Count data: Poison Models

Jan van den Broek

2020

1 Introduction

Often the measure of interest in a research is a count of some kind. This might be the number of cells of a certain type in an amount of tissue or it might be the number of events in a certain interval. The following table gives the number of lung cancer cases in 4 Danish cities between 1968 and 1971.

	City			
Age	City 1	City 2	City 3	City 4
40-54	11	13	4	5
55-59	11	6	8	7
60-64	11	15	7	10
65-69	10	10	11	14
70-74	11	12	9	8
>75	10	2	12	7

These counts have in principle no upper limit. Of course every city has a certain population size in every age group, but this size is considered to be very large as compared to the number of cases.

2 The likelihood

In order to be able to write down the likelihood a probability distribution for these counts is needed. Let's consider just the 24 counts and disregard the city-groups and the age-groups for a while. For count data one usually takes the Poisson distribution as the probability distribution. Call the count variable Y and the observed outcome y. So the first of the 24 counts can be represented by Y_1 and its observation by y_1. For the i^{th} count this is Y_i with observation y_i. The Poisson probability of observing y_i for the i^{th} count is

$$P(Y_i = y_i) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}$$

The parameter μ is the population mean of the counts. This means that if we had a large population of counts the mean of all these counts would be μ. The variance of all these counts would also be μ. So, with a Poisson distribution the mean and the variance are the same. There seems to be some logic in that: if counts have a small mean, the variance cannot be large since counts cannot be smaller than zero. For counts with a large mean, say 200, some relatively small values or large values for the counts might occur and thus the variance can be large.

One can now use the Poisson distribution to calculate the probability of the outcomes. If for instance the population mean is known to be 10 then the probability of observing no cases is $e^{-10} = 4.5 \times 10^{-5}$and that of observing 5 cases is $\frac{e^{-10}10^5}{5!} = 0.04$

So, for every of the 24 observations one can write down the probability of observing that observation. The product of these probabilities is the likelihood, which depends on the population mean μ, so

$$L(\mu) = \frac{e^{-\mu}\mu^{y_1}}{y_1!} \cdot \frac{e^{-\mu}\mu^{y_2}}{y_2!} \cdot \ldots \cdot \frac{e^{-\mu}\mu^{y_{24}}}{y_{24}!}.$$

and the log-likelihood is

$$l(\mu) = \sum_{i=1}^{24} \left[-\mu + y_i \ln(\mu) - \ln(y_i!)\right]$$

That value for μ that maximizes the log-likelihood (and thus the likelihood) is taken as an estimate for μ and is called the maximum likelihood estimate, let's say m, which in this case is the mean of the 24 counts. A measure for the peakedness of the log-likelihood is the second derivative at its maximum: $l''(m)$. For a flat function the second derivative is small, for a peaked function it is large. The amount of information is found by calculating the reverse of the second derivative:

$$Information = -l''(\hat{\mu})$$

From this the standard error can be calculated:

$$standard\,error = \sqrt{\frac{1}{Information}}$$

So, as was explained before, the first derivative gives the maximum likelihood estimator, that value of the parameter for which the log-likelihood, and thus the likelihood is maximal. The second derivative gives the standard error.

3 Evidence

Now the fact that the observations come from 4 different cities needs to be modeled. A linear model for the 4 city groups is $\beta_0 + \beta_1 city2 + \beta_2 city3 + \beta_3 city4$, where $city2$ is a indicator variable having value zero everywhere, except for city 2. $city3$ and $city4$ only have ones for city 3 and city 4. This is just a linear model for the one way anova case using indicator variables. This linear model is used not for the population mean but for the logarithm of the population mean.

$$\ln(\mu) = \beta_0 + \beta_1 city2 + \beta_2 city3 + \beta_3 city4$$

This model can also be written as a model for μ: $\mu = e^{\beta_0 + \beta_1 city2 + \beta_2 city3 + \beta_3 city4}$ giving always positive means wich makes sense for counts.

This model for μ can be plugged in in the log-likelihood. The first derivatives w.r.t. the β's give the maximum likelihood estimates b_0, b_1, b_2 and b_3.

