\pi} = (\pi_{111}, ..., \pi_{IJK})$, 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

$$R(\{n_{ijk}\}) = \{\underbrace{2\sum_{\forall i,j,k} n_{ijk}\log(\frac{n_{ijk}}{\hat{\mu}_{ijk}^{(0)}})}_{=G^2} > q\}$$

for some value q, where $\mu_{ijk} = n\pi_{ijk}$, and $\hat{\mu}_{ijk}^{(0)}$ are the MLE of μ_{ijk} under model M_0.

Solution. Well, the likelihood ratio test has rejection area

$$\frac{\sup_{\forall \mu: \text{ under } M_0} L(\boldsymbol{\pi} = n_{+++}\boldsymbol{\mu})}{\sup_{\forall \mu: \text{ under } M_1} L(\boldsymbol{\pi} = n_{+++}\boldsymbol{\mu})} < q' \iff \log(\frac{\sup_{\forall \mu: \text{ under } M_0} L(\boldsymbol{\pi} = n_{+++}\boldsymbol{\mu})}{\sup_{\forall \mu: \text{ under } M_1} L(\boldsymbol{\pi} = n_{+++}\boldsymbol{\mu})}) < q'' \quad (21)$$

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

- The likelihood is

$$L(\boldsymbol{\pi}) \propto \prod_{\forall i,j,k} \pi_{ijk}^{n_{ijk}} \propto \prod_{\forall i,j,k} \mu_{ij,k}^{n_{ij,k}}$$

- 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 $\hat{\mu}_{ijk}^{(0)}$ (this is assumed). Then from (21), I get

$$\{\log(\frac{L(\boldsymbol{\pi} = \hat{\boldsymbol{\mu}}/n_{+++})}{L(\boldsymbol{\pi} = \hat{\boldsymbol{\mu}}^{\text{satur.}}/n_{+++})}) < q''\} \Longrightarrow \{\underbrace{2\sum_{\forall i,j,k} n_{ijk}\log(\frac{n_{ijk}}{\hat{\mu}_{ijk}^{(0)}})}_{=G^2} > q\}$$

6.3 Non-nested models: The AIC

Formal/rigorous hypothesis test cannot be performed between/among non-nested models. An alternative approach is the use of Information Criteria. Such criteria do not have a very rigorous decision theory or probabilistic support, however they are reasonable measures, and often work acceptably...

- We will focus on the Akaike Information Criterion[5] (AIC), and Bayesian Information Criterion (BIC).

Akaike Information Criterion (AIC)

Definition 18. 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

$$\text{AIC}(M) = -2\sum_{i=1}^{n} \log f_M(x_i|\hat{\theta}_M) + 2p_M \tag{22}$$

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.

Procedure: Assume a collection of available models $\{M_1, M_2, ..., M_m\}$.

- For each model, compute the corresponding AIC(M).
- The model with the smallest AIC is supposed to be the preferable one.
  - I.e., if AIC(M_1) > AIC(M_2) then the model M_2 is more preferable than M_1.

In (22), the term $2p_M$ is called penalty term, and aims at penalizing models with many parameters; hence it rewards more parsimonious models.

~~**Explanation:** Pawitan Y., (2001, Chapter 13)~~

- ~~Let $g(\cdot)$ denote the real data generation process / sampling distribution (not my parametric model).~~
  - ~~I would call best model (or most desirable model) the model M whose parametric sampling distribution $f_M(\cdot|\theta_M)$ has the minimum distance (in terms of Kullback–Leibler divergence[6]) from $g(\cdot)$:~~

$$\text{KL}(g, f) = \int \log \frac{g(Z)}{f_M(Z|\theta_M)} \text{d}g(Z) = \text{E}^g(\log g(Z)) - \underbrace{\text{E}^g(\log f_M(Z|\theta_M))}_{=Q_M}$$

