An example of count data analysis

Gilbert Berdine MD, Shengping Yang PhD

Previously, we have introduced Poisson and Negative binomial regressions for modeling count data. Here we will use a real example to demonstrate how to use SAS software performing such analyses.

A new oral antibiotic drug *Gorilacillin* was developed and has had excellent effects in several clinical trials. *Gorilacillin* has two side effects, including rash and elevated liver function tests (ELFT). To evaluate whether different patient groups have different risks of having such side effects, a *Gorilacillin* side effect study was conducted. A total of 5,275 participants were recruited from five participating countries. All the patients were followed for up to one month, and the number of patients who developed rash or ELFT was recorded. The goal of the study was to investigate whether there was a significant difference in developing side effects for patients in different age groups. Data from the study were collected and stored in an Excel table (see below for a partial view of the table).

Country	Age	# Patients	# Rash	# ELFT
Great Britain	(0-4)	65	1	0
Great Britain	(5-9)	18	0	0
Great Britain	(10-14)	229	1	1
Great Britain	(15-19)	59	1	1
Great Britain	(20-24)	65	0	0
Great Britain	(25-29)	49	2	0
…				

It is typical to use a Poisson or negative binomial regression for analyzing such data since the outcome is a side effect count, and the probability for developing side effects is low – rare events. Meanwhile, because rash and ELFT are the two main side effects of Gorilacillin, an event is declared if a patient develops either rash or ELFT. Note that the numbers of patients in different age groups/countries are different; thus it makes sense to model the rate of side effect per patient, as a function of age group and country, to adjust for the differences in number of patients.

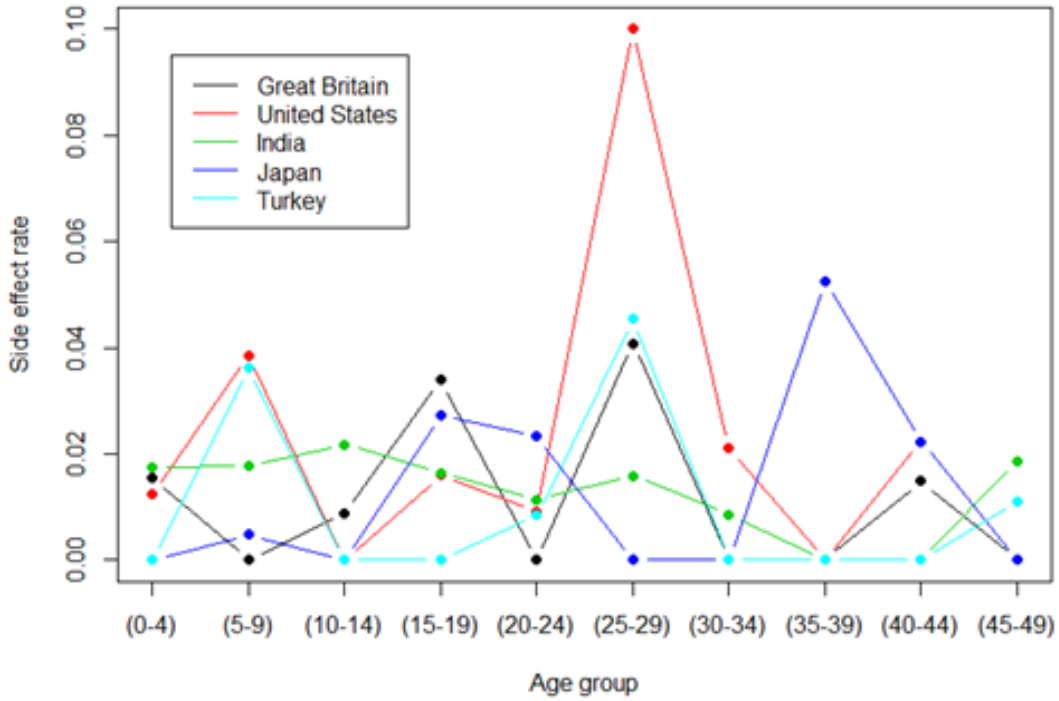
A scatter plot can usually help visualize potential relationships. As we can see from Figure 1, there does not seem to have a strong relationship between side effect rate and age group. Also the side effect rates are quite similar among countries.

To apply a Poisson regression, we have,

$$log(\mu/n) = \beta_0 + \beta_1 age_{0-4} + \beta_2 age_{5-9} + ... + \beta_{10} country_{GB} + ...$$

Corresponding author: Gilbert Berdine
Contact Information: gilbert.berdine@ttuhsc.edu.
DOI: 10.12746/swrccc2015.0311.149

Figure 1. *Side effect Rate by age group and country*



To adjust for the numbers of patients in different age groups/countries, we use the rate of side effect (dividing the expected number of events by the number of patients) as the outcome variable. Equivalently, the above equation can also be written as,

$$log(\mu) = \beta_0 + \beta_1 age_{0-4} + \beta_2 age_{5-9} + ... + \beta_{10} country_{GB} + ... + log(n),$$

where the additional term on the right-hand side, $log(n)$, is called an offset.

