MATH 60604A
Statistical modelling
§ 4g - Log-linear models with offset

Léo Belzile

HEC Montréal
Department of Decision Sciences

Offsets and comparison of counts

- Up to now, we implicitly assumed that the the count variables Y were comparable between observations.
  - In the fixation example, Y_i represented the number of times that subject i bought the product in the month following the study completion.
- What if the period of follow-up varied from one individual to another?
  - the number of work accidents seen in a business in a given time period depends on the number of its employees.
  - the number of cancer incidence per region depends on the number of inhabitants.

If the counts are not comparable, we can compare the **rates** instead (number of purchases per month, number of accident per employee, etc.)

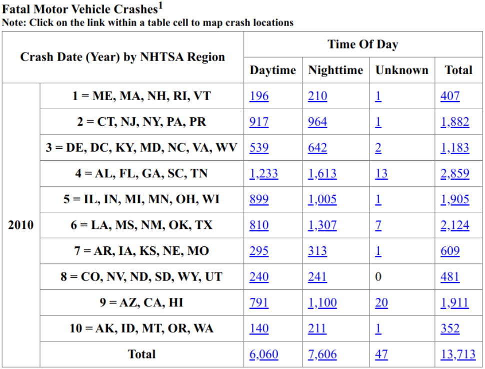
If we model rates with the Poisson model, the latter is adequate **only** if the rate is small.

Car accident example

The National Highway Traffic Safety Administration (NHTSA) compiles statistics about road traffic deaths in the Fatality Analysis Reporting System. The yearly mortality counts for 2010 and 2018 are given in `crash` according to whether the accident occured during daytime or nighttime (`time`), and according to the NHTSA-defined geographic area (`region`).

- Let Y_i denotes the number of death in a given year in region i;
- Let N_i denotes the number of inhabitants in region i.

The goal is to estimate the relationship between the number of fatal car crash and timing of the incident.



Fatal Motor Vehicle Crashes[1]

Note: Click on the link within a table cell to map crash locations

Crash Date (Year) by NHTSA Region		Time Of Day			
		Daytime	Nighttime	Unknown	Total
2010	1 = ME, MA, NH, RI, VT	196	210	1	407
	2 = CT, NJ, NY, PA, PR	917	964	1	1,882
	3 = DE, DC, KY, MD, NC, VA, WV	539	642	2	1,183
	4 = AL, FL, GA, SC, TN	1,233	1,613	13	2,859
	5 = IL, IN, MI, MN, OH, WI	899	1,005	1	1,905
	6 = LA, MS, NM, OK, TX	810	1,307	7	2,124
	7 = AR, IA, KS, NE, MO	295	313	1	609
	8 = CO, NV, ND, SD, WY, UT	240	241	0	481
	9 = AZ, CA, HI	791	1,100	20	1,911
	10 = AK, ID, MT, OR, WA	140	211	1	352
	Total	6,060	7,606	47	13,713

Car accident example

- If we ignore the size of the population, the Poisson regression model (or negative binomial) would be

$$\ln(\mu_i) = \ln\{\mathrm{E}\,(Y_i)\} = \beta_0 + \beta_1 \texttt{time} + \beta_2 \texttt{year}$$

- If we want to account for the size of the population in a given region, we would model Y_i/N_i instead of Y_i.
- This amounts to setting

$$\ln\left\{\frac{\mathrm{E}\,(Y_i)}{N_i}\right\} = \beta_0 + \beta_1 \texttt{time} + \beta_2 \texttt{year}$$

or equivalently

$$\ln\left\{\mathrm{E}\,(Y_i)\right\} = \beta_0 + \beta_1 \texttt{time} + \beta_2 \texttt{year} + \ln(N_i)$$

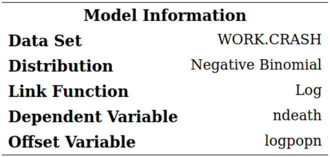
- The term $\ln(N_i)$ is called an **offset** since it is included as a covariate, but there is no β coefficient to estimate (unity).

Car accident

SAS code to include an offset term

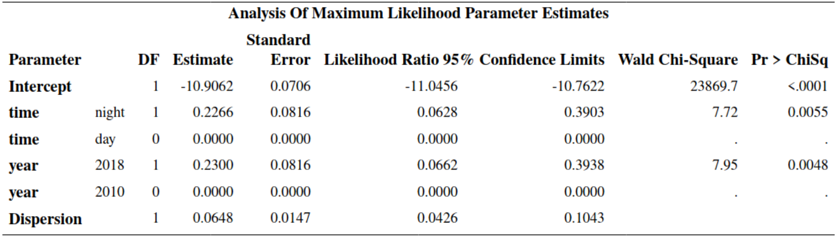
```
data crash;
set statmod.crash;
logpopn=log(popn);
run;

proc genmod data=crash;
class time(ref="day") year(ref="2010");
model ndeath=time year / dist=negbin link=log
    offset=logpopn type3 lrci;
run;
```

The option `offset` could also be included for a Poisson model.



Model Information	
Data Set	WORK.CRASH
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	ndeath
Offset Variable	logpopn

Parameter interpretation with offset



Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-10.9062	0.0706	-11.0456	-10.7622	23869.7	<.0001
time	night	1	0.2266	0.0816	0.0628	0.3903	7.72	0.0055
time	day	0	0.0000	0.0000	0.0000	0.0000	.	.
year	2018	1	0.2300	0.0816	0.0662	0.3938	7.95	0.0048
year	2010	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.0648	0.0147	0.0426	0.1043		

- Since the variable $\ln(N)$ is included as an offset, it doesn't appear in the table.
- The deviance statistic (output not shown) is 40.269 for 37 degrees of freedom (ratio of 1.0884). The corresponding p-value is 0.327, so there is no evidence that our fitted model is inadequate.
- In this setting, $\exp(\widehat{\beta}_0) = \exp(-10.9062)$ corresponds to the estimated mortality rate during daytime in 2010, which is $1.83/100000$, i.e., a rate of 1.83 per $100\,000$ inhabitants (with 95% confidence interval $[1.60 \times 10^{-5}, 2.12 \times 10^{-5}]$).
- There is a $\exp(0.23)$ change in mortality from 2010 to 2018, corresponding to a 26% increase in road casualties.