Poisson Regression: Lack of Fit is Not the Same as Overdispersion

Haim Bar, Hongyu Li

1 Introduction

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".

2 Example

In Poisson regression, the logarithm of the expected count is assumed to be a linear function of some predictors, $\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik}$ where λ_i is the expected count of the ith observation. In the case of Poisson regression, lack of fit means that the log of the expected counts cannot be predicted by $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$.

To illustrate that lack of fit should not be confused with overdispersion, we simulate $n = 100$ counts y_i from a Poisson distribution where the expected counts are related to a single predictor x_i as follows: $\log(\lambda_i) = x_i^2$. We start by fitting a (misspecified) log-linear model $\log(\lambda_i) = \beta_0 + \beta_1 x_i$. Using the Generalized Linear Models (GLM) framework, we fit this linear model with and without overdispersion. Table 2.1 displays the coefficient estimates and standard errors without accounting for overdispersion and Table 2.2 displays the estimates and standard errors while allowing for overdispersion.

Table 2.1: Coefficient estimates and standard errors for the misspecified model without allowing for overdispersion (dispersion parameter fixed at 1).

	Estimate	Std. Error	Z value	P value
Intercept	1.953	0.04	48.343	< 0.0001
x	-0.255	0.034	-7.45	< 0.0001

Table 2.2: Coefficient estimates and standard errors for the misspecified model with estimated overdispersion parameter 16.48.

	Estimate	Std. Error	Z value	P value
Intercept	1.953	0.164	11.908	< 0.0001
x	-0.255	0.139	-1.835	0.066

The slope for x in the misspecified model is significant when we do not account for overdispersion. However, the residual deviance is 1061.06 on 98 degrees of freedom, suggesting that the model does not fit the data well. Often, researchers assume that this is due to overdispersion. To account for overdispersion, one can compute an estimated overdispersion parameter equal to the sum of the squared Pearson residuals divided by the residual degrees of freedom, in this case $\phi = 1615.04/98 = 16.48$, and multiply the standard errors by $\sqrt{\phi} = 4.06$. The parameter estimates do not change, but the P-value for x is now about 0.07, far larger than the model without accounting for overdispersion. Often researchers would stop here assuming this to be the final model.

In contrast, if we fit the correct model including a quadratic term for x, we obtain the results in Table 2.3 (without adjusting for overdispersion) and Table 2.4 (including an overdispersion adjustment).

Table 2.3: Coefficient estimates and standard errors for the correctly specified model without allowing for overdispersion (dispersion parameter fixed at 1).

	Estimate	Std. Error	Z value	P value
Intercept	0.145	0.094	1.541	0.123
x	0.032	0.024	1.328	0.184
x^2	0.974	0.034	28.815	< 0.0001

Table 2.4: Coefficient estimates and standard errors for the correctly specified model with estimated overdispersion parameter 1.14.

	Estimate	Std. Error	Z value	P value
Intercept	0.145	0.382	0.38	0.704
x	0.032	0.098	0.327	0.744
x^2	0.974	0.137	7.098	< 0.0001

Once the model is correctly specified, then excess variation may be considered as overdispersion and it is possible to proceed by adjusting the standard errors. In this case the estimated overdispersion parameter is 1.14, 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 from Table 2.3 to Table 2.4. 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. Figure 2.1 shows the simulated data and the fitted regression lines for the linear (misspecified) and quadratic (correctly specified) models). Note that since the response variable can be equal to zero, 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, whereas the quadratic model fits the data well.

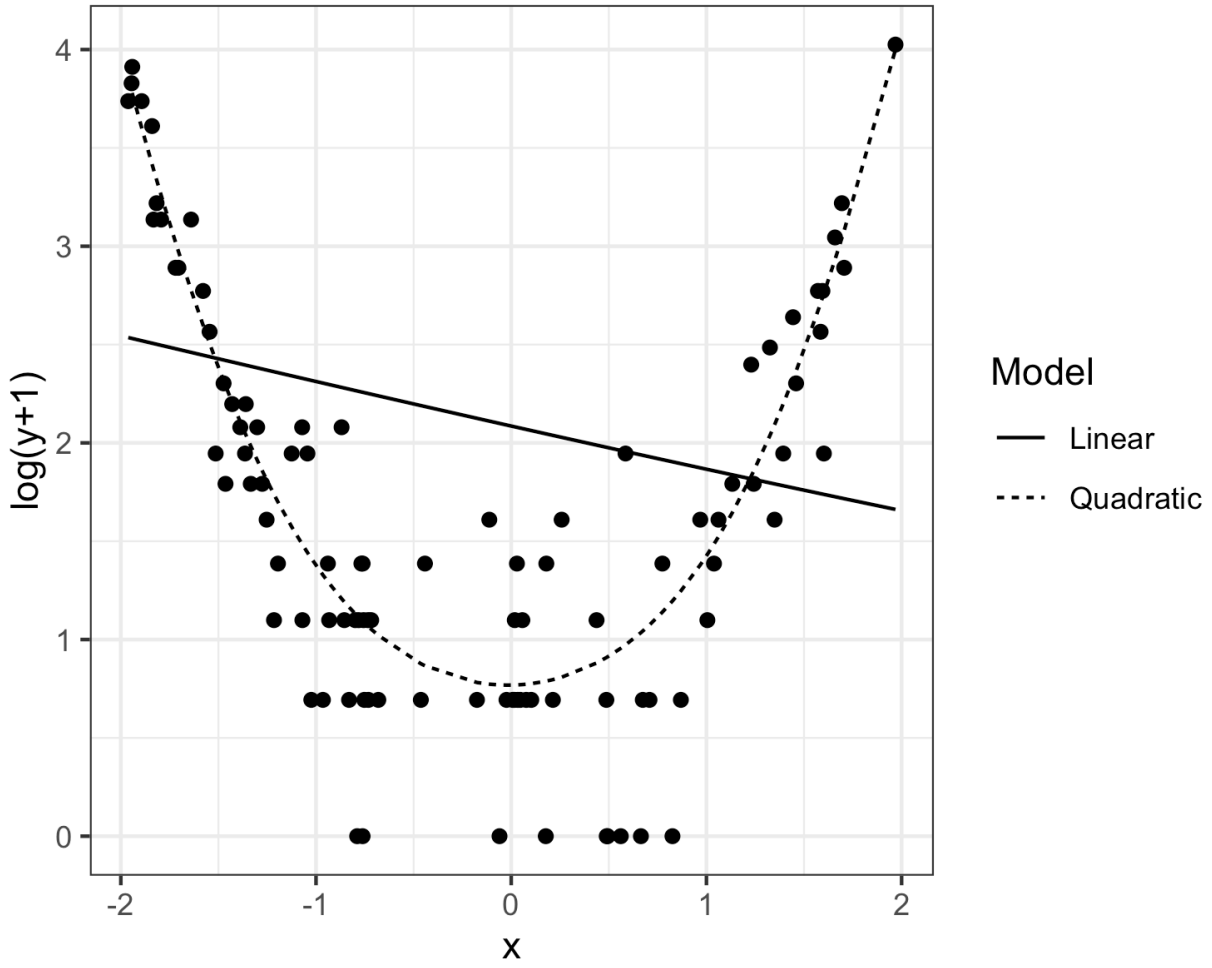

Figure 2.1: Observed data with fitted linear and quadratic models.

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

---

Created March 2013. Last updated April 2022.