Poisson regression models

If Y_i, $i = 1, \ldots, n$, are counts with means μ_i, the standard Poisson model assumes that $Y_i \sim \text{Pois}(\mu_i)$ with variance function

$$\text{Var}(Y_i) = \mu_i.$$

The canonical link function is the log

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

For different observation periods/areas/volumes:

$$Y_i \sim \text{Pois}(t_i \lambda_i)$$

Taking a log-linear model for the rates,

$$\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

results in the following log-linear model for the Poisson means

$$\log(\mu_i) = \log(t_i \lambda_i) = \log(t_i) + \mathbf{x}_i^T \boldsymbol{\beta},$$

where the $\log(t_i)$ is included as a fixed term, or *offset*, in the model.

Poisson Log-linear Model for Count Data

Example – Storing of micro-organisms

Bacterial concentrations (counts per fixed area) measured at initial freezing (-70^oC) and at 1, 2, 6, 12 months.

Time	0	1	2	6	12
Count	31	26	19	15	20

Hypothesized that

$$\text{average count} \propto \frac{1}{(\text{Time})^{\gamma}}$$

i.e. count decays over time

Model

$$\boxed{\begin{array}{c} \text{Count} \sim \text{Pois}(\mu) \\ \log \mu = \beta_0 + \beta_1 \log(\text{Time}) \end{array}}$$

Avoid problems at time 0 by using

$$\log \mu = \beta_0 + \beta_1 \log(\text{Time} + 0.1)$$

Micro-organism Data Analysis

Model	d.f.	Deviance	X^2
Constant	4	7.0672	7.1532
log(Time)	3	1.8338	1.8203

Analysis of deviance

Source	d.f.	Deviance	p-value
Linear Regression	1	5.2334	0.0222
Error	3	1.8338	
Total	4	7.0672	

