Cornell University
Cornell Statistical Consulting Unit

StatNews #86

Poisson Regression: Lack of Fit ≠ Overdispersion

March 2013

In statistical analysis of count data, it is often assumed that the dependent variable follows a Poisson distribution. This implies that the mean (the expected count) is equal to the variance. In practice, however, one often observes that the variance is much larger than the mean. This is often referred to as "overdispersion" with respect to the Poisson distribution. Statistical software packages make it very easy to specify a more flexible model that allows for the variance to be larger than the mean, for example, by adding an overdispersion parameter to model this extra variance or by assuming that the dependent variable follows a negative binomial distribution. However, this approach may be inappropriate and may lead to biased regression estimates, if the real reason for the larger-than-expected variance is a misspecified model ("lack of fit"). The objective of this newsletter is to clarify that "lack of fit" should not be confused with "overdispersion".

In Poisson regression, the logarithm of the expected count is assumed to be a linear function of some predictors, $\log(\lambda_i) = \beta_0+\beta_1 x_{1i}+...+\beta_k x_{ki}$ where λ_i is the expected count of the i^{th} observation. In the case of Poisson regression, lack of fit means that the log of the expected counts cannot be predicted by $\beta_0+\beta_1 x_{1i}+...+\beta_k x_{ki}$. To illustrate that lack of fit should not be confused with overdispersion, we use simulated data (n=100) in which the correct relationship between the expected counts and a single predictor is $\log(\lambda_i) = x_i^2$ but we fit a (misspecified) log-linear model $\log(\lambda_i) = \beta_0+\beta_1 x_i$. Using the Generalized Linear Models (GLM) framework, we fit a linear model with and without overdispersion, and obtain the following estimates:

	Poisson				**Poisson with overdispersion**			
	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	2.38456	0.03055	78.065	<0.0001	2.38456	0.1669	17.0089	<0.0001
x	0.11439	0.02507	4.563	<0.0001	0.11439	0.1370	0.8350	0.4037
Overdispersion	1				29.87			

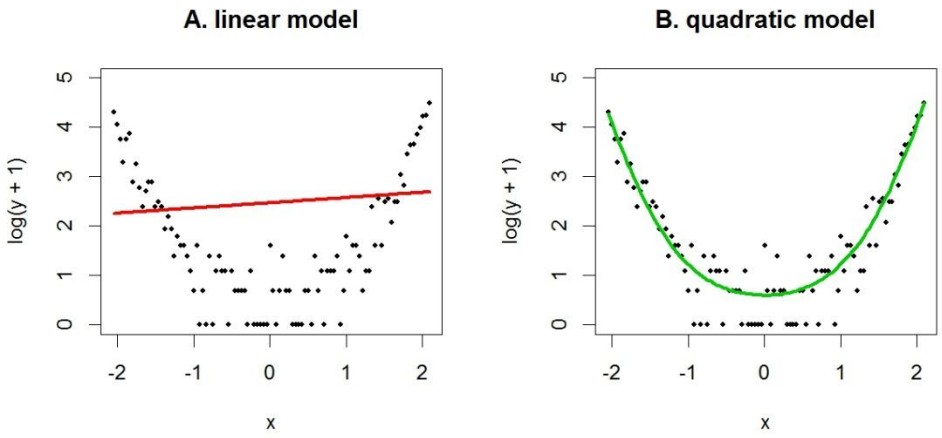
The effect of x is significant when we do not account for overdispersion. However, we observe large Pearson residuals (2927.3, on 98 degrees of freedom), which indicates that the model does not fit the data well. Often, researchers assume that this is due to overdispersion. To account for overdispersion, one computes the overdispersion parameter (ϕ=2927.3/98 = 29.87), and multiply the standard errors by sqrt(ϕ)=5.37. The parameter estimates do not change, but the predictor x is no longer significant (z value = 0.835). Often researchers would stop here assuming this to be the final model.

In contrast, if the correct model is used, then we obtain the following estimates:

	Poisson				**Poisson with overdispersion**			
	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.1849	0.0992	-1.864	0.0624	-0.1849	0.1083	-1.6613	0.0967
x	-0.0016	0.0169	-0.095	0.9245	-0.0016	0.0184	-0.0833	0.9934
x^2	1.0582	0.0288	36.624	<0.0001	1.0582	0.0315	33.594	<0.0001
Overdispersion	1				1.1916			

Once the model is correct, then excess variation may be considered as overdispersion and it is possible to proceed by adjusting the standard errors. In this case the Pearson residual statistic is 115.58 (97 degrees of freedom), so ϕ = 126.7/97=1.1916, indicating that there is hardly any overdisperion, which is also reflected by the fact that the standard error, test statistic and p-value do not change very much. Indeed, when the lack of fit statistic is not significant, it is not necessary to adjust the standard errors. What appeared to be overdispersion in the previous model was in fact due to lack of fit caused by having an important variable missing in the model.

To assess whether excess variation is due to a misspecified model, it is (as always) a good idea to plot the dependent variable versus the predictor. The following figure shows the simulated data and the fitted regression lines. Note that since we observe zero counts, we plot log(Y+1) rather than log(Y). There is a clear quadratic relationship between the predictor, x, and the logarithm of the counts (plus one). The model that assumes a linear relationship clearly does not fit the data (A), whereas the quadratic model fits the data well (B).



In summary, attributing the excess variance to overdispersion and simply adjusting the standard errors of parameter estimates may lead to the wrong conclusion. One must first check whether the larger-than-expected variance may be due to a misspecified model. Only when there is no evidence for lack of fit, can we proceed with the overdispersion adjustment.

As always if you would like assistance with this topic or any other statistical consulting question, feel free to contact statistical consultants at CSCU.
Author: Haim Bar (hyb2@cornell.edu), Hongyu Li

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).

Reference: J Quant Criminol (2008) 24: 269-284, Overdispersion and Poisson Regression. Richard Berk and John M. MacDonald