MATH 523: EXAMPLE - POISSON DATA WITH OFFSET

A Poisson GLM can also be used in situations where a binomial model might seem more appropriate. It is well known that if $X \sim Binomial(m, \pi)$, with m large and π small, then, approximately

$$X \sim Poisson(\mu)$$

with $\mu = m\pi$. Therefore, we might consider a GLM where

$$\log \mu_i = \log m_i + \log \pi_i$$

where π_i is considered a standardized Poisson rate. In this representation, however, as m_i is **fixed**, it must be included in the model with a fixed coefficient, equal to one; such a model term is referred to as an **offset.**

EXAMPLE: The numbers of lung cancer cases in four different smoking groups is recorded, and the variation in the numbers of cases across nine different age groups is inspected. In the study population, the partition into the 4×36 cross-classified subgroups involved a total of 56122 thousand individuals.

- Smoking group: `smoke`
  - 1 Does not smoke
  - 2 Smokes cigars or pipe only
  - 3 Smokes cigarettes and cigar or pipe
  - 4 Smokes cigarettes only
- Age Group: `age`
  - 1-9 for age groups 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+ respectively

			Age Group								
			40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80+
Smoking group	1	Dead	18	22	19	55	117	170	179	120	120
		Popn (thous.)	656	359	249	632	1067	897	668	361	274
	2	Dead	2	4	3	38	113	173	212	243	253
		Popn (thous.)	145	104	98	372	846	949	824	667	537
	3	Dead	149	169	193	576	1001	901	613	337	189
		Popn (thous.)	4531	3030	2267	4682	6052	3880	2033	871	345
	4	Dead	124	140	187	514	778	689	432	214	63
		Popn (thous.)	3410	2239	1851	3270	3791	2421	1195	436	113

Using a canonical log link function with offset, four models can be fitted for the linear predictor,

- Model M_0: Null model
- Model M_1: Smoking group only main effect model
- Model M_2: Age group only main effect model
- Model M_3: Smoking group plus age group additive main effects model

where all models are nested inside M_3, where for $j = 1, \ldots, 4, k = 1, \ldots, 3$

$$\log \mu_{jk} = \log m_{jk} + \alpha_j + \beta_k$$

with the usual identifiability constraints. These models are written

`1, smoke, age, smoke + age`

Fits of the four models are summarized in the following table: p is the total number of parameters in the fitted model, Dev. is residual deviance, and rdf is the residual degrees of freedom.

Model	p	Dev.	rdf
M_0	1	4056.000	35
M_1	4	3910.700	32
M_2	9	191.720	27
M_3	12	21.487	24

On deviance assessment, it is clear that the model M_3 is the most sensible model for these data.

```
smoking<-read.table("smoking.dat",header=T)

smoking$smoke<-gl(4,9)

smoking.m0<-glm(dead~offset(log(pop)),family=poisson,data=smoking)
smoking.m1<-glm(dead~offset(log(pop))+smoke,family=poisson,data=smoking)
smoking.m2<-glm(dead~offset(log(pop))+age,family=poisson,data=smoking)
smoking.m3<-glm(dead~offset(log(pop))+smoke+age,family=poisson,data=smoking)

summary(smoking.m0)
summary(smoking.m1)
summary(smoking.m2)
summary(smoking.m3)
```

Results for Model M_3:

```
Call:
glm(formula = dead ~ offset(log(pop)) + smoke + age, family = poisson,
    data = smoking)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.06055  -0.54773   0.06431   0.29963   1.48348

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.68002    0.06824 -53.929  < 2e-16 ***
smoke2       0.04781    0.04699   1.017    0.309
smoke3       0.21796    0.03869   5.633 1.77e-08 ***
smoke4       0.41696    0.03991  10.447  < 2e-16 ***
age45-59     0.55388    0.07999   6.924 4.38e-12 ***
age50-54     0.98039    0.07682  12.762  < 2e-16 ***
age55-59     1.37946    0.06526  21.138  < 2e-16 ***
age60-64     1.65423    0.06257  26.439  < 2e-16 ***
age65-69     1.99817    0.06279  31.824  < 2e-16 ***
age70-74     2.27141    0.06435  35.296  < 2e-16 ***
age75-79     2.55858    0.06778  37.746  < 2e-16 ***
age80+       2.84692    0.07242  39.310  < 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: 4055.984  on 35  degrees of freedom
Residual deviance:   21.487  on 24  degrees of freedom
AIC: 285.51
Number of Fisher Scoring iterations: 4
```