

Overcoming Overdispersion in Poisson Regression: quasi-Poisson and Negative Binomial Regression

Mark Krysan

November 8, 2023

Contents

Introduction	2
Generalized Linear Models	2
Poisson Regresssion	2
Overdispersion	3
Negative Binomial Regression	3
Quasi-Poisson Regression	5
Simulation Study	6
Comparison Between Quasi-Poisson and NB2	7
Appendix	8

Introduction

The analysis of count data has become very important in scientific research, but simple linear and multilinear regression are unable to accurately and effectively model this type of data. Researchers must use the Generalized Linear model, a family of regression models based on the Exponential family of distributions. A popular choice for modeling count data is the Poisson regression due to its simplicity and ease of implementation in R. However, because of the strong assumption that the data has equal mean and variance, Poisson regression is not always the best choice, since applications rarely have equal mean and variance. In this paper, we will look at the effects that unequal mean and variances have on Poisson regression, more specifically, overdispersion, when the variance is greater than the mean. We will also showcase Negative Binomial and Quasi-Poisson regression, two models that can account for overdispersion, and through a simulation study, how they are related to Poisson regression.

Generalized Linear Models

Before discussing Poisson regression models, we must first describe the Generalized Linear Model, which, as its name suggests, generalizes linear regression. General linear models have two parts: the random component and the systematic component. The random component is the hypothesized distribution of the response variable, assumed to be from the exponential family of distributions. For example, if Y_i is binary, we assume binomial distribution. The systematic component, denoted $\eta = \beta_0 + \sum_{j=1}^p \beta_j x_j$, is the linear combination of the explanatory variables, combined with the link function. The link function is responsible for relating the explanatory variables to the mean of the response variable. It is denoted $g(\cdot)$, is a monotonic, differentiable function such that $g(\mu) = \eta$, where $E[Y] = \mu$ [1]. There are many link functions, such as the logit link, used in logistic regression, where $g(\mu) = \ln(\frac{\mu}{1-\mu})$, or the identity link, where $\eta = g(\mu) = \mu$. For simple linear regression, we have that the random component is $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, and the systematic component using the identity link to get, $g(\mu) = \mu = \mathbf{X}\beta$.

Poisson Regression

Poisson regression is a type of generalized linear model used for modeling count data. It is a special case of the more general negative binomial regression. Examples of problems that would require Poisson regression include the number of chips in a chocolate chip cookie or the number of emails received in an hour. Poisson regression uses Poisson distribution as the random component with probability mass function, $p(y|\mu) = \frac{\exp(-\mu)\mu^y}{y!}$ for $y = 0, 1, 2, \dots$ and with expected counts $\mu > 0$. The link function used is the log link, or $g(\mu) = \ln(\mu)$. We can write Poisson regression as

$$\begin{cases} y \sim \text{Pois}(\mu) \\ \log(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \end{cases}$$

The poisson regression model estimates the maximum likelihood regression coefficients using the iteratively reweighted least squares algorithm, instead of ordinary least squares [3]. A key characteristic of poisson regression is that, following from the PMF, the mean is equal to the variance, called equidispersion. Equidispersion is one of the main assumptions required for poisson regression, including non-negative, integer counts and independent observations. As noted by Hilbe, it is rare for a 'real life' poisson dataset to be equidispersed. More often than not, count data will have a variance greater than the mean, called overdispersion[3].

Overdispersion

As stated above, overdispersion is when the variance of a model is greater than the mean. In general, there are two types of overdispersion: apparent and real. Apparent overdispersion occurs when a model does not properly correspond to the observed data. This includes when a model omits important explanatory predictors, the data includes outliers, the model lacks sufficient interaction terms, the predictors need to be transformed, or when the link function is misspecified. Apparent overdispersion is an attribute that can be fixed or treated in the model. However, real overdispersion is when the data is inherently overdispersed.

