Poisson Regression

EPI 204

Quantitative Epidemiology III

Statistical Models

Poisson Distributions

- The Poisson distribution can be used to model unbounded count data, 0, 1, 2, 3, …
- An example would be the number of cases of sepsis in each hospital in a city in a given month.
- The Poisson distribution has a single parameter λ, which is the mean of the distribution and also the variance. The standard deviation is

$$\sqrt{\lambda}$$

Poisson Regression

- If the mean λ of the Poisson distribution depends on variables $x_1, x_2, \ldots, x_p$ then we can use a generalized linear model with Poisson distribution and log link.
- We have that log(λ) is a linear function of $x_1, x_2, \ldots, x_p$.
- This works pretty much like logistic regression, and is used for data in which the count has no specific upper limit (number of cases of lung cancer at a hospital) whereas logistic regression would be used when the count is the number out of a total (number of emergency room admissions positive for C. dificile out of the known total of admissions).

The probability mass function of the Poisson distribution is

$$f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

so the log-likelihood is for a single response y is

$$L(\lambda \mid y) = y\ln(\lambda) - \lambda - \ln(y!)$$

$$L'(\lambda \mid y) = y/\lambda - 1$$

and the MLE of λ is $\hat{\lambda} = y$

In the saturated model, for each observation y,

the maximized likelihood is $y\ln(y) - y - \ln(y!)$

so the deviance when λ is estimated by $\hat{\lambda} = \exp(\eta)$ is

$$2(y\ln(y) - y - y\ln(\hat{\lambda}) + \hat{\lambda}) = 2(y\ln(y/\hat{\lambda}) - (y - \hat{\lambda}))$$

The latter term disappears when added over all data points if there is an intercept so

$$D = 2\sum y_i \ln(y_i/\hat{\lambda})$$

Each deviance term is 0 with perfect prediction.

eba1977 package:ISwR R Documentation

Lung cancer incidence in four Danish cities 1968-1971

This data set contains counts of incident lung cancer cases and population size in four neighbouring Danish cities by age group.

A data frame with 24 observations on the following 4 variables:

'city' a factor with levels 'Fredericia', 'Horsens', 'Kolding', and 'Vejle'.
'age' a factor with levels '40-54', '55-59', '60-64', '65-69', '70-74', and '75+'.
'pop' a numeric vector, number of inhabitants.
'cases' a numeric vector, number of lung cancer cases.

Details:

These data were "at the center of public interest in Denmark in 1974", according to Erling Andersen's paper. The city of Fredericia has a substantial petrochemical industry in the harbour area.

```
> library(ISwR)
> data(eba1977)
> help(eba1977)
> dim(eba1977)
[1] 24  4
> eba1977
         city   age  pop cases
1  Fredericia 40-54 3059    11
2     Horsens 40-54 2879    13
3     Kolding 40-54 3142     4
4       Vejle 40-54 2520     5
5  Fredericia 55-59  800    11
..........
20      Vejle 70-74  539     8
21 Fredericia   75+  605    10
22    Horsens   75+  782     2
23    Kolding   75+  659    12
24      Vejle   75+  619     7
```

```
> eba.glm <- glm(cases ~
  city+age+offset(log(pop)),family=poisson,data=eba1977)
> summary(eba.glm)

Call:
glm(formula = cases ~ city + age + offset(log(pop)),
  family = poisson)

Deviance Residuals:
     Min         1Q     Median         3Q        Max
-2.63573  -0.67296  -0.03436   0.37258   1.85267
```

