Lecture 8 – Program

1. Data and examples

2. The Poisson distribution

3. Over-dispersion

4. Poisson regression

5. Generalized linear models

Data structure and basic questions

As before the data have the form:

unit	response	covariates
1	y_1	$x_{11} \cdots x_{1p}$
2	y_2	$x_{21} \cdots x_{2p}$
.	.	$\cdots$
.	.	$\cdots$
.	.	$\cdots$
n	y_n	$x_{n1} \cdots x_{np}$

But now the response is now longer measured on a quantitative scale or as a proportion. The typical situation is that the response is a variable counting how many times an event has occurred.

Objective as before: Explain variation in response y by variation in $x_1, \cdots, x_p$

We will first consider the situation without covariates

Examples (no covariates)

Emission of alpha particles

Counts of number of alpha particles emitted form a source in a given time interval.

Rutherford, E. & Geiger, H. (1910) The probability variations in the distribution of alpha particles. *Philosophical Magazine*, 6. series, **20**, 698-704.

Observed and expected frequencies:

No.	0	1	2	3	4	5	6
Observed	57	203	383	525	532	408	273
Expected	54	210	407	525	509	395	255
No.	7	8	9	10	11	12	13+
Observed	139	49	27	10	4	2	0
Expected	141	68	30	11	4	1	1

We will explain below how the expected values are computed

Examples (no covariates), contd.

Horsekick deaths, ammunition accidents and bomb hits

Observed and expected frequencies for three historical sets of data (se BS page 81):

No.	Frequency					
	Horsekick dths		Ammun. acdnt.		Bomb hits	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
0	109	108.7	448	406.9	229	226.7
1	65	66.3	132	189.2	211	211.4
2	22	20.2	42	43.9	93	98.5
3	3	4.1	21	6.8	35	30.6
4	1	0.6	3	0.8	7	7.1
≥ 5			2	0.3	1	1.6
Total	200	199.9	648	647.9	576	575.9

Poisson distribution

A random variable Y is Poisson distributed with parameter λ if

$$\mathrm{P}(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y = 0, 1, 2, \ldots$$

Short: $Y \sim \mathrm{Po}(\lambda)$

We have that:

$$\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$$

The Poisson distribution arises as:

- an approximation to the distribution of $Y \sim \mathrm{bin}(n, p)$ when p is small and n is large ($\lambda = np$)

- from a Poisson process

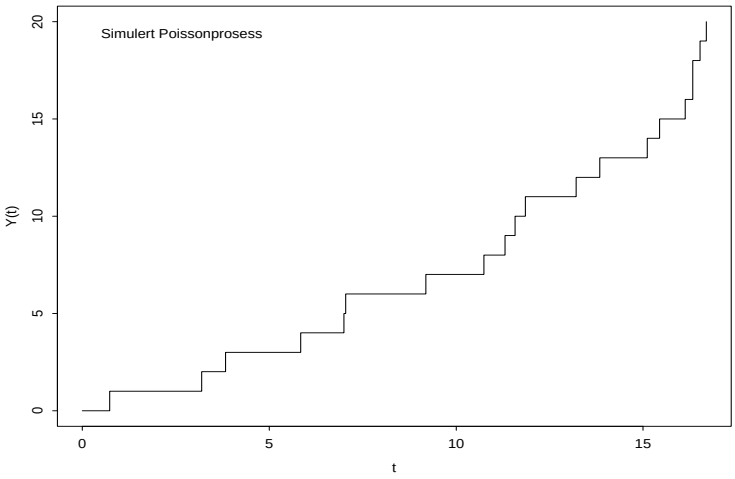
Illustration of Poisson approximation to the binomial distribution

x	Poisson $\lambda = 0.5$	Binomial $n = 500$ $p = 0.001$	Binomial $n = 50$ $p = 0.01$	Binomial $n = 5$ $p = 0.1$
0	0.6065	0.6064	0.6050	0.5905
1	0.3033	0.3035	0.3056	0.3280
2	0.0758	0.0758	0.0756	0.0729
3	0.0126	0.0126	0.0122	0.0081
4	0.0016	0.0016	0.0015	0.0005

