Ada Homework9
Wenxin Liang UNI: wl2455

Problem 20. Cancer Deaths of Atomic Bomb Survivors. The data in Display 22.13 are the number of cancer deaths among survivors of the atomic bombs dropped on Japan during World War II, catego- rized by time (years) after the bomb that death occurred and the amount of radiation exposure that the survivors received from the blast. (Data from D. A. Pierce, personal communication.) Also listed in each cell is the *person-years at risk*, in 100's. This is the sum total of all years spent by all persons in the category. Suppose that the mean number of cancer deaths 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.

DISPLAY 22.13 Cancer deaths among Japanese atomic bomb survivors, categorized by estimated exposure to radiation (in rads) and years after exposure; below the number of cancer deaths are the person-years (in 100's) at risk

		Years after exposure						
exposure (rads)		0–7	8–11	12–15	16–19	20–23	24–27	28–31
0	*deaths:*	10	12	19	31	35	48	73
	risk:	262	243	240	237	233	227	220
25	*deaths:*	17	17	17	47	50	65	71
	risk:	313	290	285	280	275	269	262
75	*deaths:*	0	2	1	5	8	7	12
	risk:	38	36	35	34	34	33	32
150	*deaths:*	1	0	4	1	6	12	11
	risk:	28	26	25	25	24	24	23
250	*deaths:*	1	1	0	4	3	7	13
	risk:	13	12	12	12	11	11	10
400	*deaths:*	0	2	5	3	2	3	5
	risk:	15	14	14	14	13	13	13

(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?

Based on R, 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 is as followed,

```
> fit1=glm(death~offset(log(risk))+year+rads+rads2,family=poisson)
> fit1

Call:  glm(formula = death ~ offset(log(risk)) + year + rads + rads2,
    family = poisson)

Coefficients:
(Intercept)   year12to15   year16to19   year20to23   year24to27   year28to31    year8to11         rads        rads2
 -3.265e+00    5.521e-01    1.249e+00    1.404e+00    1.737e+00    2.032e+00    2.332e-01    4.446e-03   -7.438e-06

Degrees of Freedom: 41 Total (i.e. Null);  33 Residual
Null Deviance:      335.7
Residual Deviance: 46.69        AIC: 214.4
```

The deviance statistics is 46.69 and the deviance residual as followed,

```
> residuals(fit1,"deviance")
           1            2            3            4            5            6            7            8            9
-0.003290809  0.080460137  0.747235652 -0.100564633 -0.209674475 -0.180333933  1.083857026  0.971426728  0.358819624
          10           11           12           13           14           15           16           17           18
-0.910999050  0.838983981  0.345150351  0.008152631 -1.558215175 -1.971522879 -0.218311559 -1.396172785 -0.444601775
          19           20           21           22           23           24           25           26           27
 0.338028840 -0.877686047 -0.138313419 -0.626055720 -2.033269089  0.715773552 -2.359822092 -0.062371014  1.099662484
          28           29           30           31           32           33           34           35           36
-0.014448344  0.052724209 -0.101707978 -1.743651686  0.518357528 -0.150059757  1.059978735  2.681643820 -1.436715003
          37           38           39           40           41           42
 0.649550597  2.072227302 -0.198794530 -0.941866119 -0.999998152 -0.733381044
```

For the extra-Poisson variation,

```
> var(fit1$fitted)
[1] 413.4749
> mean(fit1$fitted)
[1] 15.02381
```

```
> var(death)
[1] 407.6336
> mean(death)
[1] 15.02381
```

There is big difference happened when comparing the variance and the mean then we conclude extra-Poisson variation seem to be present.

```
> summary(fit1)

Call:
glm(formula = death ~ offset(log(risk)) + year + rads + rads2,
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3598  -0.8416  -0.1011   0.4785   2.6816

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.265e+00  1.890e-01 -17.276  < 2e-16 ***
year12to15   5.521e-01  2.371e-01   2.328  0.01990 *
year16to19   1.249e+00  2.132e-01   5.856 4.75e-09 ***
year20to23   1.404e+00  2.100e-01   6.687 2.28e-11 ***
year24to27   1.737e+00  2.038e-01   8.523  < 2e-16 ***
year28to31   2.032e+00  1.997e-01  10.174  < 2e-16 ***
year8to11    2.332e-01  2.528e-01   0.923  0.35614
rads         4.446e-03  1.458e-03   3.050  0.00229 **
rads2       -7.438e-06  4.016e-06  -1.852  0.06404 .
---
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: 214.44

Number of Fisher Scoring iterations: 5
```

```
> fit1_1=glm(death~offset(log(risk))+factor(year)+rads,family=poisson)
> summary(fit1_1)

Call:
glm(formula = death ~ offset(log(risk)) + factor(year) + rads,
    family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.03884  -0.79110  -0.01406   0.54891   3.06044

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.2145521  0.1868854 -17.201  < 2e-16 ***
factor(year)12to15 0.5516986  0.2371116   2.327    0.020 *
factor(year)16to19 1.2482438  0.2132413   5.854 4.81e-09 ***
factor(year)20to23 1.4038777  0.2099958   6.685 2.31e-11 ***
factor(year)24to27 1.7366566  0.2037769   8.522  < 2e-16 ***
factor(year)28to31 2.0311438  0.1997202  10.170  < 2e-16 ***
factor(year)8to11  0.2333271  0.2527737   0.923    0.356
rads               0.0018316  0.0004392   4.170 3.04e-05 ***
---
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:  50.106  on 34  degrees of freedom
AIC: 215.86

Number of Fisher Scoring iterations: 5
```

