Evaluating the Performance of Zero-Inflated Count Models in Out-of-Sample Prediction

Junhyung Park & Brice Randolph

Department of Statistics, UCLA *

March 16, 2016

Abstract

In this paper, we evaluate the out-of-sample prediction accuracy of zero-inflated (ZInf) count models compared to non-ZInf models and alternative machine learning methods. We compare the different models on two datasets of different lengths. On our smaller dataset, we found that ZInf methods provide a better prediction of zero counts, yet the overall prediction accuracy of ZInf models is not superior to non-ZInf models. This is due to non-ZInf models providing a better prediction of non-zero counts. On our larger dataset, the same pattern of prediction quality is observed, however, the type of error function used heavily influences the results. We also found that alternative machine-learning techniques can outperform traditional count models on larger datasets.

*We are grateful to Dr. Chad Hazlett, who was generous with his time in guiding us during the research process. This paper greatly benefits from the useful and illuminating course given by Dr. Hazlett. We also thank those in the audience for the presentation of this report. All errors and shortcomings of this report are our own responsibility.

1 Motivation and Hypothesis

Count data arises in many fields, ranging from marketing and public policy to the biological and physical sciences. Researchers wish to model the number of hospital visits, accidents and deaths, clicks on a website advertisement, or even the number of earthquakes in a given region. Most of these examples involve a count variable taking small non-negative integer values. Using count models, researchers are able to characterize the distribution of these counts as well as make predictions on unseen data.

Often, an increased amount of variability in the data can lead to model misspecification and poor prediction performance. This is referred to as overdispersion, and will be formally defined in the next section. One of the major causes of overdispersion is an increased number of zero counts in the data. In this setting, models such as the Zinf model propose to increase flexibility by embedding multiple probability distributions in the proposed model.

Our project is motivated by the fact that research scientists, the major consumers of these models, base model selection on coefficient interpretation and variable significance. Prediction accuracy, on the other hand, is rarely considered. The modeling routine we have observed typically begins with a simple test for overdispersion, then proceeds with a zero-inflated model or a negative binomial model for the rest of the analysis. The researcher will cite AIC and Vuong Test scores (explained later) to justify their decision. They may also show that the total predicted zero count of the Zinf model matches the actual number of observed zeros. Little attention is given to metrics for out-of-sample prediction accuracy. [1]

Our research questions are as follows: *In the presence of zero-inflation, do Zinf models outperform their non-Zinf counterparts in terms of out-of-sample prediction accuracy? In general, how useful are count models in terms of making predictions? What are some alternative methods of predicting counts?*

Prior to starting the research, we made the following two **hypotheses**:

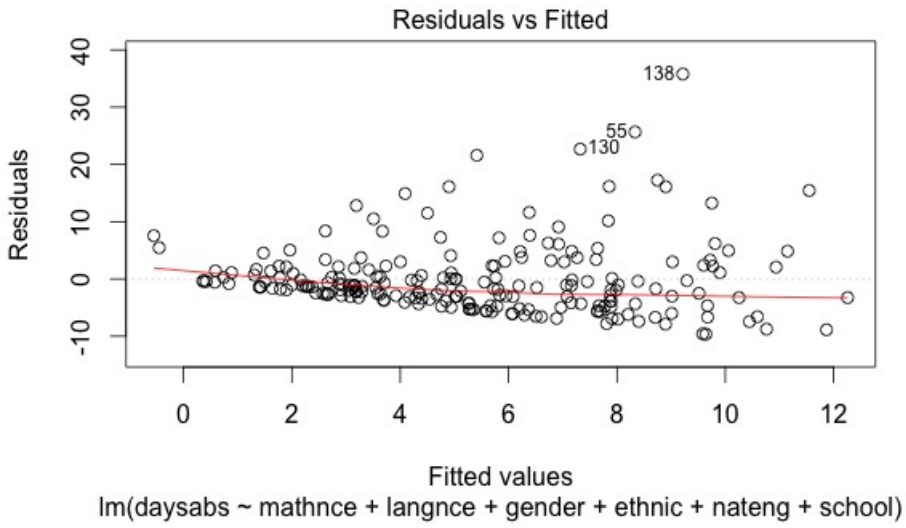
1. Although zero-inflated models (Hurdle, ZINB, ZIP) will produce more accurate in-sample predictions than standard count models (Poisson and Negative Binomial), we believe that out-of-sample prediction accuracy will not differentiate these models.

[1] See [7], [1], [3] or [4] as examples.

2. Furthermore, we believe that machine learning approaches (Random Forest, SVR) or an alternative GLM approach (Tobit) will outperform the above-mentioned GLMs in terms of out-of-sample prediction accuracy.

2 Overview of Methods

When fitting relationships between predictor and response variables, generalized linear models (GLMs) are commonly used as a regression procedure. Classical ordinary least squares (OLS) is often inappropriate to model count data due to the residual component lacking constant variance. This can be seen in the following figure, noting the "fanning out" of the residuals as the fitted values increase.