The Poisson distribution is often an appropriate model for "rare events"

Poisson process

$Y(t)$ = number of events in $[0, t]$



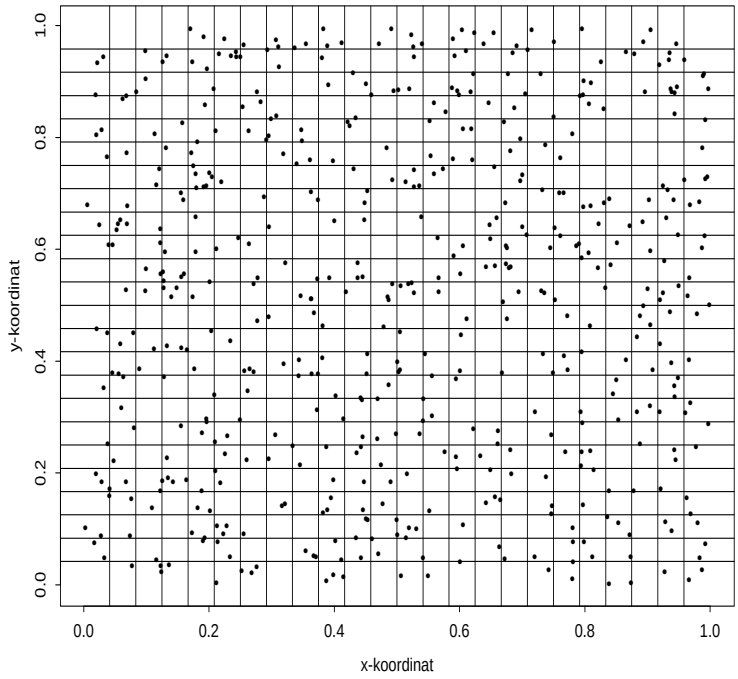
Assume that

- rate of events λ is constant over time (rate = expected number of events per unit of time)
- number of events in disjunct intervals are independent
- events do no occur together

Then $Y(t) \sim \text{Po}(\lambda t)$.

Spatial Poisson process

Points from a spatial Poisson process are "randomly" distributed over an area.



The number of points in a square is Poisson distributed

Is the Poisson distribution appropriate?

For a Poisson distribution the expected value and the variance are equal.

One way of checking whether the Poisson distribution is appropriate is to compare

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{with} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

For a Poisson distribution both $\bar{y}$ and s^2 are estimates of λ, so they should not differ too much.

We may compute the coefficient of dispersion

$$CD = \frac{s^2}{\bar{y}}$$

If CD is (substantially) larger than 1, it is a sign of *over-dispersion*.

Test of Poisson distribution

Data: $y_1, y_2, \ldots, y_n$

Null hypothesis: data are Poisson distributed

Procedure:

- Estimate (MLE): $\widehat{\lambda} = \bar{y}$
- Compute expected frequencies under the null hypothesis: $E_j = n\,(\widehat{\lambda}^j/j!)\exp(-\widehat{\lambda})$
- Compute $O_j =$ observed number of $y_i = j$
- Pearson $X^2 = \sum \frac{(O_j - E_j)^2}{E_j} \sim \chi^2_{K-2}$ under the hypothesis
- Number of groups K such that all $E_j > 5$. Aggregate smaller groups.

Examples

	$\bar{y}$	s^2	CD	K	X^2	p-value
Alpha part.	3.88	3.69	0.95	12	10.42	0.40
Horse kicks	0.610	0.611	1.002	4	0.29	0.86
Ammo-acdnts	0.465	0.691	1.49	4	62.90	0
Bomb hits	0.929	0.936	1.008	5	1.02	0.80

We have over-dispersion for the ammunition accident data.

For the other data sets, the Poisson distribution fits nicely.