The interpretation of these b's can be seen as follows. For city 1 all the indicator variables are zero so the model for city 1 is $\ln(m_1) = b_0$. So $m_1 = e^{b_0}$ is the estimated mean count for city 1. To see what the interpretations of the other b's are, consider city 4 as an example. For city 4 only the indicator for city 4 is one, so the model for city 4 is: $\ln(m_4) = b_0 + b_3$. Subtracting gives $ln(m_4) - ln(m_1) = \ln\left(\frac{m_4}{m_1}\right) = b_0 + b_3 - b_0 = b_3$. Thus e^{b_3} shows how much larger m_4 is as compared to m_1. It is a ratio. This means that b_3 is the log-ratio comparing the mean of city 4 with that of city 1.

In order to see if the mean number of cases in some city is higher as compared to other cities, one can fit a model with the city indicators in it (Model 1) and one without the cities (Model 0) and then compare these models.

Model 0 : The model is $\ln(\mu) = \beta_0$ or $\mu = e^{\beta_0}$ This model for μ is used in the log-likelihood. The derivative w.r.t. β_0 gives the maximum likelihood estimator b_0 for β_0, the second derivative gives the standard error. The maximum of the log-likelihood, l_0, is obtained by plugging in the value of the above maximum likelihood estimate. So for μ in the log-likelihood take e^{b_0}, the fitted value for μ. The maximum for the likelihood is then $L_0 = e^{l_0}$. L_0 gives the probability of the data using model 0.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.23359    0.06682   33.43   <2e-16 ***
```

Model 1 : The model is $\ln(\mu) = \beta_0 + \beta_1 city2 + \beta_2 city3 + \beta_3 city4$ or $\mu = e^{\beta_0 + \beta_1 city2 + \beta_2 city3 + \beta_3 city4}$. Plug in this value for μ in the log-likelihood. The first derivatives w.r.t the β's gives the maximum likelihood estimators for the β's, the b's, and the second derivatives gives the standard

errors for these estimates. The maximum of the log-likelihood, l_1, is obtained by plugging in the maximum likelihood estimates, so by plugging in the fitted values. The maximum for the likelihood then is $L_1 = e^{l_1}$. L_1 gives the probability of the data using model 1.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.36712    0.12500  18.937   <2e-16 ***
factor(city)2 -0.09844    0.18129  -0.543    0.587
factor(city)3 -0.22706    0.18770  -1.210    0.226
factor(city)4 -0.22706    0.18770  -1.210    0.226
```

If L_1 is larger than L_0 then the observed data is more likely using model 1, this model makes the data more probable. In that case (under model 1), one better understands why the data is observed as it is. In order to see how more likely the data is using model 1 as compared to model 0 the likelihood ratio can be calculated: $\frac{L_1}{L_0}$.

To incorporate the number of parameters a model is using one can use Akaike's information criterion (AIC): $AIC = -2\cdot$(log-likelihood)$+2\cdot$(number of parameters in the model) $= -2\cdot l + 2\cdot p$ where p is the number of parameters in the model. That model is best which has the largest likelihood and thus the largest log-likelihood. So that model is best that has the lowest AIC as compared to the others. If the difference in AIC between models is small then the model with the smallest number of parameters is chosen. One then can say that there is not much evidence in the data to keep these variables in the model. This principle is known as Occam's Razor (William Occam, 1300-1349). This principle states roughly that one should keep things as simple as possible. As a rough guide a difference in AIC's is considered small if it is smaller than 2.

As discussed before, the AIC is not a measure of how good a model fits the data, it is a measure of how good the model fits the data as compared to the other models that are fitted.

One can calculate the likelihood ratio statistic : $2 \cdot \ln\left(\frac{L_1}{L_0}\right) = 2(l_1 - l_0)$. This statistic has approximately a chi-squared distribution with *df* degrees of freedom where *df* is the difference of the number of parameters between the models. The AIC values and the likelihood ratio test are in the output below.

```
Model:
cases ~ factor(city)
             Df Deviance     AIC     LRT Pr(Chi)
<none>            25.621 130.009
factor(city)  3   27.704 126.092   2.083  0.5553
```

