R Code And Tasks Chapter 5 (MAS 6003)

Witold Wolski

December 27, 2016

Chapter 5 Poisson regression

5.1 Introduction

pdf of poisson

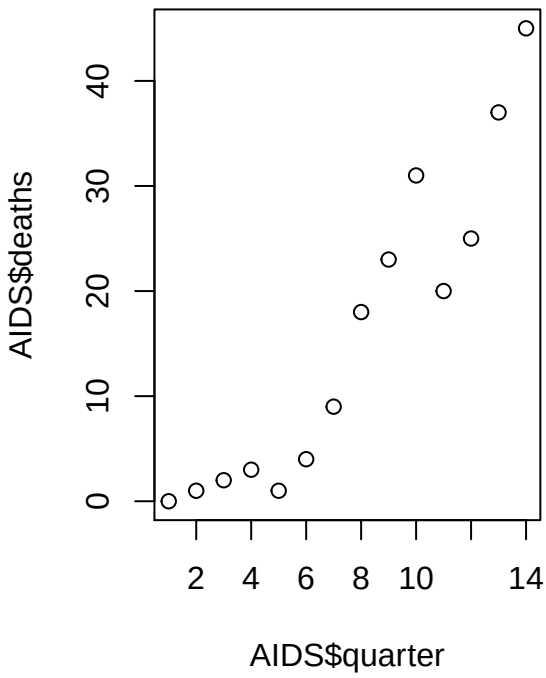
$$\frac{\lambda^k e^{-\lambda}}{k!}$$

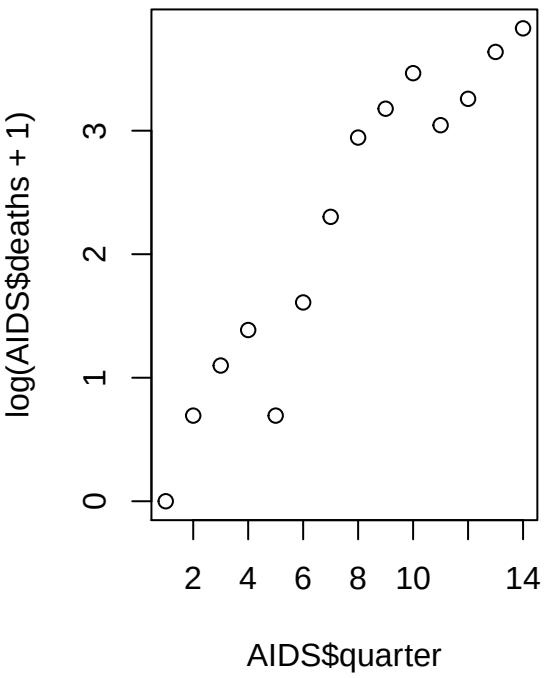
5.2.1 Example : AIDS deaths over time (Task 15)

1 plot :

```
rm(list=ls())
load("data/MAS367-GLMs.RData", envir = e <- new.env())

AIDS <- e$AIDS
par(mfrow=c(1,2))
plot(AIDS$quarter, AIDS$deaths)
plot(AIDS$quarter, log(AIDS$deaths+1))
```





2 fit poisson with log link

```
glm.lin <- glm(deaths ~ quarter, data=AIDS, family=poisson(link='log'))
qchisq(0.95,glm.lin$df.residual)
```

```
## [1] 21.02607
```

3 adding a quadratic term

```
glm.quad <- glm(deaths ~ quarter + I(quarter^2), data=AIDS, family=poisson(link='log'))
summary(glm.quad)
```

```
##
## Call:
## glm(formula = deaths ~ quarter + I(quarter^2), family = poisson(link = "log"),
##     data = AIDS)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7708  -0.9385   0.1304   0.8190   1.4421
##
```

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.713375   0.733108  -2.337 0.019432 *
## quarter       0.746031   0.153391   4.864 1.15e-06 ***
## I(quarter^2) -0.025836   0.007751  -3.333 0.000859 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 207.272  on 13  degrees of freedom
## Residual deviance:  16.371  on 11  degrees of freedom
## AIC: 75.298
##
## Number of Fisher Scoring iterations: 4
```

