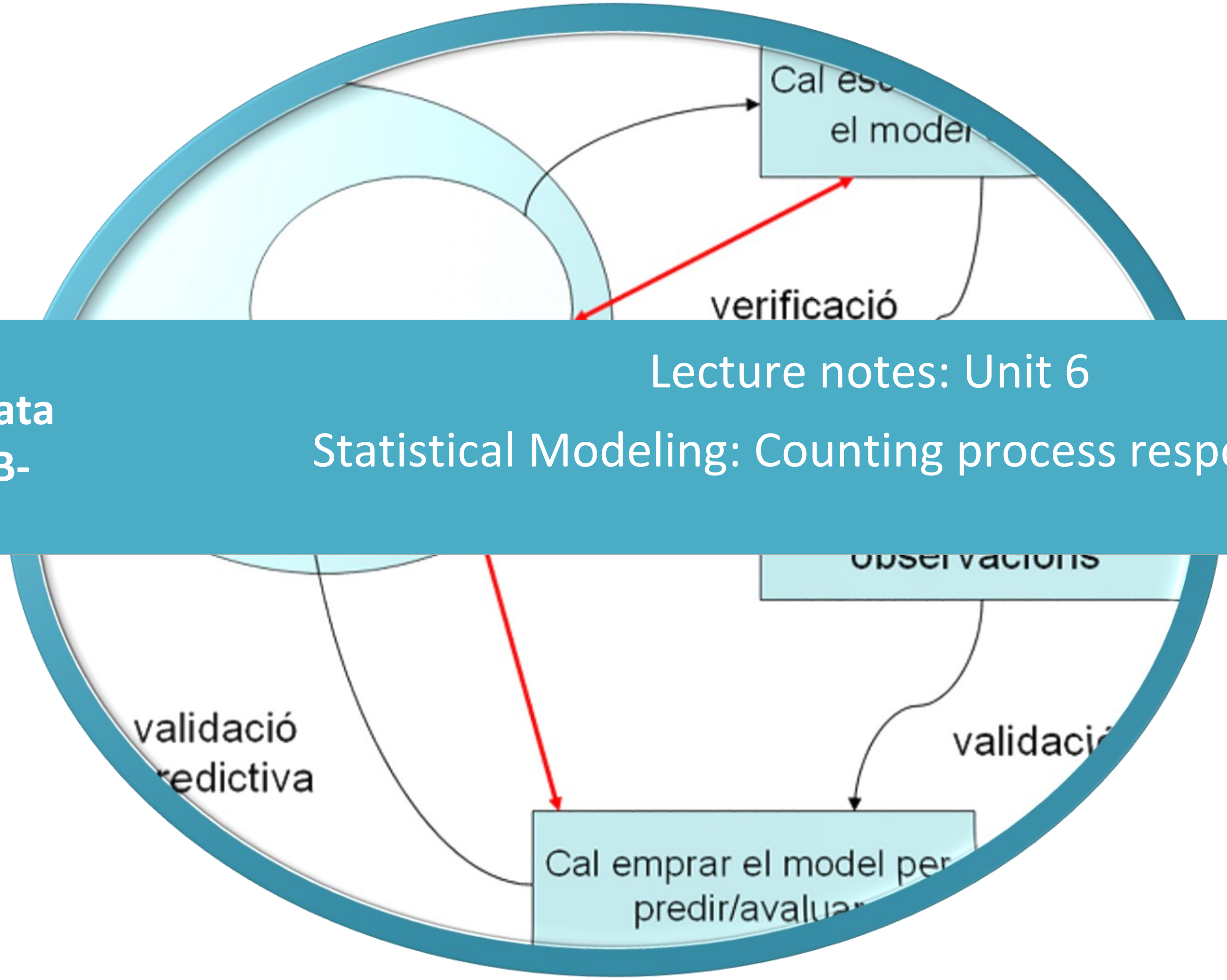
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

2024-2025 Academic Year



SIM course. Master in Data Science – FIB-UPC

Lecture notes: Unit 6

Statistical Modeling: Counting process response data

TABLE OF CONTENTS

6.1. MODELS FOR COUNTS. POISSON MODELS

6.1.1 Components of generalized linear models

Generalized linear models are extensions of classic multiple regression models.

Let $\mathbf{y}^T = (\mathbf{y}_1,\ldots,\mathbf{y}_n)$ be a vector of n components randomly drawn from vector $\mathbf{Y}^T = (\mathbf{Y}_1,\ldots,\mathbf{Y}_n)$, whose variables are statistically independent and distributed with expectation $\mu^T = (\mu_1,\ldots,\mu_n)$:

The random component assumes that mutual independence holds and each random variable in $\mathbf{Y}^T = (\mathbf{Y}_1,\ldots,\mathbf{Y}_n)$ belongs to the exponential family with one parameter distribution $Y_i|X_i \sim Pois(\mu_i), \phi = 1$ **and expected values** $\mathrm{E}(\mathrm{Y_i}|X_i) = \phi\mu_i$.

➡ **Either for grouped or individual data, the initial response model is a Poisson distribution.**

- **The systematic component in the model** specifies a vector $\boldsymbol{\eta}$. The linear predictor vector is a linear combination from a limited number of explanatory variables $\mathbf{X} = (\mathbf{X}_1,\ldots,\mathbf{X}_p)$ or regressors and parameters $\boldsymbol{\beta}^T = (\beta_1,\ldots,\beta_p)$ to be estimated. In matrix notation, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ where $\boldsymbol{\eta}$ is *nx1*, $\mathbf{X}$ is *nxp* and $\boldsymbol{\beta}$ is *px1*.

MODELS FOR COUNTS. POISSON MODELS

For each observation i, the expected value $\boxed{\mu_i}$ is related to the linear predictor $\boxed{\eta_i}$ through the scalar ***link function*, denoted *g(.)***, and thus $g(\mu_i) = \mathbf{X_i^T\beta} = \eta_i$.

The response function is $\mu_i = g^{-1}(X_i^T\beta) = g^{-1}(\eta_i)$

In ordinary least squares models for normal data, the identity link used is $\boldsymbol{\eta = \mu}$.

For counting data, several treatments are commonly used and will be presented in a later section.

Since ML estimates: $\boxed{\widehat{\boldsymbol{\beta}}\ \forall i \rightarrow \hat{\eta}_i = \mathbf{X_i^T}\widehat{\boldsymbol{\beta}}\ \ \rightarrow \hat{\mu}_i = g^{-1}(\hat{\eta}_i)}$

MODELS FOR COUNTS. POISSON MODELS

Statistical linear model classification:

Explicative Variables	Response Variable				
	Dicothomic or Binary	*Polytomous*	*Counts (discrete)*	*Continuous*	
				Normal	*Time between events*
Dicothomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	Tests for 2 subpopulation means: t.test	Survival Analysis
Polytomous	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	ONEWAY, ANOVA	Survival Analysis
Continuous (covariates)	Logistic regression	*	Log-linear models	Multiple regression	Survival Analysis
Factors and covariates	Logistic regression	*	Log-linear models	Covariance Analysis	Survival Analysis
Random Effects	Mixed models	Mixed models	Mixed models	Mixed models	Mixed models

6.2. INTRODUCTION TO COUNTING PROCESS MODELLING

- **This unit aims to cover counts as a target, proportions are not considered. The first option is considering the counting process as a Poisson variate, thus non-negative observations and unlimited large values are assumed.**

  A singular example is the one proposed by McCullagh that models the number of ship incidents (it shows an overdispersion behavior in the original analysis by the autor). See Example 5.

- Theoretically, Poisson processes account for number of independent events in a given period of time, being event rate constant by time unit. Under Poisson hypothesis: $\boxed{V[Y_i] = \mathbf{E}[Y_i] = \mu_i}$ variance is equal to expected value. Anyway, it is easy to observe in practice many situations where these restrictive hypothesis do not hold.

- Nelder and Wedderburn proposed an alternative to specify expection and variance proportional to expection that leads to maximum quasi-likelihood estimation (MQLE): $\boxed{V[Y_i] = \phi\, \mathbf{E}[Y_i] = \phi\, \mu_i}$.

INTRODUCTION TO COUNTING PROCESS MODELLING

- If $\phi = 1$ variance and expection are idential and Poisson hypothesis is satisfied.
- If $\phi > 1$, $V[Y_i] = \phi\,\mathbf{E}[Y_i] = \phi\,\mu_i$, then overdispersion is present and variance of the estimates is $\mathbf{V}[\beta] = \phi\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}$. Assuming a Poisson hypothesis variance of the estimates is conservative (whenever overdispersion holds).
- Overdispersion parameter estimate according to McCullagh for a given model consists on generalized Pearson statistic for the model divided into its degrees of freedom,

$$\hat{\phi} = \frac{X^2}{n-p}$$

➡ I would like to draw your attention to the case of a large Pearson statistic for a given model: it would lead to a $\hat{\phi} > 1$ estimate that can be confusing. It might refer either to a true overdispersion situation, or to lack of fit for the proposed model (potentially solved by including extra explanatory variables in the model).

➡ So, overdispersion parameter has to be estimated using model containing as many significant explanatory variables (colinearity avoided) as possible.

➡ All ***models are log-linear models***, a logarithmic link is considered in a such a way that the expected value for the target parameter depends on a multiplicative base from the explanatory variables …

INTRODUCTION TO COUNTING PROCESS MODELLING

(Cont.) Log-linear models: functional form

Let target observation vector have *n* components, $\mathbf{y}^T = (y_1,\ldots,y_n)$, with independent componenets and distributed with expected means $\boldsymbol{\mu}^T = (\mu_1,\ldots,\mu_n)$ and linked to the linear predictor through: $\boxed{\log(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} \quad i = 1,\ldots,n}$.

- If explanatory variables are factors, then there exists a clear analogy to analysis of variance models.
- Connections between log-linear models and multinomial response models will be highlighted at the end of the topic.
- Binomial and multinomial distributions are suitable for modelling proportions, as the ones arising in binary and polytomous target modelling, respectively, whenever the total number of observations for each covariate class is known. Counting processes modelling through is not upper bounded.
- Basic Poisson law description having $\boxed{\mu}$ parameter is:

Probability function: $\boxed{p_Y(y) = \frac{\mu^y}{y!}e^{-\mu} \quad y = 0,1,\ldots}$, $\boxed{\mathrm{E}[Y] = \mu}$ and $\boxed{V[Y] = \mu}$.

6.2.1 GLMz Poisson target

Poisson is a one parameter exponential family distribution law:

$$f_Y(y,\theta,\phi)=\mathbf{exp}\left(\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right)$$

Where *a(.), b(.) y c(.)* are scalar functions depending on common parameter θ (canonic parameter) and ϕ is known.

➡ In Poisson law $\boxed{\mu}$ parameter refers to the first order moment:

$$f_Y(y,\theta,\phi)=\frac{\mu^y}{y!}\mathbf{exp}(-\mu)=\mathbf{exp}\left(\frac{y\,\mathbf{log}(\mu)-\mu}{1}-\mathbf{log}(y!)\right)=\mathbf{exp}\left(\frac{y\,\theta-e^\theta}{1}-\mathbf{log}(y!)\right)$$

where $a(\phi)=1$, $b(\theta)=e^\theta$ (thus, $\theta=\mathbf{log}\,\mu$) and $c(y,\phi)=-\mathbf{log}(y!)$.

INTRODUCTION TO COUNTING PROCESS MODELLING

- Log-likelihood contribution of data y is:

$$\ell(\theta,\phi,y)=\log f_Y(y,\theta,\phi)=\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)=y\theta-e^{\theta}-\log(y!).$$

$$\ell(\theta,\phi,y)=y\theta-e^{\theta}-\log(y!)\cong y\log\mu-\mu$$

- Properties of the scores in this particular case:

- For Poisson law, $\mathrm{E}[Y]=\mu$ and $\mu(\theta)=b'(\theta)=\exp(\theta)$ and $\theta(\mu)=\log\mu$.
- Variance is $V[Y]=a(\phi)b''(\theta)=1\cdot\exp(\theta)=\exp(\theta)$ and $V[\mu]=\mu$.
- Canonic link $\eta=\mathrm{g}(\mu)=\theta$ (you have to understand $\theta(\mu)$). Natural logarithmic function is the canonic link: $\eta=\theta=\log\mu=\mathrm{g}(\mu)$.