$$\log(\mu) = 3.149 - 0.1261 \ \log(\text{time})$$

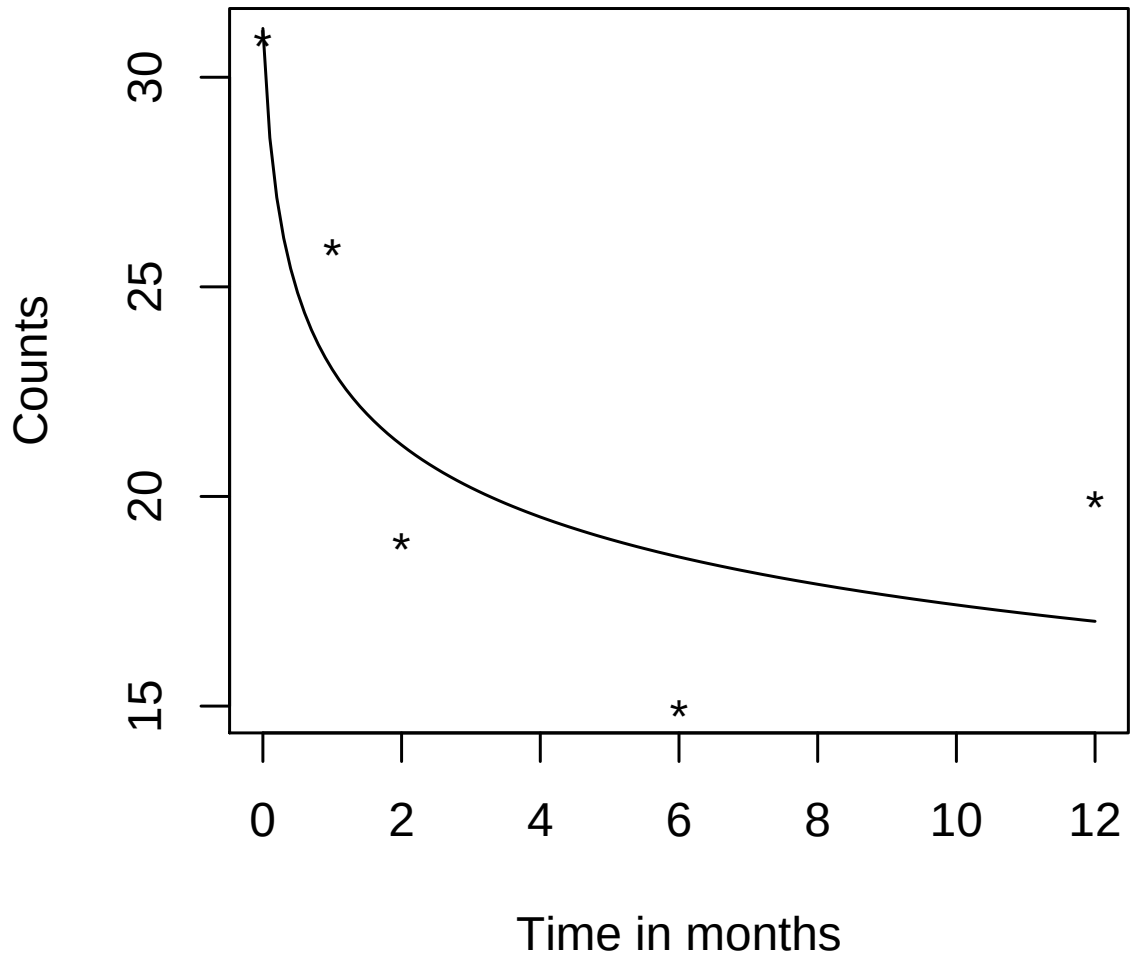

Figure 1: Bacterial Concentration: Fitted Model

Bacterial concentrations – analysis in R

```
 # *** Bacterial concentrations of stored micro-organisms
# Glim4 Manual p531
ltime <- log((tim <- c(0, 1, 2, 6, 12))+ 0.1)
lcount <- log(count <- c(31, 26, 19, 15, 20))

bacteria.dat <- data.frame(tim, count, ltime, lcount)
attach(bacteria.dat)

par(mfrow=c(1,2))
plot(tim, count, xlab="Time in months", ylab="Counts")
plot(ltime,lcount, xlab="Log(time in months)", ylab="Log(counts)")
par(mfrow=c(1,1))

modl<-glm(count ~ tim, family=poisson)
anova(modl, test="Chisq")

mod2<-glm(count ~ ltime, family=poisson)
anova(mod2, test="Chisq")

plot(c(0,12), c(15,31), type="n", xlab="Time in months", ylab="Counts")
points(tim,count,pch="*")
x<-seq(0,12,0.1)
lp<-predict(mod2,data.frame(ltime=log(x+0.1)), type="response")
lines(x,lp,lty=1)
```

Exercise:

Count of the number of plant species on plots that have different biomass (a continuous explanatory variable with three levels: high, mid and low)

Tabela 1: Number of plant species (Y), quantity of biomass (X) and levels of pH of the soil.

pH level	Y	X	Y	X	Y	X	Y	X	Y	X
Low	18	0,1008	15	2,6292	13	0,6526	8	3,6787	9	1,5079
	19	0,1385	9	3,2522	9	1,5553	2	4,8315	8	2,3259
	15	0,8635	3	4,4172	8	1,6716	17	0,2897	12	2,9957
	19	1,2929	2	4,7808	14	2,8700	14	0,0775	14	3,5381
	12	2,4691	18	0,0501	13	2,5107	15	1,4290	7	4,3645
	11	2,3665	19	0,4828	4	3,4976	17	1,1207	3	4,8705
Mid	29	0,1757	30	1,3767	21	2,5510	18	3,0002	13	4,9056
	13	5,3433	9	7,7000	24	0,5536	26	1,9902	26	2,9126
	20	3,2164	21	4,9798	15	5,6587	8	8,1000	31	0,7395
	28	1,5269	18	2,2321	16	3,8852	19	4,6265	20	5,1209
	6	8,3000	25	0,5112	23	1,4782	25	2,9345	22	3,5054
	15	4,6179	11	5,6969	17	6,0930	24	0,7300	27	1,1580
High	30	0,4692	39	1,7308	44	2,0897	35	3,9257	25	4,2667
	29	5,4819	23	6,6846	18	7,5116	19	8,1322	12	9,5721
	39	0,0866	35	1,2369	30	2,5320	30	3,4079	33	4,6050
	20	5,3677	26	6,5608	36	7,2420	18	8,5036	7	9,3909
	39	0,7648	39	1,1764	34	2,3251	31	3,2228	24	4,1361
	25	5,1371	20	6,4219	21	7,0655	12	8,7459	11	9,9817

Poisson Models for Rate Data

Data:

Counts	C_i
Exposure Period	E_i
Rate	$R_i = C_i/E_i$
Covariates	$\mathbf{x}_i$

Modelling the Counts

$$\boxed{\begin{array}{c} C_i \sim \text{Pois}(E_i\lambda_i) \\ \log\lambda_i = \mathbf{x}_i^T\boldsymbol{\beta} \end{array}}$$

$$\mathbf{E}[C_i] = \mu_i = E_i\lambda_i$$

so

$$\begin{aligned} \log\mu_i &= \log E_i + \log\lambda_i \\ &= \log E_i + \mathbf{x}_i^T\boldsymbol{\beta} \end{aligned}$$

$\log E_i$ is included as an *offset*

Modelling as a Rate

$$R_i = C_i / E_i$$

$$\mathbf{E}[R_i] = \frac{E_i \lambda_i}{E_i} = \lambda_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}} = \mu_i^R$$

$$\mathsf{Var}(R_i) = \frac{1}{E_i^2}\mathsf{Var}(C_i) = \frac{E_i \lambda_i}{E_i^2} = \frac{\lambda_i}{E_i} = \frac{\mu_i^R}{E_i}$$

Use a weighted Poisson regression with weights E_i.

Worldwide Airline Fatalities, 1976-85

Year	Fatal accidents	Passenger deaths	Passenger miles (100 million)
1976	24	734	3863
1977	25	516	4300
1978	31	754	5027
1979	31	877	5481
1980	22	814	5814
1981	21	362	6033
1982	26	764	5877
1983	20	809	6223
1984	16	223	7433
1985	22	1066	7107

Simple Models

- Passenger miles (m_i) as exposure variable
- Poisson log-linear model
- Linear time trend

$$\boxed{\begin{aligned} Y_i \sim &\quad \text{Pois}(m_i\lambda_i) \\ \log\lambda_i = &\quad \beta_0 + \beta_1 \texttt{year}_i \end{aligned}}$$

Fatal accidents:

Deviance(`time trend`)	=	20.68
Residual Deviance	=	5.46 on 8 d.f.

