MATH 60604A
Statistical modelling
§ 4d - Poisson regression

Léo Belzile

HEC Montréal
Department of Decision Sciences

Poisson regression

- Poisson regression assumes that the outcome variable Y_i follows a Poisson distribution with parameter μ_i, $Y_i \sim \mathrm{Po}(\mu_i)$, where

$$\mu_i = \mathrm{E}\left(Y_i\right) = \mathrm{Var}\left(Y_i\right).$$

- We use the natural logarithm $\ln(x)$ as link function,

$$g\{\mathrm{E}\left(Y_i\right)\} = g(\mu_i) = \ln\{\mathrm{E}\left(Y_i\right)\} = \beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}.$$

- Equivalently, we could say that the outcome for individual i, Y_i, follows a Poisson distribution with mean

$$\mathrm{E}\left(Y_i\right) = \mu_i = \exp(\beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}).$$

Coefficient interpretation for β_k in Poisson regression

- Let $\boldsymbol{x}$, $\boldsymbol{x}_+$ be two vectors which differ only in their kth components, respectively x_k and $x_k + 1$.

When $\mathbf{X} = \boldsymbol{x}$, the model linking the mean to the variable Y is

$$\mu_i(\boldsymbol{x}) = \mathsf{E}\left(Y_i \mid \mathbf{X} = \boldsymbol{x}\right) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right),$$

whereas, when $\mathbf{X} = \boldsymbol{x}_+$, we have

$$\mu_i(\boldsymbol{x}_+) = \mathsf{E}\left(Y_i \mid \mathbf{X} = \boldsymbol{x}_+\right) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \beta_k\right).$$

- The ratio between two means, $\mu_i(\boldsymbol{x}_+)/\mu_i(\boldsymbol{x})$, is $\exp(\beta_k)$.
- When X_k increases by one unit, the mean of Y is **multiplied** by $\exp(\beta_k)$.

Fitting a Poisson regression using `proc genmod`

We now consider a Poisson model for the number of items bought by participants following the advertisement.

SAS code to fit a Poisson regression

```
proc genmod data=statmod.intention;
class educ revenue;
model nitem=sex age revenue educ marital
     fixation emotion / dist=poisson link=log
     lrci type3;
run;
```

Likelihood ratio tests for global significance

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
sex	1	10.97	0.0009
age	1	1.58	0.2089
revenue	2	65.18	<.0001
educ	2	4.11	0.1281
marital	1	7.40	0.0065
fixation	1	76.25	<.0001
emotion	1	30.61	<.0001

Five explanatory variables are statistically significant according to the likelihood ratio tests.

Parameter estimates for Poisson regression

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.6305	0.6618	-2.9427	-0.3466	6.07	0.0137
sex		1	0.5361	0.1649	0.2168	0.8640	10.57	0.0011
age		1	-0.0228	0.0183	-0.0590	0.0127	1.56	0.2115
revenue	1	1	1.2463	0.2461	0.7712	1.7374	25.64	<.0001
revenue	2	1	-0.1250	0.2532	-0.6213	0.3736	0.24	0.6216
revenue	3	0	0.0000	0.0000	0.0000	0.0000	.	.
educ	1	1	0.2497	0.2226	-0.1800	0.6948	1.26	0.2620
educ	2	1	0.4040	0.2044	0.0123	0.8159	3.90	0.0482
educ	3	0	0.0000	0.0000	0.0000	0.0000	.	.
marital		1	-0.4218	0.1558	-0.7291	-0.1175	7.33	0.0068
fixation		1	0.5501	0.0614	0.4296	0.6708	80.16	<.0001
emotion		1	0.7887	0.1396	0.5133	1.0610	31.92	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

The scale parameter is unity because it is completely determined by the mean-variance relationship.

Interpretation of significant parameters

- The estimate $\widehat{\beta}_{\texttt{sex}} = 0.5361$, meaning that women made more purchases than men, on average. When the other variables remain constant, the mean number of purchases for women is $\exp(0.5361) = 1.71$ times that of men. So, the mean for women is 71% higher than for men.
- The parameter estimate for `fixation` is $\widehat{\beta}_{\texttt{fixation}} = 0.5501$ and it's significantly different from 0. The higher the value of `fixation`, the higher the number of purchases, on average. When the other variables remain constant, increasing `fixation` by one unit means the mean number of purchases is multiplied by $\exp(0.5501) = 1.73$.
- *Ceteris paribus*, the average number of items bought by people with low revenue is 3.47 higher than those with high income, a relative mean increase of 247%.

Goodness of fit

- The SAS output includes a table containing the log-likelihood (full log-likelihood) and information criteria.
- For the Poisson regression model, the deviance and Pearson X^2 statistics are two goodness-of-fit indicators used to determine if the model is adequate.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	110	203.2710	1.8479
Scaled Deviance	110	203.2710	1.8479
Pearson Chi-Square	110	216.2705	1.9661
Scaled Pearson X2	110	216.2705	1.9661
Log Likelihood		3.2104	
Full Log Likelihood		-186.1639	
AIC (smaller is better)		392.3279	
AICC (smaller is better)		394.3462	
BIC (smaller is better)		420.2028	

The first two lines are duplicated; the Poisson model has no separate scale parameter, since the variance is fully determined by the mean (unlike linear regression).