Having offset(x) in a formula is like having x in the formula except the coefficient is fixed to 1. Having offset(log(pop)) in the formula, with the log link, makes the parameter lambda proportional to the population. A similar effect would come from analyzing the ratio of cases to population, but then we would not have count data.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.6321     0.2003 -28.125  < 2e-16 ***
cityHorsens  -0.3301     0.1815  -1.818   0.0690 .
cityKolding  -0.3715     0.1878  -1.978   0.0479 *
cityVejle    -0.2723     0.1879  -1.450   0.1472
age55-59      1.1010     0.2483   4.434 9.23e-06 ***
age60-64      1.5186     0.2316   6.556 5.53e-11 ***
age65-69      1.7677     0.2294   7.704 1.31e-14 ***
age70-74      1.8569     0.2353   7.891 3.00e-15 ***
age75+        1.4197     0.2503   5.672 1.41e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 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
```

$$\text{predictor}_{ij} = \text{intercept} + \text{coef.city}_i + \log\left(\text{pop}_{ij}\right) + \text{coef.age}_j$$

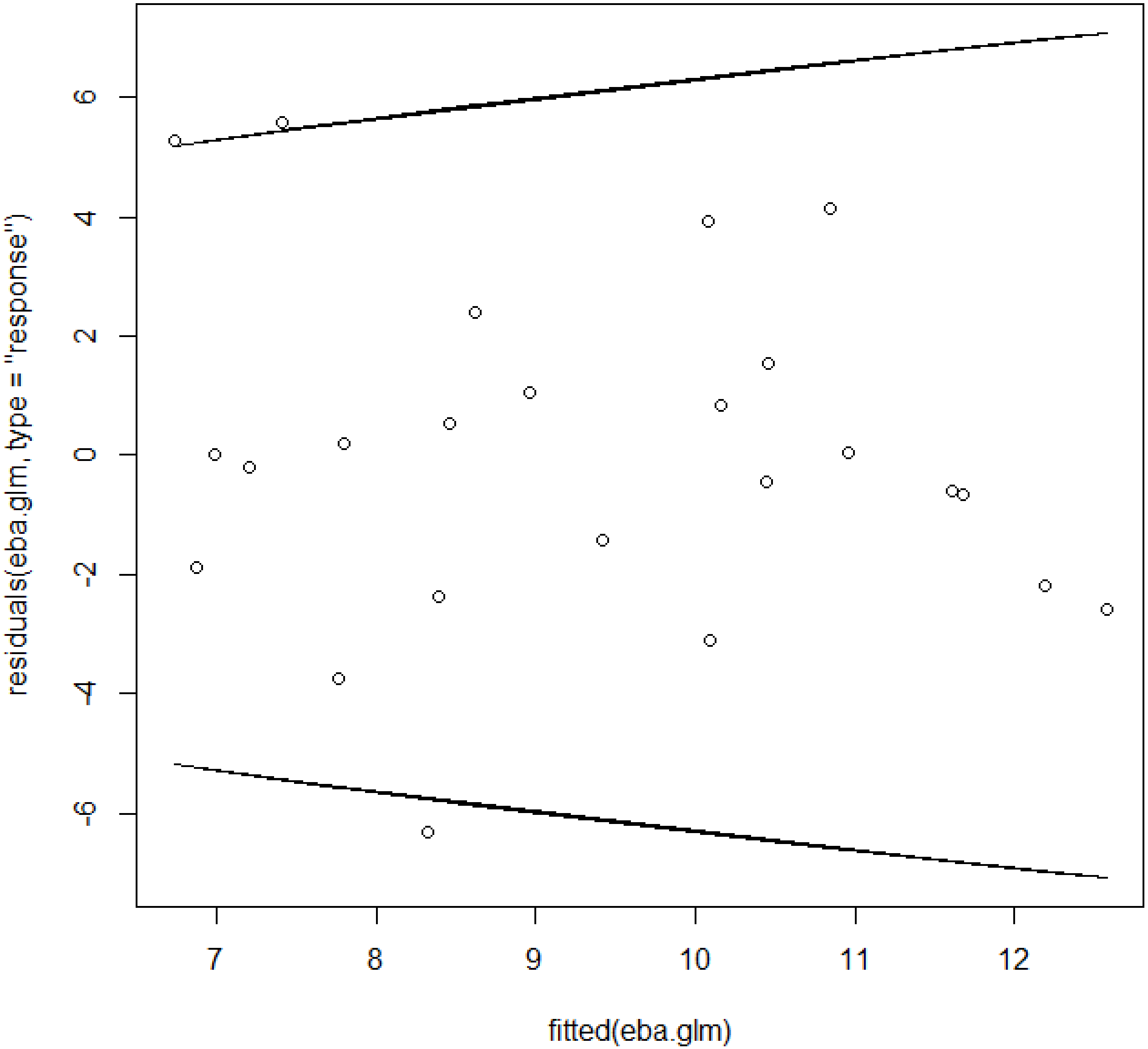
$$\lambda_{ij} = \exp\left[\text{intercept+coef.city}_i + \log\left(\text{pop}_{ij}\right) + \text{coef.age}_j\right]$$

$$= \exp[\text{intercept}]\exp\left[\text{coef.city}_i\right]\exp\left[\text{coef.age}_j\right]\text{pop}_{ij}$$

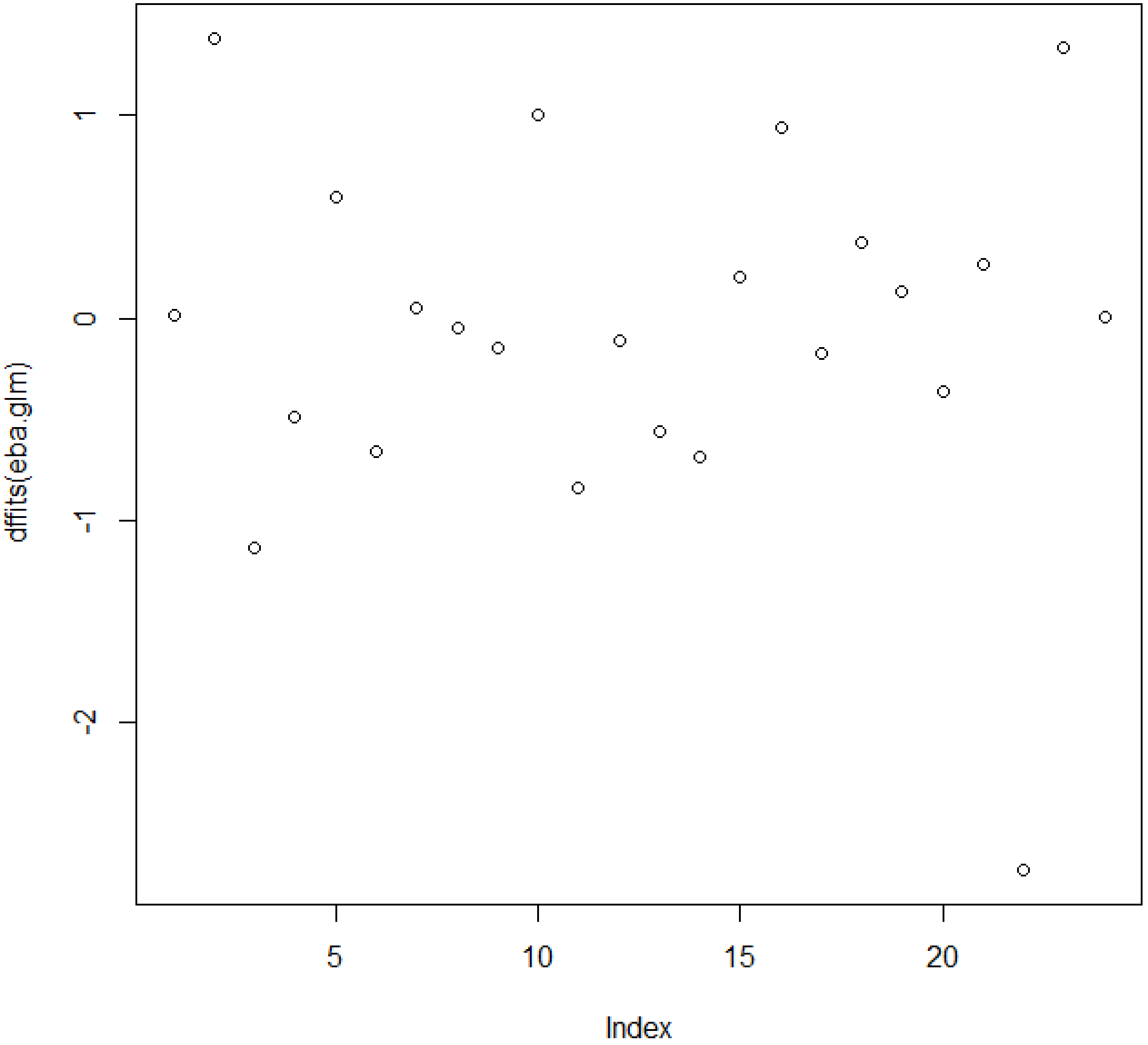
```
> plot(fitted(eba.glm),residuals(eba.glm,type="response"),ylim=c(-7,7))
> lines(fitted(eba.glm),2*sqrt(fitted(eba.glm)))
> lines(fitted(eba.glm),-2*sqrt(fitted(eba.glm)))

