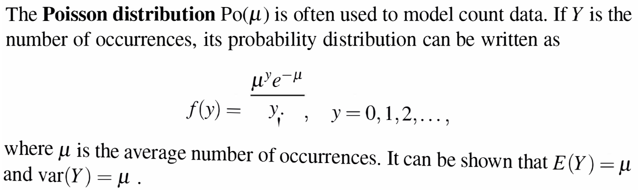
Lecture 10: Poisson Regression and Log-linear models Regression

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Poisson distribution



The **Poisson distribution** Po(μ) is often used to model count data. If Y is the number of occurrences, its probability distribution can be written as

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots,$$

where μ is the average number of occurrences. It can be shown that $E(Y) = \mu$ and $\mathrm{var}(Y) = \mu$.

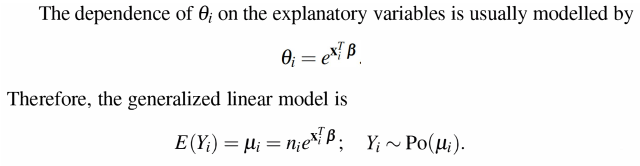
Interpretation of μ:

- The average number of customers who buy a particular product out of every 100 customers who enter the store.
- For motor vehicle crashes, the rate parameter may be defined in many different ways: crashes per 1,000 population, crashes per 1,000 licensed drivers, crashes per 1,000 motor vehicles, or crashes per 100,000 km travelled by motor vehicles.
- In insurance, the motor vehicle crash rate is usually specified as the rate per year (e.g., crashes per 100,000 km per year).
- More generally, the rate is specified in terms of units of exposure; for instance, customers entering a store are exposed to the opportunity to buy the product of interest.

Poisson regression

Let $Y_1, \dots, Y_N$ be independent random variables with Y_i denoting the number of events observed from exposure n_i for the ith covariate pattern. Then

$$E(Y_i) = \mu_i = n_i\theta_i.$$



The dependence of θ_i on the explanatory variables is usually modelled by

$$\theta_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}.$$

Therefore, the generalized linear model is

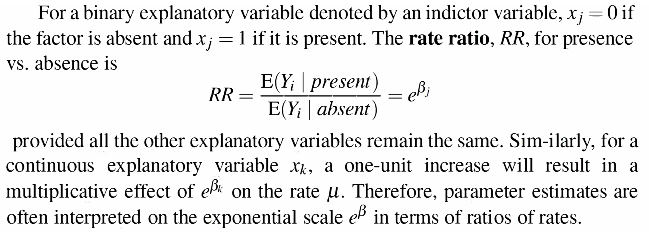
$$E(Y_i) = \mu_i = n_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}; \quad Y_i \sim \mathrm{Po}(\mu_i).$$

The natural link function is the logarithmic function

$$\log \mu_i = \log n_i + \mathbf{x}_i^T \beta.$$

The $\log n_i$, known constant, which is readily incorporated into the estimation procedure is called the is called the **offset**.

Rate ratio



For a binary explanatory variable denoted by an indictor variable, $x_j = 0$ if the factor is absent and $x_j = 1$ if it is present. The **rate ratio**, *RR*, for presence vs. absence is

$$RR = \frac{\mathrm{E}(Y_i \mid present)}{\mathrm{E}(Y_i \mid absent)} = e^{\beta_j}$$

provided all the other explanatory variables remain the same. Sim-ilarly, for a continuous explanatory variable x_k, a one-unit increase will result in a multiplicative effect of e^{β_k} on the rate μ. Therefore, parameter estimates are often interpreted on the exponential scale e^{β} in terms of ratios of rates.

Inference

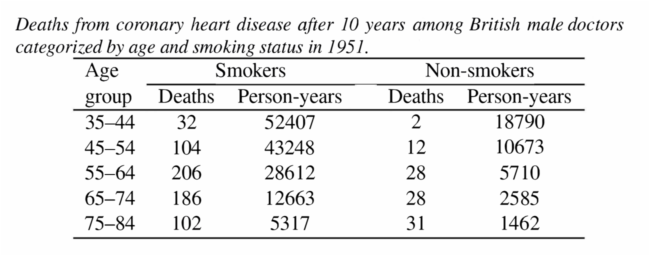
- Wald test statistic
$$\frac{b_j - \beta_j}{se(b_j)} \sim N(0, 1).$$
- Fitted values
$$\hat{Y}_i = \hat{\mu}_i = n_i \exp\left\{\mathbf{x}_i^T \boldsymbol{b}\right\}.$$
- Pearson residuals
$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}.$$
- Standardized Pearson residuals
$$sr_i = \frac{o_i - e_i}{\sqrt{e_i}\sqrt{1 - h_i}}.$$
- Pearson chi-squared statistic
$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(N - p)$$

  where p is the number of parameters that are estimated.