One can now go on and fit a model containing city groups and age groups. The age group part in the model consists of five indicator variables with their β's. One can then use AIC or likelihood ratio test statistics to see if the model can be reduced. If this model is compared with a model with only city groups in it, one then looks at the effect of age groups given that the city groups are in the model (see the line in the output below that starts with factor(age)). One can also use AIC or the likelihood ratio test to see whether the city groups are needed in the model, given that the age groups are in the model (by looking at the line in the output below that starts with factor(city)).

```
Model:
cases ~ factor(city) + factor(age)
             Df Deviance     AIC     LRT Pr(Chi)
<none>            20.673 135.061
factor(city)  3   22.756 131.144   2.083  0.5553
factor(age)   5   25.621 130.009   4.948  0.4223
```

4 Profile confidence intervals

Also here we can use the profile (log-) likelihood to calculate confidence intervals. As described above, a way to look at the (log-)likelihood as a function of one parameter, e.g. a log ratio β (i.e. the parameter for city 2), is to focus on this parameter by holding it constant. Then for each such a constant value of this parameter the log likelihood is maximized over all other parameters. For every value of the fixed parameter, a maximum value for the log-likelihood is obtained. To plot such a (log-)likelihood just put the different values of the parameter (β) on the horizontal axis and the value of the maximized (log-)likelihood on the vertical axis. This (log-)likelihood is called the profile (log-)likelihood.

This profile likelihood is used to determine confidence intervals that are based on the likelihood ratio test. The likelihood ratio test statistic was $2 \cdot \ln\left(\frac{L_1}{L_0}\right) = 2(l_1 - l_0)$. l_1 is the maximum value of the log-likelihood if the alternative hypothesis is true. It is the maximum value when the log likelihood is maximized over all the parameters in the model. This maximum value can be denoted as l_1.

l_0 is the maximum value of the log-likelihood if the null-hypothesis is true, that is it is the maximum value for the log-likelihood for a specific value for β. This maximum value of the log-likelihood can be denoted as l_0 and can thus be seen as the maximum value of the log-likelihood under the null-hypothesis.

The likelihood ratio test statistic has approximately a chi-squared distribution with one degree of freedom. So the likelihood ratio test with a

significance level of 0.05 rejects the null-hypothesis if the likelihood ratio test statistic ($2(l_1 - l_0)$) is larger then $\chi^2_{0.95}$ which is in the case of 1 degree of freedom equal to 3.84.

For all those values for β for which the likelihood ratio test statistic is smaller than 3.84, the conclusion is that the null-hypothesis is not rejected. This is an interpretation of a 95% confidence interval: all those values for the parameter of interest that, when put in the null-hypothesis would lead to not rejecting the null-hypothesis. This determines the confidence interval, see section 5 of Binary Data: Logistic Regression Models.

An other interpretation for a 95% confidence interval is: The probability that this interval contains the population value of the parameter is 0.95.

5 Model checking

Suppose a model with the city groups is fitted. Every observation has its own fitted value using this model. This fitted value for individual count i can be calculated by $m_i = e^{b_0+b_1 city2+b_2 city3+b_3 city4}$ and plugging in the values for the city indicators for that individual count. In the formula above m_i now indicates the fitted value for individual count i. The contribution for this observation to the maximum value of the likelihood is: $\frac{e^{-m_i} m_i^{y_i}}{y_i!}$. In order to see how far the fitted value for this individual count is from the observation of this individual count one calculates the likelihood contribution when the fitted value is replaced by the observation: $\frac{e^{-y_i} y_i^{y_i}}{y_i!}$. The contribution for this individual count to the deviance is $D(y_i) = 2 \cdot \left[\ln\left(\frac{e^{-y_i} y_i^{y_i}}{y_i!}\right) - \ln\left(\frac{e^{m_i} m_i^{y_i}}{y_i!}\right) \right] = 2 \cdot \left[y_i \ln \frac{m_i}{y_i} + (m_i - y_i) \right]$. If this amount is large then this individual count contributes a large amount to the deviance and, if there are more of these counts, to a bad fit. If this amount is small then this individual count contributes a small amount to the deviance, implying the model fits this individuals count observation.

As with binary data the sum of all these individual contributions is the deviance for the model used here. The deviance residuals are defined as the square root from the individual deviance contributions multiplied with a plus or a minus sign depending on whether or not the observation are larger or smaller than the fitted values:

$$res_{dev} = sign(y_i - m_i) \cdot \sqrt{D(y_i)}$$

A large residual means that this individual contributes a large individual deviance, thus contributing to a bad fit. Or, to put it differently: a large deviance residual means that the difference between the fitted value and the observation in terms of log-likelihoods is large for this individual.