Overdispersion in regression since it can drastically impact the standard errors of a model by making them much smaller than what they should be, leading to inflated p-values for the significance of predictors. Researchers must therefore be on the lookout for overdispersion in models due to their influence on inference. The main way to determine if data is overdispersed is by calculating the Pearson (χ^2) statistic for a model and dividing by the degrees of freedom [3]. The χ^2 statistic is of the form

$$T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

where E_i is the expected count for the i th observation, and O_i is the observed count for the i th observation. The χ^2 statistic measures the goodness of fit of a model. This statistic divided by the degrees of freedom is called the dispersion parameter. For moderately sized models, a model is overdispersed if the dispersion statistic is greater than 1.25, but for larger models, a statistic greater than 1.05 is considered overdispersed. If overdispersion is identified in a model, there are many different ways to handle it, both for real and apparent overdispersion. A discussion of these methods can be found in Hilbe's book [3]. We will be discussing two common methods for handling real overdispersion: negative binomial regression and quasi-poisson.

Negative Binomial Regression

A logical step to take when posed with the question of overdispersion is to relax the assumption of equidispersion. The negative binomial regression fulfills this by letting μ be a random variable, creating extra variability in the model. This adds a lot of flexibility when modeling count data [2]. To add another layer of randomness, we let the random component of our

model be

$$y_i|\lambda_i \sim \text{Pois}(\lambda_i) \text{ and } \lambda_i \sim \text{Gamma}(\mu_i, \psi), \quad (1)$$

where $\mu_i > 0$ and $\psi > 0$, and ψ is the coefficient of variability [1]. However, it is not clear that the distribution of y_i is negative binomial. We will now show that a negative binomial distribution has the same pdf as a mixed poisson-gamma distribution.

Let $X \sim NB(r, p)$. We have that the pmf of X is

$$P(X = n) = \binom{r+n-1}{r-1} p^r (1-p)^n, \quad (2)$$

with mean of $\mu = \frac{r(1-p)}{p}$. First, notice that

$$\begin{aligned} \binom{r+n-1}{r-1} &= \binom{r+n-1}{n} = \frac{(r+n-1)(r+n-2)\dots(n)}{r!} \\ &= \frac{(r+n-1)!}{n!r!} \\ &= \frac{\Gamma(r+n)}{\Gamma(n+1)\Gamma(r)}. \end{aligned}$$

Now, since $\mu = \frac{r(1-p)}{p}$, we have that $p = \frac{r}{\mu+r}$ and $1-p = \frac{\mu}{\mu+r}$. Therefore, we can rewrite (2) as

$$P(X = n) = \frac{\Gamma(r+n)}{\Gamma(n+1)\Gamma(r)} \left(\frac{r}{\mu+r}\right)^r \left(\frac{\mu}{\mu+r}\right)^n$$

Now, letting $y = n$ and $k = 1/\psi = r$, we get

$$P(y|\mu, k) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{1-\mu}{\mu+k}\right)^k. \quad (3)$$

Through messy reparameterization and using the Law of Total Probability to find the marginal distribution of y , as defined in (1), we find that it is equal to the equation in (3).

Therefore, we see that letting λ be random with the Gamma distribution, we get a negative binomial distribution. As noted by Dunn and Smyth, the variance of (1) is $\text{Var}(y_i) = \mu_i + \mu_i^2/k$. Notice that as $k \rightarrow \infty$, or $\psi \rightarrow 0$, $\text{Var}(y_i) \rightarrow \mu_i$ [1]. Therefore, as the dispersion approaches ∞ , the Negative Binomial regression approaches a Poisson regression. A consequence of the variance being defined as such is that there is a concave relationship between the weighting of observations and the mean. There is very little weight given to small means, and as the size of the mean increases, the weights level off to $\psi = 1/k$ [4].

Like Poisson regression, the estimation of the coefficients are done through the IRLS (a full description of this algorithm can be found in Hilbe)[3]. Although there is another parameter being estimated in the Negative Binomial model, the estimation of ψ is uncorrelated with the β_i s. The coefficient estimators will tend to be very similar to the Poisson model, since the dispersion of the data does not change the coefficient estimation, but the standard errors of the Negative Binomial will tend to be larger[3].