INTRODUCTION TO COUNTING PROCESS MODELLING

➡ Deviance for a sample of Poisson modelled observations is:

$$D'(\mathbf{y},\hat{\mu}) = D(\mathbf{y},\hat{\mu}) = 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}.$$

Each observation $\boxed{\theta_i = \log \mu_i}$ and the contribution of each observation to the log-likelihood function is $\boxed{y_i \log(\mu_i) - \mu_i}$.

$$\boxed{\begin{aligned} &D'(\mathbf{y},\hat{\mu}) = 2\,\ell(\mathbf{y},\phi,\mathbf{y}) - 2\,\ell(\hat{\boldsymbol{\mu}},\phi,\mathbf{y}) = \\ &= \sum_{i=1}^{n}\{2(y_i \log y_i - y_i - \log(y_i!)) - 2(y_i \log \hat{\mu}_i - \hat{\mu}_i - \log(y_i!))\} = \\ &= 2\sum_{i=1}^{n}\left\{y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right\} \end{aligned}}$$

INTRODUCTION TO COUNTING PROCESS MODELLING

- … *When a constant term is included in the model*, then it can be proved:

$$\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)=0 \quad \text{y} \quad \boxed{D(\mathbf{y},\hat{\boldsymbol{\mu}})=2\sum_{i=1}^{n}\left\{y_i\,\mathbf{log}\,\frac{y_i}{\hat{\mu}_i}-\left(y_i-\hat{\mu}_i\right)\right\}=2\sum_{i=1}^{n}\left\{y_i\,\mathbf{log}\,\frac{y_i}{\hat{\mu}_i}\right\}}$$

- An approximation, proposed by Pearson,

$$\boxed{D(\mathbf{y},\hat{\boldsymbol{\mu}})\cong \mathrm{X}^2=\sum_{i=1}^{n}\frac{\left(y_i-\hat{\mu}_i\right)^2}{\hat{\mu}_i}}$$

- Maximum likelihood estimates are assymptotically normal and consistent and variance-covariance can be assympotically proved to be $\boxed{\hat{\phi}\,\mathfrak{I}_\beta^{-1}}$, where $\boxed{\mathfrak{I}_\beta}$ is the information matrix and dispersion parameter can be estimated by $\boxed{\hat{\phi}=\frac{\mathbf{X}^2}{n-p}=\sum_{i=1}^{n}\frac{\left(y_i-\hat{\mu}_i\right)^2}{\hat{\mu}_i}\Big/\left(n-p\right)}$.

- Assymptotic normal distribution might not hold when the fitted number of observations is low (less than 1). Degrees of freedom for the assymptotical distribution is usually less than $\boxed{\boldsymbol{n-p}}$ in these situations (it depends on 4th order moment, formula is omitted).

INTRODUCTION TO COUNTING PROCESS MODELLING

6.2.2 GLMz for negative binomial response

An alternative approximation to model overdispersion data considers: $\boxed{V\left[Y_i|X_i\right]=\mu_i+\alpha h(\mu_i)}$ with $\alpha>0$. Overdispersion parameter can be estimated using an auxiliary OLS regression (t.test or z-test can be used to test significant > 1 situations). $h(\mu_i)$ function can be defined as:

- Model NB1 - $h(\mu_i)=\mu_i \rightarrow V\left[Y_i|X_i\right]=(1+\alpha)\mu_i$. Quasi-Poisson models relying on quasi-likehood paradigm.
- Modelo NB 2- $h(\mu_i)=\mu_\mathrm{i}^2 \rightarrow V\left[Y_i|X_i\right]=\mu_i+\alpha\mu_\mathrm{i}^2=(1+\alpha\mu_i)\mu_i$. Negative binomial models.

AER package in R contains a dispersión test (`dispersiontest()`) to contrast values for alpha parameter either in quasi-Poisson (trafo=1) or NB2 (trafo=2) situations.

Negative binomial can be derived from a mixture between a Poisson target modelling where canonic parameter is affected by a random effect (gamma distributed) to model non-observed heterogeneity.

INTRODUCTION TO COUNTING PROCESS MODELLING

Assuming this proposal, the conditional distribution of target Yᵢ given θ_i, $Y_i|\theta_i$, is a true Poisson distributed variate with mean $\boxed{\theta_i \mu_i}$ and variance $\boxed{\theta_i \mu_i}$.

If θ_i were observed then Yᵢ targets would be Poisson distributed. Since θ_i is not observed, then a gamma distribution is assumed with shape and scale parameters $1/\alpha = \beta = \theta$ (leading to an expected value of $\alpha\beta = 1$, and variance $\alpha\beta^2 = \theta$ and a probability density function $P(\{Y = y\}) = \frac{(y/\beta)^{\alpha-1}}{\beta\Gamma(\alpha)} e^{-y/\beta}$).

Under these hypothesis, response model for target is NB2 (unconditional distribution for Y), negative binomial distribution, having probability function parameters $1/\alpha = \theta$

$$P(\{Y = y\}) = \frac{\Gamma(\theta + y)}{y!\,\Gamma(\theta)} \frac{\theta^\theta}{(\mu + \theta)^\theta} \frac{\mu^y}{(\mu + \theta)^y} \quad y = 0,1,2,\dots \quad \mu > 0 \quad \theta > 0$$ …

… expected mean is $\mathrm{E}[Y] = \mu$ and $V[Y] = \mu + \frac{1}{\theta}\mu^2$.

Poisson distribution with parameter μ holds whenever $\theta \to \infty$. Geometric distribution is another particular case that arises when $\theta = 1$.

INTRODUCTION TO COUNTING PROCESS MODELLING

In R, MASS package allows to estimate GLMz models with unknown θ parameter using glm.nb(), once θ is estimated glm() method indicating family=negative.binomial(theta=value) can be used. Logarithmic link is assumed by default.

$$V[\mu] = \mu + \frac{1}{\theta}\mu^2 .$$

Negative binomial distribution is discussed on basics Bachelor courses on Probability and Statistics. Negative binomial is linked in basic courses to repeated Bernoulli processes each one having π as the positive outcome and to model the number of repetition of a binary Bernoulli experiment required to obtain **r** positive outcomes, if $\alpha \; and \; \beta$ parameters for the generalized negative binomial density are set as,

$$\alpha = r \text{ and } \beta \;\; s.t. \;\; \pi = \beta / (\mu + \beta)$$

Then the well-known basic formula, more intuitive is obtained.

6.3. LOG-LINEAR AND MULTINOMIAL MODELS CONNECTION

Log-linear models and multinomial models are connected because a multinomial law can be derived from a set of Poisson variates conditioned to a fix total number of observations known.

This analysis is interesting to justify the equivalence between some log-linear models and multinomial models: if the analyst is interested in mean poisson variates quocients, then log-likelihood from conditional log-linear models is equivalent to a multinomial variate. Log-linear models linked to multinomial models include some *nuisance paremeters,* τ , related to multinomial totals.

➡ Not all log-linear models are equivalent to multinomial and the reverse is also false.

LOG-LINEAR AND CONTINGENCY TABLES

A first approximation to contingency table analysis using log-linear models shows an intuitive connection to ANOVA models and log-likehood functions depending on $\mu_1,\dots,\mu_n$ parameters, instead of τ, β parameters.

Let $Y_1,\dots,Y_L$, L be independent random variates Poisson distributed with expected values $\mu_1,\dots,\mu_L$, for $\boxed{l=1, \dots, L.}$

- ➡ Two dimension contingency tables having a row factor A profile with I levels and J levels representing factor B in columns representando would be indexes as corren $\boxed{i=1, \dots , I}$ (rows) and $\boxed{j=1,\dots, J}$ (columns).
- ➡ Two dimension contingency tables having a row factor A profile with I levels, J levels representing factor B in columns and K levels representing subtable factor C, would have indexes $\boxed{i=1, \dots , I}$ (rows), $\boxed{j=1,\dots, J}$ (columns) and $\boxed{k=1,\dots,K}$ (subtable)
- ➡ Let us clarify the terminology and marginal total notation …

LOG-LINEAR AND CONTINGENCY TABLES

	FACTOR C											
	FACTOR B				FACTOR B				FACTOR B			
FACTOR A	C_1				...				C_K			
	B_1	...	B_J	TOTAL	B_1	...	B_J	TOTAL	B_1	...	B_J	TOTAL
A_1	Y_{111}	...	Y_{1J1}	Y_{1+1}	...	...	...	...	Y_{11K}	...	Y_{1JK}	Y_{1+K}
A_2	Y_{211}	...	Y_{2J1}	Y_{2+1}	...	...	...	...	Y_{21K}	...	Y_{2JK}	Y_{2+K}
...	...	...	...	...	...	...	...	...	...	...	...	...
A_I	Y_{I11}	...	Y_{IJ1}	Y_{I+1}	...	...	...	...	Y_{I1K}	...	Y_{IJK}	Y_{I+K}
TOTAL	Y_{+11}	...	Y_{+J1}	Y_{++1}	...	...	...	...	Y_{+1K}	...	Y_{+JK}	Y_{++K}

Univariant Marginal Total for factor A: $Y_{i++} = \sum_j \sum_k Y_{ijk}$.	Bivariant Marginal Totals for A and C factors: $Y_{i+k} = \sum_j Y_{ijk}$
Univariant Marginal Total for factor B: $Y_{+j+} = \sum_i \sum_k Y_{ijk}$	Bivariant Marginal Totals for B and C factors: $Y_{+jk} = \sum_i Y_{ijk}$
Univariant Marginal Total for factor C: $Y_{++k} = \sum_i \sum_j Y_{ijk}$	Trivariant Marginal Totals for A, B and C factors: Y_{ijk} .
Bivariant Marginal Totals for A and B factors: $Y_{ij+} = \sum_k Y_{ijk}$	Total: $Y_{+++} = \sum_i \sum_j \sum_k Y_{ijk}$.

LOG-LINEAR AND CONTINGENCY TABLES

6.3.1 Constraint on total counts

Let $Y_1, \ldots, Y_L$, L be independent random variates Poisson distributed with expected values $\mu_1, \ldots, \mu_L$, for $\boxed{l=1, \ldots, L.}$

➡ $Y_1, \ldots, Y_L$ represent multivariant totals, counts per cell, in a contingency table rewritten in list form. For example, a contingency table of 3 dimensions would be rewritten as $\boxed{L=IxJxK}$ Poisson variates.

➡ Joint likelihood function would be in terms $\mu_1, \ldots, \mu_L$,

$$f_{\mathbf{Y}}(\mathbf{y}, \boldsymbol{\mu}) = \prod_{l=1}^{L} \frac{\mu_l^{y_l}}{y_l!} e^{-\mu_l} \quad \text{y} \quad L(\boldsymbol{\mu}, \mathbf{y}) = \prod_{l=1}^{L} \frac{\mu_l^{y_l}}{y_l!} e^{-\mu_l}$$

➡ Given a fix total number of observations $m = y_+ = \sum_l y_l$, then according to the additive property that holds for independent Poisson variates adding up $Y_1, \ldots, Y_L$ is Poisson distributed with expected mean $\mu_+ = \mu_1 + \ldots + \mu_L$ ...

LOG-LINEAR AND CONTINGENCY TABLES

➡ … Then joint probability function for $Y_1,\ldots,Y_L$ conditional to total $\boxed{m}$ is,

$$f_{\mathbf{Y}/m}(\mathbf{y},\boldsymbol{\mu})=\prod_{l=1}^{L}\frac{\mu_l^{y_l}}{y_l!}e^{-\mu_l}\Bigg/\frac{\mu_+^m}{m!}e^{-\mu_+}=m!\prod_{l=1}^{L}\frac{\pi_l^{y_l}}{y_l!}\quad \text{donde}\quad \pi_l=\frac{\mu_l}{\mu_+}.$$

➡ And directly, $f_{\mathbf{Y}/m}(\mathbf{y},\boldsymbol{\mu})$ multinomial law $\boxed{m}$ and $\boxed{\boldsymbol{\pi}^T=(\pi_1,\ldots,\pi_L)}$ parameters where $\pi_l=\frac{\mu_l}{\mu_+}$ and thus it is satisfied :

1. $\sum_l \pi_l=1$ 2. $0\le\pi_l\le 1 \quad l=1,\ldots,L$ 3. $\mathrm{E}[Y_l]=m\pi_l \quad l=1,\ldots,L$.

LOG-LINEAR AND CONTINGENCY TABLES

6.3.2 Row total constraints on tables of dimension 2

Let $Y_{11},\ldots Y_{1J},\ldots,Y_{21},\ldots Y_{2J},\ldots,Y_{I1},\ldots,Y_{IJ}$, L=IxJ be independent Poisson variates with expected values $\mu_1,\ldots,\mu_L$, and indexed (row > column, row ordering) l =1, …, L .

- ➡ $Y_1,\ldots,Y_L$ model cell frequencies (row ordering) in a contingency table of 2 dimensions and Poisson expected parameters $\mu_1,\ldots,\mu_L$.
- ➡ Joint likelihood function on $\mu_1,\ldots,\mu_L$ is,

$$f_{\mathbf{Y}}(\mathbf{y},\boldsymbol{\mu})=\prod_{l=1}^{L}\frac{\mu_l^{y_l}}{y_l!}e^{-\mu_l}=\prod_{i=1}^{I}\prod_{j=1}^{J}\frac{\mu_{ij}^{y_{ij}}}{y_{ij}!}e^{-\mu_{ij}} \quad .$$

- ➡ If univariant row totals are known and fixed, univariant marginal totals for factor A, $m_i=Y_{i+}=\sum_j Y_{ij}$, then adding up by rows $Y_{1+},\ldots,Y_{I+}$ are Poisson distributed with expected means $\mu_{i+}=\mu_{i1}+\ldots+\mu_{iJ}$ *(i-th row).*

LOG-LINEAR AND CONTINGENCY TABLES

- …Then the joint probability function for $Y_1,\dots,Y_L$ $(Y_{11},\dots Y_{1J},\dots,Y_{21},\dots Y_{2J},\dots,Y_{I1},\dots,Y_{IJ})$ given row univariant totals $\mathbf{m}^{\mathrm{T}}=(m_1,\dots,m_I)$ is,

$$f_{\mathbf{Y}/\mathbf{m}}(\mathbf{y},\boldsymbol{\mu})=\prod_{i=1}^{I}\left(\prod_{j=1}^{J}\frac{\mu_{ij}^{y_{ij}}}{y_{ij}!}e^{-\mu_{ij}}\Big/\frac{\mu_{i+}^{m_i}}{m_i!}e^{-\mu_{i+}}\right)=\prod_{i=1}^{I}\left(m_i!\prod_{j=1}^{J}\frac{\pi_{ij}^{y_{ij}}}{y_{ij}!}\right)\quad \text{where } \pi_{ij}=\frac{\mu_{ij}}{\mu_{i+}}.$$

- Directly, $f_{\mathbf{Y}/\mathbf{m}}(\mathbf{y},\boldsymbol{\mu})$ can be shown to be the joint probability function for the product of multinomial laws each one belonging to a row level i.