  ~~or equivalently the model M with maximum quantity~~

$$Q_M = \text{E}^g(\log f_M(Z|\theta_M))$$

[5] which Akaike called "An Information Criterion"

[6] ~~Kullback–Leibler divergence between g and f measures the "distance" between functions g and f.~~

– However, I do not know the real value of θ_M and instead I use the MLE, i.e.

$$\hat{Q}_M = \mathrm{E}^g(\log f_M(Z|\hat{\theta}_M))$$

– As a consequence $\hat{Q}_M$ is a biased estimator of Q_M, with bias it tends equal to $2p_M$. Hence $\hat{Q}_M$ tends to get larger values when the dimension of the parameter θ_M increases; hence it prefers models with more parameters.

– To correct this bias, as the best model, I choose the one that minimizes

$$\mathrm{AIC}(M) = \underbrace{-2\sum_{i=1}^{n} \log f_M(x_i|\hat{\theta}_M)}_{\text{big}} + 2\underbrace{p_M}_{\text{small}}; \tag{23}$$

I.e., it has big likelihood (aka supported by the data) and it is parsimonious (aka simple).

Comments:

- AIC penalizes the complexity of the model where complexity refers to the number of parameters in the model.

Example 19. Assume a $I \times J \times K$ table with counts (n_{ijk}), generated by a Poisson sampling scheme. Assume a model $M : (XY, Z)$. Compute the AIC for the model M.

Solution.

- The MLE of μ_{ijk} is

$$\hat{\mu}_{ijk}(\boldsymbol{\lambda}) = n_{+++}\hat{\pi}_{ij+}\hat{\pi}_{++k} = \frac{n_{ij+}n_{++k}}{n_{+++}}$$

- The number of free parameters is

$$\begin{aligned} p_M &= 1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) \\ &= IJ + K - 1 \end{aligned}$$

- The log likelihood at $\hat{\mu}_{ijk}(\boldsymbol{\lambda})$ is

$$\sum_{ijk} \log(f(n_{ijk}|\hat{\mu}_{ijk})) = -\sum_{ijk} \hat{\mu}_{ijk} + \sum_{ijk} n_{ijk} \log(\hat{\mu}_{ijk}) + \sum_{ijk} \log(n_{ijk}!)$$

- So

$$\mathrm{AIC}(M) = 2\sum_{ijk} \frac{n_{ij+}n_{i++}}{n} - 2\sum_{ijk} n_{ijk} \log(\frac{n_{ij+}n_{i++}}{n}) - 2\sum_{ijk} \log(n_{ijk}!) + 2[IJ + K - 1]$$

Well, if since we need AIC for comparison with other models, e.g. $M' : (XZ, YZ)$ we may ignore the constant term $\log(n_{ijk}!)$ as it is ...constant.

Bayesian Information Criterion (BIC)

Definition 20. 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

$$\text{BIC}(M) = -2\sum_{i=1}^{n} \log f_M(x_i|\hat{\theta}_M) + \log(N)p_M \tag{24}$$

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.

Procedure: Assume a collection of available models $\{M_1, M_2, ..., M_m\}$.

- For each model, compute the corresponding BIC(M).
- The model with the smallest BIC is supposed to be the preferable one.
  - I.e., if BIC(M_1) > BIC(M_2) then the model M_2 is more preferable than M_1.

~~**Explanation:** (For the interested reader (Konishi and Kitigawa (2008, p. 217))~~

- ~~Let $\pi(\theta_M|M)$ be a prior on the parameter θ_M given model M.~~
- ~~Considering regularity conditions (omitted here), we get~~

  ~~$$f_M(x) = \int f_M(x|\theta_M)\pi(\theta_M|M)\mathrm{d}\theta_M \approx \exp(-0.5\text{BIC}(M)) \tag{25}$$~~

  ~~where f_M is the marginal likelihood.~~
- ~~Then according to Maximum Likelihood methods, the desired model is the one maximizing $f_M(x)$ and hence minimizing BIC(M).~~
- ~~... strange to see Frequentists, using priors...~~

Comments:

- It can measure the efficiency of the parameterized model in terms of predicting the data.
- It penalizes the complexity of the model where complexity refers to the number of parameters in the model.
- Approximation (25) is only valid for sample size n is much larger than the number k of parameters in the model.
- BIC cannot handle large complex collections of models as in the variable selection (or feature selection) problem in high-dimension with many variables.