Inference

- Deviance
$$D = 2\sum o_i \log(o_i/e_i) \sim \chi^2(N-p).$$
- Residual deviance
$$D = \mathrm{sign}(o_i - e_i)\sqrt{2o_i \log(o_i/e_i)}.$$
- The likelihood ratio chi-squared statistic
$$C = 2[l(\boldsymbol{b}) - l(\boldsymbol{b})_{\min}] \sim \chi^2(p-1).$$
- Pseudo R^2
$$R^2 = 1 - \frac{l(\boldsymbol{b})}{l(\boldsymbol{b}_{\min})}.$$

Example: British doctors’smoking and coronary death



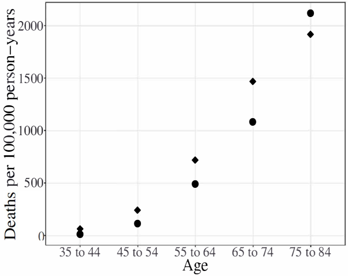
Deaths from coronary heart disease after 10 years among British male doctors categorized by age and smoking status in 1951.

Age	Smokers		Non-smokers	
group	Deaths	Person-years	Deaths	Person-years
35–44	32	52407	2	18790
45–54	104	43248	12	10673
55–64	206	28612	28	5710
65–74	186	12663	28	2585
75–84	102	5317	31	1462

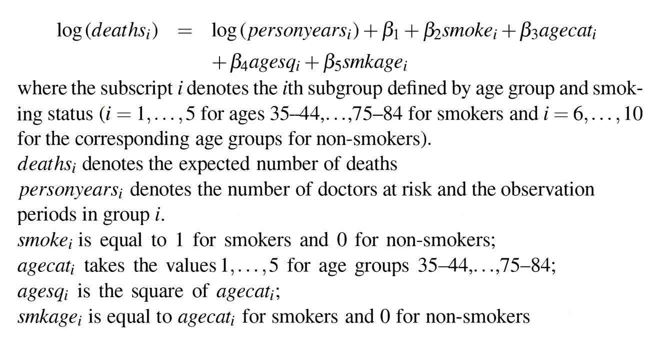
The questions of interest are

- Is the death rate higher for smokers than non-smokers?
- If so, by how much?
- Is the differential effect related to age?

Example: British doctors’smoking and coronary death



Deaths rates from coronary heart disease per 100,000 person-years for smokers (diamonds) and non-smokers (dots).



$$\begin{aligned}\log(deaths_i) &= \log(personyears_i) + \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i \\ &\quad + \beta_4 agesq_i + \beta_5 smkage_i\end{aligned}$$

where the subscript i denotes the ith subgroup defined by age group and smoking status ($i = 1, \ldots, 5$ for ages 35–44,…,75–84 for smokers and $i = 6, \ldots, 10$ for the corresponding age groups for non-smokers).

$deaths_i$ denotes the expected number of deaths

$personyears_i$ denotes the number of doctors at risk and the observation periods in group i.

$smoke_i$ is equal to 1 for smokers and 0 for non-smokers;

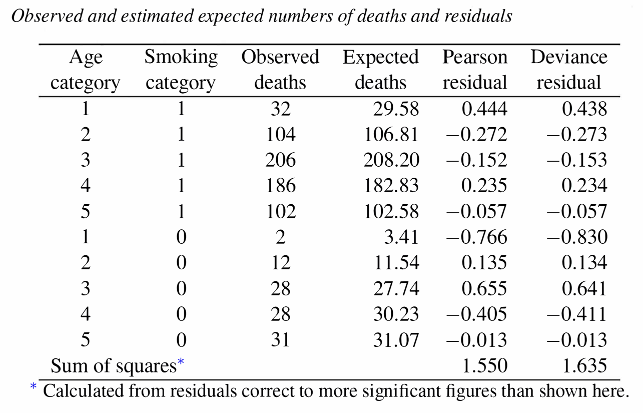
$agecat_i$ takes the values $1, \ldots, 5$ for age groups 35–44,…,75–84;

$agesq_i$ is the square of $agecat_i$;

$smkage_i$ is equal to $agecat_i$ for smokers and 0 for non-smokers

Table *Parameter estimates obtained by fitting Model to the data*

Term	*agecat*	*agesq*	*smoke*	*smkage*
$\widehat{\beta}$	2.376	−0.198	1.441	−0.308
$s.e.(\widehat{\beta})$	0.208	0.027	0.372	0.097
Wald statistic	11.43	−7.22	3.87	−3.17
p-value	< 0.001	< 0.001	< 0.001	0.002
Rate ratio	10.77	0.82	4.22	0.74
95% confidence interval	7.2, 16.2	0.78, 0.87	2.04, 8.76	0.61, 0.89