Corresponding to the above two models, there are two equivalent SAS statements:

```
proc genmod data=data;
class country age;
model incident/n = age country / dist= poisson link=log;
lsmeans age / ilink diff cl;
run;
```

Or equivalently,

```
proc genmod data=data;
class country age;
model incident = age country / offset= logn dist=poisson
          link=log;
lsmeans age / ilink diff cl;
run;
```

Note that, in the second equation, $log\ n = log(n)$, where n is the number of patients in a specific group. The lsmeans statement can be used to obtain the side effect rate estimates for the 10 age groups, averaged over countries. The ilink option specifies the inverse link function to be used for calculating the rate estimates, and the cl option produces the confidence intervals. In addition, the diff option provides all pairwise comparisons of side effect rates among age groups.

The above SAS output table shows that age group was not significantly associated with *Gorilacillin* side effect rate and there was no significant difference among the 5 countries.

From the lsmeans estimates, we see that the estimated side effect rate for patents 0-4 years old was 1.1% (the Mean column; table above) with a confidence interval (0.6%, 2.2%; the Lower Mean

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.8135	0.3599	-4.5189	-3.1081	112.28	<.0001
Age	**(0-4)**	1	-0.5275	0.4444	-1.3984	0.3435	1.41	0.2352
Age	**(10-14)**	1	-0.8328	0.6796	-2.1647	0.4992	1.50	0.2204
Age	**(15-19)**	1	0.0258	0.3575	-0.6749	0.7266	0.01	0.9424
Age	**(20-24)**	1	-0.3447	0.4277	-1.1830	0.4936	0.65	0.4203
Age	**(25-29)**	1	0.6658	0.4733	-0.2618	1.5934	1.98	0.1595
Age	**(30-34)**	1	-0.4745	0.5735	-1.5986	0.6495	0.68	0.4080
Age	**(35-39)**	1	-0.3067	0.6396	-1.5603	0.9469	0.23	0.6316
Age	**(40-44)**	1	-0.4021	0.5016	-1.3853	0.5810	0.64	0.4227
Age	**(45-49)**	1	-0.9098	0.5660	-2.0192	0.1995	2.58	0.1080
Age	**(5-9)**	0	0.0000	0.0000	0.0000	0.0000	.	.
country	**Great Britain**	1	-0.1291	0.4409	-0.9931	0.7350	0.09	0.7697
country	**India**	1	-0.2636	0.3141	-0.8792	0.3520	0.70	0.4013
country	**Japan**	1	-0.1973	0.3572	-0.8974	0.5027	0.31	0.5806
country	**Turkey**	1	-0.1805	0.4629	-1.0877	0.7268	0.15	0.6967
country	**United States**	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

and Upper Mean columns), and for the 5-9 years old it was 1.9% with a confidence interval (1.1%, 3.3%), etc.

The diff option does provide all pairwise comparisons should such comparisons be of interest (table below shows part of the comparisons).

Now, recall that we previously explained that a negative binomial regression model might be more appropriate should data overdispersion exist. To test overdispersion, an easy way is to apply a negative binomial regression with scale=0 and noscale options in the model statement. These options test whether overdispersion of the form $\mu+k\mu^2$ exists by testing whether the dispersions parameter equals to 0.

```
proc genmod data=data;
class country age;
model incident/n = age country / dist= nb link=log;
scale=0 noscale;
run;
```

Age Least Squares Means											
Age	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Mean	Standard Error of Mean	Lower Mean	Upper Mean
(0-4)	-4.4951	0.3556	-12.64	<.0001	0.05	-5.192	-3.798	0.01116	0.003970	0.00556	0.0224
(5-9)	-3.9676	0.2778	-14.28	<.0001	0.05	-4.512	-3.423	0.01892	0.005256	0.01098	0.0326
(10-14)	-4.8004	0.6035	-7.95	<.0001	0.05	-5.983	-3.617	0.00822	0.004965	0.00252	0.0268
(15-19)	-3.9418	0.2602	-15.15	<.0001	0.05	-4.451	-3.432	0.01941	0.005051	0.01166	0.0323
(20-24)	-4.3123	0.3369	-12.80	<.0001	0.05	-4.972	-3.652	0.01340	0.004516	0.00692	0.0259
(25-29)	-3.3018	0.3809	-8.67	<.0001	0.05	-4.048	-2.555	0.03682	0.014020	0.01745	0.0776
(30-34)	-4.4421	0.5144	-8.64	<.0001	0.05	-5.450	-3.434	0.01177	0.006055	0.00429	0.0322
(35-39)	-4.2743	0.5877	-7.27	<.0001	0.05	-5.426	-3.122	0.01392	0.008182	0.00440	0.0440
(40-44)	-4.3698	0.4182	-10.45	<.0001	0.05	-5.189	-3.550	0.01265	0.005292	0.00557	0.0287
(45-49)	-4.8775	0.5061	-9.64	<.0001	0.05	-5.869	-3.885	0.00762	0.003854	0.00282	0.0205