Example 21. Continuing Example 19. Find the BIC of M.

Solution. It is

$$\text{BIC}(M) = 2\sum_{ijk} \frac{n_{ij+}n_{i++}}{n} - \sum_{ijk} n_{ijk} \log(\frac{n_{ij+}n_{i++}}{n}) - \sum_{ijk} \log(n_{ijk}!) + \log(n)[IJ + K - 1]$$

Well, since we need BIC for comparison with other models, e.g. $M' : (XZ, YZ)$, we may ignore the constant term $\log(n_{ijk}!)$ as it is ...constant.

Exercise 22. Continue the Example 19, and consider that you may want to compare Model $M : (XY, Z)$ with model $M' : (XZ, YZ)$

1. Compute the AIC and BIC for model M'.
2. Which condition the counts should satisfy in order for M to be more desirable than M' based on AIC?

Solution.

1.

- The MLE of μ_{ijk} is

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

- The number of free parameters is

$$\begin{aligned} p_M &= 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) + (J-1)(K-1) \\ &= (I + J - 1)K \end{aligned}$$

- The log likelihood at $\hat{\mu}_{ijk}(\lambda)$ is

$$\sum_{ijk} \log(f(n_{ijk}|\hat{\mu}_{ijk})) = -\sum_{ijk} \hat{\mu}_{ijk} + \sum_{ijk} n_{ijk} \log(\hat{\mu}_{ijk}) - \sum_{ijk} \log(n_{ijk}!)$$

- So

$$\begin{aligned} \text{AIC}(M') =& 2\sum_{ijk} \frac{n_{i+k}n_{+jk}}{n_{++k}} - 2\sum_{ijk} n_{ijk} \log(\frac{n_{i+k}n_{+jk}}{n_{++k}}) - 2\sum_{ijk} \log(n_{ijk}!) \\ &+ 2(I + J - 1)K \\ \text{BIC}(M') =& 2\sum_{ijk} \frac{n_{i+k}n_{+jk}}{n_{++k}} - 2\sum_{ijk} n_{ijk} \log(\frac{n_{i+k}n_{+jk}}{n_{++k}}) - 2\sum_{ijk} \log(n_{ijk}!) \\ &+ \log(n)(I + J - 1)K \end{aligned}$$

2. It has to be

$$\text{AIC}(M) < \text{AIC}(M')$$

etc...

6.4 Sequential search for Best model

In applications with large contingency tables with many many classification variables, Frequentist statistics has three generally accepted sequential procedures (recipes) to find the 'Best' model among many: the Backward elimination ; the Forward selection ; and the Stepwise selection.

- These sequential procedures move between nested models in each step.
- There is no general guaranty that any of them converges to the 'Real' model.
- None of the three is commonly accepted against the others.

When I was a student I felt that the Backward elimination (which starts from the complex model and progressively simplifies it) was more reliable ; I felt that Forward selection and Stepwise selection (which start from the simplest model and extend it) may get be prone to get trapped in poor models.

Assume a 3 way contingency table with classification variables X, Y, Z.