Observed and estimated expected numbers of deaths and residuals

Age category	Smoking category	Observed deaths	Expected deaths	Pearson residual	Deviance residual
1	1	32	29.58	0.444	0.438
2	1	104	106.81	−0.272	−0.273
3	1	206	208.20	−0.152	−0.153
4	1	186	182.83	0.235	0.234
5	1	102	102.58	−0.057	−0.057
1	0	2	3.41	−0.766	−0.830
2	0	12	11.54	0.135	0.134
3	0	28	27.74	0.655	0.641
4	0	28	30.23	−0.405	−0.411
5	0	31	31.07	−0.013	−0.013
Sum of squares*				1.550	1.635

* Calculated from residuals correct to more significant figures than shown here.

Contingency tables: design

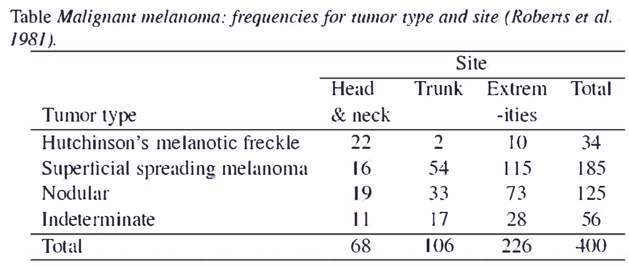
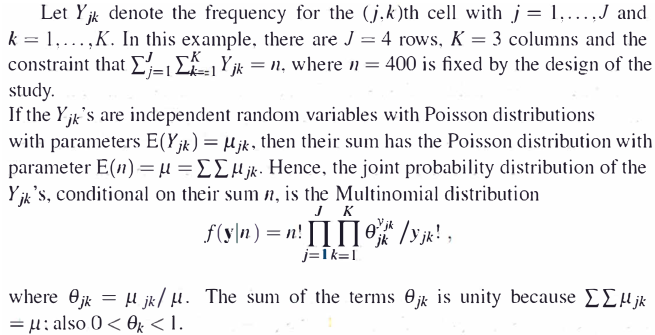


Table *Malignant melanoma: frequencies for tumor type and site (Roberts et al. 1981).*

	Site			
	Head	Trunk	Extrem	Total
Tumor type	& neck		-ities	
Hutchinson's melanotic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

The question of interest is whether there is any association between tumor type and site?

Contingency tables: design



Let Y_{jk} denote the frequency for the (j,k)th cell with $j = 1,\dots,J$ and $k = 1,\dots,K$. In this example, there are $J = 4$ rows, $K = 3$ columns and the constraint that $\sum_{j=1}^{J} \sum_{k=1}^{K} Y_{jk} = n$, where $n = 400$ is fixed by the design of the study.

If the Y_{jk}'s are independent random variables with Poisson distributions with parameters $\mathrm{E}(Y_{jk}) = \mu_{jk}$, then their sum has the Poisson distribution with parameter $\mathrm{E}(n) = \mu = \sum\sum \mu_{jk}$. Hence, the joint probability distribution of the Y_{jk}'s, conditional on their sum n, is the Multinomial distribution

$$f(\mathbf{y}|n) = n! \prod_{j=1}^{J} \prod_{k=1}^{K} \theta_{jk}^{y_{jk}} / y_{jk}!\,,$$

where $\theta_{jk} = \mu_{jk}/\mu$. The sum of the terms θ_{jk} is unity because $\sum\sum \mu_{jk} = \mu$; also $0 < \theta_k < 1$.