Passenger deaths:

Deviance(`time trend`)	=	202.1
Residual Deviance	=	1051.5 on 8 d.f.

$\Longrightarrow$ compounding with aircraft size

Dilution assays

Example – These data are counts of virus particles at 5 different dilutions. There are 4 replicate counts at each dilution except the last for which there are 5 counts. The aim is to estimate the number of virus particles per unit volume(Ridout, 1990).

Table 2: Counts of virus particles at 5 different dilutions

Dilution	Counts				
0.3162	13	14	17	22	
0.1778	9	14	6	14	
0.1000	4	4	3	5	
0.0562	3	2	1	3	
0.0316	2	1	3	2	2

Aim of dilution assays: to estimate density of particles of bacteria, viruses, fungi

Suppose

λ: density of particles by unit volume in an initial solution

v_i: volume of the i-th dilution (amount of the original suspension)

$Y_{i,j}$: number of particles from i-th dilution, j-th replicate

$$Y_{ij} \sim \mathsf{P}(\lambda v_i)$$

$\mu_{ij} = \lambda v_i \Rightarrow \log(\mu_{ij}) = \log \lambda + \log(v_i) \Rightarrow \log(\mu_{ij}) = \beta + \text{offset}$

Then $\hat{\lambda} = \exp(\hat{\beta})$ and a confidence interval for λ is

$$\exp[\hat{\beta} \mp 1.96\text{s.e.}(\hat{\beta})]$$

Virus concentration – analysis in R

```
# Dilution assay
count <- c(2,1,3,2,2, 3,2,1,3, 4,4,3,5, 9,14,6,14, 13,14,17,22)
dilui <- c(.0316,.0316,.0316,.0316,.0316, .0562,.0562,.0562,.0562, .1,.1,.1,.1,
.1778,.1778,.1778,.1778, .3162,.3162,.3162,.3162)
dilui.dat <- data.frame(dilui, count)
attach(dilui.dat)
plot(dilui, count, xlab="Dilution", ylab="Counts")

# Fitting
dilui.fit1<-glm(count ~ 1 + offset(log(dilui)), family=poisson)
anova(dilui.fit1, test="Chisq")
1-pchisq(deviance(dilui.fit1), df.residual(dilui.fit1))

# estimated density  of  particles of virus
# and confidence interval
conc <- exp(dilui.fit1$coef[1])
InfL <- exp(dilui.fit1$coef[1] - 1.96*summary(dilui.fit1)$coef[2])
InfL
SupL <- exp(dilui.fit1$coef[1] + 1.96*summary(dilui.fit1)$coef[2])
SupL

# Plotting
# Plotting
plot(c(0,0.32), c(0,25), type="n", xlab="Dilution", ylab="Counts")
points(dilui,count,pch="*")
x<-seq(0.01,0.32,0.01)
lp<-exp(predict(dilui.fit1,data.frame(dilui=x)))
lines(x,lp,lty=1)
```

2-way Contingency Table

Habitat study of lizards

		Perch Diameter (in)		
		≤ 4.0	> 4.0	
Perch	≥ 4.75	61	41	102
Height	< 4.75	73	70	143
		134	111	245

Are diameter and height classifications independent?

Association measured by odds-ratio

Independence $\Longrightarrow$ odds-ratio $= 1$

$$\text{Observed Odds-ratio} = \frac{61 \times 70}{41 \times 73} = 1.43$$

Is this significantly different from 1?

Log-linear models for 2×2 table

B

A	1	2	
1	m_{11}	m_{12}	$m_{1.}$
2	m_{21}	m_{22}	$m_{2.}$
	$m_{.1}$	$m_{.2}$	$m_{..}$

Model Counts m_{ij} as Poisson random variables with A and B as explanatory factors, log-link gives log-linear model.

In general, marginal distributions of A and B are not of interest.

Two models of interest:

(i) A+B (independence model)

(ii) A∗B ≡ A+B+A.B (saturated model).

(i) A+B (independence model)

$$\log \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B \qquad i, j = 1, 2$$

with $\lambda_1^A = \lambda_1^B = 0$, i.e linear preditor

	B	
A	1	2
1	λ	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B$

$$\hat{\lambda} = \log(\frac{m_{1.}m_{.1}}{m_{..}})$$
$$\hat{\lambda}_2^A = \log(\frac{m_{2.}}{m_{1.}})$$
$$\hat{\lambda}_2^B = \log(\frac{m_{.2}}{m_{.1}})$$

Fitted log-odds-ratio

$\log \psi =$
$(\lambda + \lambda_2^A + \lambda_2^B) + \lambda - (\lambda + \lambda_2^B) - (\lambda + \lambda_2^A) = 0,$
i.e, odds ratio $\psi = 1 \Longrightarrow$ independence model.