```
qchisq(0.95,glm.lin$df.residual)
```

```
## [1] 21.02607
```

4 a line predictor on log(x)

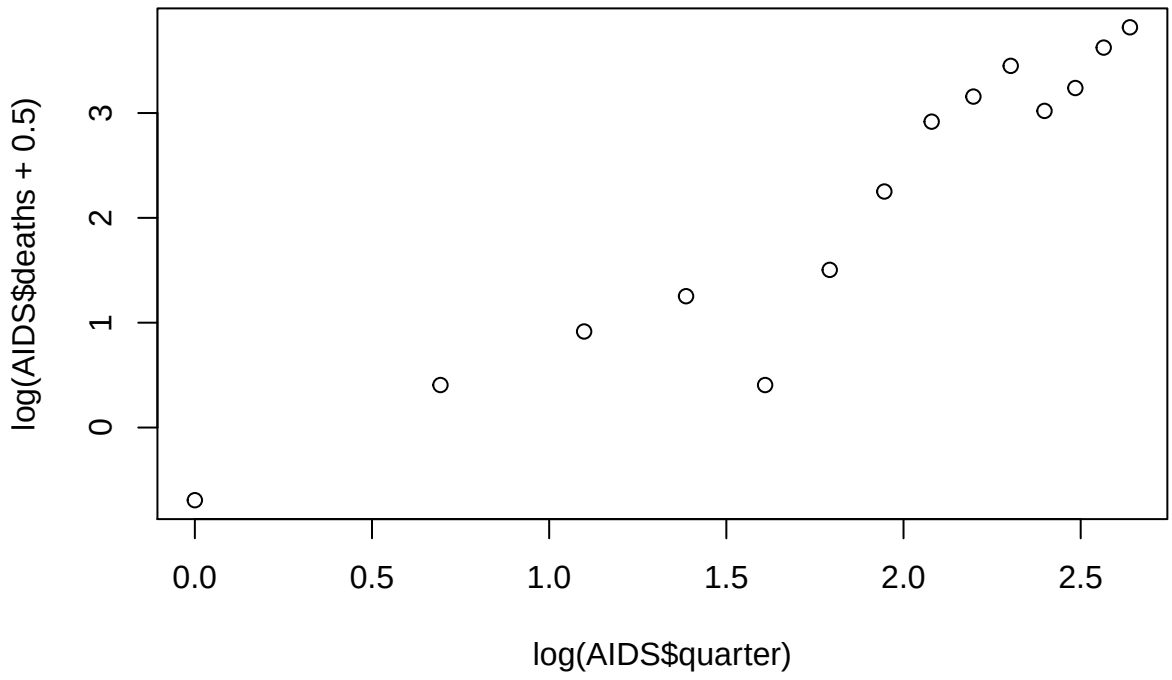
```
glm.logline <- glm(deaths ~ I(log(quarter)), data=AIDS, family=poisson(link='log'))
summary(glm.logline)
```

```
##
## Call:
## glm(formula = deaths ~ I(log(quarter)), family = poisson(link = "log"),
##     data = AIDS)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.08992  -1.07141  -0.04657   0.38956   1.94311
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.9442     0.5116   -3.80 0.000145 ***
## I(log(quarter))   2.1748     0.2150   10.11  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 207.272  on 13  degrees of freedom
## Residual deviance:  17.092  on 12  degrees of freedom
## AIC: 74.019
##
## Number of Fisher Scoring iterations: 4
```

```
qchisq(0.95,glm.logline$df.residual)
```

```
## [1] 21.02607
```

```
plot(log(AIDS$quarter), log(AIDS$deaths+0.5))
```



5

Thus possible simple models are a line in logx or a quadratic in x, but there are reservations about both.

5.3 Adjusting for exposure : offset (Task 16)

An explanation of offset which is brief and clear can be found here (but not in the lecture notes of MAS 6003).

5.3.1 Example: Smoking and heart disease

1,2,3,4 death rates

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(gridExtra)
```

