Offsets and Overdispersion

Patrick Breheny

April 11
The meaning of \( \lambda \) often requires additional thought.

When we employ a Poisson model, what we are modeling is the rate of events.

We need to be careful about specifying what we are estimating: a rate per what?

For example, if we are modeling motor vehicle crashes, we may be estimating a rate per 1,000 population, a rate per 1,000 licensed drivers, a rate per 1,000 registered motor vehicles, or a rate per 100,000 miles traveled.
A kind of rate that is particularly common in epidemiological studies is a rate per person-years of follow-up.

For example, consider the classic study by Doll et al. in which all British male doctors were sent a questionnaire about their age and whether they smoked tobacco.

The doctors were then followed up for a number of years to see whether or not they had died from coronary heart disease.
Suppose, then, that we wish to model \( \lambda(x) \), the rate per 1,000 person-years of follow-up, given the explanatory variables Age and Smoking.

Now, \( E(Y_i) = t_i \lambda_i \),

where \( t_i \) denotes the person-years of follow-up for observation \( i \).

This implies that

\[
\log(\mu_i) = \log(t_i) + \log(\lambda_i) = \log(t_i) + \eta_i;
\]

thus, the usual relationship between \( \mu_i \) and the linear predictor is offset by the amount \( \log(t_i) \).
Both R and SAS allow you to specify an offset.

In SAS, one simply adds the option OFFSET= to the model statement.

Similarly, in R, one specifies the offset= option in the glm function.

Note: In SAS, one must compute the offset in a separate DATA step, while in R, one can submit code such as offset=log(PersonYears/1000).
Estimating linear combinations

- We can then estimate the rate per 1,000 person-years of follow-up for any category we choose using either the `ESTIMATE` statement in SAS or the `predict` function in R.

- For example, with SAS’s default coding of class variables, the following statement estimates the rate of CHD deaths for smokers aged 45–54:

  ```
  ESTIMATE '45-54 smokers' Intercept 1
  Age 0 1 0 0 0
  Smoking 0 1;
  ```

- In R, we can set up a data frame consisting of all the linear combinations we are interested in, and then submit

  ```
  predict(fit,df,type="response")
  ```

- Note: In SAS, the offset is set to zero; in R, you specify the offset variable.
Estimated rates

The estimated rates from our Poisson regression model:

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Smokers</th>
<th>Non-smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>35–44</td>
<td>0.52</td>
<td>0.36</td>
</tr>
<tr>
<td>45–54</td>
<td>2.29</td>
<td>1.60</td>
</tr>
<tr>
<td>55–64</td>
<td>7.17</td>
<td>5.03</td>
</tr>
<tr>
<td>65–74</td>
<td>14.78</td>
<td>10.37</td>
</tr>
<tr>
<td>75–84</td>
<td>20.97</td>
<td>14.71</td>
</tr>
</tbody>
</table>

Note that, by fitting a model with no interaction between age and smoking, we enforce that the rate ratio (RR) between smokers and non-smokers are the same in each age group (0.52/0.36 = ⋯ = 20.97/14.71 = 1.43)
If we allow an interaction, we obtain

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Smokers</th>
<th>Non-smokers</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>35–44</td>
<td>0.61</td>
<td>0.11</td>
<td>5.5</td>
</tr>
<tr>
<td>45–54</td>
<td>2.40</td>
<td>1.12</td>
<td>2.1</td>
</tr>
<tr>
<td>55–64</td>
<td>7.20</td>
<td>4.90</td>
<td>1.5</td>
</tr>
<tr>
<td>65–74</td>
<td>14.69</td>
<td>10.83</td>
<td>1.4</td>
</tr>
<tr>
<td>75–84</td>
<td>19.18</td>
<td>21.20</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Poisson regression is an adequate tool for analyzing cohort studies; however, if one has detailed individual-level data, one can apply the more sophisticated approaches that have been developed in the field of survival analysis.
One of the defining characteristics of Poisson regression is its lack of a scale parameter: $E(Y) = \text{Var}(Y)$, and no parameter is available to adjust that relationship.

In practice, when working with Poisson regression, it is often the case that the variability of $y_i$ about $\hat{\lambda}_i$ is larger than what $\hat{\lambda}_i$ predicts.