One can now make a plot of the deviance residuals and the fitted values

and then check whether or not there are large residuals, and for these, try to find out what is going on there.

6 Modeling risks and rates

6.1 Population size: risk

In the lung cancer research every city has a certain population size in every age group, but this size was considered to be very large as compared to the number of cases. The population sizes of the age groups were:

	City			
Age	City 1	City 2	City 3	City 4
40-54	3059	2879	3142	2520
55-59	800	1083	1050	878
60-64	710	923	895	839
65-69	581	834	702	631
70-74	509	634	535	539
>75	605	782	659	619

Instead of assuming that the population sizes are large enough, one can try to get them in the model. One way to do that is to model the mean number of cases per individual in the population, that is: model $\frac{\mu}{population\ size}$. So, now a risk is modeled: the mean number of cases per individual. The model then becomes $\ln\left(\frac{\mu}{population\ size}\right) = \beta_0 + \beta_1 age + \beta_2 city2 + \cdots$ in the short notation mentioned above. This is the same as $\ln(\mu) = \beta_0 + \ln(population\ size) + \beta_1 age + \beta_2 city2 + \cdots$. That is, $ln(population\ size)$ is in the model without a β-parameter. To put it differently, the β for the variable $ln(population\ size)$ is taken to be one. One calls a variable like this an offset variable. Putting this in the model might be important if the population sizes differ a lot among the different groups.

6.2 Counts per period: rate

A common feature of count data is a period of observation over which these counts occur. In the urinary track disease example for each patient the number of urinary track episodes was determined. Besides this the follow-up time for each patient was recorded. An individual with a long follow-up time might have more chance of showing the event of interest as compared to a patient with a shorter follow-up time. Using $ln(followuptime)$ as an offset in the model means that the mean number of events per time unit is modeled, which is a rate: $\ln(\mu) = \beta_0 + \ln(followup) + \cdots$ or $\ln(\frac{\mu}{followup}) = \beta_0 +$

Sometimes the data can be more aggregated. Suppose that in a research one is interested in the age effect on the number of deaths due to a certain

cause. Suppose further that age is a factor with say 5 levels. In the first age group there were 3 subjects of which one died. The first subject was observed 3 weeks, the second for 5 weeks and the third for 7 weeks. Then there are 15 person-weeks observed. This is an example of what is called person-time. Person-time shows how long several person were at risk.

The death rate then is $\frac{1}{15} = 0.067$ per person-week. Because this is the number of new cases per person-time this is also called the incidence rate.

7 Frequencies: Log-linear (Poisson) models[1]

In the years 1995 until 1998 a research was done among 1243 Dalmatian pups. It was determined whether or not they were deaf in at least one ear. The research question was if deafness was related to pigmentation. In order to answer this question it was measured whether or not there were many spots on the skin, whether or not the pup had a spot on the head and whether or not the pup had blue eyes.

Here, as an example, we look at the spot measurement and will treat it as the variable of interest, that is to say as the dependent variable. The number of spots was measured in three classes: light, moderate and heavy. A research question might be: are the number of spots different between males and females. First the data:

	Spots			
Gender	light	moderate	severe	total
male	86	458	75	619
female	105	468	51	624
total	191	926	126	1243

This table is called a frequency table. The six numbers in the table are the frequencies. These frequencies count the number of males with light spots (86), the number of males with moderate spots etc. So these frequencies can be seen as counts and thus the Poisson distribution might be taken as the probability distribution for the frequencies and the model used above for the count can be used for the frequencies. In the literature this is called the log-linear model.

But, there is more to it than that.

7.1 Log-linear models

The dependent variable in this example is the number of spots divided in three classes. It's a variable with three possible outcomes. Such a variable is called a categorical variable. Binary variables are also categorical variables,

[1]This section may be skipped

but with only two possible outcomes. If the categories of a categorical variable cannot be ordered then it is called a nominal variable. If these categories can be ordered then the variable is called ordinal. So the variable spots in this research is an ordinal variable: light is less than moderate which in turn is less than heavy. For the time being let's treat this variable as a nominal one, so let's forget about the ordering.