> plot(dffits(eba.glm))
> which(dffits(eba.glm) < -2)
22
22
> eba1977[22,]
      city age pop cases
22 Horsens 75+ 782     2

> eba1977[eba1977$age=="75+",]
         city age pop cases
21 Fredericia 75+ 605    10
22    Horsens 75+ 782     2
23    Kolding 75+ 659    12
24      Vejle 75+ 619     7
```



residuals(eba.glm, type = "response")
6
4
2
0
-2
-4
-6
7
8
9
10
11
12
fitted(eba.glm)


dffits(eba.glm)
1
0
-1
-2
5
10
15
20
Index

Goodness of Fit

- If the model fits well, the residual deviance should be in the neighborhood of the df of the residual deviance.
- 23.447 on 15 df
- Under the null hypothesis that the model fits, and if the smallest fitted value is > 5, then the null distribution is approximately chi-squared

```
> min(fitted(eba.glm))
[1] 6.731286
> pchisq(deviance(eba.glm),
         df.residual(eba.glm),lower=F)
[1] 0.07509017
```

```
> drop1(eba.glm,test="Chisq")
Single term deletions

Model:
cases ~ city + age + offset(log(pop))
       Df Deviance    AIC     LRT Pr(Chi)
<none>      23.447 137.84
city    3   28.307 136.69   4.859  0.1824
age     5  126.515 230.90 103.068  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
  ’ 1
```

The test of the city effect would not be correct if we had individual patient data, since it then would be a characteristic of a group of patients, not of a patient. This would require a hierarchical model as in glmer() or PROC GLIMMIX

```
> cf <- coef(summary(eba.glm))
> cf
              Estimate Std. Error    z value      Pr(>|z|)
(Intercept) -5.6320645  0.2002545 -28.124529 4.911333e-174
cityHorsens -0.3300600  0.1815033  -1.818479  6.899094e-02
cityKolding -0.3715462  0.1878063  -1.978348  4.788946e-02
cityVejle   -0.2723177  0.1878534  -1.449629  1.471620e-01
age55-59     1.1010140  0.2482858   4.434463  9.230223e-06
age60-64     1.5186123  0.2316376   6.555985  5.527587e-11
age65-69     1.7677062  0.2294395   7.704455  1.314030e-14
age70-74     1.8568633  0.2353230   7.890701  3.004950e-15
age75+       1.4196534  0.2502707   5.672472  1.407514e-08
```

```
> est <- cf[,1]
> se <- cf[,2]
> rr <- exp(cbind(est, est-se*qnorm(.975),
                       est+se*qnorm(.975)))
colnames(rr) <- c("RateRatio","LowerCL","UpperCL")
> rr
              RateRatio     LowerCL     UpperCL
(Intercept) 0.003581174 0.002418625  0.005302521
cityHorsens 0.718880610 0.503687146  1.026012546
cityKolding 0.689667168 0.477285856  0.996553318
cityVejle   0.761612264 0.527026991  1.100613918
age55-59    3.007213795 1.848515376  4.892215085
age60-64    4.565884929 2.899710957  7.189442499
age65-69    5.857402508 3.735990951  9.183417356
age70-74    6.403619032 4.037552548 10.156236043
age75+      4.135686847 2.532309969  6.754270176
```

These are rates per 4 person years.
The confidence intervals use an asymptotic approximation. A more accurate method in some cases is

```
> exp(cbind(coef(eba.glm),confint(eba.glm)))
Waiting for profiling to be done...
                              2.5 %       97.5 %
(Intercept) 0.003581174 0.002373629  0.005212346
cityHorsens 0.718880610 0.502694733  1.025912422
cityKolding 0.689667168 0.475568043  0.995045687
cityVejle   0.761612264 0.525131867  1.098950868
age55-59    3.007213795 1.842951851  4.901008833
age60-64    4.565884929 2.907180919  7.236296972
age65-69    5.857402508 3.748295295  9.248885425
age70-74    6.403619032 4.043044796 10.211923083
age75+      4.135686847 2.522891909  6.762422572
```

Is $\mathrm{Var}(x|\lambda) = \Theta\lambda$