- A+B (margins ($\Rightarrow$ total) reproduced)

Linear preditor

	B	
A	1	2
1	λ	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B$

$\hat{\lambda}_2^A = \log(\frac{m_{2.}}{m_{1.}})$ and $\hat{\lambda}_2^B = \log(\frac{m_{.2}}{m_{.1}})$

$$\begin{aligned} m_{..} &= e^{\hat{\lambda}} + e^{\hat{\lambda}+\hat{\lambda}_2^A} + e^{\hat{\lambda}+\hat{\lambda}_2^B} + e^{\hat{\lambda}+\hat{\lambda}_2^A+\hat{\lambda}_2^B} \\ &= e^{\hat{\lambda}}(1+e^{\hat{\lambda}_2^A})(1+e^{\hat{\lambda}_2^B}) \end{aligned}$$

$$\begin{aligned} \log m_{..} &= \hat{\lambda} + \log(1+\frac{m_{2.}}{m_{1.}}) + \log(1+\frac{m_{.2}}{m_{.1}}) \\ &= \hat{\lambda} + \log(\frac{m_{..}}{m_{1.}}) + \log(\frac{m_{..}}{m_{1.}}) \end{aligned}$$

Then,

$\hat{\lambda} = \log(m_{..}\frac{m_{1.}}{m_{..}}\frac{m_{.1}}{m_{..}}) = \log(\frac{m_{1.}m_{.1}}{m_{..}})$

i.e,

$e^{\hat{\lambda}} = \text{Fitted}(1,1) = \hat{\mu}_{11} = m_{..}\frac{m_{1.}}{m_{..}}\frac{m_{.1}}{m_{..}} = np_{1.}p_{.1}$ (usual independence form)

$$\begin{aligned} \text{Fitted}(1,2) = \hat{\mu}_{12} &= e^{\hat{\lambda}+\hat{\lambda}_2^B} = e^{\hat{\lambda}}e^{\hat{\lambda}_2^B} = m_{..}\frac{m_{1.}}{m_{..}}\frac{m_{.1}}{m_{..}}\frac{m_{.2}}{m_{.1}} \\ &= m_{..}\frac{m_{1.}}{m_{..}}\frac{m_{.2}}{m_{..}} = np_{1.}p_{.2} \end{aligned}$$

(ii) A∗B ≡ A+B+A.B (Saturated model)

$$\log \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad i, j = 1, 2$$

with $\lambda_1^A = \lambda_1^B = \lambda_{1j}^{AB} = \lambda_{i1}^{AB} = 0$,
i.e linear preditor

	B	
A	1	2
1	λ	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB}$

i.e $\hat{\lambda} = \log(m_{11})$
$\hat{\lambda}_2^A = \log(\frac{m_{21}}{m_{11}})$
$\hat{\lambda}_2^B = \log(\frac{m_{12}}{m_{11}})$
$\hat{\lambda}_{22}^{AB} = \log(\frac{m_{22}m_{11}}{m_{12}m_{21}}) = \log($observed odds-ratio$)$

Lizard Data Analysis

Model	d.f.	Deviance X^2
Height+Diameter	1	1.8477
Height*Diameter	0	0

Estimates for the interaction model

estimate	s.e.	parameter	
1	4.1111	0.1280	1
2	0.1796	0.1735	Height(2)
3	-0.3973	0.2019	Diameter(2)
4	0.3553	0.2622	Height(2).Diameter(2)

Note that this model reproduces the data and also the fitted log-odds ratio is 0.3553 giving a fitted odds-ratio of

$$exp(0.3553) = 1.427$$

Estimates for the main effects, or independence, model

estimate	s.e.	parameter	
1	4.022	0.1148	1
2	0.3379	0.1296	Height(2)
3	-0.1883	0.1283	Diameter(2)

The fitted odds ratio is now zero and the change in deviance can be used to assess the significance of the null hypothesis that the log-odds ratio is zero, i.e. the odds ratio is 1

change in deviance $= 1.848$ $(p = 0.1740)$

So this change in deviance is not significant, so we cannot reject the hypothesis that the odds-ratio is 1, i.e. diameter and height are independent.

Lizard – analysis in R

```
 # 2 by 2 table - lizard habitat data
y <- c(61, 41, 73, 70)
height <- factor(c(1, 1, 2, 2))
diam <- factor(c(1, 2, 1, 2))

lizard.dat <- data.frame(height, diam, y)
attach(lizard.dat)

# calculate  observed odds ratio
61*70/(41*73)

# now set up log linear model
mod1<-glm(y ~ height*diam, family=poisson)
anova(mod1, test="Chisq")
summary(mod1)

# note that this model reproduces the data
# also the fitted log-odds ratio is 0.3553 giving a
# fitted odds-ratio of

exp(mod1$coef[4])

# the interaction term is not significant, so we
# cannot reject the hypothesis that the odds-ratio is 1,
# i.e. diameter and height are independent.

# refit simplest adequate model
# the main effects, or independence, model
mod2<-glm(y ~ height + diam, family=poisson)
anova(mod2, test="Chisq")
summary(mod2)
```

2-way Contingency Table

Table 3: Números de insetos coletados em armadilhas adesivas e sexados

Armadilha	Machos	Fêmeas	Totais
Alaranjada	246	17	263
Amarela	458	32	490
Totais	704	49	753

Fonte: Silveira Neto *et al.* (1976)

Pergunta: Há influência da cor da armadilha sobre a atração de machos e fêmeas dessa espécie?

Association measured by odds-ratio

Independence $\Longrightarrow$ odds-ratio $= 1$

$$\text{Observed Odds-ratio} = \frac{246 \times 32}{458 \times 17} = 1.01$$

Is this significantly different from 1?