The usual link function for a Poisson model gives

$$\log \mu_{jk} = \log n + \log \theta_{jk}.$$

Contingency tables: design

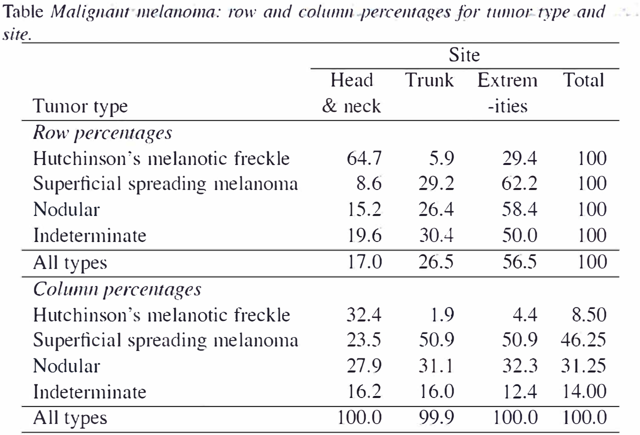


Table *Malignant melanoma: row and column percentages for tumor type and site.*

	Site			
Tumor type	Head & neck	Trunk	Extrem -ities	Total
Row percentages				
Hutchinson's melanotic freckle	64.7	5.9	29.4	100
Superficial spreading melanoma	8.6	29.2	62.2	100
Nodular	15.2	26.4	58.4	100
Indeterminate	19.6	30.4	50.0	100
All types	17.0	26.5	56.5	100
Column percentages				
Hutchinson's melanotic freckle	32.4	1.9	4.4	8.50
Superficial spreading melanoma	23.5	50.9	50.9	46.25
Nodular	27.9	31.1	32.3	31.25
Indeterminate	16.2	16.0	12.4	14.00
All types	100.0	99.9	100.0	100.0

Thus, θ_{jk} can be interpreted as the probability of an observation in the (j, k)th cell of the table.

Example: Randomized controlled trial of influenza vaccine

- In a prospective study of a new living attenuated recombinant vaccine for influenza, patients were randomly allocated to two groups, one of which was given the new vaccine and the other a saline placebo.
- The responses were titre levels of hemagglutinin inhibiting antibody found in the blood six weeks after vaccination; they were categorized as “small," “medium" or “large."
- We want to know if the pattern of responses is the same for each treatment group.

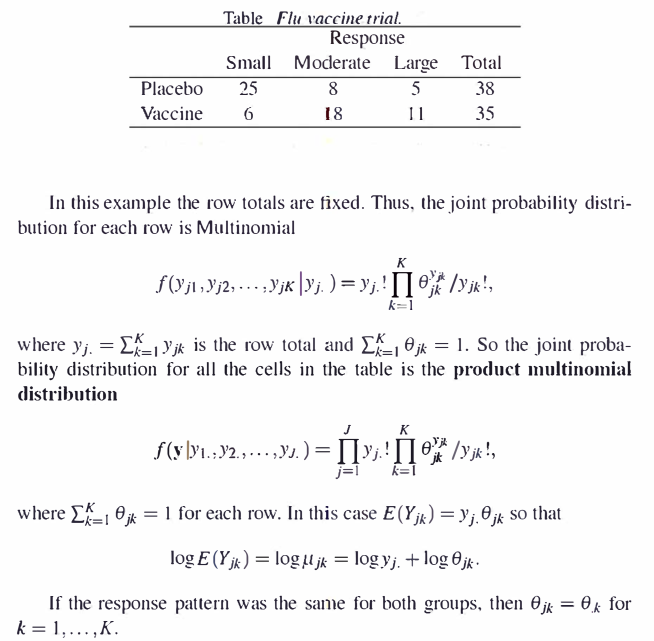


Table *Flu vaccine trial.*

	Response			
	Small	Moderate	Large	Total
Placebo	25	8	5	38
Vaccine	6	18	11	35

In this example the row totals are fixed. Thus, the joint probability distribution for each row is Multinomial

$$f(y_{j1},y_{j2},\ldots,y_{jK}|y_{j.})=y_{j.}!\prod_{k=1}^{K}\theta_{jk}^{y_{jk}}/y_{jk}!,$$

where $y_{j.}=\sum_{k=1}^{K}y_{jk}$ is the row total and $\sum_{k=1}^{K}\theta_{jk}=1$. So the joint probability distribution for all the cells in the table is the **product multinomial distribution**

$$f(\mathbf{y}|y_{1.},y_{2.},\ldots,y_{J.})=\prod_{j=1}^{J}y_{j.}!\prod_{k=1}^{K}\theta_{jk}^{y_{jk}}/y_{jk}!,$$

where $\sum_{k=1}^{K}\theta_{jk}=1$ for each row. In this case $E(Y_{jk})=y_{j.}\theta_{jk}$ so that

$$\log E(Y_{jk})=\log\mu_{jk}=\log y_{j.}+\log\theta_{jk}.$$

If the response pattern was the same for both groups, then $\theta_{jk}=\theta_{.k}$ for $k=1,\ldots,K$.

Example: Case-control study of gastric and duodenal ulcers and aspirin use

In this retrospective case-control study, a group of ulcer patients was compared with a group of control patients not known to have peptic ulcer, but who were similar to the ulcer patients with respect to age, sex and socioeconomic status.

The ulcer patients were classified according to the site of the ulcer: gastric or duodenal. Aspirin use was ascertained for all subjects.

Some questions of interest are:

- Is gastric ulcer associated with aspirin use?
- Is duodenal ulcer associated with aspirin use?
- Is any association with aspirin use the same for both ulcer sites?