$\boxed{\mathbf{m_i}}$ and $\boxed{\boldsymbol{\pi}_i^T=(\pi_{i1},\dots,\pi_{iJ})}$ with $\pi_{ij}=\dfrac{\mu_{ij}}{\mu_{i+}}$ satisfying :

1. $\sum_j \pi_{ij}=1 \quad \forall i=1,\dots,I$
2. $0\le \pi_{ij}\le 1$
3. $\mathrm{E}[Y_{ij}]=m_i\pi_{ij}$
4. $m_i=\sum_j y_{ij}$

- Thus, $\boxed{\pi_{ij}=\dfrac{\mu_{ij}}{\mu_{i+}}=\pi_{j/i}}$

LOG-LINEAR AND CONTINGENCY TABLES

6.3.3 Subtable total constraints on tables of dimension 3

Let $\boxed{Y_{ijk}}$'s, L=IxJxK, be independent Poisson variates with expected values $\mu_1,\ldots,\mu_L$, l =1, …, L .

- Joint probability function on $\mu_1,\ldots,\mu_L$ is,

$$f_{\mathbf{Y}}(\mathbf{y},\boldsymbol{\mu})=\prod_{l=1}^{L}\frac{\mu_l^{y_l}}{y_l!}e^{-\mu_l}=\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\mu_{ijk}^{y_{ijk}}}{y_{ijk}!}e^{-\mu_{ijk}}\quad .$$

- Given, $Y_{++k}=\sum_i\sum_j Y_{ijk}$ univariant marginal total for factor C, fixed by design, $m_k=Y_{++k}=\sum_i\sum_j Y_{ijk}$, then according to the additive property that holds for independent Poisson variates, adding up rows and columns for each subtable level $\boxed{k,}$ $Y_{++1},\ldots,Y_{++k}$, are Poisson distributed with expected mean $\mu_{++k}=\mu_{11k}+\ldots+\mu_{IJk}$ …

LOG-LINEAR AND CONTINGENCY TABLES

- Then, the joint probability function for $\boxed{Y_{ijk}}$ s conditioned to univariant marginal totals for levels of factor C, $\mathbf{m}^{\mathrm{T}} = (\ldots, m_k, \ldots)$ is,

$$f_{\mathbf{Y}/\mathbf{m}}(\mathbf{y},\boldsymbol{\mu}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\mu_{ijk}^{y_{ijk}}}{y_{ijk}!}e^{-\mu_{ijk}} \Bigg/ \frac{\mu_{++k}^{m_k}}{m_k!}e^{-\mu_{++k}} = \prod_{k=1}^{K} m_k! \prod_{i=1}^{I}\prod_{j=1}^{J}\frac{\pi_{ijk}^{y_{ijk}}}{y_{ijk}!} \quad \text{where} \quad \pi_{ijk} = \frac{\mu_{ijk}}{\mu_{++k}}.$$

- Directly, $f_{\mathbf{Y}/\mathbf{m}}(\mathbf{y},\boldsymbol{\mu})$ is the joint probability function of K multinomial laws each one with parameters $\boxed{m_k}$ and $\boxed{\boldsymbol{\pi}_{\mathbf{k}}^{\mathrm{T}} = (\pi_{11k}, \ldots, \pi_{IJk})}$ with $\pi_{ijk} = \dfrac{\mu_{ijk}}{\mu_{++k}}$ satisfying :

- $\sum_i\sum_j \pi_{ijk} = 1 \;\; \forall\; k$ 2. $0 \le \pi_{ijk} \le 1$ 3. $\mathrm{E}\left[Y_{ijk}\right] = m_k\pi_{ijk}$ 4. $m_k = \sum_{i,j} y_{ijk}$

- Thus $\boxed{\pi_{ijk} = \dfrac{\mu_{ijk}}{\mu_{++k}} = \pi_{ij/k}}$

LOG-LINEAR AND CONTINGENCY TABLES

6.3.4 Row and Table total constraints on tables of dimension 3

Let $\boxed{Y_{ijk}}$'s be, *L=IxJxK* independent Poisson variates with expected values $\mu_1,\ldots,\mu_L$, indexed *l* =1, …, *L* .

- $Y_1,\ldots,Y_L$ represent cell frecuencies (order defined as table > row > column) in a contingency table having 3 dimensions and they are modelled as independent Poisson variates with expected mean parameters $\mu_1,\ldots,\mu_L$. Order used is: $Y_{111},\ldots Y_{1J1},\ldots,Y_{I1K},\ldots,Y_{IJK}$.

- Joint probability function on $\mu_1,\ldots,\mu_L$ is,

$$f_{\mathbf{Y}}(\mathbf{y},\boldsymbol{\mu})=\prod_{l=1}^{L}\frac{\mu_l^{y_l}}{y_l!}e^{-\mu_l}=\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\mu_{ijk}^{y_{ijk}}}{y_{ijk}!}e^{-\mu_{ijk}}\quad .$$

- Bivariant marginal totals for rows and tables are given, $Y_{i+k}=\sum_j Y_{ijk}$ and fixed by design …

LOG-LINEAR AND CONTINGENCY TABLES

- … $m_{ik} = Y_{i+k} = \sum_j Y_{ijk}$, then according to the additive property that holds for independent Poisson variates the total number of observations for each joint row-table $Y_{1+k},\ldots,Y_{I+k}$ is Poisson distributed with expection $\mu_{i+k} = \mu_{i1k} + \ldots + \mu_{iJk}$ (i-th row and k-th table).
- Then the joint probabily of $\boxed{Y_{ijk}}$ s condicionadal to bivariant A and C totals, $\mathbf{m}^{\mathrm{T}} = (\ldots, m_{ik}, \ldots)$ is,

$$f_{\mathbf{Y}/\mathbf{m}}(\mathbf{y},\boldsymbol{\mu}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\mu_{ijk}^{y_{ijk}}}{y_{ijk}!}e^{-\mu_{ijk}} \Bigg/ \frac{\mu_{i+k}^{m_{ik}}}{m_{ik}!}e^{-\mu_{i+k}} = \prod_{i=1}^{I}\prod_{k=1}^{K} m_{ik}! \prod_{j=1}^{J}\frac{\pi_{ijk}^{y_{ijk}}}{y_{ijk}!}$$

where $\pi_{ijk} = \dfrac{\mu_{ijk}}{\mu_{i+k}}$.

- Thus $\boxed{\pi_{ijk} = \dfrac{\mu_{ijk}}{\mu_{i+k}} = \pi_{j/ik}}$

LOG-LINEAR AND CONTINGENCY TABLES

➡ Directly, $f_{\mathbf{Y}/\mathbf{m}}(\mathbf{y},\boldsymbol{\mu})$ can be proved to be the joint probability function of IxK multinomial laws each one defined with parameters $\boxed{(ik)}$ $\boxed{m_{ik}}$ and $\boxed{\boldsymbol{\pi}_{ik}^{T}=\left(\pi_{i1k},\ldots,\pi_{iJk}\right)}$ with

➡ $$\pi_{ijk}=\frac{\mu_{ijk}}{\mu_{i+k}}\quad\left(=\pi_{j/ik}\right)$$

satisfying:

1. $\sum_j \pi_{ijk}=1 \quad \forall\, i,k$
2. $0\le \pi_{ijk}\le 1$
3. $\mathrm{E}\left[Y_{ijk}\right]=m_{ik}\pi_{ijk}$
4. $m_{ik}=\sum_j y_{ijk}$

6.4. TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

Contingency tables are used to determine association between defining factors. All common hypothesis to be checked between factors defining contingency tables of dimensions 2 or 3 can be assessed through multiplicative models where cell frequencies can be written from univariant, bivariant marginal probabilities.

6.4.1 Independency between row and column in dimension 2 contingency tables

- In dimension 2 contingency tables, independence hypothesis can be stated as a null hypothesis (H0) where joint probability is set as the product of univariant marginal probabilities (Factors A and C). If the total number of observations in the table is fixed and known **m**, the expected observations is:

$\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$, where $\sum_i \pi_{i\bullet} = 1$ $\sum_j \pi_{\bullet j} = 1$ and expected number of observations under H0 would be $\mathrm{E}\left[Y_{ij}\right] = m\pi_{i\bullet}\pi_{\bullet j}$ that it is equivalent from the Poisson point of view to $\mathrm{E}\left[Y_{ij}\right] = \mu_{i+}\mu_{+j}/m$.

- Log-linear equivalent model would be A+B

$$\boxed{\log\left(\mu_{ij}\right) = \eta_{ij} = \mu + \alpha_i + \beta_j \quad i = 1,\ldots,I \qquad j = 1,\ldots,J}$$

I+J-1 independent parameter where μ is a fixed offset

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

- The saturated log-linear model would be (use the analogy to analysis of variance) as A*B

$$\boxed{\log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{i\,j} \quad i = 1,\ldots,I \quad j = 1,\ldots,J}$$

$\boxed{IJ}$ parameters (some of them colinear and 1 fixed)

- $\boxed{\mu}$ is an offset related to the total number of observations **m (fixed)** in fact, $\mu = \log(m)$.

- ***Independence null hypothesis can be assessed by comparing interactive A*B log-linear model vs additive A+B log-linear model.***
- ***A*B and A+B are nested models (A+B parameters included into A*B) and deviance test or Wald test can be used to determine pvalue for the null hypothesis.***
- ***In R, anova(A+B, A*B, test="Chisq") or wald.test(A+B, A*B).***
- ***Parameters related to fixed constants are introduced in the models using offset parameter in glm() method.***

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

6.4.2 Homogeneity hypothesis testing in dimension 3 tables.

➡ In dimension 2 contingency tables, homogeneity hypothesis can be stated as marginal column probabilities being common for all table rows. Using conditional probability definition and fixed parameters for row marginal totals, it holds ***fijados como constantes los totales univariantes del factor A (filas).***

$\boxed{\pi_{j/i} = \pi_{\bullet j}}$ or $\boxed{\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}}$, where $\sum_j \pi_{\bullet j} = 1$ and expection in cells are $\mathrm{E}\left[Y_{ij}\right] = m_i \pi_{\bullet j} = y_{i+}\pi_{\bullet j}$.

➡ The equivalent log-linear model would be A+B:

$$\boxed{\log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j \quad i = 1,\ldots,I \qquad j = 1,\ldots,J}$$

$\boxed{I+J-1}$ independent parameters, some of them fixed.

➡ Alternative hypothesis would be $\mathrm{E}\left[Y_{ij}\right] = m\pi_{ij}$ that can be stated as log-linear A*B model:

$$\boxed{\log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} \quad i = 1,\ldots,I \quad j = 1,\ldots,J}$$

$\boxed{IJ}$ independent parameters (some of them fixed)

➡ Parameters linked to fixed row totals are $\boxed{\mu + \alpha_i \quad i = 1,\ldots,I}$.

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

6.4.3 Independency hypothesis in dimension 3 contingency tables and total observations fixed

➡ **Null hypothesis of full independency in dimension 3 contingency tables** can be stated as (total fixed),

$\boxed{\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k}}$ and expected counts in cells be

$$\mathrm{E}\left[Y_{ijk}\right] = m\pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k} = y_{+++}\pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k} .$$

➡ This statement is consistent to A+B+C log-linear Poisson model

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k}$$

$\boxed{\textit{I+J+K-2}}$ independent parameters (reparametrization needed and 1 parameter fixed)

➡ Null hypothesis assessment would compare additive model against saturated model $\mathrm{E}\left[Y_{ijk}\right] = m\pi_{ijk} = y_{+++}\pi_{ijk}$ stated as A*B*C log-linear Poisson model:

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

$\boxed{\textit{IJK}}$ independent parameters (reparametrization needed and 1 parameter fixed)

➡ Fixed parameter set as an offset term is: $\boxed{\mu}$.

➡ So full independency is checked using a $\boxed{\textit{log-linear model without any order 2 interaction}}$ $\boxed{\textit{A+B+C}}$.