Quasi-Poisson Regression

Another model that accounts for overdispersion is the quasi-poisson regression. In practice, the quasi-poisson model is a poisson model, but rescaled so the dispersion estimate is 1. The quasi-poisson model is an instance of the quasi-likelihood family of models, which only requires the specification of the first and second moment of the distribution of data [2]. However, quasi-poisson regression can be framed in the form of a GLM. To do this, we let Y be a random variable with $E[Y] = \mu$ and $\text{Var}(Y) = \theta\mu$ where $\mu, \theta > 0$, and we use the log link. In the GLM framework, the only difference between the Poisson and quasi-Poisson is the difference in the variance of the random component. This definition of variance will lead observation weights being directly proportional to the mean [3]. Since the quasi-Poisson model is not defined by a complete probability density function, just by the first two moments, there are many statistical tools, such as AIC and BIC which are unable to be calculated [3]. There are some statistics which have been created for quasi-likelihood models, but they cannot be compared to standard statistics, such as R-squared, AIC, or BIC [4].

The addition of a new parameter for the variance of Y allows for an increased flexibility in the modeling of the data, specifically for overdispersed data. In the quasi-poisson model, the parameter θ will be estimated as the inverse square root of the dispersion estimate [3]. This choice of θ will lead to a dispersion estimate of 1, transforming the data in a way to be equidispersed. However, this first requires a poisson model to be created, then a dispersion estimate to be calculated, then a new quasi-poisson model to be run with the chosen value of θ . These steps are all taken care of in the MASS function **glm(formula, family = quasipoisson)**.

Like Poisson and Negative Binomial regression, the coefficients for quasi-Poisson regression are calculated using IRLS. However, due to the rescaling of the variance, there is one more iteration of the algorithm, except with the weight matrix multiplied by the inverse square root of the dispersion parameter (more details of estimation of quasi-Poisson coefficients are outlined in [3]).

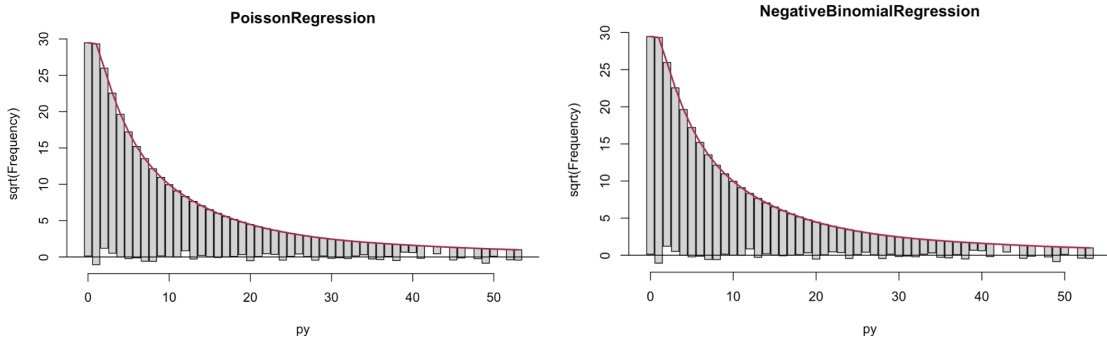
As mentioned above, overdispersion does not effect the estimated coefficients in the regression model, but the standard errors and the resulting statistics of those coefficients, such as confidence intervals and p-values. Therefore, for the quasi-poisson model, the only difference from the poisson model will be the standard errors and p-value of the coefficients of the model. These values will be larger than the Poisson model, accurately accounting for the amount of variance in the model. An example of this will be demonstrated in the simulation study.

Simulation Study

To demonstrate the relationships between the Poisson, quasi-Poisson, and Negative Binomial regression models and how they react to overdispersion, we have chosen to simulate count data. We have created two data sets, both based on the Poisson distribution, but one data set is equidispersed and the other is overdispersed. For the overdispersed data set, a Gamma distribution is used to simulate dispersion parameter. We used a function from [3] which takes in the number of observations and regression coefficients desired for our simulation,