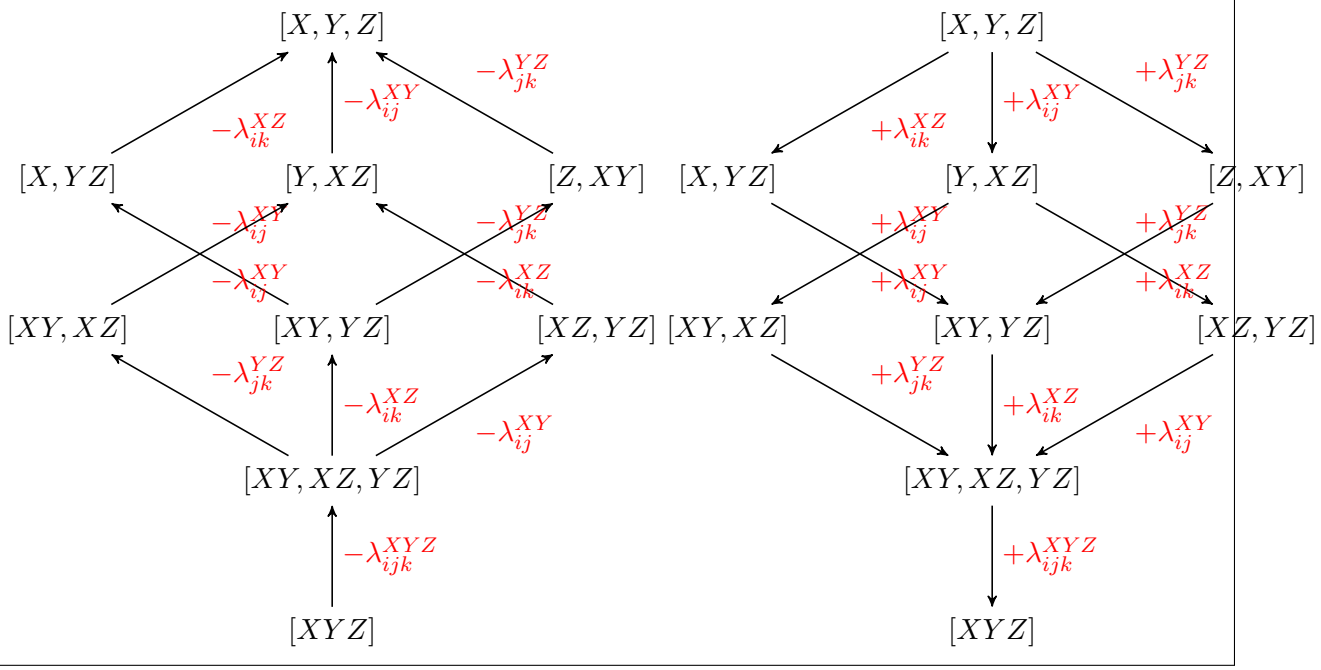
A schematic of how the Backward elimination, and Forward selection procedures evolve

Backward Elimination process

The Backward elimination proceeds with direction from $[XYZ]$ to $[X, Y, Z]$.

Forward Selection process

The Forward selection proceeds with direction from $[X, Y, Z]$ to $[XYZ]$.



6.4.1 Backward elimination

It starts from the most complex model, aka the saturated model, and sequentially tries to exclude one term based on pre-specified criteria. It goes as follows:

1. [Initialization]

   SET M_1 as the saturated model;

   I.E.; SET $M_1 \leftarrow$ SATURATED MODEL

2. [Selection]

   From all the interaction terms in model M_1, SELECT for removal the one term whose removal causes the least damage effect in fitting and leads to hierarchical model.

   I.E., among the models which result by removing one interaction term and which are still hierarchical, SELECT that, say M^*, with the

   - smallest $G^2(M^*)$, or
   - smallest $X^2(M^*)$, or
   - largest AIC(M^*), or
   - largest BIC(M^*).