Traps Data Analysis

Model	d.f.	Deviance	X^2
Cor da armadilha + sexo	1	0.001254	0.001252
Cor da armadilha * sexo	0	0	

Estimates for the interaction model

	estimate	s.e.	parameter
1	5.505	0.0638	1
2	0.622	0.0790	armcor(2)
3	-2.672	0.2508	sexo(2)
4	0.011	0.3104	armcor(2).sexo(2)

Note that this model reproduces the data and also the fitted log-odds ratio is 0.01098 giving a fitted odds-ratio of

$$exp(0.01098) = 1.01$$

Estimates for the main effects, or independence, model

	estimate	s.e.	parameter
1	5.505	0.0624	1
2	0.622	0.0764	armcor(2)
3	-2.665	0.1478	sexo(2)

The fitted odds ratio is now zero and the change in deviance can be used to assess the significance of the null hypothesis that the log-odds ratio is zero, i.e. the odds ratio is 1

change in deviance = 0.00125 ($p = 0.9117$)

So this change in deviance is not significant, so we cannot reject the hypothesis that the odds-ratio is 1, i.e. diameter and height are independent.

Traps – analysis in R

```
# 2 by 2 table - Traps y <- c(246, 17, 458, 32) armcor <-
factor(c(1, 1, 2, 2)) sexo <- factor(c(1, 2, 1, 2))

count.dat <- data.frame(armcor, sexo, y) attach(count.dat)

# calculate  observed odds ratio 246*32/(17*458)

# now set up log linear model mod1<-glm(y ~ armcor*sexo,
family=poisson) print(sum(residuals(mod1, 'pearson')^2)) anova(mod1,
test="Chisq") summary(mod1)

# note that this model reproduces the data # also the fitted
log-odds ratio is 0.01098 giving a # fitted odds-ratio of

exp(mod1$coef[4])

# the interaction term is not significant, so we # cannot reject the
hypothesis that the odds-ratio is 1, # i.e. traps colour and sex
are independent.

# refit simplest adequate model # the main effects, or independence,
model mod2<-glm(y ~ armcor+sexo, family=poisson)
print(sum(residuals(mod1, 'pearson')^2)) anova(mod2, test="Chisq")
1-pchisq(deviance(mod2), df.residual(mod2)) summary(mod2)
```

Simple 3-way Table

Random sample of diabetic patients classified by

- family history
- insulin dependence
- age at onset

		Family history of diabetes			
		Yes		No	
		Insulin Dep		Insulin Dep	
		Yes	No	Yes	No
Age at	< 45	6	1	16	2
Onset	≥ 45	6	36	8	48

Interested in the inter-relationship between the three classifying variables.

Model Fitting

Model	Scaled Deviance	df
A+H+I	51.93	4
A*I+H	1.85	3
A*H+I	50.03	3
I*H+A	51.02	3
A*I+I*H	1.04	2
A*I+A*H	0.05	2
I*H+A*H	49.12	2
A*I+I*H+A*H	0.04	1
A*I*H	0	0

Conclusions

Age and Insulin are independent of family history $\Longrightarrow$ can collapse table over History

		Insulin Dep		
		Yes	No	Total
Age at	< 45	22	3	25
Onset	≥ 45	14	84	98

$$\text{Odds-ratio} = \frac{(22/25)/(3/25)}{(14/98)/(84/98)} = \frac{22 \times 84}{3 \times 14} = 44$$

i.e. the odds of having insulin in the group younger than 45 years is 44 times that for the group older than 45 years.

Note that odds ratios in the individual tables are 36 and 48.

This collapsing over history does not distort the relationship between Insulin dependence and Age at onset.

Cannot collapse over classifications that have significant interactions – Simpson's Paradox

Study of diabetic patients – analysis in R

```
# Study of diabetic patients
y <- c(6, 1, 16, 2, 6, 36, 8, 48)
age <- factor(c(1, 1, 1, 1, 2, 2, 2, 2))
hist <- factor(c(1, 1, 2, 2, 1, 1, 2, 2))
ins <-  factor(c(1, 2, 1, 2, 1, 2, 1, 2))
lizard.dat <- data.frame(age, hist, ins, y)
attach(lizard.dat)

# now set up the log-linear model
mod1<-glm(y ~ age*hist*ins, family=poisson)
anova(mod1, test="Chisq")
summary(mod1)

# refit simplest adequate model
mod2<-glm(y ~ age + hist + ins + ins:age, family=poisson)
anova(mod2, test="Chisq")
summary(mod2)

# odds-ratio for age and insulin is
exp(mod2$coef[5])

# all other fitted odds-ratios are 1.
```

Log-linear Models for 3-way Tables

Suppose classifying factors are A, B and C.

There are the following classes of models:

1. `A*B*C`
   – completely reproduces response pattern

2. `A*B+B*C+C*A`
   – no 3-factor interaction model

3. `A*B+B*C`
   A and C conditionally independent given B,

   $$A \perp C | B$$

   (a) *fitted odds-ratio for A,B same for all levels of C*
   (b) *fitted odds-ratio for B,C same for all levels of A*
   (c) *fitted odd-ratio for A,C is 1 for all levels of B, but marginal odds ratio is not 1*

4. `A*B+C`
   – joint distribution of A and B is the same for all levels of C $\Rightarrow$ can collapse over C.

