ADA HW9

Liangquan Zhou lz2377

November 06, 2014

Cancer Deaths of Atomic Bomb Survivors.

The data is the number of cancer deaths among survivors of the atomic bombs dropped on Japan during World War II, categorized by time (years) after the bomb that death occurd and the amount of radiation exposure that the survivors received from the blast. Also listed in each cell is the `person-years at risk`, in 100s. This is the sum total of all years spent by all persons in the category.

Suppose that the mean number of cancer death in each cell is Poisson with mean $\mu = risk \times rate$, where `risk` is the `person-years at risk` and `rate` is the rate of cancer deaths per person per year.

It is desired to describe this `rate` in terms of the amount of radiation, adjusting for the effects of time after exposure.

(a). Using log(`risk`) as an offset, fit the Poisson log-linear regression model with time after blast treated as a factor (with seven levels) and with `rads` and `rads-squared` treated as covariates. Look at the deviance statistic and the deviance residuals. Does extra-Poisson variation seem to be present? Is the `rads-squared` term necessary?

Fit the poisson regression model with the data.

```
fit.a = glm(Deaths ~ Years + Exposure + I(Exposure^2) + offset(log(Risk)), data = dat,
  family = poisson)
anova(fit.a, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Deaths
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                             41        336
## Years          6    270.7        35         65  < 2e-16 ***
## Exposure       1     14.9        34         50  0.00011 ***
## I(Exposure^2)  1      3.4        33         47  0.06454 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit.a)
```

```
##
## Call:
## glm(formula = Deaths ~ Years + Exposure + I(Exposure^2) + offset(log(Risk)),
```

```
##     family = poisson, data = dat)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.360  -0.842  -0.101   0.478   2.682
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.26e+00   1.89e-01  -17.28  < 2e-16 ***
## Years8-11      2.33e-01   2.53e-01    0.92   0.3561
## Years12-15     5.52e-01   2.37e-01    2.33   0.0199 *
## Years16-19     1.25e+00   2.13e-01    5.86  4.8e-09 ***
## Years20-23     1.40e+00   2.10e-01    6.69  2.3e-11 ***
## Years24-27     1.74e+00   2.04e-01    8.52  < 2e-16 ***
## Years28-31     2.03e+00   2.00e-01   10.17  < 2e-16 ***
## Exposure       4.45e-03   1.46e-03    3.05   0.0023 **
## I(Exposure^2) -7.44e-06   4.02e-06   -1.85   0.0640 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 335.75  on 41  degrees of freedom
## Residual deviance:  46.69  on 33  degrees of freedom
## AIC: 214.4
##
## Number of Fisher Scoring iterations: 5
```

The deviance residuals is 46.6896 and the mean of `Deaths` is 15.0238. Since the sample mean substantially smaller than deviance residuals, there is a extra-Poisson variation (Overdispersion).

The coefficient of `rads`-squared is small, and it corresponded p-value is greater than 0.05, thus we think `rads`-squared term is not necessary.

(b). Try the same model in part (a); but insted of treating time after bomb as a factorwith seven levels, ocmpute the midpoint of each interval and include log(`time`) as a numerical explanatory variable. Is the deviance statistic substantially larger in this model, or does it appear that time can adequately be represented through this single term?

Since the `rads`-squared term is not necessary in the model, we drop this term. Fit the model in (a) again but treat the time as a numeric variable, then compare the anova table with the anova table in (a):

```r
fit.a = glm(Deaths ~ Years + Exposure + offset(log(Risk)), data = dat, family = poisson)
levels(dat$Years) = as.character(c(3.5, 9.5, 13.5, 17.5, 21.5, 25.5, 29.5))
dat$Years = as.numeric(as.character(dat$Years))
fit.b = glm(Deaths ~ Years + Exposure + offset(log(Risk)), data = dat, family = poisson)
#compare the deviance residuals
anova(fit.a)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
```

```
## Response: Deaths
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev
## NULL                         41        336
## Years     6    270.7         35         65
## Exposure  1     14.9         34         50
```

```
anova(fit.b)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Deaths
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev
## NULL                         41        336
## Years     1    264.8         40         71
## Exposure  1     14.9         39         56
```

The deviance residuals in (a) is 50.1062, and the deviance residuals in (b) is 56.0443, which is not substantially larger the deviance redisuals in (a). So we conclude that time can adequately be represented through this single term.

(c). Try fitting a model that includes the interaction of log(`time`) and exposure. Is the interaction significant?

```
fit.c = glm(Deaths ~ Years + Exposure + I(log(Years)*Exposure) + offset(log(Risk)),
  data = dat, family = poisson)
summary(fit.c)
```

```
##
## Call:
## glm(formula = Deaths ~ Years + Exposure + I(log(Years) * Exposure) +
##     offset(log(Risk)), family = poisson, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9555  -1.0526   0.0047   0.5746   2.8292
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -3.547236   0.143265  -24.76   <2e-16 ***
## Years                    0.080448   0.006160   13.06   <2e-16 ***
## Exposure                -0.000800   0.003121   -0.26     0.80
```

```
## I(log(Years) * Exposure)  0.000868   0.001011    0.86     0.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 335.750  on 41  degrees of freedom
## Residual deviance:  55.243  on 38  degrees of freedom
## AIC: 213
##
## Number of Fisher Scoring iterations: 5
```

The inteaction term has a coefficient 8.6787 × 10-4 with p-value greater than 0.05, thus it is not significant.

(d). Based on a good-fitting model, make a statement about the effect of radiation exposure on the number of cancer deaths per person per year (and include a confidence interval if you supply an estimate of a parameter).

Based on the previous questions, we choose the model in (b) as the best good-fitting model. We can get the coefficients and their 95 percent confidence intervals:

```
fit.b$coef
```

```
## (Intercept)       Years    Exposure
##   -3.603193    0.082960    0.001833
```

```
confint(fit.b)
```

```
##                  2.5 %    97.5 %
## (Intercept) -3.8621533 -3.354612
## Years        0.0723246  0.093840
## Exposure     0.0009382  0.002662
```

So we the model is

$$log(Death) = -3.603192703 + 0.082959982 * Years + 0.001832762 * Exposure$$

Since we set the `Exposure` as a numeric variable, then it can be interpreted as: When the radiation exposure increase 1, the number of death per person per year would likely to increase $e^{0.001832762} = 1.0018$.