   [just choose the criterion of your preference, among (G^2, X^2, AIC, BIC, etc...) and stick with it during the whole procedure]

   and SET $M_0 \leftarrow M^*$

3. [Termination criterion]

   If model M_0 does not give a proper fit then choose M_1 ; I.E.,

   - rejects hypothesis {H_0 : model M_0} against {H_1 : model M_1} based nested hypothesis test with statistic $G(M_0|M_1)$, aka p-value less than the specified sig. level.
   - AIC(M_0) > AIC(M_1)
   - BIC(M_0) > BIC(M_1)

     [just choose the criterion of your preference, among (G^2, X^2, AIC, BIC, etc...) and stay with it during the whole procedure]

   then

   SET M_1 as the 'Best' model;

   otherwise

   SET $M_1 \leftarrow M_0$,

   GOTO Step (2)

6.4.2 Forward selection

It starts from the most simple model, aka the model of independence, and sequentially tries to include one interaction term based on pre-specified criteria. It goes as follows:

1. [Initialization]

   SET M_0 as the model of independency;

   I.E.; SET $M_0 \leftarrow$ INDEPENDENCYMODEL

2. [Selection]

   From all the terms that result as interactions of existing terms in model M_0, SELECT for inclusion the one term whose inclusion gives the greatest improvement in fit and leads to a hierarchical model.

   I.E., among the models which result by including an interaction term associated to the existing terms in model M_0 and which are still hierarchical, SELECT that, say M^*, with the

   - smallest $G^2(M^*)$, or
   - smallest $X^2(M^*)$, or
   - smallest AIC(M^*), or
   - smallest BIC(M^*).

   [just choose the criterion of your preference, among (G^2 , X^2, AIC, BIC, etc...) and stick with it during the whole procedure]

   and SET $M_1 \leftarrow M^*$

3. [Termination criterion]

   If model M_1 does not significantly / importantly improve fit then choose M_0 ; I.E.,

   - does not reject hypothesis {H_0 : model M_0} against {H_1 : model M_1} based nested hypothesis test with statistic $G(M_0|M_1)$, aka p-value less than the specified sig. level.
   - AIC(M_0) < AIC(M_1)
   - BIC(M_0) < BIC(M_1)

     [just choose the criterion of your preference, among (G^2 , X^2, AIC, BIC, etc...) and stay with it during the whole procedure]

   then

   SET M_0 as the 'Best' model;

   otherwise

   SET $M_0 \leftarrow M_1$,

   GOTO Step (2)

6.4.3 Stepwise selection

It is a slight (but important) variation of the forward selection, which after adding an interaction term aka after : ' SET $M_0 \leftarrow M_1$ ' in Step 3. [Termination criterion], it tests whether any of the existing terms in the current model M_0 can be removed.

Practice

Exercise sheet

See the exercises in the Exercise sheet.

References

[1] Agresti, A. (2013). *Categorical data analysis*. Wiley.

[2] Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.

[3] Bishop, Y. M., S. E. Fienberg, and P. W. Holland (2007). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

[4] Burden, R. L. and J. D. Faires (1989). *Numerical Analysis* (Fourth ed.). The Prindle, Weber and Schmidt Series in Mathematics. Boston: PWS-Kent Publishing Company.

[5] Kateri, M. (2014). Contingency table analysis. *Statistics for Industry and Technology 525*.

[6] Lauritzen, S. L. (1996). *Graphical models*, Volume 17. Clarendon Press.

Appendix

A Miscellaneous: Computer Practical class theory [7, 8]

A.1 The Newton's method

Newton's method is a general purpose procedure to compute numerically the solution of a system of non-linear equations given that a number of assumptions are satisfied. For more info see

In general.

- Consider a function $f : \mathbb{R}^n \to \mathbb{R}^n$.
- Assume you need to find the solution x^* of the equation
$$f(x^*) = 0 \tag{26}$$
- Newtons's method for solving the system (26) is the recursion
$$x^{(t+1)} = x^{(t)} - [\nabla_x f(x^{(t)})]^{-1} f(x^{(t)}) \tag{27}$$
for $t \in \mathbb{N}$ and for a pre-specified seed value $x^{(0)} \in \mathbb{R}^n$.
- In theory, Newton's method converges to the solution quadratically; i.e.
$$\lim_{t\to\infty} \frac{|x^{(t+1)} - x^*|_\infty}{|x^{(t)} - x^*|_\infty^2} = 0$$
under regularity conditions discussed in (Numerical analysis / R. L. Burden, J. D. Faires.)
- In practice, we run Newton's recursion several times starting from a different seed each time to avoid local trapping.