5. `A+B+C` – mutual independence model

Binomial Logit & Poisson Log-linear Models

	Response		
Gender	No	Yes	
Male	20	12	32
Female	14	30	44
	34	42	76

Interested in the gender difference for the proportion responding "Yes".

Tabelas com Resposta Binomial

Duas categorias de resposta: sim/não, sucesso/falha etc

O interesse está em saber como a proporção que respondeu SIM depende de outras covariáveis.

Especificação:

1. Dados: $r_{i2}, m_i, \mathbf{x}_i, \quad i = 1, \ldots, n$, onde r_{i2} representa o número de sucessos (SIM), $m_i = r_{i+}$ é o número de indivíduos, e $\mathbf{x}_i$ são variáveis explanatórias (por exemplo, fatores de classificação etc).

2. Modelo probabilístico:

$$R_{i2} \sim \text{Bin}(m_i, \pi_i).$$

3. Função de ligação: logit

$$logit(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = \eta_i = \boldsymbol{\beta}'\mathbf{x}_i$$

Resultados do Modelo Binomial

(i) Modelo minimal = modelo nulo: $\eta_i = logit(\pi_i) = \theta_0$

Deviance = 7,131 com 1g.l.

$$logit(\hat{\pi}) = 0,2113 \Rightarrow \hat{\pi} = 0,5526 = \frac{42}{76}$$

(ii) Modelo para efeito de sexo:
$\eta_i = logit(\pi_i) = \beta_0 + sexo_i$

$logit(\pi_1) = \theta$

$logit(\pi_2) = \theta + \theta_1$

Deviance = 0 com 0g.l.

Parâmetro	Estimativa	Erro padrão
1	-0,5108	0,3651
sexo(2)	1,273	0,4879

$$logit(\hat{\pi}_H) = -0,5108$$

$$\Rightarrow \hat{\pi}_H = 0,375 = \frac{e^-0,5108}{1 + e^-0,5108} = \frac{12}{32}$$

$$logit(\hat{\pi}_M) = -0,5108 + 1,273 = 0,7622$$

$$\Rightarrow \hat{\pi}_M = 0,682 = \frac{30}{44}$$

Tabelas com Resposta Poisson log-linear

Tratar as caselas da tabela como variáveis Poisson independentes, isto é $R_{ij} \sim$ Pois(μ_{ij}), fixados os totais marginais de sexo (reproduz os m_i da binomial) e com classificação cruzada para tipo de resposta.

Especificação:

1. Dados: $r_{ij}, \mathbf{x}_{ij}$, em que r_{ij} representa a contagem na célula ij e $\mathbf{x}_{ij}$ são variáveis explanatórias (por exemplo, fatores de classificação etc).

2. Modelo probabilístico:

$$R_{ij} \sim \mathsf{P}(\mu_{ij}).$$

3. Função de ligação: logarítmica

$$log(\mu_{ij}) = \eta_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij}$$

Resultados do Modelo Poisson

(i) Modelo minimal (reproduz as margens):
$\eta_{ij} = log(\mu_{ij}) = \alpha_0 + sexo_i + resp_j$

	Resposta	
Sexo	1	2
1	α_0	$\alpha_0 + \beta_0$
2	$\alpha_0 + \gamma_0$	$\alpha_0 + \beta_0 + \gamma_0$

$\theta_0 = logit(\pi_1) = \log \mu_{12} - \log \mu_{11} = \beta_0$

$\theta_0 = logit(\pi_2) = \log \mu_{22} - \log \mu_{21} = \beta_0$

Deviance = 7,131 com 1g.l.

(ii) $\eta_{ij} = log(\mu_{ij}) = \alpha_0 + sexo_i + resp_j + sexo.resp_{ij}$

	Resposta	
Sexo	1	2
1	α	$\alpha + \beta$
2	$\alpha + \gamma$	$\alpha + \beta + \gamma + \delta$

$\theta = logit(\pi_1) = \log \mu_{12} - \log \mu_{11} = \beta$

$\theta_1 = logit(\pi_2) - logit(\pi_1) =$
$(\log \mu_{22} - \log \mu_{21}) - (\log \mu_{12} - \log \mu_{11}) = \delta$

Deviance = 0 com 0g.l.

```
# Fitting Binomial Logit model
y1 <- c(12, 30)
m <- c(32, 44)
sex1 <- factor(c(1, 2))
Bin.Pois1 <- data.frame(sex1, m, y1)
attach(Bin.Pois1)

# now define binomial logit model
resp1 <- cbind(y1, m-y1)
modBin <- glm(resp1 ~ sex1, family=binomial)
anova(modBin, test="Chisq")
summary(modBin)

# calculate  observed odds ratio
exp(modBin$coef[2])

# now fit the data using Poisson log-linear models

y2 <- c(20, 14, 12, 30)
resp2 <- factor(c(1, 1, 2, 2))
sex2 <- factor(c(1, 2, 1, 2))
Bin.Pois2 <- data.frame(resp2, sex2, y2)
attach(Bin.Pois2)

#now define the Poisson model - independence model
modPois<-glm(y2 ~ resp2+sex2, family=poisson)
anova(modPois, test="Chisq")
summary(modPois)
# notice that the marginal totals have been reproduced

# interaction model
modPois<-glm(y2 ~ resp2*sex2, family=poisson)
anova(modPois, test="Chisq")
summary(modPois)

# calculate  observed odds ratio
20*30/(12*14)
exp(modPois$coef[4])
```