Differences of Age Least Squares Means								
Age	_Age	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper
(0-4)	(10-14)	0.3053	0.7116	0.43	0.6679	0.05	-1.0894	1.7000
(0-4)	(15-19)	-0.5533	0.4237	-1.31	0.1916	0.05	-1.3837	0.2771
(0-4)	(20-24)	-0.1828	0.4870	-0.38	0.7074	0.05	-1.1372	0.7716
(0-4)	(25-29)	-1.1933	0.5264	-2.27	0.0234	0.05	-2.2250	-0.1616
(0-4)	(30-34)	-0.0529	0.6021	-0.09	0.9299	0.05	-1.2330	1.1271
(0-4)	(35-39)	-0.2208	0.6694	-0.33	0.7415	0.05	-1.5329	1.0912
(0-4)	(40-44)	-0.1253	0.5363	-0.23	0.8152	0.05	-1.1764	0.9257
(0-4)	(45-49)	0.3824	0.6129	0.62	0.5327	0.05	-0.8189	1.5837
(0-4)	(5-9)	-0.5275	0.4444	-1.19	0.2352	0.05	-1.3984	0.3435
(10-14)	(15-19)	-0.8586	0.6681	-1.29	0.1988	0.05	-2.1681	0.4509
…								
…	…		…	…	…	…	…	…

From the above test of overdispersion result, we can see that the p value is less than 0.0001, and thus it is appropriate to use a negative binomial regression.

Lagrange Multiplier Statistics		
Parameter	Chi-Square	Pr > ChiSq
Dispersion	8309.4881	<.0001*
** One-sided p-value*		

The result from the negative binomial regression (table above) is similar to that from the Poisson regression. We did not detect any difference in side effect rate between the reference and other age groups. Looking at the raw data in the scatter plot (Figure 1), one might think that Americans of age 25-29 were at risk for adverse effects of the drug, but the statistical analysis shows the result to be within the 95% confidence limit for purely random effect compared to the reference group. The American 25-29 data point appears, at first glance, to be an outlier with some non-random effect, but, in fact, it is a purely random

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.6458	0.3249	-4.2826	-3.0091	125.94	<.0001
Age	(0-4)	1	-0.5332	0.4053	-1.3276	0.2612	1.73	0.1883
Age	(10-14)	1	-0.6927	0.5926	-1.8542	0.4688	1.37	0.2424
Age	(15-19)	1	-0.1166	0.3365	-0.7761	0.5429	0.12	0.7289
Age	(20-24)	1	-0.4504	0.3971	-1.2287	0.3279	1.29	0.2567
Age	(25-29)	1	0.6895	0.4348	-0.1627	1.5417	2.51	0.1128
Age	(30-34)	1	-0.4385	0.5165	-1.4509	0.5738	0.72	0.3959
Age	(35-39)	1	0.4003	0.4448	-0.4715	1.2721	0.81	0.3681
Age	(40-44)	1	-0.3026	0.4460	-1.1768	0.5715	0.46	0.4974
Age	(45-49)	1	-0.3803	0.4186	-1.2009	0.4402	0.83	0.3636
Age	(5-9)	0	0.0000	0.0000	0.0000	0.0000	.	.
country	Great Britain	1	-0.1721	0.4080	-0.9717	0.6275	0.18	0.6731
country	India	1	-0.2972	0.2858	-0.8573	0.2629	1.08	0.2983
country	Japan	1	0.0361	0.3020	-0.5558	0.6280	0.01	0.9048
country	Turkey	1	-0.1599	0.4131	-0.9695	0.6498	0.15	0.6987
country	United States	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	1.0492	0.0141	1.0219	1.0773		

walk from the other data points.

The statistical analysis is consistent with the reality of the situation. *Gorilacillin* does not exist and the data was simulated by sampling rare occurrence events from an online game in which the game developers assure us that the events are, indeed, random. The game has generated all sorts of "theories" about how to elicit these rare events more often, but the statistical analysis shows the "theories" to be no more substantial than Americans age 25-29. This example illustrates how rare events can seem to generate "outliers" that are merely results of small samples and rare occurrence rates.

Many times, rare events are hard to observe, and it might take quite some time before one event is observed. If feasible, one alternative strategy of studying an association between a rare event and potential risk factors is to collect data retrospectively. For example, identify the list of patients who had the event, match them with those who did not have the event, then collect all the necessary data and perform data analysis.

Author affiliation : Gilbert Berdine is a pulmonary physician in the Department of Internal Medicine at TTUHSC. Shengping Yang is a biostatistician in the Department of Pathology at TTHUSC.
Published electronically: 7/15/2015