An intuitive explanation why it works From the Taylor expansion, and assuming that $\nabla_x^2 f(x)$ is continuous, I get

$$f(x_{t+1}) = f(x_t) + \nabla_x f(x_t)(x_{t+1} - x_t) + O(|x_{t+1} - x_t|^2)$$

and by ignoring the error term and rearranging the quantities I get

$$x_{t+1} \approx x_t + \nabla_x f(x_t)(f(x_{t+1}) - f(x_t))$$

If x_{t+1} was the solution, or close to that, then $f(x_{t+1}) = 0$, and hence

$$x_{t+1} \approx x_t - \nabla_x f(x_t) f(x_t)$$

We can imagine that the gradient $\nabla_x f$ times the value of f at x_t leads the sequence towards locations where f is zero.

So, it may work, eventually ...

[7] CP1: http://www.maths.dur.ac.uk/~mffk55/Teaching_Topics_in_statistics_term_1_2018-2019/Computer_practical_1/Computer_practical_1.nb.html

[8] CP2: http://www.maths.dur.ac.uk/~mffk55/Teaching_Topics_in_statistics_term_1_2018-2019/Computer_practical_2/Computer_practical_2.nb.html

Pseudo-algorithm of Newton's method:

Aim: Approximate the solution of $f(x) = 0$

Input: number of equations n; initial approximation $x^{(0)} = (x_1^{(0)}, ..., x_n^{(0)}) \in \mathbb{R}^n$; tolerance τ ; maximum number of iterations T

Output: Approximate solution $x^* \in \mathbb{R}^n$; number of iterations performed t^*; relative error $\tau^* = |x^{(t^*)} - x^{(t^*-1)}|_\infty$;

Step 1: Set $x_{\text{opt}} = x^{(0)}$

Step 2: Set $t = 1$

Step 3: While $(t \leq T)$ do:

Step 3.1: Compute $n \times 1$ vector $F \in \mathbb{R}^n$ whose i-th element is $F_i = f(x_{\text{opt},i})$

Step 3.2: Compute $n \times n$ vector $J \in \mathbb{R}^{n \times n}$ whose (i,j)-th element is $J_{i,j} = \frac{\mathrm{d}}{\mathrm{d}x_j} f_i(x_{\text{opt}})$ for $(i,j) \in \{1, ..., n\}^2$

Step 3.3: Solve the $n \times n$ linear system $Jy = -F$ and compute $y \in \mathbb{R}^n$

Step 3.4: Update $x_{\text{opt}} = x_{\text{opt}} + y$

Step 3.5: Compute $\epsilon^* = |y|_\infty$

Step 3.6: If $\epsilon^* < \tau$, then escape from the loop

Step 3.7: Increase the time step $t = t + 1$

Step 4: Set $x^* = x_{\text{opt}}$

Step 5: Return as output: x^*, t^*, and ϵ^*.

Example 23. Solve the system of non-linear equations

$$\begin{cases} \cos(x_2 x_3) + \frac{1}{2} & = 3x_1 \\ 81(x_2 + 0.1)^2 & = x_1^2 + (x_3 + 0.1)^2 + \sin(x_3) + 1.06 \\ -\frac{10\pi - 3}{3} & = \exp(-x_1 x_2) + 20x_3 \end{cases}$$