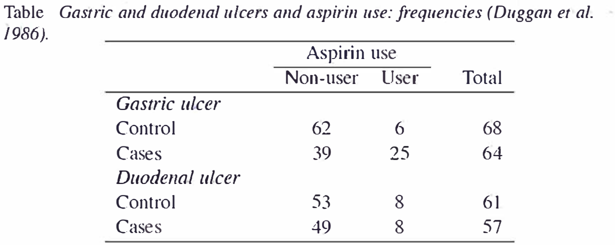


Table *Gastric and duodenal ulcers and aspirin use: frequencies (Duggan et al. 1986).*

	Aspirin use		
	Non-user	User	Total
Gastric ulcer			
Control	62	6	68
Cases	39	25	64
Duodenal ulcer			
Control	53	8	61
Cases	49	8	57

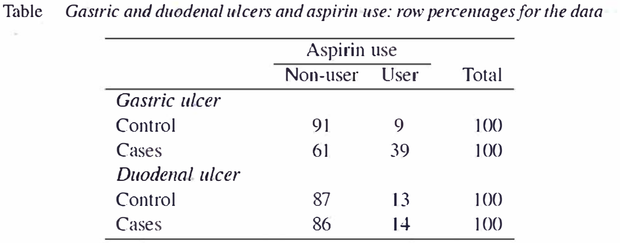
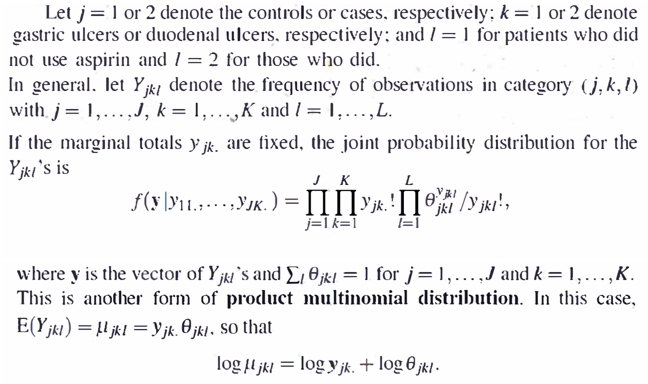


Table *Gastric and duodenal ulcers and aspirin use: row percentages for the data*

	Aspirin use		
	Non-user	User	Total
Gastric ulcer			
Control	91	9	100
Cases	61	39	100
Duodenal ulcer			
Control	87	13	100
Cases	86	14	100



Let $j = 1$ or 2 denote the controls or cases, respectively; $k = 1$ or 2 denote gastric ulcers or duodenal ulcers, respectively; and $l = 1$ for patients who did not use aspirin and $l = 2$ for those who did.
In general, let Y_{jkl} denote the frequency of observations in category (j, k, l) with $j = 1, \dots, J$, $k = 1, \dots, K$ and $l = 1, \dots, L$.

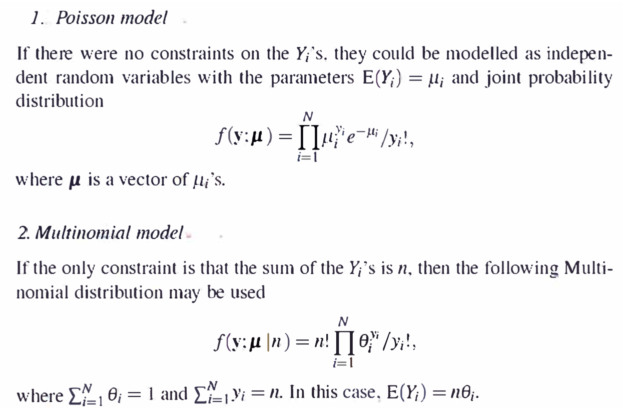
If the marginal totals $y_{jk.}$ are fixed, the joint probability distribution for the Y_{jkl}'s is

$$f(\mathbf{y}\,|\,y_{11.}, \dots, y_{JK.}) = \prod_{j=1}^{J} \prod_{k=1}^{K} y_{jk.}! \prod_{l=1}^{L} \theta_{jkl}^{y_{jkl}} / y_{jkl}!,$$

where $\mathbf{y}$ is the vector of Y_{jkl}'s and $\sum_l \theta_{jkl} = 1$ for $j = 1, \dots, J$ and $k = 1, \dots, K$. This is another form of **product multinomial distribution**. In this case, $\mathrm{E}(Y_{jkl}) = \mu_{jkl} = y_{jk.}\theta_{jkl}$, so that

$$\log \mu_{jkl} = \log \mathbf{y}_{jk.} + \log \theta_{jkl}.$$

Probability models for contingency tables



1. Poisson model

If there were no constraints on the Y_i's, they could be modelled as independent random variables with the parameters $\mathrm{E}(Y_i) = \mu_i$ and joint probability distribution

$$f(\mathbf{y}:\boldsymbol{\mu}) = \prod_{i=1}^{N} \mu_i^{y_i} e^{-\mu_i}/y_i!,$$

where $\boldsymbol{\mu}$ is a vector of μ_i's.