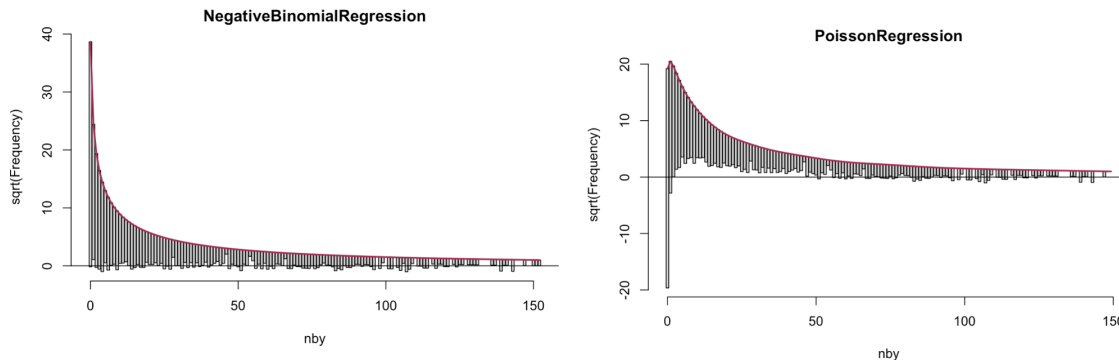
and creates a dataframe with the explanatory and response variables (The code can be found in the Appendix: Figure 3 and 4). The explanatory values are normally distributed with mean of the desired coefficient and variance of 1. To get the response values function then takes the exponentiation of the explanatory variables and uses this value as the mean of either a Poisson or Negative Binomial distribution. For the Negative Binomial distribution, the function also incorporates a variance parameter, which is incorporated into a Gamma distribution. The values of the Gamma distribution are then multiplied by the exponentiation of the explanatory values, and that is then used as the mean of a Poisson distribution.

For both data sets, we ran Poisson, quasi-Poisson, and Negative Binomial regression, and compared their results. For the equidispersed dataset, we simulated that data with simulation coefficients of $\beta_0 = 1$, $\beta_1 = -.5$, and $\beta_2 = 1$ and 5000 observations. We then ran the corresponding regressions using the `glm` and `glm.nb` functions from the **MASS** package in R. As expected the Poisson and quasi-Poisson had the same coefficient estimates, with $\hat{\beta}_0 = .9937$, $\hat{\beta}_1 = -.5061$, and $\hat{\beta}_2 = 1.0089$. The Negative Binomial model had slightly, but negligibly different estimates of $\hat{\beta}_0 = .9936$, $\hat{\beta}_1 = -.5061$, and $\hat{\beta}_2 = 1.0090$. The standard errors of all three models were all different from each other, but again, in a negligible amount (The full models can be found in the Appendix: Figures 1 and 2). The two plots below compare the fit of the Poisson and Negative Binomial models. The red line is the expected distribution of the data and it is overlaid on the observed counts. The boxes are placed to show the expected count, and the measure of the distance is displayed in how far the bottom of the box is from $y = 0$. We see that both models fit the data very well. However, since the quasi-Poisson model is not defined by a probability distribution, we are unable to create this type of plot.



Regarding the overdispersed dataset, we start to see differences between the models. We created the dataset with simulation coefficients of $\beta_0 = 2$, $\beta_1 = .75$, and $\beta_2 = -1.25$ and a dispersion parameter of .5, this will lead to a variance of $\mu + \mu^2 \cdot 2$ as a function of μ . We ran the same three models as above. The model results were very similar to the equidispersed model, with a few notable differences (Full Models in Appendix: Figures 1 and 2). First off, the standard errors for the Negative Binomial and quasi-Poisson were much larger than those for the Poisson model. For the Negative Binomial, the standard errors were: .02, .02, and .02, for the respective coefficients, .03, .02, and .02 for the quasi-Poisson, and .006, .003, and .003 for the Poisson model. We see here that the Poisson model does not account for the true variability in the model, resulting in very small error terms. These terms would then create very small p-values, resulting in incorrect inference. The following plots show how the

Negative Binomial model fit the data very well, but the Poisson model underfit values very close to 0 and overfit values around 25.



Comparison Between Quasi-Poisson and NB2

Although both quasi-Poisson and Negative Binomial regression can model overdispersed count data, there are differences in the two which help determine which to use in a given situation. First off, the quasi-Poisson model is overall more simple than the Negative Binomial. Since the quasi-Poisson model acts as a transformed Poisson, it makes both the model and interpretation easier to work with. This simplicity does come with a drawback though. Since quasi-Poisson regression is a pseudo-transformation of a Poisson model, a Poisson model must be created first in order to determine the dispersion parameter. Therefore, question of p-hacking could be raised, depending on the goal of the model. If that is of no worry, then the quasi-Poisson model is best used for data that is not known to be overdispersed *a priori*.