Solution. This is equivalent to solving the system $f(x_1, x_2, x_3) = 0$ where

$$f(x) = \begin{bmatrix} 3x_1 - \cos(x_2 x_3) - \frac{1}{2} \\ x_1^2 - 81(x_2 + 0.1)^2 + \sin(x_3) + 1.06 \\ \exp(-x_1 x_2) + 20x_3 + \frac{10\pi - 3}{3} \end{bmatrix}$$

with Jacobian

$$\nabla_x f(x) = \begin{bmatrix} 3 & x_3 \sin(x_2 x_3) & x_2 \sin(x_2 x_3) \\ 2x_1 & -162(x_2 + 0.1) & \cos(x_3) \\ -x_2 \exp(-x_1 x_2) & -x_1 \exp(-x_1 x_2) & 20 \end{bmatrix}$$

I will feed the above to the Newton's algorithm, and consider a tolerance, e.g. $1e-4$ (meaning 10^{-4}), seed value e.g., $x^{(0)} = (0.1, 0.1, -0.1)^T$.

- Possibly after 5 iterations it 'll produce something close to $x^* = (0.5, 0, -0.523359...)^T$.
- Double check with R packages
  - nleqslv {nleqslv} available from https://cran.r-project.org/web/packages/nleqslv/index.html, or
  - optim {stats} a core package in R (it contains other procedures more general)

Implementation to the Log linear model.

- I wish to solve non-linear equation $\boldsymbol{X}^T\boldsymbol{n} = \boldsymbol{X}^T\hat{\boldsymbol{\mu}}(\boldsymbol{\beta})$ in (18), e.g., for the model (XY, XZ, YZ).
- Equivalently, I want to find $\hat{\boldsymbol{\beta}}$ for $f(\hat{\boldsymbol{\beta}}) = 0$, where $f(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}^T(\boldsymbol{n} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$
- The Jacobean is

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \boldsymbol{X}^T(\boldsymbol{n} - \boldsymbol{\mu}(\boldsymbol{\beta})) \ &= \nabla_{\boldsymbol{\beta}}[\boldsymbol{X}^T\boldsymbol{\mu}(\boldsymbol{\beta})] \\ &= \boldsymbol{X}^T \mathrm{diag}(\boldsymbol{\mu}(\boldsymbol{\beta}))\boldsymbol{X} \end{aligned}$$

  Because the (j,k)th element of $\nabla_{\boldsymbol{\beta}}[\boldsymbol{X}^T\boldsymbol{\mu}(\boldsymbol{\beta})]$ is

$$\begin{aligned} \left[\nabla_{\boldsymbol{\beta}}[\boldsymbol{X}^T\boldsymbol{\mu}(\boldsymbol{\beta})]\right]_{j,k} &= -\frac{\mathrm{d}}{\mathrm{d}\beta_k} \sum_i X_{i,j} \exp(\sum_j X_{i,j}\beta_j) \\ &= -\sum_i X_{i,j} \exp(\sum_j X_{i,j}\beta_j) X_{i,k} \end{aligned}$$

  since $\mu_i(\boldsymbol{\beta}) = \exp(\sum_j X_{i,j}\beta_j)$.
- Then the Newton's recursion (27) becomes

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + [\boldsymbol{X}^T \mathrm{diag}(\boldsymbol{\mu}(\boldsymbol{\beta}_t))\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{n} - \boldsymbol{\mu}(\boldsymbol{\beta}_t))$$

- It is proven that $\boldsymbol{\beta}_t \to \hat{\boldsymbol{\beta}}$.

A.2 The Iterative Proportions Fitting method

The iterative proportional fitting (IPF) algorithm is a simple method for calculating μ_{ijk} for log-linear models. The main idea of the procedure is the following:

- Start with $\mu^{(0)}_{ijk...}$ satisfying a model no more complex than the one being fitted. E.g., $\mu^{(0)}_{ijk...} = 1.0$ should be ok.
- For $t = 1, ...,$
  - adjust $\mu^{(t)}_{ijk}$ to match by multiplying each marginal table in the set of minimal sufficient statistics, by appropriate factors
  - escape the loop, when the maximum difference between the sufficient statistics and their fitted values is sufficiently close to zero.