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

- **Null hypothesis of block independency in dimension 3 contingency tables**, can be stated as one dimension (row/column/table) being independent from the other 2 dimensions. For example, Factor A independency (row) from Factors B and C (columns and subtables) (fixed table total) can be written as: $\boxed{\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet jk}}$ and expected cell frequencies $\mathrm{E}[Y_{ijk}] = m\pi_{i\bullet\bullet}\pi_{\bullet jk} = y_{+++}\pi_{i\bullet\bullet}\pi_{\bullet jk}$.
- This statement is consistent to **A+B*C** log-linear Poisson model,

$$\boxed{\log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \beta\gamma_{jk}}$$

I+JK-1 independent parameters (reparametrization needed and 1 parameter fixed)

To be assessed against the alternative hypothesis stated by $\mathrm{E}[Y_{ijk}] = m\pi_{ijk} = y_{+++}\pi_{ijk}$ and the equivalent saturated log-linear Poisson model would be A*B*C:

$$\boxed{\log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

IJK independent parameters (reparametrization needed and 1 parameter fixed)

- Fixed parameter is related to total counts μ . Set as an offset equal to log(m)
- **Block independency** corresponds to log-linear Poisson models ***1order 2 interaction: A+B*C o B+A*C or C+A*B.***

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

➡ In dimension 3 contingency tables, **partial independency hypothesis** can be assessed by comparing 2 second order interactions model against saturated log-linear Poisson models once the total number of observations is fixed.

1. **A*B+B*C** $(I+K-1)J$ parameters.
2. **A*C+B*C** $(I+J-1)K$ parameters.
3. **A*B+A*C** $(J+K-1)I$ parameters.

➡ In dimension 3 contingency tables, **uniform association hypothesis** can be assessed by comparing 3 second order interactions model against saturated log-linear Poisson models once the total number of observations is fixed:

1. **A*B+A*C+B*C** $IJK-(I-1)(J-1)(K-1)$ parameters.

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

6.4.4 Homogeneity hypothesis in dimension 3 tables given a fixed univariant total

➨ In dimension 3 contingency tables, **homogeneity association hypothesis between rows and columns (dimensions 1 and 2) for each subtable (dimension 3, Factor C univariant total is fixed).**

$\boxed{\pi_{ij/k} = \pi_{ij\bullet}}$ or $\boxed{\pi_{ijk} = \pi_{\bullet\bullet k}\pi_{ij\bullet}}$, where $\sum_{ij}\pi_{ij\bullet} = 1$ and $\mathrm{E}\left[Y_{ijk}\right] = m_k\pi_{ij\bullet} = y_{++k}\pi_{ij\bullet}$.

➨ The equivalent log-linear Poisson model would be C+A*B

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij}}$$

IJ+K-1 independent parameters (reparametrization needed and several parameters fixed)

➨ To be assessed against $\mathrm{E}\left[Y_{ijk}\right] = m_k\pi_{ij/k}$ the saturated model A*B*C:

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

IJK independent parameters (reparametrization needed and several parameters fixed)

➨ Fixed parameters corresponding to univariant subtable totals are: $\boxed{\mu + \gamma_k}$.

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

- In dimension 3 contingency tables, independency between rows and columns (A and B Factors) in each subtable (Factor C) once univariant factor C totals are fixed can be stated as,

$\boxed{\pi_{ijk} = \pi_{i\bullet k}\pi_{\bullet jk} / \pi_{\bullet\bullet k}}$, where $\sum_{j,k} \pi_{\bullet jk} = 1$ $\sum_{i,k} \pi_{i\bullet k} = 1$, and expected cell counts

$$\mathrm{E}\left[Y_{ijk}\right] = m_k \pi_{i\bullet k}\pi_{\bullet jk} = y_{++k}\pi_{i\bullet k}\pi_{\bullet jk} .$$

- The equivalent log-linear Poisson model would be C+A*C+B*C,

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk}}$$

K(I+J-1) independent parameters (reparametrization needed and several parameters fixed)

- To be assessed against $\mathrm{E}\left[Y_{ijk}\right] = m_k \pi_{ij/k}$ the saturated model A*B*C:

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

IJK independent parameters (reparametrization needed and several parameters fixed)

- Fixed parameters corresponding to univariant subtable totals are: $\boxed{\mu + \gamma_k}$.

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

6.4.5 Homogeneity hypothesis in dimension 3 tables given fixed bivariant totals

- In dimension 3 contingency tables, null hypothesis of homogeneity, this is identical marginal column probabilities in all subtables given fixed counts for bivariant A and C totals can be stated as

$\boxed{\pi_{j/ik} = \pi_{\bullet j \bullet}}$ or $\boxed{\pi_{ijk} = \pi_{i \bullet k} \pi_{\bullet j \bullet}}$, where $\sum_j \pi_{\bullet j \bullet} = 1$ y $\mathrm{E}[Y_{ijk}] = m_{ik} \pi_{\bullet j \bullet} = y_{i+k} \pi_{\bullet j \bullet}$.

- The equivalent log-linear Poisson model would be A*C+B (B independent from A and C),

$$\boxed{\log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik}}$$

_IK+J-1_ independent parameters (reparametrization needed and several parameters fixed)

- To be assessed against $\mathrm{E}[Y_{ijk}] = m_k \pi_{ij/k}$ the saturated model A*B*C:

$$\boxed{\log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

IJK independent parameters (reparametrization needed and several parameters fixed)

- Fixed parameters corresponding to bivariant row-subtable totals are: $\boxed{\mu + \alpha_i + \gamma_k + \alpha\gamma_{ik}}$.

- Pay attention to the fact that A, B and C have the same role, it is not defined any dimension as the response and the rest as the explanatory factors.

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

- In dimension 3 contingency tables, null hypothesis of column homogeneity, this is identical marginal column probabilities depending on subtable given fixed counts for bivariant A and C totals can be stated as:

$$\boxed{\pi_{ijk} = \pi_{i\bullet k}\pi_{\bullet jk}} \text{, where } \sum_{i,j}\pi_{\bullet jk} = 1 \text{ and } \mathrm{E}\left[Y_{ijk}\right] = m_{ik}\pi_{\bullet jk} = y_{i+k}\pi_{\bullet jk}\,.$$

- Response variable is column (Factor B) and explanatory variables are A and C factors (joint probability function is the product of multinomial probability functions) once bivariant totals for A and C are fixed.
- The equivalent log-linear Poisson model would be A*C+B*C (conditional to C, A and B are independent)

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \gamma_k + \alpha\gamma_{ik} + \beta_j + \beta\gamma_{jk}}$$

K(I+J-1) independent parameters (reparametrization needed and several parameters fixed)

- To be assessed against $\mathrm{E}\left[Y_{ijk}\right] = m_k\pi_{ij/k}$ the saturated model A*B*C:

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

IJK independent parameters (reparametrization needed and several parameters fixed)

- Fixed parameters corresponding to bivariant row-subtable totals are: $\boxed{\mu + \alpha_i + \gamma_k + \alpha\gamma_{ik}}$.

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

- **A null hypothesis involving homogeneity indicating asociation between C and B factors is the same for all bivariant A - B levels** (bivariant marginal probability C-D identical for all pairs of A-B levels) (response variable is Factor B, explanatory factors are A and C and bivariant totals for A-C are fixed) can be stated as

$\boxed{\pi_{ijk} = \pi_{i\bullet k}\pi_{\bullet jk}\pi_{ij\bullet}}$ and expected cell counts $\mathrm{E}\left[Y_{ijk}\right] = m_{ik}\pi_{\bullet jk}\pi_{ij\bullet} = y_{i+k}\pi_{\bullet jk}\pi_{ij\bullet}$.

- The equivalent log-linear Poisson model would be A*C+B*C+A*B,

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \gamma_k + \alpha\gamma_{ik} + \beta_j + \beta\gamma_{jk} + \alpha\beta_{ij}}$$

IJK-(I-1)(J-1)(K-1) independent parameters (reparametrization needed and several parameters fixed)

- To be assessed against $\mathrm{E}\left[Y_{ijk}\right] = m_k \pi_{ij/k}$ the saturated model A*B*C:

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

IJK independent parameters (reparametrization needed and several parameters fixed)

- Fixed parameters corresponding to bivariant row-subtable totals are: $\boxed{\mu + \alpha_i + \gamma_k + \alpha\gamma_{ik}}$.

TESTING MULTINOMIAL HYPOTHESIS USING LOG-LINEAR MODELS

6.4.6 Equivalence Nominal Response and Poisson log-linear models

➡ **Let us focus on 3 dimension contingency tables:** Factor B is the polytomous response and Factors A and C are explanatory factors (A and C bivariant totals are fixed),

LOG-LINEAR MODELS	NOMINAL POLYTHOMIC MODELS
A*C+B	**Minimal**
A*C+A*B	**A**
A*C+B*C	**C**
A*C+A*B+B*C	**A+C**
A*B*C	**A*C (Maximal)**

➡ **NOMINAL and POISSON LOG-LINEAR models relationship (general setting on explanatory variables, index *i*). Reference level is 1 for response factor B.**

$\boxed{\log(\boldsymbol{\mu}_{ij}) = \boldsymbol{\eta}_{ij} = \boldsymbol{\mu} + \theta_i + \boldsymbol{\alpha}_j + x_{\mathrm{i}}^{\mathrm{T}} \boldsymbol{\beta}_j}$ and $\boxed{\log(\mu_{i1}) = \eta_{i1} = \mu + \theta_i + \alpha_1 + x_{\mathrm{i}}^{\mathrm{T}} \beta_1}$ then

$$\boxed{\log(\mu_{ij}) - \log(\mu_{i1}) = \log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = (\alpha_j - \alpha_1) + x_{\mathrm{i}}^{\mathrm{T}}(\beta_j - \beta_1)}$$

6.5. MODELS FOR COUNTS. DIAGNOSTICS

- Residual Deviance for model (M) is $D = 2\sum y_l \log \frac{y_l}{\hat{\mu}_l}$.
- Pearson Statistic for model (M) is $\boxed{D(\mathbf{y}, \hat{\mu}) \cong \mathrm{X}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}}$.
- Assuming that model (M) is consistent to data, for large samples both statistics are assymptotically chi squared distributed with as many degrees of freedom non-null cells minus the number of model parameters (n-p).
- Pearson residuals are the ones more commonly found in Statistical Software (SPSS, MINITAB, etc). Standarized Pearson residuals in absolute value less than 2 or 3 (depending of the number of groups).
- Car package offers generic diagnostic for glm() models and can be applied.

6.6. LOG-LINEAR MODELS. EXAMPLES

6.6.1 Example 1: Melanomas (Dobson)

Data from n=400 pacients relative to **Melanoma type (Factor A)** and **place of appearance (Factor B)** . Data is formatted as presented in this course.

FACTOR A Type	FACTOR B - Place			
	B_1	B_2	$B_{J=3}$	*TOTAL*
A_1	y_{11}	y_{12}	y_{1J}	y_{1+}
A_2	y_{21}	y_{22}	y_{2J}	y_{2+}
A_3	y_{31}	y_{32}	y_{3J}	y_{3+}
$A_{I=4}$	y_{I1}	y_{I2}	y_{IJ}	y_{I+}
TOTAL	y_{+1}	y_{+2}	y_{+J}	y_{++}

FACTOR A	FACTOR B - Place			
Type	B_1	B_2	$B_{J=3}$	*TOTAL*
A_1	22	2	10	34
A_2	16	54	115	185
A_3	19	33	73	125
$A_{I=4}$	11	17	28	56
TOTAL	68	106	226	400

➡ **Is there any relationship between Tumor Type and Tumor Place?**

➡ Null hypothesis is stated as '**Independency between Type and Place**"; i.e. A and B Factors.

LOG-LINEAR MODELS. EXAMPLE 1

- This is a particular case of a dimension 2 contingency table where total table count is given (m=400) and H0 assessment (row and column independency).
- From a probabilistic point of view, null hypothesis can be rewritten as joint probability (i,j) equal to the product of marginal probabilities given $\boxed{m=400}$, total number of observaciones:

$\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$, where $\sum_i \pi_{i\bullet} = 1$ $\sum_j \pi_{\bullet j} = 1$ and expected cell counts under H0 $\mathrm{E}\left[Y_{ij}\right] = m\pi_{i\bullet}\pi_{\bullet j}$ or equivalently the log-linear Poisson model is $\mathrm{E}\left[Y_{ij}\right] = \mu_{i+}\mu_{+j}/m$.

- The log-linear Poisson model consistent to H0 is A+B

$$\boxed{\log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j \quad i = 1,\ldots,I \qquad j = 1,\ldots,J} \quad \boxed{I = 4,\ J = 3}$$

$\boxed{I+J-1=4+3-1=6}$ independent parameters and $\mu = \log(m)$

- H0 Assessment should involve A+B model comparison to the saturated model A*B,

$$\boxed{\log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} \quad i = 1,\ldots,I \quad j = 1,\ldots,J} \quad \boxed{I = 4,\ J = 3}$$

$\boxed{IJ=4x3=12}$ independent parameters and $\mu = \log(m)$

- Parameter affected by total count m fixed is μ .

LOG-LINEAR MODELS. EXAMPLE 2

6.6.2 Example 2: High Education Follow-up intention (Secondary students)

A sample of 4991 secondary school students at Wisconsin is given in the following contingency table (3 dimensions). Factors defining each dimension are Factor A- STATUS socio-economic status (4 levels, low, medium-low, medium-high and high), Factor B – Follow-up considered? (2 levels, Yes-No) and Factor C-Motivation (support from the family) (2 levels, high-low). Initially, all 3 variables are equally considered (no target is defined). Data from Fienberg (1977).