This implies that there is more variability around the model’s fitted values than is consistent with the Poisson distribution.
The term for this phenomenon is *overdispersion*. Data for which this phenomenon manifests itself are often called “overdispersed”, although as we will see, it is perhaps better to refer to the model as overdispersed, not the data. There are two common approaches to correcting for overdispersion:

- Quasi-likelihood
- Negative binomial regression
Recall that the score arising from a Poisson regression model is

$$\frac{\partial \ell}{\partial \theta} = \sum_i \{y_i - \hat{\lambda}_i\}$$

where $\theta = \log(\lambda)$, the canonical parameter.

Note, of course, that there is no scale parameter, which would show up in the denominator on the right hand side.

Now suppose we add one:

$$\frac{\partial \ell}{\partial \theta} = \sum_i \frac{y_i - \hat{\lambda}_i}{\phi}$$
Recall that $\text{Var}(Y) = \phi V(\mu)$; thus, we now have a parameter that allows the variance to be larger or smaller than the mean by a multiplicative factor $\phi$.

This will not change $\hat{\beta}$, of course.

However, it will affect inference, since

$$\hat{\beta} \sim N(\beta, \phi(X^T WX)^{-1})$$
So what distribution is this, that gives rise to this score?

There isn’t one (at least, not one for which you can write down the distribution in closed form)

This approach, where you modify the score directly and never actually specify a distribution, is known as quasi-likelihood
Typically, the scale parameter $\phi$ is estimated using the method of moments estimator

$$\hat{\phi} = \frac{X^2}{n - p}$$

To use this approach in $\mathbb{R}$, one can specify `family=quasipoisson`; in SAS, one can add a `PSCALE` option to the model statement.
For our Belgian AIDS data, \( \hat{\phi} = 6.7 \), implying that the variance was nearly 7 times larger than that implied by the Poisson distribution.

Again, the fit is the same.

However, our standard errors are \( \sqrt{6.7} \approx 2.6 \) times larger.
Quasi-likelihood: Belgian AIDS data (cont’d)
The quasi-Poisson approach is attractive for several reasons, but its big drawback is that lacks a log-likelihood. This prevents you from using any of the likelihood-based tools we have discussed for GLMs: likelihood ratio tests, AIC/BIC, deviance explained, deviance residuals. An alternative approach that allows all those maximum likelihood tools is based on the negative binomial distribution.
The negative binomial distribution

- The negative binomial distribution has other uses in probability and statistics, but for our purposes we can think about it as arising from a two-stage hierarchical process:

\[ Z \sim \text{Gamma}(\theta, \theta) \]
\[ Y|Z \sim \text{Poisson}(\lambda Z) \]

- The marginal distribution of \( Y \) is then negative binomial, with

\[ \text{E}(Y) = \lambda \]
\[ \text{Var}(Y) = \lambda + \frac{\lambda^2}{\theta} \]

- Thus, like the Poisson distribution, the negative binomial has support only on the positive integers, but unlike the Poisson, its variance is larger than its mean
Note, however, that the negative binomial distribution is not a member of the exponential family.

Thus, the theory and fitting procedures we have developed for GLMs do not directly apply here.

For example, there is no “canonical link”; however, it is customary to employ a log link to make negative binomial regression look like Poisson regression.

Regardless, PROC GENMOD in SAS allows the choice of DIST=NB for negative binomial models; in R, one must use the glm.nb function in the MASS package.
For the Belgian AIDS data, $\hat{\theta} = 19.2$, implying the following mean-variance relationship:
This leads to the following:
Arguably, the negative binomial estimates are even worse than the Poisson estimates, and certainly drastically worse than the quadratic Poisson model.

However, its “goodness of fit” measures are much better.

This is why I remarked earlier that it’s wrong to think of the data as overdispersed – if the data show more variability than the model can explain, the most likely explanation is a bad model.

The quadratic Poisson fit shows no overdispersion (the residuals are actually slightly “underdispersed”.)
Accounting for overdispersion is a good idea – if the model doesn’t fit the data, this should be reflected with larger standard errors and wider confidence intervals.

However, many analysts have the view that quasi-Poisson or negative binomial regression automatically “fixes” the overdispersion problem.

This is a potentially dangerous misconception – surely, accurately modeling the mean is of greater priority than modeling the variance.

While quasi-Poisson and negative binomial approaches are useful, they are certainly no substitute for careful consideration of the systematic component of the model.