Illustration: Consider 3-way, $I \times J \times K$ tables, and with classifiers X, Y,Z. Given the model (XY, XZ, YZ) design a IPF recursion producing estimates for μ_{ijk}'s

- I know that the minimal sufficient statistics are $\{n_{ij+}\}$, $\{n_{i+k}\}$, $\{n_{+jk}\}$.
- Assume that the approximated μ_{ijk}'s from the $(t-1)$-th cycle is $\mu^{(t-1)}_{ijk}$. Then the t-th cycle of the IPF algorithm has the following steps:
  1. Set $m^{(0)}_{ijk} = \mu^{(t-1)}_{ijk}$
  2. Compute

$$m^{(1)}_{ijk} = m^{(0)}_{ijk} \frac{n_{ij+}}{m^{(0)}_{ij+}};\ \forall i, j, k$$
$$m^{(2)}_{ijk} = m^{(1)}_{ijk} \frac{n_{i+k}}{m^{(1)}_{i+k}};\ \forall i, j, k$$
$$m^{(3)}_{ijk} = m^{(2)}_{ijk} \frac{n_{+jk}}{m^{(2)}_{+jk}};\ \forall i, j, k$$

  - Set $\mu^{(t)}_{ijk} = m^{(3)}_{ijk}$, $\forall i, j, k$

  ...and produces $\mu^{(t)}_{ijk}$ as approximation.

Example 24. Consider 4-way, $I \times J \times K \times L$ tables, and with classifiers X, Y,Z,D. Given the model (XY, XZ, YZ, DX, DY, DZ) design a IPF recursion producing estimates for μ_{ijkc}'s

Solution. I know that the minimal sufficient statistics are $\{n_{ijk+}\}$, $\{n_{ij+c}\}$, $\{n_{i+kc}\}$, $\{n_{+jkc}\}$. Assume that the approximated μ_{ijkc}'s from the $(t-1)$-th cycle is $\mu^{(t-1)}_{ijkc}$. Then the t-th cycle of the IPF algorithm has the following steps:

1. Set $m_{ijkc}^{(0)} = \mu_{ijkc}^{(t-1)}$
2. Compute

$$m_{ijkc}^{(1)} = m_{ijkc}^{(0)} \frac{n_{ijk+}}{m_{ijk+}^{(0)}};\ \forall i,j,k,c$$
$$m_{ijkc}^{(2)} = m_{ijkc}^{(1)} \frac{n_{ij+c}}{m_{ij+c}^{(1)}};\ \forall i,j,k,c$$
$$m_{ijkc}^{(3)} = m_{ijkc}^{(2)} \frac{n_{i+kc}}{m_{i+kc}^{(2)}};\ \forall i,j,k,c$$
$$m_{ijkc}^{(4)} = m_{ijkc}^{(3)} \frac{n_{+jck}}{m_{+jck}^{(3)}};\ \forall i,j,k,c$$

Set $\mu_{ijkc}^{(t)} = m_{ijkc}^{(4)}$

...and produces $\mu_{ijkc}^{(t)}$ as approximation.

Comments

- IPF produces MLE's
- If IPF is applied to log-linear models where an exact solution can be calculated, they the produce this exact solution.
  - E.g., A recursion applied to the independent model (X, Y) in a 2-way, $I \times J$ table with classifiers X, Y; produces $\mu_{ij}^{(t)} = n_{i+}n_{+j}/n_{++}$ for $t > 2$. Double check ....

A.3 IPF method Vs. Newton method

IPF method pros:

- When the levels $I, J, K, ...$ are too many, IPF is faster.
- It cannot be applied directly
- when the table has cells with zero counts.
- It is easy to be implemented.

IPF method cons:

- It cannot be applied directly when the table has cells with zero counts.
- Newton's method can produce directly the MLE's of $\boldsymbol{\lambda}$'s and their covariance matrices (this will be discussed later in detail.). IPF needs the user to express $\lambda's$ with respect to μ_{ijk}'s in order to estimate them, because it directly estimates the μ_{ijk}'s.