	FACTOR C-Motivation					
	FACTOR B – Follow-Up?			***FACTOR B Follow-Up?***		
FACTOR A Status	C_1 - Low			$C_{K=2}$ High		
	B_1 No	$B_{J=2}$ Yes	*TOTAL*	B_1 No	$B_{J=2}$ Yes	*TOTAL*
A_1 Low	749	35	**784**	233	133	**366**
A_2 Medium-Low	627	38	**665**	330	303	**633**
A_2 Medium-High	420	37	**457**	374	467	**841**
$A_{I=4}$ High	153	26	**179**	266	800	**1066**
TOTAL	**1949**	**136**	**2085**	**1203**	**1703**	**2906**

LOG-LINEAR MODELS. EXAMPLE 2

Several log-linear Poisson models are calculated, from more simple to more complex ones. Total counts m is fixed:

MODEL	DEVIANCE	d.f.	
A+B+C	2714	10	Motivation, Follow-up and Status are independent
A+B*C	1092	9	Social Status is independent from Motivation and Follow-up
B+A*C	1877.4	7	Follow-up is independent from Motivation and Status
C+A*B	1920.4	7	Motivation is independent from Status and Follow-up
A*B+A*C	1083.8	4	Conditional to Status, Motivation and Follow-up are independient
A*B+B*C	298.5	6	Conditional to Follow-up, Status and Motivation are independient
A*C+B*C	255.5	6	Conditional to Motivation, Status and Follow-up are independient
A*B+A*C+B*C	1.575	3	Interpretation of log-linear model without 3rd order interaction ¿??

LOG-LINEAR MODELS. EXAMPLE 2

MODEL A+B+C: The simplest model from the multinomial point of view, where the joint probability is the product of univariant marginal probabilities.

- **Total independency model**, rows, columns and subtables are independent **given total count m.**

  H0: $\boxed{\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k}}$ and cell counts $\mathrm{E}[Y_{ijk}] = m\pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k} = y_{+++}\pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k}$.

- Equivalent log-linear Poisson is A+B+C ,

$$\boxed{\log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k = \log y_{+++} + \log \pi_{i\bullet\bullet} + \log \pi_{\bullet j\bullet} + \log \pi_{\bullet\bullet k}}$$

  I+J+K-2=4+2+2-2=6 independent parameters

- Alternative hypothesis relies on the saturated log-linear Poisson model $\mathrm{E}[Y_{ijk}] = m\pi_{ijk} = y_{+++}\pi_{ijk}$ or A*B*C: ***IJK=16*** independent parameters.

$$\boxed{\log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

- Offset is set to represent total count **m:** $\mu = \log(m)$.

- ***It can be shown that A+B+C gives ML estimates leading to:*** $\boxed{\hat{\mu}_{ijk} = y_{i++}y_{+j+}y_{++k} / m^2}$.

- **D(A+B+C)=2714 with 10 d.f. and D(A*B*C)=0 have to compared. H0 is rejected.**

LOG-LINEAR MODELS. EXAMPLE 2

Block Independence Models: From the multinomial point of view 2 dimensions are dependent, but the third dimension is independent from the other two. For example, A+B*C, Motivation from the family and Follow-Up intention are associated, but they are independent of Social Status (it seems not realistic according to visual inspection of the table, but this would be the model)

- Block Independency models, as for example factor A (rows) independent from the other 2 dimensions (columns and subtables) given a fixed total count.

H0: $\boxed{\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet jk}}$ and expected cell counts $\mathrm{E}[Y_{ijk}] = m\pi_{i\bullet\bullet}\pi_{\bullet jk} = y_{+++}\pi_{i\bullet\bullet}\pi_{\bullet jk}$ since,

$$P\left(\{B=j\}\cap\{C=k\} / \{A=i\}\right) = P(\{B=j\}\cap\{C=k\}) = \frac{P(\{A=i\}\cap\{B=j\}\cap\{C=k\})}{P(\{A=i\})}$$

- Equivalent log-linear Poisson is A+B*C,

$$\boxed{\log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \beta\gamma_{jk} = \log y_{+++} + \log \pi_{i\bullet\bullet} + \log \pi_{\bullet jk}}$$

_I+JK-1=4+2*2-1=7_ independent parameters

- To be assessed against saturated log-linear Poisson model $\mathrm{E}[Y_{ijk}] = m\pi_{ijk} = y_{+++}\pi_{ijk}$ A*B*C:

$$\boxed{\log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

IJK=16 independent parameters and 1 fixed parameter (μ)

LOG-LINEAR MODELS. EXAMPLE 2

- These *models have 1 interaction (2nd order) : A+B*C or B+A*C or C+A*B: In the case, A+B*C*

  ML estimates give expected cell values $\hat{\mu}_{ijk} = y_{i++} y_{+jk} / m$.

- Block Independence models are usefull for assessing H0 against the saturated model A*B*C.
- D(A+B*C)=1092 with 9 =16-7 d.f. pvalue << 0.05 and thus H0s are rejected (all block independence hypothesis are rejected).

MODELS for PARTIAL INDEPENDENCE: From the multinomial perspective, 2 factors are not associated conditioned to the third one. For example, A*C+B*C, indicates that conditioned to C, A and B are independent.

H0: $\pi_{ijk} = \pi_{i\bullet k}\pi_{\bullet jk} / \pi_{\bullet\bullet k}$ and expected cell counts are

$$\mathrm{E}\left[Y_{ijk}\right] = m\pi_{i\bullet k}\pi_{\bullet jk} / \pi_{\bullet\bullet k} = y_{+++}\pi_{i\bullet k}\pi_{\bullet jk} / \pi_{\bullet\bullet k}\,.$$

- Equivalent log-linear Poisson is A*C+B*C, to multinomial probabilities satisfying $\pi_{ij/k} = \pi_{i\bullet k}\pi_{\bullet jk}$,

$$\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} = \log m + \log \pi_{i\bullet k} + \log \pi_{\bullet jk} - \log \pi_{\bullet\bullet k}$$

(I+J-1)K* *=(4+2-1)*2=10 independent parameters

LOG-LINEAR MODELS. EXAMPLE 2

➡ To assess the null hypothesis, A*C+B*C is compared to A*B*C, $\mathrm{E}\left[Y_{ijk}\right] = m\pi_{ijk} = y_{+++}\pi_{ijk}$:

$$\boxed{\log\left(\mu_{ijk}\right) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}}$$

$\boxed{IJK=16}$ independent parameters

➡ Fixed parameter is μ (offset set to log(m)). **D(A*C+B*C)=255.5 with 6 =16-10 d.f., p_value << 0.05 and thus H0 is rejected.**

From the probability point of view, partial independence is stated as:

$$P\left(\{A=i\}\cap\{B=j\} \big/ \{C=k\}\right) = P\left(\{A=i\} \big/ \{C=k\}\right) P\left(\{B=j\} \big/ \{C=k\}\right) = \frac{P\left(\{A=i\}\cap\{C=k\}\right)P\left(\{B=j\}\cap\{C=k\}\right)}{P\left(\{C=k\}\right)P\left(\{C=k\}\right)}$$

Thus, trivariant probabilities can be rewritten from marginal probabilities as ,

$$P\left(\{A=i\}\cap\{B=j\} \big/ \{C=k\}\right) = \frac{P\left(\{A=i\}\cap\{B=j\}\cap\{C=k\}\right)}{P\left(\{C=k\}\right)}$$

➡ Partial Independence hypothesis can be addressed using log-linear Poisson models having 2 interactions:

1. A*B+B*C	*(I+K-1)J* parameters.	$\boxed{\hat{\mu}_{ijk} = y_{ij+}y_{+jk} / y_{+j+}}$
2. A*C+B*C	*(I+J-1)K* parameters.	$\boxed{\hat{\mu}_{ijk} = y_{i+k}y_{+jk} / y_{++k}}$
3. A*B+A*C	*(J+K-1)I* parameters.	$\boxed{\hat{\mu}_{ijk} = y_{ij+}y_{i+k} / y_{i++}}$

LOG-LINEAR MODELS. EXAMPLE 2

UNIFORM ASSOCIATION MODELS: These models include 3 pairs of interactions A*C+B*C+A*B, It can not be formulated in terms of multinomial probabilities.

- The log-linear Poisson is A*C+B*C+A*B,

$$\log\left(\mu_{ijk}\right)=\eta_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\beta\gamma_{jk}+\alpha\gamma_{jk}$$

IJK-(I-1)(J-1)(K-1)=16-3*1*1=13 independent parameters

- To be assesses against the saturated log-linear Poisson model $\mathrm{E}\left[Y_{ijk}\right]=m\pi_{ijk}=y_{+++}\pi_{ijk}$ A*B*C:

$$\log\left(\mu_{ijk}\right)=\eta_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+\alpha\beta_{ij}+\alpha\gamma_{ik}+\beta\gamma_{jk}+\alpha\beta\gamma_{ijk}$$

IJK=16 independent parameters

- Fixed parameter is μ (offset set to log(m)).
- D(A*C+B*C+A*B)=1.575 with 3 =16-13 g.l., p_value > 0.05 and thus H0 can not be rejected. This model is consistent to data.

LOG-LINEAR MODELS. EXAMPLE 2

```
MTB > BLogistic 'Uni_SI' 'Uni_NO' = StatusS Estimul;
SUBC>   SF;
SUBC>   Factors 'StatusS' 'Estimul';
SUBC>   Logit;
SUBC>   Brief 3.
```

Binary Logistic Regression: Uni_SI; Uni_NO versus StatusS; Estimul

```
Link Function: Logit

Response Information

Variable  Value    Count
Uni_SI    Success   1839
Uni_NO    Failure   3152
          Total     4991
Factor Information

Factor   Levels  Values
StatusS       4  1Baix; 2Mig-baix; 3Mig-alt; 4Alt
Estimul       2  1Baix; 2Alt

Logistic Regression Table
                                                Odds     95% CI
Predictor       Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant    -3,19497   0,118491  -26,96  0,000
StatusS
 2Mig-baix  0,420133   0,117675    3,57  0,000   1,52   1,21   1,92
 3Mig-alt   0,738511   0,113821    6,49  0,000   2,09   1,67   2,62
 4Alt        1,59311   0,115270   13,82  0,000   4,92   3,92   6,17
Estimul
 2Alt        2,68292  0,0986602   27,19  0,000  14,63  12,06  17,75

Log-Likelihood = -2346,837
Test that all slopes are zero: G = 1875,806, DF = 4, P-Value = 0,000
Goodness-of-Fit Tests

Method                 Chi-Square  DF      P
Pearson                   1,57281   3  0,666
Deviance                  1,57547   3  0,665
Hosmer-Lemeshow           0,89577   4  0,925               Somers' D              0,66
```

LOG-LINEAR MODELS. EXAMPLE 2

Model Log-Lineal in R

```
> wisconsin
   Estimul   StatusS Plans_Uni Y_ijk
1    1Baix     1Baix       1No    35
2    1Baix 2Mig-baix       1No    38
3    1Baix  3Mig-alt       1No    37
4    1Baix      4Alt       1No    26
5     2Alt     1Baix       1No   133
6     2Alt 2Mig-baix       1No   303
7     2Alt  3Mig-alt       1No   467
8     2Alt      4Alt       1No   800
9    1Baix     1Baix       2Si   749
10   1Baix 2Mig-baix       2Si   627
11   1Baix  3Mig-alt       2Si   420
12   1Baix      4Alt       2Si   153
13    2Alt     1Baix       2Si   233
14    2Alt 2Mig-baix       2Si   330
15    2Alt  3Mig-alt       2Si   374
16    2Alt      4Alt       2Si   266
> wis.ordre1 <-glm(Y_ijk~Estimul+StatusS+Plans_Uni, family=poisson(link=log))
> wis.ordre2 <-glm(Y_ijk~Estimul+StatusS+Plans_Uni+Estimul*StatusS+Estimul*Plans_Uni+StatusS*Plans_Uni,
family=poisson(link=log))
> summary(wis.ordre1)
Call:  glm(formula = Y_ijk ~ Estimul + StatusS + Plans_Uni, family = poisson(link = log))
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       5.17624    0.03863 133.997  < 2e-16 ***
Estimul2Alt       0.33201    0.02870  11.568  < 2e-16 ***
StatusS2Mig-baix  0.12106    0.04050   2.989  0.00279 **
StatusS3Mig-alt   0.12106    0.04050   2.989  0.00279 **
StatusS4Alt       0.07937    0.04090   1.941  0.05230 .
Plans_Uni2Si      0.53882    0.02934  18.362  < 2e-16 ***
---
(Dispersion parameter for poisson family taken to be 1)
```