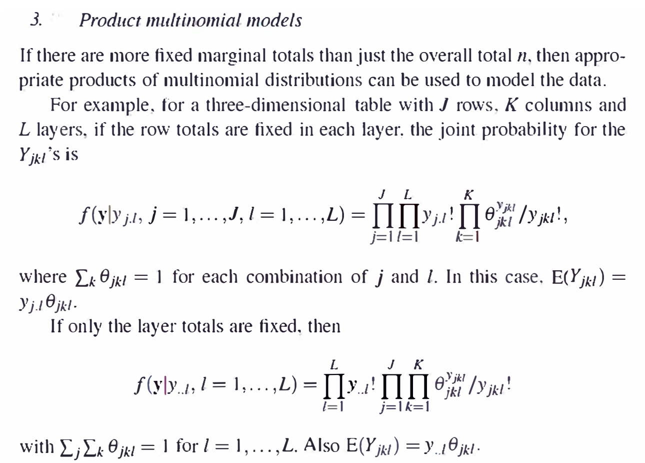
2. Multinomial model

If the only constraint is that the sum of the Y_i's is n, then the following Multinomial distribution may be used

$$f(\mathbf{y}:\boldsymbol{\mu}\,|n) = n!\prod_{i=1}^{N} \theta_i^{y_i}/y_i!,$$

where $\sum_{i=1}^{N}\theta_i = 1$ and $\sum_{i=1}^{N} y_i = n$. In this case, $\mathrm{E}(Y_i) = n\theta_i$.

Probability models for contingency tables



3. Product multinomial models

If there are more fixed marginal totals than just the overall total n, then appropriate products of multinomial distributions can be used to model the data.

For example, for a three-dimensional table with J rows, K columns and L layers, if the row totals are fixed in each layer, the joint probability for the Y_{jkl}'s is

$$f(\mathbf{y}|y_{j.l},\ j=1,\ldots,J,\ l=1,\ldots,L)=\prod_{j=1}^{J}\prod_{l=1}^{L}y_{j.l}!\prod_{k=1}^{K}\theta_{jkl}^{y_{jkl}}/y_{jkl}!,$$

where $\sum_k \theta_{jkl}=1$ for each combination of j and l. In this case, $\mathrm{E}(Y_{jkl})=y_{j.l}\theta_{jkl}$.

If only the layer totals are fixed, then

$$f(\mathbf{y}|y_{..l},\ l=1,\ldots,L)=\prod_{l=1}^{L}y_{..l}!\prod_{j=1}^{J}\prod_{k=1}^{K}\theta_{jkl}^{y_{jkl}}/y_{jkl}!$$

with $\sum_j\sum_k \theta_{jkl}=1$ for $l=1,\ldots,L$. Also $\mathrm{E}(Y_{jkl})=y_{..l}\theta_{jkl}$.

Log-linear models

The natural link function for the Poisson distribution, the logarithmic function, yields a linear component

$$\log E(Y_i) = C + \mathbf{x}_i^T \beta.$$

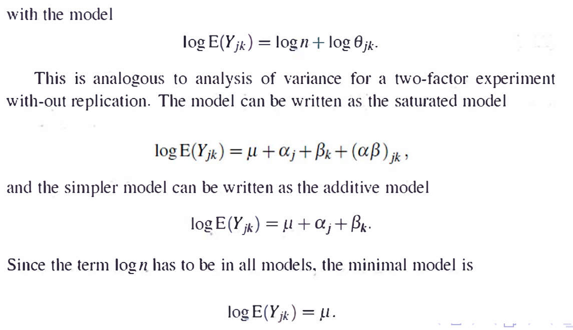
The term log-linear model is used to describe all these generalized linear models.

