MATH 60604A
Statistical modelling
§ 4f - Overdispersed count data

Léo Belzile

HEC Montréal
Department of Decision Sciences

Extensions to Poisson to deal with overdispersion

- The Poisson distribution is not very flexible, because it only includes one parameter, which is equal to both the mean and the variance.
- In most cases, this assumption is not valid. In the previous output, the deviance divided by the degrees of freedom was $203.2710/110 = 1.85$, suggesting the Poisson model is **not adequate** (p-value less than 10^{-5}).
- The underlying reason is that the observed variability in counts is much larger than the mean in this example, a phenomenon termed **overdispersion**.
- The negative binomial model if often used as replacement for overdispersed count data.

Negative binomial distribution

- The negative binomial distribution is a probability distribution for integer random variables with two parameters.
- We restrict attention the most common parametrization used in modelling. The probability mass function is

$$\mathsf{P}\left(Y=y\right)=\frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)}\left(\frac{1/k}{1/k+\mu}\right)^{1/k}\left(\frac{\mu}{1/k+\mu}\right)^{y}$$

  for $y=0,1,2,3,\ldots$, where Γ denotes the gamma function. Both parameters are positive, meaning $\mu>0$ and $k>0$.
- The mean and the variance are

$$\mathsf{E}\left(Y\right)=\mu, \qquad \mathsf{Var}\left(Y\right)=\mu+k\mu^2.$$

- The variance of the negative binomial distribution is always larger than its mean.

Negative binomial regression

- Negative binomial regression usually assumes that the response variable Y follows a negative binomial distribution and that the link function is the logarithmic function

$$g\{\mathsf{E}\left(Y_i\right)\} = \log\{\mathsf{E}\left(Y_i\right)\} = \beta_0 + \beta_1 \mathrm{X}_{i1} + \ldots + \beta_p \mathrm{X}_{ip}.$$

- Equivalently, we assume that each observation Y_i follows a negative binomial distribution with mean

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

- The interpretation of the parameters is the same as for Poisson regression.
- There is a second parameter, k, which is assumed to be the same for every observation and therefore doesn't depend on the predictor variables.

Mathematical aside: The negative binomial model is not a generalized linear model per say because it is part of exponential-dispersion family, but we can use maximum likelihood and the GLM machinery to fit the model.

Negative binomial regression with `proc genmod`

The only difference from the Poisson model is that we specify `dist=negbin`.

SAS code to fit a negative binomial model

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

In R, the parametrization of `MASS::glm.nb` is such that $\theta = 1/k$.

Goodness-of-fit diagnostics for negative binomial

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	110	118.2310	1.0748
Scaled Deviance	110	118.2310	1.0748
Pearson Chi-Square	110	119.5504	1.0868
Scaled Pearson X2	110	119.5504	1.0868
Log Likelihood		14.7494	
Full Log Likelihood		-174.6250	
AIC (smaller is better)		371.2501	
AICC (smaller is better)		373.6945	
BIC (smaller is better)		401.9125	

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
sex	1	3.80	0.0513
age	1	2.23	0.1350
revenue	2	19.68	<.0001
educ	2	2.11	0.3481
marital	1	2.61	0.1061
fixation	1	35.54	<.0001
emotion	1	12.15	0.0005

The deviance over degrees of freedom is closer to unity. Only `revenue`, `fixation` and `emotion` are statistically significant.

Parameter estimates for the negative binomial model

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.1761	0.9729	-3.1103	0.7640	1.46	0.2267
sex		1	0.5077	0.2550	-0.0029	1.0155	3.96	0.0465
age		1	-0.0415	0.0281	-0.0990	0.0130	2.18	0.1395
revenue	1	1	1.1053	0.3521	0.4124	1.8148	9.86	0.0017
revenue	2	1	-0.1617	0.3535	-0.8660	0.5377	0.21	0.6473
revenue	3	0	0.0000	0.0000	0.0000	0.0000	.	.
educ	1	1	0.3645	0.3441	-0.3263	1.0500	1.12	0.2895
educ	2	1	0.4386	0.3041	-0.1624	1.0494	2.08	0.1492
educ	3	0	0.0000	0.0000	0.0000	0.0000	.	.
marital		1	-0.3873	0.2369	-0.8593	0.0850	2.67	0.1021
fixation		1	0.6316	0.1056	0.4338	0.8581	35.81	<.0001
emotion		1	0.7570	0.2127	0.3401	1.1902	12.66	0.0004
Dispersion		1	0.5840	0.2119	0.2564	1.1193		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

The scale parameter $\widehat{k} = 0.584$. Note that the likelihood-ratio based 95% confidence interval may lead to different inference than the Wald tests and their p-values; prefer the former as they are more reliable.

Model selection

- The deviance indicates that the negative binomial model is preferable to the Poisson, but this is informal.
- Another to answer this would be to look at information criteria (smaller is better): the negative binomial model is selected by both AIC and BIC.

Model	Poisson	neg. binom.
AIC	392.33	371.25
BIC	420.20	301.91

Negative binomial distribution versus Poisson

- As k approaches zero, we recover the Poisson distribution.
- We can actually compare these two models using the likelihood ratio test since they are nested.
- We can test the hypotheses $\mathscr{H}_0 : k = 0, \mathscr{H}_1 : k \neq 0$ using a likelihood ratio test
  - beware! the null distribution is **non-regular** because when $n \to \infty$, there is a 0.5 probability that the deviance will be exactly zero and 0.5 that it follows a χ^2_1 under $\mathscr{H}_0$.
- The asymptotic null distribution is

$$2\{\ell_{\text{negbin}}(\widehat{\mu}_{\text{negbin}}, \widehat{k}) - \ell_{\text{pois}}(\widehat{\mu}_{\text{pois}})\} \dot{\sim} \frac{1}{2}\chi^2_1 + \frac{1}{2}\delta_0;$$

Practical aspect: if we do not observe $\widehat{k} = 0$, we calculate the p-value as usual using the χ^2_1 distribution and **divide it by two** to get the **correct result**.

Likelihood ratio test (non-regular)

This shows how to to the calculations by hand using the output from the tables.

- The "**Full Log Likelihood**" give the fitted likelihood of the model, -174.6250 for the negative binomial model and -186.1639 for the Poisson model.
- The difference is 11.5389 and the likelihood ratio statistic is 23.08.
- The probability that a χ^2_1 is larger than 23.08 is 1.55×10^{-7}.
- Since the problem is non-regular, we halve this probability and so our p-value is 7.7×10^{-8}.

SAS code for likelihood ratio test (non-regular)

```
data pval;
pval=(1-CDF('CHISQ',23.08,1))/2;
run;
proc print data=pval;
run;
```

There is overwhelming evidence that the negative binomial model is preferable.