Poisson regression

Exercises

The table below presents dose-response data concerning the relationship between smoking and lung cancer taken from a publication by Doll. The data in the table were put in the data file DOLL.DAT.

[Man-years at risk, number of cases of lung cancer (in parentheses)]

Cigarettes/day0		**1-9**	**10-14**	**15-19**	**20-24**	**25-34**	**35 +**	*total*
(mean)	(Non-smokers)	(5.2)	(11.2)	(15.9)	(20.4)	(27.4)	(40.8)	
Yrs of smoking								
15-19	10366 (1)	3121	3577	4317	5683	3042	670	30776 (1)
20-24	8162	2937	3286 (1)	4214	6385 (1)	4050 (1)	1166	30200 (3)
25-29	5969	2288	2546 (1)	3185	5483 (1)	4290 (4)	1482	25243 (6)
30-34	4496	2015	2219 (2)	2560 (4)	4687 (6)	4268 (9)	1580 (4)	21825 (25)
35-39	3512	1648 (1)	1826	1893	3646 (5)	3529 (9)	1336 (6)	17390 (21)
40-44	2201	1310 (2)	1386 (1)	1334 (2)	2411(12)	2424(11)	924(10)	11990 (38)
45-49	1421	927	988 (2)	849 (2)	1567 (9)	1409(10)	556 (7)	7717 (30)
50-54	1121	710 (3)	684 (4)	470 (2)	857 (7)	663 (5)	255 (4)	4760 (25)
55-59	826 (2)	606	449 (3)	280 (5)	416 (7)	284 (3)	104 (1)	2965 (21)
total	38074 (3)	15562 (6)	16961 (14)	19102 (15)	31135 (48)	23959 (52)	8073 (32)	152866 (170)

Question 1

Confirm the number of person years and the number of events using the data set `doll`. For this we need to weight the data.

Question 2

Now we are going to estimate a Poisson model. First create the term to use as an offset variable. Now estimate a Poisson model using only an intercept (and the offset). Call the model glm1. What is the interpretation of the estimated coefficient? Also estimate the coefficient by hand using the table.

Question 3

Now fit the model with age category. Call the model glm2. Verify that you can calculate these coefficients also by hand. What can you say about the goodness of fit?

Question 4

Now add the variable smoke. Call the model glm3. What can you say about the goodness of fit?

Question 5

Now we check if the effect of age can be better modeled by a linear trend. Use `as.numeric` to convert the factor to a numeric value. Call the model glm4. Does this model do better than the previous model (use a likelihood ratio test)?

Question 6

Now use `log(as.numeric(age))` to model the effect of age. What about the goodness-of-fit? Does this do better than the previous model?