Melanoma: if there are no associations between site and type of tumor so that these two variables are independent, their joint probability θ_{jk} is the product of the marginal probabilities

$$\theta_{jk} = \theta_{j\cdot}\theta_{\cdot k}.$$

The hypothesis of independence can be tested by comparing the additive model

$$\log E(Y_{jk}) = \log n + \log \theta_{j\cdot} + \log \theta_{\cdot k}$$



with the model

$$\log \mathrm{E}(Y_{jk}) = \log n + \log \theta_{jk}.$$

This is analogous to analysis of variance for a two-factor experiment with-out replication. The model can be written as the saturated model

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk},$$

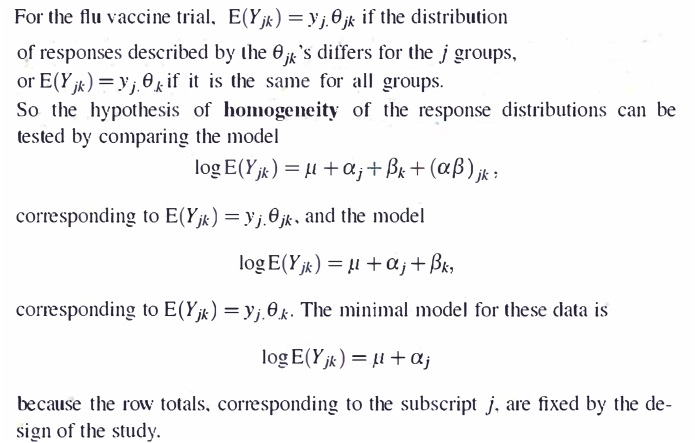
and the simpler model can be written as the additive model

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j + \beta_k.$$

Since the term $\log n$ has to be in all models, the minimal model is

$$\log \mathrm{E}(Y_{jk}) = \mu.$$

Example: Flu vaccine trial



For the flu vaccine trial, $\mathrm{E}(Y_{jk}) = y_{j.}\theta_{jk}$ if the distribution of responses described by the θ_{jk}'s differs for the j groups, or $\mathrm{E}(Y_{jk}) = y_{j.}\theta_{.k}$ if it is the same for all groups. So the hypothesis of **homogeneity** of the response distributions can be tested by comparing the model

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk},$$

corresponding to $\mathrm{E}(Y_{jk}) = y_{j.}\theta_{jk}$, and the model

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j + \beta_k,$$

corresponding to $\mathrm{E}(Y_{jk}) = y_{j.}\theta_{.k}$. The minimal model for these data is

$$\log \mathrm{E}(Y_{jk}) = \mu + \alpha_j$$

because the row totals, corresponding to the subscript j, are fixed by the design of the study.

Inference for log-linear models

For any log-linear model the maximum likelihood estimators are the same for all these distributions provided that the parameters which correspond to the fixed marginal totals are always included in the model. **This means that for the purpose of estimation, the Poisson distribution can always be assumed.**

Hypothesis tests can be conducted by comparing the difference in goodness of fit statistics between a general model corresponding to an alternative hypothesis and a nested, simpler model corresponding to a null hypothesis.

Example: Melanoma

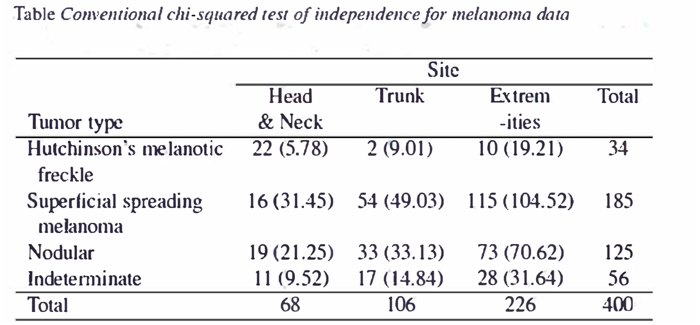


Table *Conventional chi-squared test of independence for melanoma data*

	Site			
	Head	Trunk	Extrem	Total
Tumor type	& Neck		-ities	
Hutchinson's melanotic freckle	22 (5.78)	2 (9.01)	10 (19.21)	34
Superficial spreading melanoma	16 (31.45)	54 (49.03)	115 (104.52)	185
Nodular	19 (21.25)	33 (33.13)	73 (70.62)	125
Indeterminate	11 (9.52)	17 (14.84)	28 (31.64)	56
Total	68	106	226	400