However, to use Negative Binomial regression, the researcher should already know that the data is overdispersed before modeling. Since the dispersion is estimated at the same time as the mean, there is no worry of p-hacking. Another notable difference regarding selection of model is the relationship between the mean and variance. If it is known that a data set is overdispersed before modeling, there are some cases where the approximate relationship between the mean and variance can be determined. If there is a linear relationship, then the quasi-Poisson method will result in a better fitting model. However, if there is a quadratic relationship, the Negative Binomial model will be better. This is due to the definition of the variance in the two models. In addition, VerHoef and Boveng state that this difference in definition of variance, specifically the ways in which the observations are weighted, are very helpful for choosing which model to use (See [4] for greater explanation with real world data).

References

- [1] Peter K Dunn and Gordon K Smyth. "Chapter 10: Models for Counts: Poisson and Negative Binomial GLMs". In: *Generalized Linear Models With Examples in R*. New York, NY: Springer New York, 2018, pp. 371–424. ISBN: 978-1-4419-0118-7. DOI: 10.

1007/978-1-4419-0118-7_10. URL: https://doi.org/10.1007/978-1-4419-0118-7_10.

- [2] William Gardner, Edward P Mulvey, and Esther C Shaw. “Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models”. English. In: *Psychological Bulletin* 118.3 (Nov. 1995), pp. 392–404. ISSN: 0033-2909, 0033-2909. DOI: <https://doi.org/10.1037/0033-2909.118.3.392>.
- [3] Joseph M Hilbe. 2nd ed. Cambridge University Press, 2011. DOI: 10.1017/CB09780511973420.008.
- [4] Jay M Ver Hoef and Peter L Boveng. “Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?” eng. In: *Ecology* 88.11 (Nov. 2007), pp. 2766–2772. ISSN: 0012-9658 (Print). DOI: 10.1890/07-0043.1.

Appendix

Figure 1: Models of Equidispersed Data

	Poisson	quasi-Poisson	Negative Binomial
(Intercept)	0.993734	0.993734	0.993677
	(0.009323)	(0.009419)	(0.009330)
x1	−0.506117	−0.506117	−0.506107
	(0.006196)	(0.006260)	(0.006204)
x2	1.008970	1.008970	1.009050
	(0.005835)	(0.005895)	(0.005854)
Num.Obs.	5000	5000	5000
AIC	18772.9		18774.9
BIC	18792.4		18801.0

Figure 2: Models of Overdispersed Data

	Poisson	quasi-Poisson	Negative Binomial
(Intercept)	2.067111	2.067111	1.987917
	(0.005437)	(0.033657)	(0.021418)
x1	0.666306	0.666306	0.740751
	(0.003148)	(0.019486)	(0.021714)
x2	-1.212119	-1.212119	-1.257935
	(0.003389)	(0.020982)	(0.022579)
Num.Obs.	5000	5000	5000
AIC	153329.1		29856.7
BIC	153348.6		29882.8

The values in parantheses are the standard errors of the coefficient estimate.

Figure 3

Equidisperesed Simulation Function

```
poisson_syn <- function(nobs = 5000, xv = c(1, -.5, 1)) {
  p <- length(xv) - 1
  X <- cbind(1, matrix(rnorm(nobs * p), ncol = p))
  xb <- X %*% xv
  exb <- exp(xb)
  py <- rpois(nobs, exb)
  out <- data.frame(cbind(py, X[, -1]))
  names(out) <- c("py", paste("x", 1:p, sep = ""))
  return(out)
}
```

Figure 4

Overdispersed Simulation Function

```
nb2_syn <- function(nobs = 5000,
                    alpha = 1,
                    xv = c(1, 0.75, -1.5)) {
  p <- length(xv) - 1
  X <- cbind(1, matrix(rnorm(nobs * p), ncol = p))
  xb <- X %*% xv
  a <- alpha
  ia <- 1/a
  exb <- exp(xb)
  xg <- rgamma(nobs, a, a)
  xbg <- exb*xg
  nby <- rpois(nobs, xbg)
  out <- data.frame(cbind(nby, X[, -1]))
  names(out) <- c("nby", paste("x", 1:p, sep=""))
  return(out)
}
```