Poisson regression

The observed responses y_i are realizations of independent Poisson distributed random variables

$$Y_i \sim \text{Po}(\lambda_i) \quad i = 1, \ldots, n$$

We will consider models of the form:

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip})$$

Interpretation: Suppose individuals 1 and 2:

- Have the same values $x_{2j} = x_{1j}$ for covariates no. $j = 1, \ldots, p-1$
- Differ with one unit on covariate p, i.e. $x_{2p} = x_{1p} + 1$.

Then the *rate ratio* (RR) of individual 2 vs. individual 1 is given by

$$RR = \frac{\lambda_2}{\lambda_1} = \exp(\beta_p)$$

Maximum likelihood estimation

We have

$$P(Y_i = y) = \frac{\lambda_i^y}{y!} \exp(-\lambda_i)$$

The likelihood is the simultaneous distribution of the random variables considered as a function of the parameters (i.e. the β_j s) for the observed y_i values:

$$L = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i)$$

MLE $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$, maximizes L or equivalently the log-likelihood function $l = \log L$.

Approximately: $\widehat{\beta}_j \sim \mathrm{N}(\beta_j, se_j^2)$ where the estimated standard error $\widehat{se}_j$ is computed by the statistical software.

95% c.i. for β_j: $\widehat{\beta}_j \pm 1.96\widehat{se}_j$

95% c.i. for RR_j: $\exp(\widehat{\beta}_j \pm 1.96\widehat{se}_j)$

Wald test statistic for $\mathsf{H}_{0j} : \beta_j = 0$

$$\frac{\widehat{\beta}_j}{\widehat{se}_j} \sim \mathsf{N}(0,1) \text{ under } \mathsf{H}_{0j}$$

Test based on deviance. We will test the null hypothesis H_0 that q of the β_j s are equal to zero. (Equivalently that there are q linear restrictions among the β_j s.)

Procedure:

- $\widehat{l}$ is log-likelihood under the full Poisson regression model
- l^* is log-likelihood under H_0
- $\tilde{l}$ is log-likelihood for saturated model (with $\tilde{\lambda}_i = y_i$)
- Deviances $\hat{D} = 2(\tilde{l} - \widehat{l})$ and $D^\star = 2(\tilde{l} - l^*)$
- Test statistic $G = D^* - \hat{D} = 2(\widehat{l} - l^*) \sim \chi^2_q$ under H_0

Often the following specification is reasonable: $Y_i \sim \text{Po}(T_i\lambda_i)$ where

- $\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$
- T_i is known.

Examples:

- $Y_i \sim \text{bin}(T_i, \lambda_i)$ and λ_i small
  $\Rightarrow Y_i$ approximately $\text{Po}(T_i\lambda_i)$
- $Y_i =$ no. events in a Poisson process with rate λ_i observed over $[0, T_i]$
- $Y_i =$ no. of deaths among persons with rate of death λ_i observed in T_i person-years.

The situation is treated as follows:

We write the model as:

$$\begin{aligned} \mathrm{E}(Y_i) &= T_i \, \lambda_i \\ &= \exp(1 \cdot \log(T_i) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \end{aligned}$$

Formally $\log(T_i)$ is a "covariate" where the regression coefficient is known to equal 1.

Such a "covariate" is called an "offset" (cf. below)

Example: Lung cancer in Denmark

Number of lung cancer cases in four Danish cities from 1968 to 1971

Number of lung cancer cases.

	City				
Age	Fredericia	Horsens	Kolding	Vejle	Total
40–54	11	13	4	5	33
55–59	11	6	8	7	32
60–64	11	15	7	10	43
65–69	10	10	11	14	45
70–74	11	12	9	8	40
> 75	10	2	12	7	31
Total	64	58	51	51	224

Population of the four cities for different age groups.

	City				
Age	Fredericia	Horsens	Kolding	Vejle	Total
40–54	3059	2879	3142	2520	11600
55–59	800	1083	1050	878	3811
60–64	710	923	895	839	3367
65–69	581	834	702	631	2748
70–74	509	634	535	539	2217
> 75	605	782	659	619	2665

Lung cancer in Denmark, contd.