Example: Melanoma

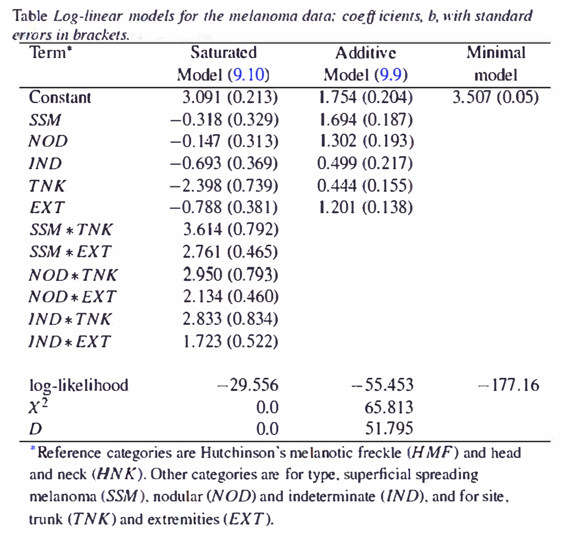
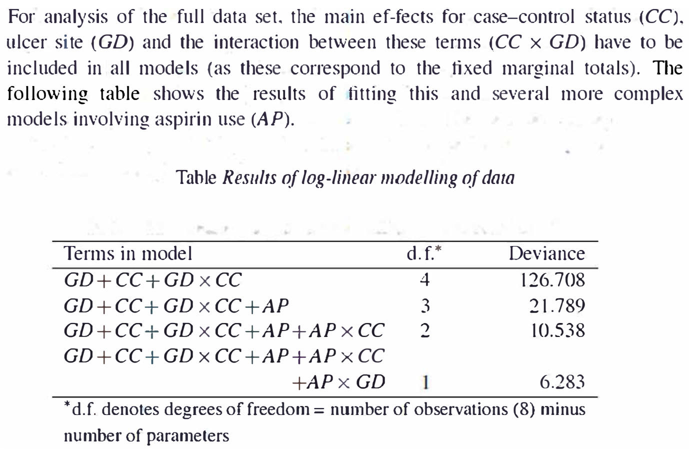


Table *Log-linear models for the melanoma data; coefficients, b, with standard errors in brackets.*

Term*	Saturated	Additive	Minimal
	Model (9.10)	Model (9.9)	model
Constant	3.091 (0.213)	1.754 (0.204)	3.507 (0.05)
SSM	−0.318 (0.329)	1.694 (0.187)	
NOD	−0.147 (0.313)	1.302 (0.193)	
IND	−0.693 (0.369)	0.499 (0.217)	
TNK	−2.398 (0.739)	0.444 (0.155)	
EXT	−0.788 (0.381)	1.201 (0.138)	
$SSM * TNK$	3.614 (0.792)		
$SSM * EXT$	2.761 (0.465)		
$NOD * TNK$	2.950 (0.793)		
$NOD * EXT$	2.134 (0.460)		
$IND * TNK$	2.833 (0.834)		
$IND * EXT$	1.723 (0.522)		
log-likelihood	−29.556	−55.453	−177.16
X^2	0.0	65.813	
D	0.0	51.795	

*Reference categories are Hutchinson's melanotic freckle (HMF) and head and neck (HNK). Other categories are for type, superficial spreading melanoma (SSM), nodular (NOD) and indeterminate (IND), and for site, trunk (TNK) and extremities (EXT).

Example:Case-control study



For analysis of the full data set, the main ef-fects for case–control status (CC), ulcer site (GD) and the interaction between these terms ($CC \times GD$) have to be included in all models (as these correspond to the fixed marginal totals). The following table shows the results of fitting this and several more complex models involving aspirin use (AP).

Table *Results of log-linear modelling of data*

Terms in model	d.f.*	Deviance
$GD+CC+GD\times CC$	4	126.708
$GD+CC+GD\times CC+AP$	3	21.789
$GD+CC+GD\times CC+AP+AP\times CC$	2	10.538
$GD+CC+GD\times CC+AP+AP\times CC$ $+AP\times GD$	1	6.283

*d.f. denotes degrees of freedom = number of observations (8) minus number of parameters

Example:Case-control study

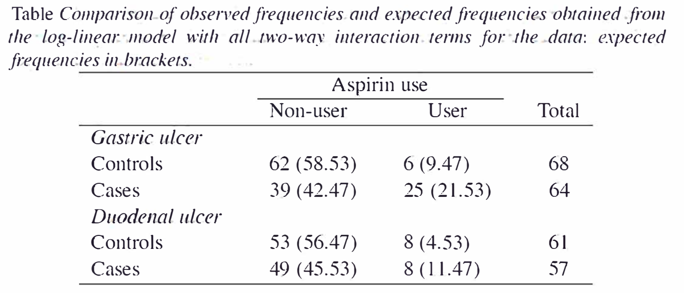


Table *Comparison of observed frequencies and expected frequencies obtained from the log-linear model with all two-way interaction terms for the data: expected frequencies in brackets.*

	Aspirin use		
	Non-user	User	Total
Gastric ulcer			
Controls	62 (58.53)	6 (9.47)	68
Cases	39 (42.47)	25 (21.53)	64
Duodenal ulcer			
Controls	53 (56.47)	8 (4.53)	61
Cases	49 (45.53)	8 (11.47)	57