The question is: are the number of spots measured in three classes different for males as compared to females. This is a main effect question: is the outcome of the dependent variable (spots) different for males as compared to females? What is the effect of sex on the outcome variable spots? No instead of looking at the dependent variable spots we take the frequency in the spot-sex classes as the dependent variable. If one looks at the frequencies as the dependent variable, the spot effect translates to: are the frequencies for the three spot-classes the same for males and for females? To put it differently: are the spot-category effects with respect to the frequencies, the log-rates, the same for males and for females? This is a question about an interaction. If the frequencies are taken as the dependent variable, where it actually is the spots variable, then the question if the males in general have a different outcome for the categorical variable as compared to the females translates to the question whether the log-rates for the categories depend on gender. In other words, a main effect question for the categorical variable translates to an interaction question for the frequencies. If the categorical variable depends on gender then the distribution of the frequencies over the three categories should depend on gender and thus the frequencies for the spot categories should be different for male as compared to females.

The frequency for row i and column j is denoted as y_{ij}. There are six of these frequencies. They are assumed to have a Poisson distribution with mean μ_{ij}. As with counts a linear model is taken for the logarithm of the mean: $\ln(\mu_{ij}) = \beta_0 + SPOT + GENDER + SPOT \cdot GENDER$. The term $SPOT \cdot GENDER$ is the interaction. It measures whether the log-rates between the categories depend on gender.

This model uses $1 + 1 + 2 + 2 = 6$ degrees of freedom, the same as the number of observations. So six parameters are used to fit six data points. This means that the model will fit the data exactly. Such a model is called a saturated model. Fitting this model gives:

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)            4.4543     0.1078  41.308   <2e-16 ***
gender                 0.1996     0.1454   1.373   0.1699
factor(spot)2          1.6725     0.1175  14.232   <2e-16 ***
factor(spot)3         -0.1369     0.1580  -0.866   0.3864
gender:factor(spot)2  -0.1780     0.1596  -1.115   0.2647
gender:factor(spot)3  -0.5853     0.2326  -2.516   0.0119 *
```

So if gender is male (coded as a zero) then the estimated mean in class 1 is $e^{4.4543} = 86$. The rate between class 2 and 1 is $e^{1.6725} = 5.315$. So for males class 2 occurs 5.315 times more often than class 1. The rate between class 3 and 1 is $e^{-0.1369} = .872$. For males class 3 appears 0.872 less often than class 1. For females the estimated mean for class 1 is $e^{4.4543+0.1996} = 104$. The rate between class 2 and 1 is $e^{1.6725-0.178} = 4.46$. So for females class 2 occurs 4.46 times more often than class 1. The rate between class 3 and 1 is $e^{-0.1369-0.5853} = 0.486$. For females class 3 appear 0.486 less often than class 1. This rate is much lower than for males. This means that for males the class "severe" occurs (relatively) more often.

7.2 conditioning

In the example we are looking at, the data can be obtained by different study designs. The first possibility is that one sample is taken of 1243 dogs and that for each dog two things are measured: the gender and the outcome of the spot variable. Is that case the number of males and females is not known before the data is gathered. It is the outcome of a random process. Another possibility is that two samples are taken, one of males and one of females, and that only one thing is measured: the number of spots. In that case the number of males and females is known in advance and a model should be used that fits the total number of males and females exactly. These total numbers are called the gender marginals. It is also possible that one sample was taken but that one wants to act as if two samples were taken, as if the gender marginals were known in advance. This is called conditioning. One wants to use a model that conditions on the gender marginal, whether they were known in advance or not. This is because the number of males and females is not informative for the question at hand. For the question whether or not males and females differ in number of spots it is not informative how many males and females there are but how they were divided over the spot categories.

Let's consider a log-linear model with only gender in it: $\ln(\mu_{ij}) = \beta_0 + GENDER$. The estimates are:

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 5.329493   0.040193 132.596   <2e-16 ***
gender      0.008045   0.056728   0.142    0.887
```

The three fitted values for the males are $e^{5.3295} = 206.333$ and for the females $e^{5.3295+0.008} = 208$:

	Spots			
Gender	light	moderate	severe	total
male	206.333	206.333	206.333	619
female	208	208	208	624
total	414.333	414.333	414.333	1243

The fitted values for the gender marginals are exactly the same as the observed marginals. This is not true for the spot marginal which seems logical because spot is not in the model: the model does not know about the spot-categories.