Since the p-value of *rads*-squared term is 0.06404, which is not significant for critical level 0.05 and it is significant when critical level is 0.1. Also we showed the Poisson log-linear regression model without radiation exposure squared term and do the F-test to check whether the square term should exsit.

```
> 1-pchisq(deviance(fit1_1)-deviance(fit1),df.residual(fit1_1)-df.residual(fit1))
[1] 0.06454215
```

We obtain the p-value 0.06454215>0.05 then we conclude that the square term is not necessary.

(b) Try the same model as in part (a); but instead of treating time after bomb as a factor with seven levels, compute the midpoint of each inter- val 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?

The Poisson log-linear regression model for part b as followed,

```
> time=rep(c(3.5,9.5,13.5,17.5,21.5,25.5,29.5),6)
> fit2=glm(death~offset(log(risk))+log(time)+rads+rads2,family=poisson)
> fit2

Call:  glm(formula = death ~ offset(log(risk)) + log(time) + rads + 
    rads2, family = poisson)

Coefficients:
(Intercept)    log(time)         rads        rads2  
 -5.515e+00    1.223e+00    4.438e-03   -7.417e-06  

Degrees of Freedom: 41 Total (i.e. Null);  38 Residual
Null Deviance:      335.7 
Residual Deviance: 73.68        AIC: 231.4
```

Since the deviance statistic is equal to 73.68 so the deviance statistic is larger in this model.

```
> pchisq(73.68,38,lower.tail=F)
[1] 0.000458099
> pchisq(46.69,33,lower.tail=F)
[1] 0.05754219
```

From R, we observe the p-value when comparing the deviance statistic to a chi-squared distribution on 33 degrees of freedom, 0.05754219 which is for the Poisson log-linear regression model in part a is lager than 0.000458099, the p-value when comparing the deviance statistic to a chi-squared distribution on 38 degrees of freedom which is for the Poisson log-linear regression model in part b. Based on deviance goodness-of-fit, the lager p-value when comparing the deviance statistic to a chi-squared distribution on 33 degrees of freedom which is for the Poisson log-linear regression model in part a indicates that the model in problem a is adequate. Therefore, time cannot 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?

The Poisson log-linear regression model for part c as followed,

```
> interaction=log(time)*rads
> fit3=glm(death~offset(log(risk))+log(time)+rads+rads2+interaction,family=poisson)
> fit3

Call:  glm(formula = death ~ offset(log(risk)) + log(time) + rads +
    rads2 + interaction, family = poisson)

Coefficients:
(Intercept)    log(time)         rads        rads2  interaction
 -5.338e+00    1.164e+00    4.742e-04   -7.577e-06    1.324e-03

Degrees of Freedom: 41 Total (i.e. Null);  37 Residual
Null Deviance:      335.7
Residual Deviance: 72.43        AIC: 232.2
```

```
> summary(fit3)

Call:
glm(formula = death ~ offset(log(risk)) + log(time) + rads +
    rads2 + interaction, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7737  -0.8527  -0.1982   0.5410   3.2844

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.338e+00  3.252e-01 -16.413   <2e-16 ***
log(time)    1.164e+00  1.070e-01  10.879   <2e-16 ***
rads         4.742e-04  3.958e-03   0.120   0.9046
rads2       -7.577e-06  4.025e-06  -1.882   0.0598 .
interaction  1.324e-03  1.232e-03   1.075   0.2823
---
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:  72.426  on 37  degrees of freedom
AIC: 232.18

Number of Fisher Scoring iterations: 5
```

```
> anova(fit3,test='Chisq')
Analysis of Deviance Table

Model: poisson, link: log

Response: death

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           41     335.75
log(time)    1  243.779        40      91.97 < 2.2e-16 ***
rads         1   14.895        39      77.08 0.0001137 ***
rads2        1    3.397        38      73.68 0.0653116 .
interaction  1    1.254        37      72.43 0.2628348
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From R, we can conclude that the interaction significant 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).