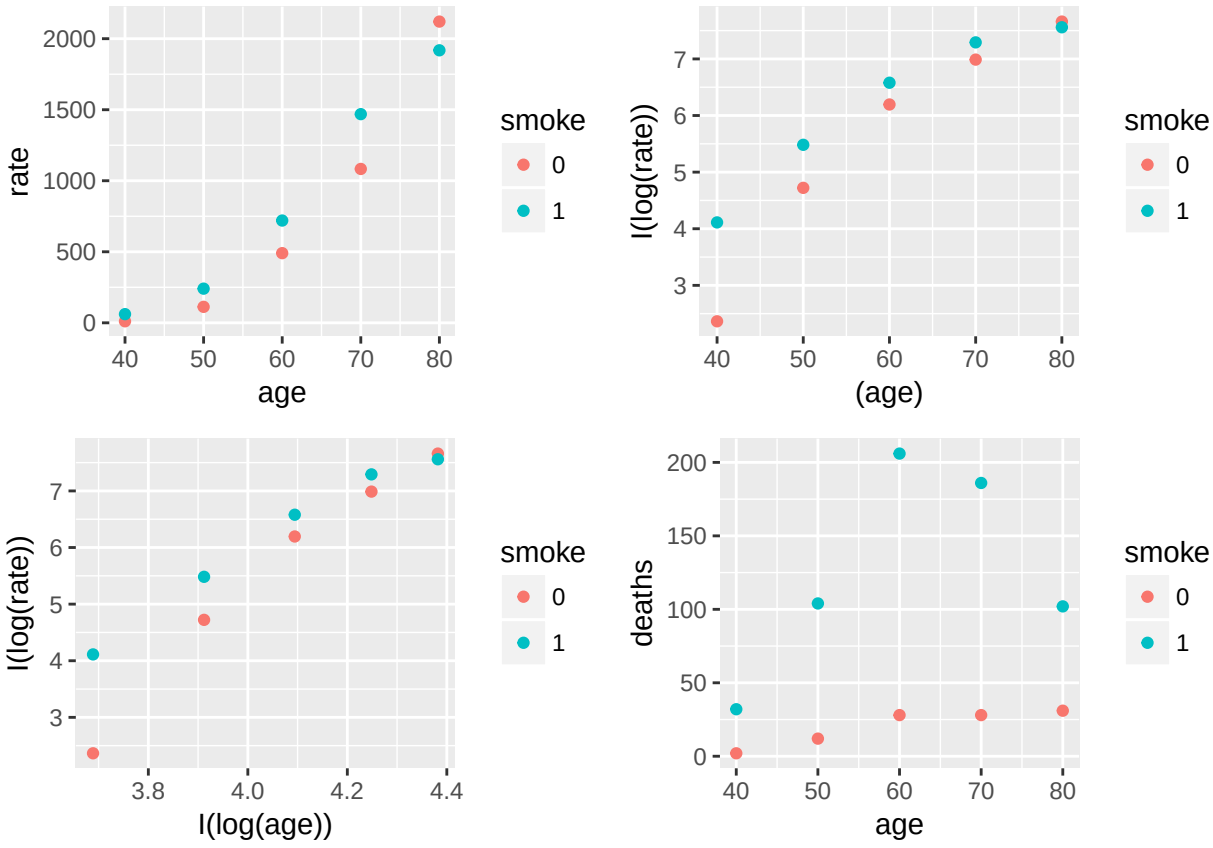
```
## Warning: package 'gridExtra' was built under R version 3.3.2
```

```
smoking <- e$smoking
smoking$rate <- smoking$deaths/smoking$person.years * 1e5
lapply(smoking,class)
```

```
## $age
## [1] "integer"
##
## $smoke
## [1] "integer"
##
## $deaths
## [1] "integer"
##
## $person.years
## [1] "integer"
##
## $rate
## [1] "numeric"
```

```
smoking$smoke <- as.factor(smoking$smoke)
p1 <- ggplot(smoking, aes(age, rate, colour=smoke)) + geom_point()
p2 <- ggplot(smoking, aes((age), I(log(rate)), colour=smoke)) + geom_point()
p3 <- ggplot(smoking, aes(I(log(age)), I(log(rate)), colour=smoke)) + geom_point()
p4 <- ggplot(smoking, aes(age, deaths, colour=smoke)) + geom_point()


grid.arrange(p1,p2,p3,p4, ncol = 2)
```



5 The model

```
mod.offset <- glm(deaths~ offset(log(person.years)) + smoke * age + I(age^2), family = poisson, data=sm
summary(mod.offset)
```

```
##
## Call:
## glm(formula = deaths ~ offset(log(person.years)) + smoke * age +
##     I(age^2), family = poisson, data = smoking)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
## -0.83049   0.43820   0.13404  -0.27329   0.64107  -0.15265  -0.41058
##        8         9        10
##  0.23393  -0.01275  -0.05700
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.970e+01  1.253e+00 -15.717  < 2e-16 ***
## smoke1       2.364e+00  6.562e-01   3.602 0.000316 ***
## age          3.563e-01  3.632e-02   9.810  < 2e-16 ***
## I(age^2)    -1.977e-03  2.737e-04  -7.223 5.08e-13 ***
## smoke1:age  -3.075e-02  9.704e-03  -3.169 0.001528 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 935.0673  on 9  degrees of freedom
## Residual deviance:   1.6354  on 5  degrees of freedom
## AIC: 66.703
##
## Number of Fisher Scoring iterations: 4
```

With smokers = 1 and 0 for nonsmokers: for non-smokers:

$-19.7 + 0.36x^2 - 0.02x^2$

for smokers:

$-17.34 + 0.33x2 - 0.02x2$

5.4 Non negative data with variance $\propto$ means