LOG-LINEAR MODELS. EXAMPLE 2

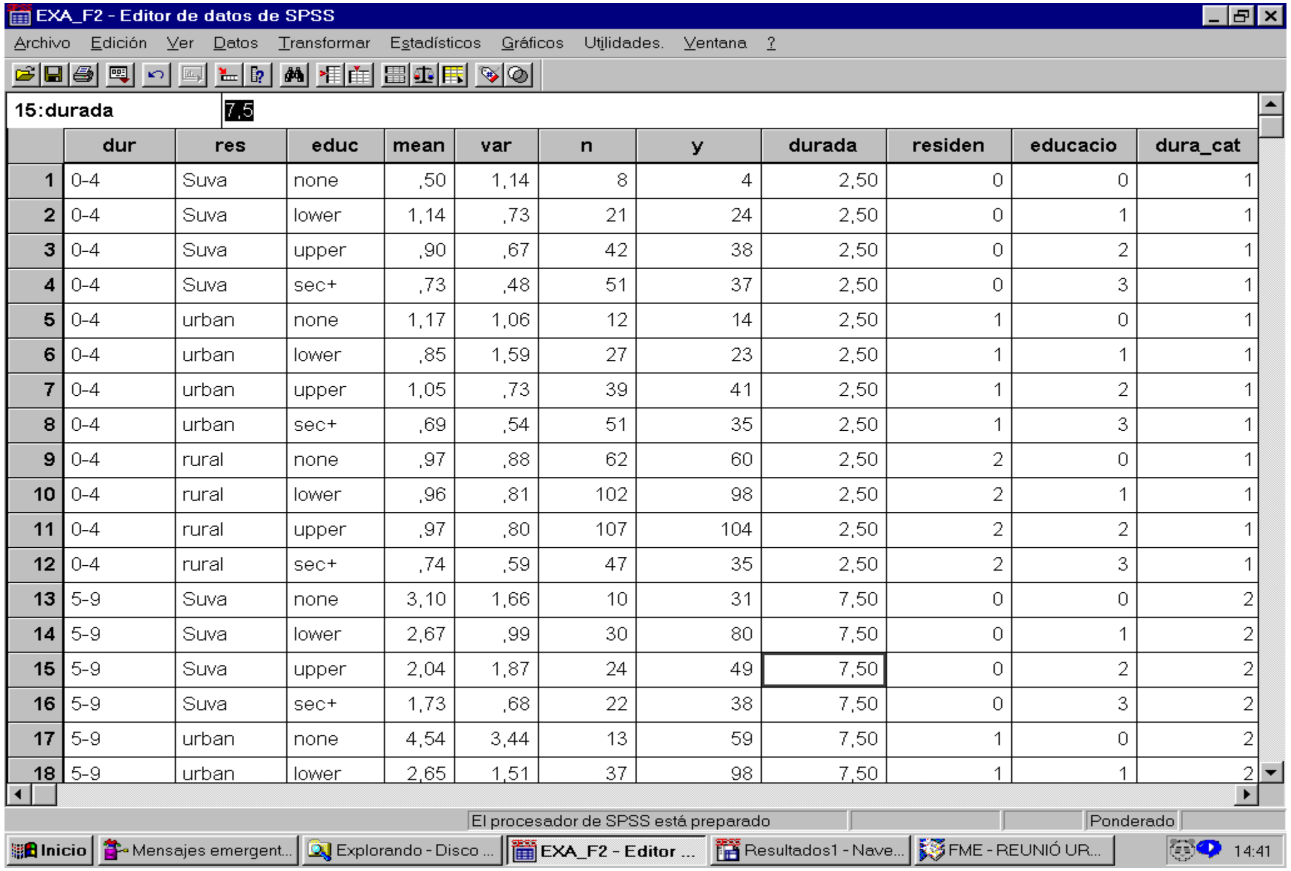
```
    Null deviance: 3211.0  on 15  degrees of freedom
Residual deviance: 2714.0  on 10  degrees of freedom
AIC: 2839.8

> summary(wis.ordre2)

Call: glm(formula = Y_ijk ~ Estimul + StatusS + Plans_Uni + Estimul * StatusS + Estimul *
Plans_Uni + StatusS * Plans_Uni, family = poisson(link = log))

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.42929    0.11931  28.743  < 2e-16 ***
Estimul2Alt                    1.49175    0.11148  13.381  < 2e-16 ***
StatusS2Mig-baix               0.23517    0.12390   1.898 0.057697 .
StatusS3Mig-alt                0.15668    0.12266   1.277 0.201493
StatusS4Alt                   -0.02735    0.13388  -0.204 0.838132
Plans_Uni2Si                   3.19497    0.11850  26.962  < 2e-16 ***
Estimul2Alt:StatusS2Mig-baix   0.55410    0.09469   5.852 4.87e-09 ***
Estimul2Alt:StatusS3Mig-alt    1.07056    0.09649  11.095  < 2e-16 ***
Estimul2Alt:StatusS4Alt        1.78588    0.11444  15.606  < 2e-16 ***
Estimul2Alt:Plans_Uni2Si      -2.68292    0.09867 -27.191  < 2e-16 ***
StatusS2Mig-baix:Plans_Uni2Si -0.42013    0.11768  -3.570 0.000357 ***
StatusS3Mig-alt:Plans_Uni2Si  -0.73851    0.11382  -6.488 8.69e-11 ***
StatusS4Alt:Plans_Uni2Si      -1.59311    0.11527 -13.820  < 2e-16 ***
---
 (Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3211.0014  on 15  degrees of freedom
Residual deviance:    1.5755  on  3  degrees of freedom
AIC: 141.39
```

LOG-LINEAR MODELS. EXAMPLE 2

```
> anova(wis.ordre2,test="Chi")
Analysis of Deviance Table
Model: poisson, link: log
Response: Y_ijk
Terms added sequentially (first to last)
                  Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                15     3211.0
Estimul            1    135.7       14     3075.3  2.360e-31
StatusS            3     11.9       11     3063.5  7.856e-03
Plans_Uni          1    349.5       10     2714.0  5.406e-78
Estimul:StatusS    3    836.6        7     1877.4 5.062e-181
Estimul:Plans_Uni  1   1621.9        6      255.5        0.0
StatusS:Plans_Uni  3    253.9        3        1.6  9.418e-55
```

LOG-LINEAR MODELS. EXAMPLE 3

6.6.3 Example 3: Number of children ever born (Little '78, G. Rodríguez '00)


EXA_F2 - Editor de datos de SPSS
Archivo Edición Ver Datos Transformar Estadísticos Gráficos Utilidades. Ventana ?
Inicio
Mensajes emergent...
Explorando - Disco ...
EXA_F2 - Editor ...
Resultados1 - Nave...
FME - REUNIÓ UR...
14:41

15:durada 7.5

	dur	res	educ	mean	var	n	y	durada	residen	educacio	dura_cat
1	0-4	Suva	none	,50	1,14	8	4	2,50	0	0	1
2	0-4	Suva	lower	1,14	,73	21	24	2,50	0	1	1
3	0-4	Suva	upper	,90	,67	42	38	2,50	0	2	1
4	0-4	Suva	sec+	,73	,48	51	37	2,50	0	3	1
5	0-4	urban	none	1,17	1,06	12	14	2,50	1	0	1
6	0-4	urban	lower	,85	1,59	27	23	2,50	1	1	1
7	0-4	urban	upper	1,05	,73	39	41	2,50	1	2	1
8	0-4	urban	sec+	,69	,54	51	35	2,50	1	3	1
9	0-4	rural	none	,97	,88	62	60	2,50	2	0	1
10	0-4	rural	lower	,96	,81	102	98	2,50	2	1	1
11	0-4	rural	upper	,97	,80	107	104	2,50	2	2	1
12	0-4	rural	sec+	,74	,59	47	35	2,50	2	3	1
13	5-9	Suva	none	3,10	1,66	10	31	7,50	0	0	2
14	5-9	Suva	lower	2,67	,99	30	80	7,50	0	1	2
15	5-9	Suva	upper	2,04	1,87	24	49	7,50	0	2	2
16	5-9	Suva	sec+	1,73	,68	22	38	7,50	0	3	2
17	5-9	urban	none	4,54	3,44	13	59	7,50	1	0	2
18	5-9	urban	lower	2,65	1,51	37	98	7,50	1	1	2

El procesador de SPSS está preparado | Ponderado

Table shows data from Little (1978) from the World Fertility Report about the number of children ever born from Indian mothers in Fiji. Included factors are: Residential Area (R, Suva, urban and rural), Years since the first marriage (D, in years grouped into 6 levels) and Education level (E, 4 levels, none, primary-low, primary-high and secondary and more) .

➡ *Target response is 'number of children ever born per woman'.*

LOG-LINEAR MODELS. EXAMPLE 3

Data is presented grouped into total number of women and total number of children in each defined by categories in factors (D), (R) i (E), n_{ijk} is the total number of women belonging to each grous. It is assumed that total number of children per woman is Poisson distributed with expected value μ_{ijk} and thus, the total number of children for each group can be modelled as a Poisson variate with expected mean $n_{ijk}\mu_{ijk}$ (Yijk)

1. Can available grouped data be valid for modelling 'the number of children per woman' ? Do we need individual data?

Individual data is not needed.

- Let Y_{ijkl} be the number of children from l-th woman in group *ijk*, (*i* for D, *j* for R, *k* for E). It is Poisson distributed with expection μ_{ijk} . Independence between women in the same group is assumed.
- Let Y_{ijk} be the total number of children from women in group *ijk*, (*i* for D, *j* for R, *k* for E). It is Poisson distributed with expection $n_{ijk}\mu_{ijk}$.

Base-line reparametrization is assumed: *i=j=k=1*, so D *0-4*, R *Suva* and E *none.*

The **additive model** can be stated as:

$$\log \mathrm{E}\left[Y_{ijkl}\right] = \log \mu_{ijk} = \mathrm{x}_{\mathrm{ijk}}^{\mathrm{T}}\beta = \eta + \alpha_i + \beta_j + \gamma_k \ \ \textit{where} \ \begin{array}{c} i=1,\dots,6 \ \ j=1,\dots,3 \ \ k=1,\dots,4 \\ \textit{with} \ \alpha_1 = \beta_1 = \gamma_1 = 0 \end{array}$$

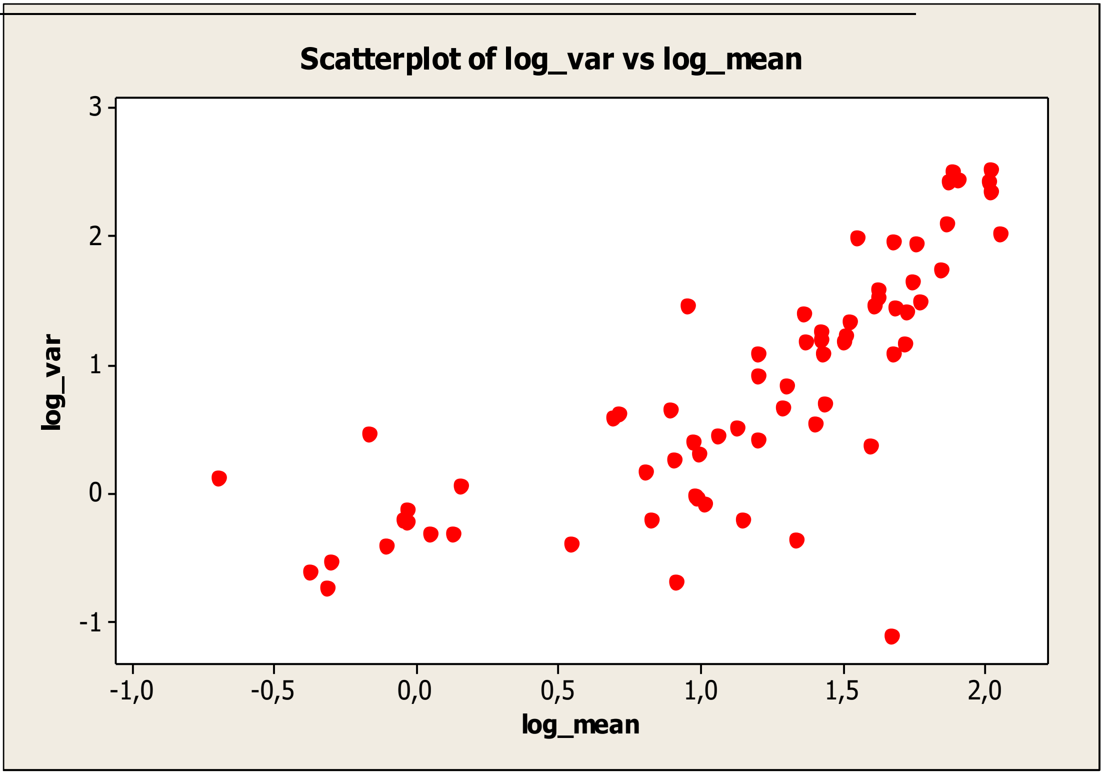
LOG-LINEAR MODELS. EXAMPLE 3

The **additive model** can be stated for grouped data with offset:

offset

$$\log \mathrm{E}\left[Y_{ijk}\right] = \log n_{ijk}\mu_{ijk} = \log n_{ijk} + \mathrm{x}_{\mathrm{ijk}}^{\mathrm{T}}\beta = \eta' + \alpha_i + \beta_j + \gamma_k \;\; where \;\; \begin{array}{c} i = 1,\ldots,6 \;\; j = 1,\ldots,3 \;\; k = 1,\ldots,4 \\ with\ \alpha_1 = \beta_1 = \gamma_1 = 0 \end{array}$$

MODEL	DEVIANCE	d.g. (ν)	$x\ t.q.\ P(\chi^2_\nu < x) = 0.95$
Nul	3731.52	69	89.3912
D	165.84	64	83.6753
R	3659.23	67	87.1081
E	2661.00	66	85.9649
D+R	120.68	62	81.3810
D+E	100.01	61	80.2321
DR	108.84	52	69.8322
DE	84.46	46	62.8296
D+R+E	70.65	59	77.9305
D+RE	59.89	53	70.9935
E+DR	57.06	49	66.3386
R+DE	54.91	44	60.4809
DR+RE	44.27	43	59.3035
DE+RE	44.60	38	53.3835
DR+DE	42.72	34	48.6024
DR+DE+RE	30.95	28	41.3371



LOG-LINEAR MODELS. EXAMPLE 3

2. What is the most significant main effect? Are all factors statistically significant? Address gross and net effects. Justify with formal tests your answer.