```
> summary(glm(cases ~ city+age+offset(log(pop)),family=quasipoisson,data=eba1977))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -5.6321     0.2456 -22.932 4.31e-13 ***
cityHorsens  -0.3301     0.2226  -1.483  0.15884    
cityKolding  -0.3715     0.2303  -1.613  0.12756    
cityVejle    -0.2723     0.2304  -1.182  0.25561    
age55-59      1.1010     0.3045   3.616  0.00254 ** 
age60-64      1.5186     0.2841   5.346 8.17e-05 ***
age65-69      1.7677     0.2814   6.282 1.47e-05 ***
age70-74      1.8569     0.2886   6.434 1.13e-05 ***
age75+        1.4197     0.3069   4.625  0.00033 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.504109)

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

Number of Fisher Scoring iterations: 5
```

Specific test vs. Fredericia

```
> is.fred <- as.numeric(eba1977$city == "Fredericia")
> summary(glm(cases ~ is.fred+age+offset(log(pop)),family=poisson,data=eba1977))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9589     0.1809 -32.947  < 2e-16 ***
is.fred       0.3257     0.1481   2.200   0.0278 *
age55-59      1.1013     0.2483   4.436 9.17e-06 ***
age60-64      1.5203     0.2316   6.564 5.23e-11 ***
age65-69      1.7687     0.2294   7.712 1.24e-14 ***
age70-74      1.8592     0.2352   7.904 2.71e-15 ***
age75+        1.4212     0.2502   5.680 1.34e-08 ***
---
    Null deviance: 129.91  on 23  degrees of freedom
Residual deviance:  23.70  on 17  degrees of freedom
AIC: 134.09
> drop1(glm(cases ~ is.fred+age+offset(log(pop)),family=poisson,data=eba1977),test="Chisq")

Model:
cases ~ is.fred + age + offset(log(pop))
        Df Deviance    AIC     LRT Pr(>Chi)
<none>       23.700 134.09
is.fred  1   28.307 136.69   4.606  0.03185 *
age      5  127.117 227.50 103.417  < 2e-16 ***
```

```
bcmort                  package:ISwR                  R
  Documentation

Breast cancer mortality

     Danish study on the effect of screening for breast
     cancer.

Format:

     A data frame with 24 observations on 4 variables.

      ‘age’ a factor with levels ‘50-54’, ‘55-59’,
           ‘60-64’, ‘65-69’, ‘70-74’, and ‘75-79’

      ‘cohort’ a factor with levels ‘Study gr.’,
           ‘Nat.ctr.’, ‘Hist.ctr.’, and ‘Hist.nat.ctr.’.

      ‘bc.deaths’ numeric, number of breast cancer deaths.

      ‘p.yr’ a numeric vector, person-years under study.
```

Details:

Four cohorts were collected. The "study group" consists of the population of women in the appropriate age range in Copenhagen and Frederiksberg after the introduction of routine mammography screening. The "national control group" consisted of the population in the parts of Denmark in which routine mammography screening was not available. These two groups were both collected in the years 1991-2001. The "historical control group" and the "historical national control group" are similar cohorts from 10 years earlier (1981-1991), before the introduction of screening in Copenhagen and Frederiksberg. The study group comprises the entire population, not just those accepting the invitation to be screened.

A.H. Olsen et al. (2005), Breast cancer mortality in Copenhagen after introduction of mammography screening. British Medical Journal, 330: 220-222.

Exercise

- In the bcmort data set, the four-level factor cohort can be considered the product of two two-level factors, say “period” (1981–1991 or 1991–2001) and “area” (Copenhagen/Fredriksberg and National). Generate those two factors.
- Fit a Poisson regression model to the data with age, period, and area as descriptors, as well as the three two-factor interaction terms. The interaction between period and area can be interpreted as the effect of screening (explain why). How should person-years under study be used in the model?
- Check the model with plots and quasipoisson fit.