```
> pchisq(46.69,33,lower.tail=F)
[1] 0.05754219
> pchisq(73.68,38,lower.tail=F)
[1] 0.000458099
> pchisq(72.43,37,lower.tail=F)
[1] 0.0004421376
```

Comparing the p-value for three models above we conclude that the model in part a is the best good-fitting model within the three models, however based on the conclusion we obtain in part a and part c the radiation exposures square term and the interaction term should not be included in the model so we choose the Poisson log-linear regression model with time after blast treated as a factor (with seven levels) and with *rads* (fit1_1) as our good-fitting model.

```
> summary(fit1_1)

Call:
glm(formula = death ~ offset(log(risk)) + factor(year) + rads,
    family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.03884  -0.79110  -0.01406   0.54891   3.06044

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.2145521  0.1868854 -17.201  < 2e-16 ***
factor(year)12to15  0.5516986  0.2371116   2.327    0.020 *
factor(year)16to19  1.2482438  0.2132413   5.854 4.81e-09 ***
factor(year)20to23  1.4038777  0.2099958   6.685 2.31e-11 ***
factor(year)24to27  1.7366566  0.2037769   8.522  < 2e-16 ***
factor(year)28to31  2.0311438  0.1997202  10.170  < 2e-16 ***
factor(year)8to11   0.2333271  0.2527737   0.923    0.356
rads                0.0018316  0.0004392   4.170 3.04e-05 ***
---
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:  50.106  on 34  degrees of freedom
AIC: 215.86

Number of Fisher Scoring iterations: 5
```

Since the coefficient of Exposure is 0.0018316, thus the effect of Radiation Exposure to the Rate (the number of cancer deaths per person per year) should be

```
> exp(0.0018316)
[1] 1.001833
```

It indicates that one unit increase in radiation exposure, which is the variable rads, would increase the mean number of cancer deaths per person per year by 1.001833.

For the confidence interval of the parameter in my good-fitting model which are radiation exposure and radiation exposure squared are as followed, the formula we use is $e^{\widehat{\beta} \pm 1.96 \times Standard\ Error}$ at 95% significant level.

```
> exp(confint(fit1_1))
Waiting for profiling to be done...
                        2.5 %      97.5 %
(Intercept)         0.0272123  0.05676607
factor(year)12to15 1.0975107  2.79206154
factor(year)16to19 2.3250849  5.38147858
factor(year)20to23 2.7365041  6.25301713
factor(year)24to27 3.8704892  8.62970413
factor(year)28to31 5.2431017 11.50461886
factor(year)8to11  0.7699860  2.08460641
rads                1.0009374  1.00266485
```

Therefore, the confidence interval for the parameter radiation exposure is [1.0009374,1.00266485].

The code followed,

```
# Ada Homework 9

library("Sleuth3")

data=ex2220

year1=data$YearsAfter

year=data$YearsAfter

risk=data$AtRisk

rads=data$Exposure

rads2=rads^2

death=data$Deaths



#a

fit1=glm(death~offset(log(risk))+year+rads+rads2,family=poisson)

fit1_1=glm(death~offset(log(risk))+factor(year)+rads,family=poisson)

summary(fit1_1)

1-pchisq(deviance(fit1_1)-deviance(fit1),df.residual(fit1_1)-df.residual(fit1))

fit1

summary(fit1)

anova(fit1,test='Chi')
```

```
qchisq(0.95,33)

var(fit1$fitted)

mean(fit1$fitted)

var(death)

mean(death)

var(fit1$fitted)\mean(fit1$fitted)


#b

time=rep(c(3.5,9.5,13.5,17.5,21.5,25.5,29.5),6)

fit2=glm(death~offset(log(risk))+log(time)+rads+rads2,family=poisson)

fit2

anova(fit2,test='Chi')

qchisq(0.95,38)


pchisq(73.68,38,lower.tail=F)

pchisq(46.69,33,lower.tail=F)




#c

interaction=log(time)*rads

fit3=glm(death~offset(log(risk))+log(time)+rads+rads2+interaction,family=poisson)

fit3

anova(fit3,test='Chisq')

summary(fit3)
```

```
#d

summary(fit1)

pchisq(46.69,33,lower.tail=F)

pchisq(73.68,38,lower.tail=F)

pchisq(72.43,37,lower.tail=F)

pchisq(50.106,34,lower.tail=F)

exp(0.0018316-1.96*0.0004392)

exp(0.0018316+1.96*0.0004392)

coef(fit1_1)

exp(confint(fit1_1))
```