Most significant effect is Factor D (Years from the first marriage): 5 degrees of freedom allow to reduce null model deviance from 3731.52 to D(D)=165.84 ; i.e. 3731.52-165.84 units. Residential and Education factors are less significant ; i.e., gross effect tests show a larger pvalue for Residential and Education factors than Factor D.

Contrast E Net-effect : $\boxed{\mathbf{D}(D+R) - \mathbf{D}(D+R+E)}$=120.68-70.65=50.03 $\approx \chi^2_3 > \chi^2_{3,\alpha=0.05} = 7.815$ thus, once Facotors D and R are already in the model, adding Education Factor E is worth.

Contrast R Net-effect: $\boxed{\mathbf{D}(D+E) - \mathbf{D}(D+R+E)}$=100.01-70.65=29,36 $\approx \chi^2_2 > \chi^2_{2,\alpha=0.05} = 5.992$ thus, once Facotors D and E are already in the model, adding Residential Factor R is worth.

LOG-LINEAR MODELS. EXAMPLE 3

3. Discuss goodness of fit for the additive model D+R+E.

Assimptotically **D**(D+R+E) =70.65 $\approx \chi^2_{59}$ < $\chi^2_{59,\alpha=0.05} = 77.931$ thus, H0 "A+B+C model is consistent to data", can not be rejected.

Let us use available residual data for several hierarchical models. Contrast additive model to 1 interaction models (3), 2 interactions (3) and 3 second order interactions (1).

Contrast R*E : $\boxed{\textbf{D}\text{(D+R+E)- }\textbf{D}\text{ (D+R*E)}}$=70.65-59.89=10.76 < $\chi^2_{6,\alpha=0.05} = 12.59$ and thus H0 can not be rejected.

Contrast D*R : $\boxed{\textbf{D}\text{(D+R+E)- }\textbf{D}\text{ (E+D*R)}}$=70.65-57.06=13.59 < $\chi^2_{10,\alpha=0.05} = 18.31$ and thus H0 can not be rejected.

Contrast D*E : $\boxed{\textbf{D}\text{(D+R+E)- }\textbf{D}\text{ (R+D*E)}}$=70.65-54.91=15.74 < $\chi^2_{15,\alpha=0.05} = 24.996$ and thus H0 can not be rejected.

Any of the 1 interaction models is better than the additive model. Thus, it makes no sense to discuss more complex models. Additive model is consistent to data.

LOG-LINEAR MODELS. EXAMPLE 3.

```
# Offset is log( n ) ¡!!
> summary(ceb.ordre1)  # Additive Model D+R+E

Call: glm(formula = y ~ offset(offset) + dur + res + educ, family = poisson(link =
 log), data = ceb, na.action = na.exclude, control = list(epsilon = 0.0001,
     maxit = 50, trace = F))

Coefficients:
                  Value  Std. Error     t value
(Intercept)  1.164222457 0.015789343  73.7346981
       dur1  0.685266039 0.025537522  26.8336934
       dur2  0.309655001 0.011620359  26.6476277
       dur3  0.197641936 0.007459102  26.4967483
       dur4  0.156846343 0.004877973  32.1540010
       dur5 -0.058626491 0.005206364 -11.2605443
       res1  0.075608641 0.014162783   5.3385440
       res2  0.012216806 0.008369488   1.4596838
      educ1 -0.011540169 0.011327745  -1.0187526
      educ2 -0.107041521 0.017783621  -6.0191070
      educ3 -0.001541016 0.007702372  -0.2000703
(Dispersion Parameter for Poisson family taken to be 1 )
    Null Deviance: 3731.525 on 69 degrees of freedom
Residual Deviance: 70.65262 on 59 degrees of freedom
```

LOG-LINEAR MODELS. EXAMPLE 3.

```
> anova( ceb.ordre2, test="Chi" ) #  ceb.ordre2 interactions order 2
Analysis of Deviance Table
Poisson model
Response: y
Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev   Pr(Chi)
    NULL                    69   3731.525
     dur  5 3565.685        64    165.840 0.0000000
     res  2   45.158        62    120.681 0.0000000
    educ  3   50.029        59     70.653 0.0000000
 dur:res 10   13.594        49     57.058 0.1923126
dur:educ 15   14.339        34     42.719 0.4999782
res:educ  6   11.765        28     30.954 0.0674285

> summary(cebquasi.ordre1)
Call: glm(formula = y ~ offset(oset) + dur + res + educ, family = quasi(link = log,
     variance = "mu"), data = ceb)
Coefficients:
                   Value  Std. Error     t value
(Intercept)  1.164222457 0.017381927  66.9789056
       dur1  0.685266039 0.028113351  24.3751105
       dur2  0.309655001 0.012792440  24.2060928
       dur3  0.197641936 0.008211460  24.0690374
       dur4  0.156846343 0.005369987  29.2079557
       dur5 -0.058626491 0.005731501 -10.2288197
       res1  0.075608641 0.015591305   4.8494107
       res2  0.012216806 0.009213672   1.3259433
       educ1 -0.011540169 0.012470312  -0.9254114
      educ2 -0.107041521 0.019577357  -5.4676185
      educ3 -0.001541016 0.008479267  -0.1817393
```

LOG-LINEAR MODELS. EXAMPLE 3.

```
(Dispersion Parameter for Quasi-likelihood family taken to be 1.211903 )

    Null Deviance: 3731.525 on 69 degrees of freedom
Residual Deviance: 70.65262 on 59 degrees of freedom

> anova( cebquasi.ordre2, test="Chi")

Analysis of Deviance Table
Quasi-likelihood model
Response: y

Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev   Pr(Chi)
    NULL                    69   3731.525
     dur  5 3565.685        64    165.840 0.0000000
     res  2   45.158        62    120.681 0.0000000
    educ  3   50.029        59     70.653 0.0000000
 dur:res 10   13.594        49     57.058 0.1923126
dur:educ 15   14.339        34     42.719 0.4999782
res:educ  6   11.765        28     30.954 0.0674285
```

LOG-LINEAR MODELS. EXAMPLE 4

6.6.4 Example 4: Car insurance –Classification of sinister risk (Ll.Bermúdez, M.Denuit, J.Dhaene)

Data from '*Exponential Bonus-Malus Systems Integrating a priori Risk Classification*' (2000), the first author was a student in my Generalized Linear Models course at the FME.

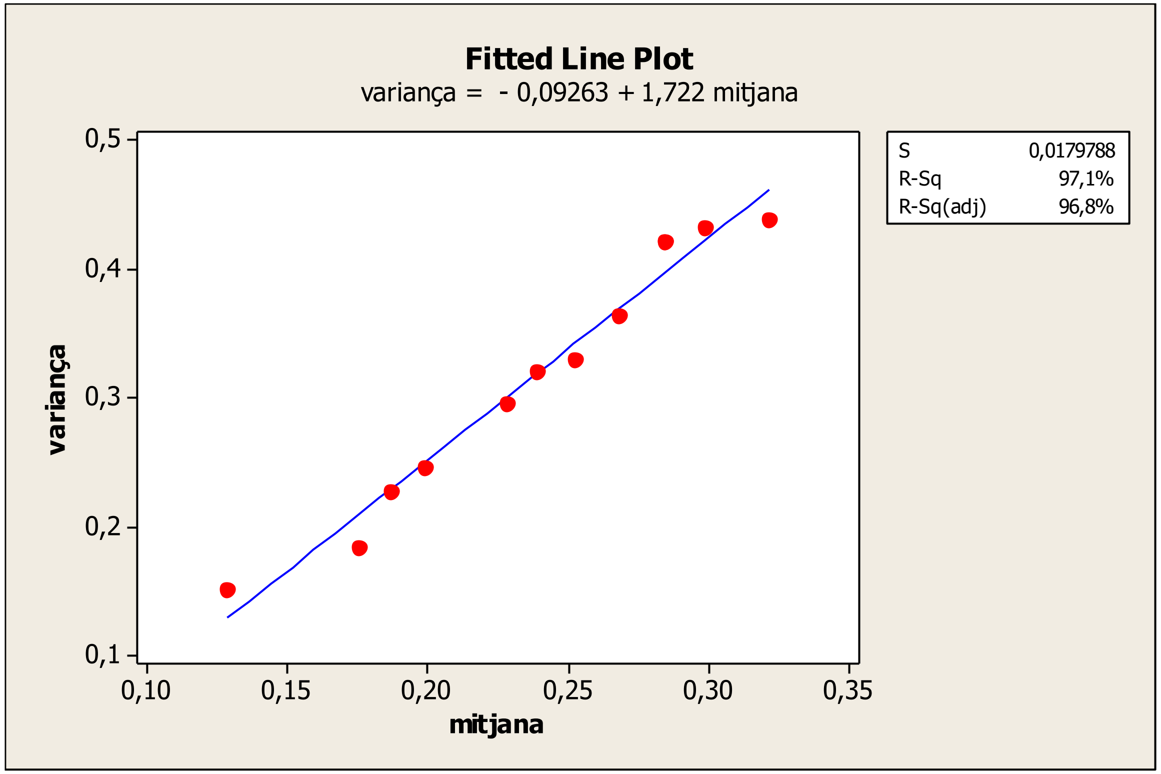
Fare system for car insurance companies has been studied in greater depth. Sinister risk is analyzed as a function of gender, age group, occupancy, colour of the vehicle, etc. Annual kilometers are not taken into account.

A Bonus-Malus system is applied. Insurance keepers are segmented into potential risk groups and fare is determined based on this classification.

An Spanish car insurance Company models sinister risk from Factor Age-group (<36, 36 a 49, >49) i del Factor Potència, vehicle thrust has been categorized 4 levels (<54, 54-75, 76-118 and >118). Number of sinisters for each group is shown in the table below.

LOG-LINEAR MODELS. EXAMPLE 4

F_Age	F_Thrust	m_k	y_k	mean	variance
<36	<54	3945	736	0.1866	0.2270
36-49	<54	9023	1418	0.1751	0.1828
>50	<54	11758	1509	0.1283	0.1501
<36	54-75	11947	3208	0.2685	0.3635
36-49	54-75	25719	5862	0.2279	0.2946
>50	54-75	27287	5420	0.1986	0.2451
<36	76-118	8447	2527	0.2992	0.4322
36-49	76-118	19609	4953	0.2526	0.3288
>50	76-118	18688	4459	0.2386	0.3200
<36	>119	1486	478	0.3217	0.4376
36-49	>119	5762	1640	0.2846	0.4214



```
> summary(bm.ordre1)

Call: glm(formula = y ~ offset(bm$logn) + edat + pot, family = poisson(link = log))
Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
 (Intercept) -1.72185    0.01980  -86.97   <2e-16 ***
edate2      -0.16338    0.01472  -11.10   <2e-16 ***
edate3      -0.28004    0.01492  -18.77   <2e-16 ***
potp2        0.39874    0.01850   21.55   <2e-16 ***
potp3        0.53238    0.01891   28.16   <2e-16 ***
potp4        0.61495    0.02355   26.11   <2e-16 ***
---
 (Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1413.850  on 11  degrees of freedom
Residual deviance:   18.604  on  6  degrees of freedom
AIC: 144.78
> summary(bmquasi.ordre1)

Call: glm(formula = y ~ offset(bm$logn) + edat + pot, family = quasi(link = log,
    variance = "mu"))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
 (Intercept) -1.72185    0.03488 -49.368 4.63e-09 ***
edate2      -0.16338    0.02593  -6.301 0.000745 ***
edate3      -0.28004    0.02628 -10.654 4.03e-05 ***
potp2        0.39874    0.03260  12.233 1.82e-05 ***
potp3        0.53238    0.03330  15.985 3.81e-06 ***
potp4        0.61495    0.04149  14.821 5.93e-06 ***
---
 (Dispersion parameter for quasi family taken to be 3.103063)

    Null deviance: 1413.850  on 11  degrees of freedom
Residual deviance:   18.604  on  6  degrees of freedom
AIC: NA
```

```
> anova(bm.ordre1,test="Chi")
Analysis of Deviance Table
Model: poisson, link: log
Response: y
Terms added sequentially (first to l

     Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                    11    1413.85
edat  2   374.45         9    1039.40  4.888e-82
pot   3  1020.80         6      18.60 5.542e-221
```