For age group no. i and city no. j let

- y_{ij} = number of lung cancer cases
- T_{ij} = number of persons

A reasonable model is to consider the observed number of lung cancer cases y_{ij} to be realizations of random variables $Y_{ij} \sim \text{Po}(T_{ij}\lambda_{ij})$, where λ_{ij} is given (e.g.) by

$$\lambda_{ij} = \exp(\alpha + \beta_{\text{agegr}(i)} + \gamma_{\text{city}(j)})$$

Lung cancer in Denmark, contd.

```
mod2<-glm(cancer~offset(log(pop))
                 +factor(age)+factor(city),family=poisson)
summary(mod2)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.6321     0.2003 -28.125  < 2e-16 ***
factor(age)2    1.1010     0.2483   4.434 9.23e-06 ***
factor(age)3    1.5186     0.2316   6.556 5.53e-11 ***
factor(age)4    1.7677     0.2294   7.704 1.31e-14 ***
factor(age)5    1.8569     0.2353   7.891 3.00e-15 ***
factor(age)6    1.4197     0.2503   5.672 1.41e-08 ***
factor(city)2  -0.3301     0.1815  -1.818   0.0690 .
factor(city)3  -0.3715     0.1878  -1.978   0.0479 *
factor(city)4  -0.2723     0.1879  -1.450   0.1472
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 129.908  on 23  degrees of freedom
Residual deviance:  23.447  on 15  degrees of freedom
AIC: 137.84

Number of Fisher Scoring iterations: 5
```

Lung cancer in Denmark, contd.

```
mod0<-glm(cancer~offset(log(pop)),family=poisson)
mod1<-glm(cancer~offset(log(pop))+factor(age),
                                    family=poisson)
mod2<-glm(cancer~offset(log(pop))
       +factor(age)+factor(city),family=poisson)
mod3<-glm(cancer~offset(log(pop))
       +factor(age)*factor(city),family=poisson)
anova(mod0,mod1,mod2,mod3, test="Chisq")


Analysis of Deviance Table

Model 1: cancer~offset(log(pop))
Model 2: cancer~offset(log(pop))+factor(age)
Model 3: cancer~offset(log(pop))+factor(age)+factor(city)
Model 4: cancer~offset(log(pop))+factor(age)*factor(city)

  Resid. Df Resid. Dev Df Deviance  P(>|Chi|)
1        23    129.908
2        18     28.307  5  101.601  2.429e-20
3        15     23.447  3    4.859      0.182
4         0 -1.113e-25 15   23.447      0.075
```

There is not a clear effect of city. But there is an indication that the lung cancer risk in Fredericia is larger than in the other cities.

Generalized linear models (GLM)

The models for:

- Multiple linear regression
- Logistic regression
- Poisson regression

are the most common GLMs.

A GLM consists of 3 parts

- A family of distributions
- A link function
- A linear predictor

GLM, contd.

Families of distributions are e.g.

- Normal
- Binomial
- Poisson
- Gamma (incl. exponential distributions)

The linear predictor is a linear expression in regression coefficients and covariates

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Link function

Let $\mu_i = \mathrm{E}(Y_i)$ be the mean of Y_i

The link function g connects the mean μ_i and the linear predictor η_i :

$$g(\mu_i) = \eta_i$$

- Linear regression: $\eta_i = g(\mu_i) = \mu_i$
- Logistic regression: $\eta_i = g(p_i) = \log(\frac{p_i}{1-p_i})$
- Poisson regression: $\eta_i = g(\mu_i) = \log(\mu_i)$

Link function, contd.

Other link functions can be specified.

For binomial responses:

- Complementary log-log link:
  $\eta_i = g(p_i) = \log(-\log(1 - p_i))$

- Probit link: $\eta_i = \Phi^{-1}(p_i)$
  where $\Phi(z) =$ c.d.f. for N(0, 1).

For Poisson responses:

- Identity link: $\eta_i = g(\mu_i) = \mu_i$

- Square root link: $\eta_i = g(\mu_i) = \sqrt{\mu_i}$

Statistical inference in GLM

Estimation:

- Maximum likelihood

Testing and confidence intervals

- Wald test
- Deviance
- More generally: likelihood based