This illustrates an important point:
For the variables in the model the fitted marginals are the same as the observed marginals.
If there is a need to condition on the marginals of a certain variable then that variable should be in the model. For a model with only a constant, the fitted value for the total number of observations equals the total number of observations. In a model with one variable the fitted values of the marginals of this variable equal the observed marginals. For a model with an interaction between two, the fitted values for this marginal table equal the observed marginal table. A model with fixed marginals for some variable (the observed marginals are the same as the fitted) is also called a multinomial model because the probability distribution of the frequencies with some fixed marginals is the multinomial distribution.

Summary:

1. If the dependent variable is categorical and you model the frequencies from a frequency table then the effect of other variables is measured by the interaction of these other variables with the dependent categorical variable.

2. If you want to condition on the marginals of a certain variable, then that variable should be in the model.

7.3 Tables in more dimensions

In the Dalmatian research it was also determined whether or not the dogs had blue eyes. One can now make a table with three variables, a table in three dimensions:

```
, , blueeye = 0
gender
spot   0   1
1  81  97
2 448 447
3  75  50
```

```
, , blueeye = 1
gender
spot   0   1
1   5   8
2  10  21
3   0   1
```

One can fit the log-linear model: $\ln(\mu_{ijk}) = \beta_0 + BLUEEYE + GENDER + SPOT + BLUEEYE \cdot SPOT + BLUEEYE \cdot GENDER + GENDER \cdot SPOT + BLUEEYE \cdot GENDER \cdot SPOT$). The interaction $BLUEEYE \cdot SPOT$ shows whether or not the number of spots differs for those with blue eyes as compared to those without. The interaction $GENDER \cdot SPOT$ shows whether the number of spot for the males are different from that of the females. The interaction $BLUEEYE \cdot GENDER \cdot SPOT$ shows whether the effect of having blue eyes on the number of spots depends on the gender. It is the $BLUEEYE \cdot GENDER$-interaction for the dependent categorical variable spot.

The estimates are (be is blue eyes, ge is gender and sp is spots):

```
Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.394e+00  1.111e-01  39.550  < 2e-16 ***
be2         -2.785e+00  4.608e-01  -6.044 1.51e-09 ***
ge2          1.803e-01  1.505e-01   1.198   0.2311
sp2          1.710e+00  1.207e-01  14.166  < 2e-16 ***
sp3         -7.696e-02  1.603e-01  -0.480   0.6310
be2:ge2      2.897e-01  5.896e-01   0.491   0.6231
be2:sp2     -1.017e+00  5.609e-01  -1.814   0.0697 .
be2:sp3     -2.384e+01  4.225e+04  -0.001   0.9995
ge2:sp2     -1.825e-01  1.647e-01  -1.108   0.2678
ge2:sp3     -5.857e-01  2.366e-01  -2.475   0.0133 *
be2:ge2:sp2  4.544e-01  7.069e-01   0.643   0.5203
be2:ge2:sp3  2.242e+01  4.225e+04   0.001   0.9996
```

Note the huge standard deviation of be2:sp3 and be2:ge2:sp3, indicating some estimating difficulties. This is not surprising since the number of observations is very low there. There is only one dog with blue eyes and a severe number of spots.

This model contains the variable blueeyes so the marginal fitted values for blueeyes are the same as the observed ones. The same is true for gender. In this model there is also a conditioning on the marginal gender-blueeyes table, because there is a blueeyes-gender interaction in the model. If one wants to condition on the four numbers (males with and without blue eyes, and females with and without blue eyes), then this interaction should be in the model.