Chi squared based tests, under Poisson distribution with dispersion parameter of 1

```
> anova(bmquasi.ordre1,test="Chi")
Analysis of Deviance Table
Model: quasi, link: log
Response: y
Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    11    1413.85
edat  2   374.45         9    1039.40 6.260e-27
pot   3  1020.80         6      18.60 5.348e-71
>
> bm.nb1<-glm.nb(formula = y ~ offset(bm$logn) + edat + pot)
> summary(bm.nb1)
Call: glm.nb(formula = y ~ offset(bm$logn) + edat + pot, init.theta = 4559.012502,
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.71973    0.02305 -74.612   <2e-16 ***
edate2      -0.16121    0.01896  -8.501   <2e-16 ***
```

```
edate3      -0.28277    0.01909 -14.810   <2e-16 ***
potp2        0.39721    0.02229  17.818   <2e-16 ***
potp3        0.53030    0.02263  23.430   <2e-16 ***
potp4        0.61211    0.02680  22.844   <2e-16 ***
---
 (Dispersion parameter for Negative Binomial(4559.012) family taken to be 1)
    Null deviance: 948.577  on 11  degrees of freedom
Residual deviance:  11.423  on  6  degrees of freedom
AIC: 144.99
              Theta:  4559
          Std. Err.:  4892
 2 x log-likelihood:  -130.986
>
> bm.gnb1<-glm(formula = y ~ offset(bm$logn) + edat + pot, family=neg.bin(4559.013),
data = bm )
> summary(bm.gnb1)

Call: glm(formula = y ~ offset(bm$logn) + edat + pot, family = neg.bin(4559.013),
    data = bm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.71973    0.03182 -54.051 2.69e-09 ***
edate2      -0.16121    0.02618  -6.158 0.000841 ***
edate3      -0.28277    0.02636 -10.729 3.87e-05 ***
potp2        0.39721    0.03077  12.908 1.33e-05 ***
potp3        0.53030    0.03124  16.973 2.67e-06 ***
potp4        0.61211    0.03699  16.548 3.10e-06 ***
---
 (Dispersion parameter for Negative Binomial family taken to be 1.905539)
```

```
    Null deviance: 948.577  on 11  degrees of freedom
Residual deviance:  11.423  on  6  degrees of freedom
AIC: 142.99

Number of Fisher Scoring iterations: 3

>> anova(bm.gnb1,test="F")
Analysis of Deviance Table

Model: Negative Binomial, link: log
Response: y
Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                    11     948.58
edat  2   225.46         9     723.12  59.159 0.0001124 ***
pot   3   711.69         6      11.42 124.496 8.595e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>> summary(bm.ga1)

Call:
glm(formula = y ~ offset(bm$logn) + edat + pot, family = Gamma(link = log))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.71132    0.02604 -65.726 8.34e-10 ***
edate2      -0.15682    0.02604  -6.023 0.000945 ***
edate3      -0.28981    0.02604 -11.130 3.14e-05 ***
potp2        0.39024    0.03007  12.980 1.29e-05 ***
potp3        0.52217    0.03007  17.368 2.34e-06 ***
```

Contrasts based on Fisher distribution and considering dispersion parameter > 1

```
potp4         0.59897    0.03007  19.922 1.04e-06 ***
---
(Dispersion parameter for Gamma family taken to be 0.001355877)

    Null deviance: 0.7606494  on 11  degrees of freedom
Residual deviance: 0.0081576  on  6  degrees of freedom
AIC: 144.76

Number of Fisher Scoring iterations: 3

> alfa<-1/0.001355877;alfa # Gamma shape parameter
[1] 737.53
> anova(bm.ga1,test="F")
Analysis of Deviance Table
Model: Gamma, link: log
Response: y
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                    11    0.76065
edat  2  0.15690         9    0.60375  57.86 0.0001198 ***
pot   3  0.59559         6    0.00816 146.42 5.325e-06 ***
```

Contrasts based on Fisher distribution and considering dispersion parameter > 1

```
>
> summary(bm.lg1lm)
Call:
lm(formula = log(y) ~ offset(bm$logn) + edat + pot)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.71188    0.02609 -65.614 8.43e-10 ***
edate2      -0.15676    0.02609  -6.008 0.000958 ***
```

```
edate3       -0.29019    0.02609 -11.122 3.15e-05 ***
potp2         0.39090    0.03013  12.975 1.29e-05 ***
potp3         0.52229    0.03013  17.337 2.36e-06 ***
potp4         0.59952    0.03013  19.900 1.04e-06 ***
Residual standard error: 0.0369 on 6 degrees of freedom
Multiple R-squared: 0.9989,     Adjusted R-squared: 0.9979
F-statistic:  1060 on 5 and 6 DF,  p-value: 9.474e-09

> summary(bm.lg1glm)
Call: glm(formula = log(y) ~ offset(bm$logn) + edat + pot, family = gaussian)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.71188    0.02609 -65.614 8.43e-10 ***
edate2       -0.15676    0.02609  -6.008 0.000958 ***
edate3       -0.29019    0.02609 -11.122 3.15e-05 ***
potp2         0.39090    0.03013  12.975 1.29e-05 ***
potp3         0.52229    0.03013  17.337 2.36e-06 ***
potp4         0.59952    0.03013  19.900 1.04e-06 ***
---
 (Dispersion parameter for gaussian family taken to be 0.001361379)

    Null deviance: 0.8157702  on 11  degrees of freedom
Residual deviance: 0.0081683  on  6  degrees of freedom
AIC: -39.454

> 0.0369^2
[1] 0.00136161


> anova(bm.lg1lm)
Analysis of Variance Table
```

```
Response: log(y)
          Df  Sum Sq  Mean Sq F value    Pr(>F)    
edat       2 0.16878 0.084389  61.988 9.837e-05 ***
pot        3 0.63882 0.212941 156.416 4.381e-06 ***
Residuals  6 0.00817 0.001361                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(bm.lg1glm,test="F")
Analysis of Deviance Table
Model: gaussian, link: identity
Response: log(y)
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev       F    Pr(>F)    
NULL                    11    0.81577                      
edat  2  0.16878         9    0.64699  61.988 9.837e-05 ***
pot   3  0.63882         6    0.00817 156.416 4.381e-06 ***
---
>
```

Contrasts based on Fisher distribution and considering dispersion parameter > 1

Overdispersion hypothesis implies an scaled deviance affected by ϕ, dispersion parameter estimate and scaled deviance test leads to be assymptotically Fisher distributed instead of Chi-squared distributed.

LOG-LINEAR MODELS. EXAMPLE 5

6.6.5 Example 5: Ship incidents (McCullagh, 1989)

A data frame containing 40 observations on 5 ship types in 4 vintages and 2 service periods. The models are fit only to those observations with service > 0.

Variable	Description
type	factor with levels "A" to "E" for the different ship types,
construction	factor with levels "1960-64", "1965-69", "1970-74", "1975-79" for the periods of construction,
operation	factor with levels "1960-74", "1975-79" for the periods of operation,
service	aggregate months of service,
incidents	number of damage incidents.

```
> summary(df)
 type   construction   operation       service         incidents    
 A:7   1960-64: 8    1960-74:14   Min.    :   45   Min.    : 0.00  
 B:7   1965-69:10    1975-79:20   1st Qu.:  371   1st Qu.: 1.00  
 C:7   1970-74:10                 Median : 1095   Median : 4.00  
 D:7   1975-79: 6                 Mean   : 4811   Mean   :10.47  
 E:6                              3rd Qu.: 2223   3rd Qu.:11.75  
                                  Max.   :44882   Max.   :58.00  
```

LOG-LINEAR MODELS. EXAMPLE 5

```
> m3 <- glm(incidents ~ type + construction + operation, family = poisson,
+            data = df, offset = log(service))
> summary(m3)
Call:glm(formula = incidents ~ type + construction + operation, family = poisson,
    data = df, offset = log(service))

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.40288    0.21752 -29.435  < 2e-16 ***
typeB               -0.54471    0.17761  -3.067  0.00216 **
typeC               -0.68876    0.32903  -2.093  0.03632 *
typeD               -0.07431    0.29056  -0.256  0.79815
typeE                0.32053    0.23575   1.360  0.17396
construction1965-69  0.69585    0.14966   4.650 3.33e-06 ***
construction1970-74  0.81746    0.16984   4.813 1.49e-06 ***
construction1975-79  0.44497    0.23324   1.908  0.05642 .
operation1975-79     0.38386    0.11826   3.246  0.00117 **
---
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  38.963  on 25  degrees of freedom
AIC: 154.83
```

LOG-LINEAR MODELS. EXAMPLE 5

```
> Anova(m3)
Analysis of Deviance Table (Type II tests)
Response: incidents
             LR Chisq Df Pr(>Chisq)
type           23.573  4  9.725e-05 ***
construction   31.401  3  6.998e-07 ***
operation      10.628  1   0.001114 **
---
```

Model interpretation for factor operation in (m3). For (m3) model, ships operating in 1975-59, show a 46.8% increase in the expected number of incidents per month with respect to the base operation category 1960-74, all else being equal (*ceteris paribus*).

```
> 100*(exp(coef(m3)[9])-1)
operation1975-79
        46.79386
```

Expected number of incidents per month in the reference group (Type A ships operating during 1960-74 and contructed in 1960-64):

```
> exp(coef(m3)[1]) # Expected nb of incidents per month in reference group

(Intercept)
0.001656784
```

LOG-LINEAR MODELS. EXAMPLE 5

Expected number of incidents per year in the reference group

```
> 12*exp(coef(m3)[1]) # Expected nb of incidents per year in reference group
(Intercept)
  0.0198814
```

Probability of having 1 accident in a year

```
> # Prob 1 accident in 1 year  . Poisson( mu = 0.0198814, k=1) =( mu^k)*exp(-mu)/k!
> 0.0198814*exp(-0.0198814 )
[1] 0.01949003
```

Probability of having 1 or more accidents in a year

```
> # Prob 1 or more accidents in 1 year: 1-Poisson( mu = 0.0198814, k=0) =(
mu^k)*exp(-mu)/k!
> 1-exp(-0.0198814 )
[1] 0.01968507
```

Probability of having 0 accidents in a year

```
> (1)*exp(-12*exp(coef(m3)[1]))  # 0 Incidents per year
(Intercept)
  0.9803149
```

Goodness of fit test: H0: (m3) Model fits data

```
> 1-pchisq(m3$deviance,m3$df.residual)
[1] 0.03715884
```

LOG-LINEAR MODELS. EXAMPLE 5

Are there any interactions needed?

```
> Anova(m3ac,test="LR")
Analysis of Deviance Table (Type II tests)
Response: incidents
                  LR Chisq Df Pr(>Chisq)
type                23.573  4  9.725e-05 ***
construction        31.401  3  6.998e-07 ***
operation           10.621  1   0.001118 **
type:construction   24.216 11   0.011852 *
---
> Anova(m3ad,test="LR")
Analysis of Deviance Table (Type II tests)
Response: incidents
               LR Chisq Df Pr(>Chisq)
type            23.5733  4  9.725e-05 ***
operation       10.6284  1   0.001114 **
construction    30.5565  3  1.054e-06 ***
type:operation   5.0451  4   0.282699
---
> Anova(m3cd,test="LR")
Analysis of Deviance Table (Type II tests)
Response: incidents
                       LR Chisq Df Pr(>Chisq)
type                    23.6008  4  9.602e-05 ***
construction            31.4012  3  6.998e-07 ***
operation               10.6284  1   0.001114 **
construction:operation   1.7666  2   0.413407
```

LOG-LINEAR MODELS. EXAMPLE 5

```
> m3ac <- glm(incidents ~ type * construction + operation, family = poisson,
+   data = df, offset = log(service))  # Best Model
> summary(m3ac)# Some parameters can not be estimated
```

Overdispersion test: Negative binomial

```
> dispersiontest(m3, trafo = 2)
  Overdispersion test data:  m3
z = -0.6129, p-value = 0.73
alternative hypothesis: true alpha is greater than 0
sample estimates:
     alpha
-0.0111868
```

Negative binomial model (m3.nb)(not needed)

```
> library(MASS)
> m3.nb<-glm.nb(incidents ~ type+ construction + operation+ offset(log(service)),
data = df)
Warning messages:
1: In theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace >  :
  iteration limit reached
> summary(m3.nb)
Call:glm.nb(formula = incidents ~ type + construction + operation +
    offset(log(service)), data = df, init.theta = 52521.06565, link = log)
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.40286    0.21757 -29.429  < 2e-16 ***
typeB              -0.54471    0.17764  -3.066  0.00217 **
```

```
typeC                -0.68875    0.32905  -2.093  0.03634 *
typeD                -0.07431    0.29058  -0.256  0.79816
typeE                 0.32057    0.23578   1.360  0.17395
construction1965-69   0.69584    0.14971   4.648 3.35e-06 ***
construction1970-74   0.81743    0.16988   4.812 1.50e-06 ***
construction1975-79   0.44493    0.23328   1.907  0.05649 .
operation1975-79      0.38387    0.11830   3.245  0.00117 **
---
 (Dispersion parameter for Negative Binomial(52521.07) family taken to be 1)

    Null deviance: 146.247  on 33  degrees of freedom
Residual deviance:  38.958  on 25  degrees of freedom
AIC: 156.83

              Theta:  52521
          Std. Err.:  565839
Warning while fitting theta: iteration limit reached
 2 x log-likelihood:  -136.832

> m3.nb1<-glm(incidents ~ type+ construction + operation, offset=log(service),
family=neg.bin(52521.07),data = df)
> Anova(m3.nb1,test="F")
Analysis of Deviance Table (Type II tests)
Response: incidents
Error estimate based on Pearson residuals
             Sum Sq Df F value   Pr(>F)
type         23.569  4  3.4712 0.021852 *
construction 31.378  3  6.1618 0.002774 **
operation    10.622  1  6.2578 0.019276 *
Residuals    42.436 25
```