Theoretical Relationship between Binomial and Poisson

Denote table entries by R_{ij},

($i =$ gender, $j =$ response)

If $R_{ij} \sim$ Pois(μ_{ij}), independently and we constrain the gender totals, i.e. consider

$$R_{i+} = R_{i1} + R_{i2}, \quad i = 1, 2$$

to be fixed, then

$$R_{i2} | R_{i+} \sim \text{Binomial}(R_{i+}, p_i)$$

where

$$p_i = \frac{\mu_{i2}}{\mu_{i1} + \mu_{i2}}$$

Note

$$\begin{aligned} \text{logit}(p_i) &= \log\left(\frac{\mu_{i2}}{\mu_{i1}}\right) \\ &= \log(\mu_{i2}) - \log(\mu_{i1}) \end{aligned}$$

i.e. logit of YES response is the response effect on the Poisson log-linear scale

Contingency tables and log-linear models

Pneumoconiosis in Coalminers

From J.A.Ashford, "An approach to the Analysis of Data for Semi-quantal Responses in Biological Response". Biometrics 15: 573-581 1959

Coalminers classified by radiological examination into 3 categories of pneumoconiosis (normal, mild, severe) by period of time spent at the coalface (mid-points of class interval).

Years	None	Mild	Severe	Total
5.8	98	0	0	98
15.0	51	2	1	54
21.5	34	6	3	43
27.5	35	5	8	48
33.5	32	10	9	51
39.5	23	7	8	38
46.0	12	6	10	28
51.5	4	2	5	11

Relationship between Multinomial and Poisson

If $N_j \sim \text{Pois}(\mu_j), \quad j = 1, \ldots, k$, independently, then,

$$N_1 + N_2 + \ldots + N_k = N \sim \text{Pois}(\mu_1 + \mu_2 + \ldots + \mu_k)$$

and

$$\begin{aligned}
& P(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k \mid N = n) \\
= \quad & \frac{P(N_1 = n_1, N_2 = n_2, \ldots + N_k = n_k)}{P(N = n)} \\
= \quad & \frac{\prod_{j=1}^{k} P(N_j = n_j)}{P(N = n)} = \frac{n!}{\mu^n e^{-\mu}} \prod_{j=1}^{k} \frac{\mu_j^{n_j} e^{-\mu_j}}{n_j!} \\
= \quad & \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}
\end{aligned}$$

with $p_j = \dfrac{\mu_j}{\mu}$.

Multinomial Response Data

Categorical response variable with k categories, labeled $1, 2 \ldots, k$.

Multinomial distribution with n_i outcomes type i and $\sum_{i=1}^{k} n_i = n$ has probabilities

$$P(n_1, n_2 \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots, n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

where $\sum_{i=1}^{k} p_i = 1$

Multinomial Logits - use 1st category as reference

$$\theta_j = \log\left(\frac{p_j}{p_1}\right) = \log\left(\frac{\mu_j}{\mu_1}\right) \quad j = 2, \ldots, k$$
$$\theta_1 = 0$$

No real restriction in using 1st category since

$$\theta_j - \theta_i = \log\left(\frac{p_j}{p_1}\right) - \log\left(\frac{p_i}{p_1}\right) = \log\left(\frac{p_j}{p_i}\right)$$

Multinomial Logit Models

Expl. var.	1	...	k	Total
x_1	n_{11}	...	n_{k1}	$n_{.1}$
x_2	n_{12}	...	n_{k2}	$n_{.2}$
...	...	...	...	
x_m	n_{1m}	...	n_{km}	$n_{.m}$

For m multinomial observations

$$\mathbf{n}_i = (n_{1i}, n_{2i} \ldots n_{ki})$$

with explanatory variables $\mathbf{x}_i$ we have multinomial logit models

$$\theta_{ji} = \mathbf{x}_i^T \boldsymbol{\beta}_j$$

with $j = 2, \ldots, k$ (categories) and $i = 1, \ldots, m$ (observations).

Poisson Log-linear Model

Modelling the individual counts n_{ij} as Poisson random variables with means μ_{ij} the Poisson/multinomial relationship gives us

$$\theta_{ji} = \log\left(\frac{p_{ji}}{p_{1i}}\right) = \log\left(\frac{\mu_{ji}}{\mu_{1i}}\right) = \mathbf{x}_i^T \boldsymbol{\beta}_j$$

or

$$\phi_{ji} = \log(\mu_{ji}) = \psi_i + \mathbf{x}_i^T \boldsymbol{\beta}_j$$

where $j = 2, \ldots, k$, $\boldsymbol{\beta}_1 = 0$ and $\psi_i = \log(\mu_{1i})$ are a set of nuisance parameters.

Fitting via Log-Linear Models

To fit these Poisson models we need to

- include a full model in the classifying factors to give the nuisance parameters ψ_i and reproduce the marginal multinomial totals
- include a response factor (k levels) to reproduce the overall response pattern;
- $\boldsymbol{\beta}_j$ are the coefficients of the *interaction* of the jth level of the response with the explanatory variables $\mathbf{x}$.

Advantages:

(i) a multivariate problem is converted to a univariate one.