So the only dependent variable in the model is spot. Is the effect of blue

eyes on the number of spots dependent on gender: do we need the blueeyes-gender-spot interaction?

```
Model:
y ~ be + ge + sp + +be:ge + be:sp + ge:sp + be:ge:sp
Df  Deviance    AIC    LRT Pr(Chi)
<none>      4.123e-10 82.910
be:ge:sp  2     1.572 80.481  1.572  0.4558
```

The model without the 3-way interaction has a lower AIC, so we do not need that interaction. We can now fit a model with only the 2-way interactions:

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.40686    0.10914  40.378   <2e-16 ***
be2         -3.02335    0.36301  -8.329   <2e-16 ***
ge2          0.15738    0.14711   1.070   0.2847
sp2          1.69665    0.11848  14.321   <2e-16 ***
sp3         -0.09517    0.15893  -0.599   0.5493
be2:ge2      0.65758    0.32269   2.038   0.0416 *
be2:sp2     -0.72144    0.34150  -2.113   0.0346 *
be2:sp3     -2.12362    1.04566  -2.031   0.0423 *
ge2:sp2     -0.15703    0.16031  -0.980   0.3273
ge2:sp3     -0.54840    0.23349  -2.349   0.0188 *
```

Are there gender and blue eyes main effects on the number of spots?

```
Model:
y ~ be + ge + sp + be:ge + be:sp + ge:sp
Df Deviance    AIC    LRT Pr(Chi)
<none>       1.572 80.481
be:ge   1    5.949 82.858  4.377 0.03642 *
be:sp   2    9.628 84.538  8.057 0.01780 *
ge:sp   2    7.369 82.279  5.798 0.05509 .
```

Note that we do not look at the blue eyes-gender interaction since we are conditioning on the blue eyes gender-marginal table so this interaction is left in the model. The difference in AIC's between leaving everything in the table or deleting the gender-spot interaction is 1.8 (less than 2), so we do not need the gender main effect on the dependent variable (number of spots) although it is close.

Exercises

R-commands:
The Poisson model can be fitted with read in the data and call it ca for instance:

```
fit <- glm(cases~factor(age)+factor(city)+offset(log(pop)),
           family=poisson,data=ca)
```

The estimates and standard errors can be obtained and with

```
summary(fit)
```

The AIC's and the likelihood ratio test are obtained with:

```
drop1(fit,test="Chisq")
```

and

```
confint(fit)
```

gives the profile confidence intervals.

Exercises:

1. (a) To the lung cancer count data from the text, fit the model $\ln(\mu) = \beta_0 + \ln(population\ \ size/1000) + CITY + AGE$. The offset is population size in thousands.
   (b) Make a plot of the deviance residuals against the fitted values and discuss this plot.
   (c) See which terms are needed in the model.
   (d) Give the estimates of the best fitted model and give their profile likelihood intervals and discuss these.

2. The table below gives the number of coronary deaths for smokers and nonsmokers per age group.

Age	Smokes	deaths	Person years
<40	1	32	52407
<40	0	2	18790
41-50	1	104	43248
41-50	0	12	10673
51-60	1	206	28612
51-60	0	28	5710
61-70	1	186	12663
61-70	0	28	2585
>70	1	102	5317
>70	0	31	1462

(a) Read in the data.

(b) Use a poisson model to analyze the data. Use the likelihood ratio tests to see which terms are needed in the model.

(c) Give a careful interpretation of the estimates

(d) Are there age and smoke effects for the log(person years). What can you say about the age effects and the smoke effects on the log(person years)

3. In the data set grouseticks in the lme4 library the number of ticks on the heads of red grouse chicks sampled in the field is recorded. Other variables in this data set are:
INDEX: (factor) chick number (observation level)
TICKS: number of ticks sampled
BROOD: (factor) brood number
HEIGHT: height above sea level (meters)
YEAR: year
LOCATION: (factor) geographic location code
cHEIGHT: centered height, derived from HEIGHT

(a) Fit a model for TICKS with and with HEIGHT and YEAR as fixed effect.

(b) Give the ant-log of the estimates of the YEAR variable and explain them.

(c) Give a plot of the residuals against the fitted values.

4. In an experiment from 1996 the effects of crowding on reproductive properties of a certain species of leave beetle was examined. Cages of fixed size could contain either 1 male and 1 female or 5 males and 5 females. The temp variable contains the temperature which could be either 21 or 24. The TRT measures the crowding which can be either "I" for cages with 1 female and ""G" for cages with 5 females. The variable of interest was the number of eggs (NumEggs). A complicating feature is that in cages with 5 females, there is no easy way to identify which females led a given eggmass. The variable unit can be disregarded. The data is in the file BeetleEggCrowding.txt .