(ii) standard log-linear modelling software can be used.

Disadvantages:

(i) requires estimation of many nuisance parameters;

(ii) can be slow for large tables.

Fitting in R

- Define a response factor with k levels corresponding to the multinomial response categories;
- Model the cell counts as Poisson variables with log link function;
- Set up model in classifying variables to index multinomial observations – reproduce the marginal multinomial totals

  η=var
- Fit the "null" multinomial model – overall pattern

  η=var+resp
- Fit models of interest

  η=var+resp+resp:var

  - β_j is the coefficient of the interaction of the j-th category response with the explanatory variables x

  - standard errors and deviances are also correct.

Ordinal Logistic Regression

Aim: modelling an ordered categorical variable

- response variable Y – ordered categories
- use linear function of explanatory variables
- can include categorical variables
- model **cumulative logits** of response probabilities

$$\text{logit}(\gamma_{ij}) = \log\left(\frac{Pr(Y_i \leq j|\mathbf{x}_i)}{Pr(Y_i > j|\mathbf{x}_i)}\right)$$

- the cumulative logit model is given by

$$\text{logit}(\gamma_{ij}) = \theta_j - \mathbf{x}_i^T\boldsymbol{\beta}$$

where θ_j may be thought of as category cut-points for an underlying continuous latent variable.
$\mathbf{x}_i$ are the covariates and $\boldsymbol{\beta}$ the parameter vector.

- gives proportional odds model;
  e.g. for a simple model

$$\text{logit}(\gamma_{ij}) = \theta_j - x_i^T \beta$$

  the difference in two logits is a *cumulative odds ratio* and

$$\begin{aligned}
&\text{logit}(\gamma_{rj}) - \text{logit}(\gamma_{sj}) \\
&= \log\left(\frac{Pr(Y_r \leq j|x_r)/Pr(Y_r > j|x_r)}{Pr(Y_s \leq j|x_s)/Pr(Y_s > j|x_s)}\right) \\
&= -\beta(x_r - x_s)
\end{aligned}$$

- negative sign for $\boldsymbol{\beta}$ associates increasing values of Y with increasing values of $\mathbf{x}$.

Log-Linear models for Ordered Data

Linear by Linear Association

Another possible approach is to assign scores to any ordered response. For a 2-way table ($A \times B$, A with a levels, B with b levels), assigning scores u_i and v_j to the rows and columns gives an ordered interaction model

$$\log(m_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \gamma u_i v_j$$

with $ab - a - b$ degrees of freedom $\gamma = 0$ corresponds to the independence model.
Special case of the full interaction (saturated) model.
Odds-ratio interpretation:

$$\log\left(\frac{m_{hj}m_{ik}}{m_{hk}m_{ij}}\right) = \gamma(u_i - u_h)(v_k - v_j)$$

Considering adjacent cells gives local odds ratios

$$\theta_{ij} = \frac{m_{i,j} m_{i+1,j+1}}{m_{i,j+1} m_{i+1,j}}$$

and so

$$\log \theta_{ij} = \gamma(u_{i+1} - u_i)(v_{j+1} - v_j)$$

Taking integer scores gives

$$\log \theta_{ij} = \gamma$$

the *uniform association model*.

Easily fitted as a log-linear model.

```
Miners.dat<-scan("Miners.dat", what=list(N=0, M=0, S=0))

#      Pneumoconiosis in Coalminers
#     ============================
# From J.A.Ashford, "An approach to the Analysis of Data for Semi-quantal
# Responses in Biological Response". Biometrics 15: 573-581 1959
# Coalminers classified by radiological examination into 3 categories
# of pneumoconiosis by period of time spent at the coalface.

# Response Category Data
#        N = No of normal respondents
#        M = No with mild pneumoconiosis
#        S = No with severe pneumoconiosis
#        P = Period at coalface (8 level factor)
#    YEARS = mid-point of interval corresponding to period



attach(Miners.dat)
NM<-N+M
T<-NM+S

Period<-factor(c(1:8))
Period<-factor(c(Period,Period,Period))

Year1<-c(5.8,15,21.5,27.5,33.5,39.5,46,51.5)
Years<-c(Year1, Year1, Year1)

count<-c(N,M,S)
T<-c(T,T,T)
Po<-count/T

plot(Years,Po, pch=c(rep("*",8),rep("+",8),rep("#",8)),
col=c(rep("green",8), rep("red",8), rep("blue",8)),
main="Figura 1. Proporoes observadas")
legend(50,1,c("N","M", "S"), #lty=c(-1,2,1),
pch=c("*","+","#"), col=c("green","red","blue"),cex=.6)
Resp<-factor(rep(c("N","M","S"), c(8,8,8)))

mod1<-glm(count~Period, family=poisson)
mod2<-glm(count~Period+Resp, family=poisson)
mod3<-glm(count~Period+Resp+Resp:Years, family=poisson)
anova(mod1, mod2, mod3, test="Chisq")

ly<-log(Years)
mod4<-glm(count~Period+Resp+Resp:ly, family=poisson)
anova(mod1, mod2, mod4, test="Chisq")

r23<-factor(rep(c("N","MS"), c(8,16)))
mod5<-glm(count~Period+Resp+r23:ly, family=poisson)
anova(mod1, mod2, mod5, mod4, test="Chisq")
```