In light of the drawbacks of OLS, the standard parametric model for count data is the Poisson distribution. However, some of its assumptions might not be appropriate for the given data set. For example, the Poisson distribution assumes equidispersion ($\text{Var}(Y_i \mid \mathbf{X}) = \mathbf{E}[Y_i \mid \mathbf{X}]$), an assumption that is often violated. It will also perform poorly when the data is characterized by an excessive number of zeros, a common cause of overdispersion. In this paper, we will explore models that attempt to remedy these shortcomings. We will compare model performance on a dataset of school absence count from two LA

high schools. Our goal is to predict student absence count based on several demographic factors and two exams, measuring academic performance in mathematics and English.

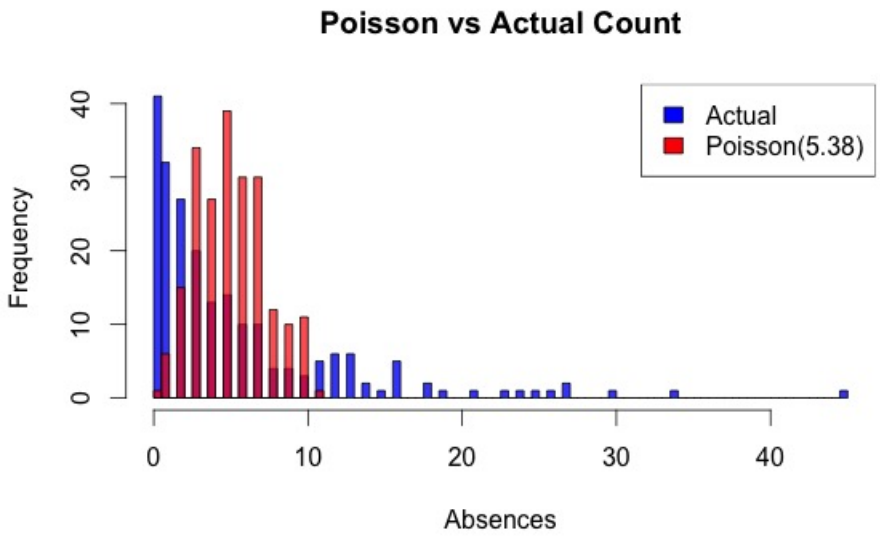
Poisson Model

The first distribution we used to model the dataset is the Poisson distribution, given by the probability mass function

$$f(Y_i \mid \lambda) = \frac{exp(-\lambda)\lambda^{Y_i}}{Y_i!}$$

Poisson regression is a special case of the generalized linear model, where the canonical link function is given by $g(\mathbf{E}[Y_i \mid X_i]) = g(\lambda_i) = exp(X_i'\beta)$

As mentioned earlier, this model assumes $\text{Var}(Y_i \mid \mathbf{X}) = \mathbf{E}[Y_i \mid \mathbf{X}]$. In order to relax this assumption, we could use a quassi-Poisson model, which specifies a variance function in terms of the mean and an additional parameter ϕ. In this case, $Var(Y_i \mid X_i) = \phi\lambda_i$. Although a quassi-Poisson model will have more robust standard errors and confidence intervals, this function ϕ will not change the estimates for the conditional mean $\mathbf{E}[Y_i \mid X_i]$, and thus, will not improve prediction accuracy over the Poisson model.



This plot demonstrates the lack of fit that usually deters researchers from using the Poisson model in instances of overdispersion. This leads us to the negative binomial model, the standard solution to this fitting problem.

Negative Binomial

The negative binomial model relaxes the assumption that the conditional mean and conditional variance are equal. Thus, the negative binomial model is a natural substitute to the Poisson model in cases of overdispersion.

P.m.f.(NB2)

$$f(Y_i \mid \mu_i, \alpha) = \frac{\Gamma(Y_i + \alpha^{-1})}{Y_i!\Gamma(\alpha^{-1})}(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i})^{\alpha^{-1}}(\frac{\mu_i}{\alpha^{-1} + \mu_i})^{Y_i}$$

This model has mean μ and variance function $\mu+\alpha\mu^2$. The exponent on μ can be varied to produce different models, however, the NB2 model is commonly chosen because of its robustness to distributional misspecification [2]. A common tip for researchers attempting to model count data is to start with a negative binomial model, then check the estimated dispersion parameter α. If α is significantly greater than zero, then the data demonstrates overdispersion and the negative binomial model is appropriate. There are more formal overdispersion tests[2], however, looking at the α parameter is one of the most common approaches.

Generalized count models

Often, a single distribution may lead to an inappropriate fit of the data. If the researcher believes that the observed counts come from two subpopulations, a generalized count model could provide a better fit. In our data, we believe that some students will have perfect attendance (zero absence count) because their parents force them to attend school everyday, regardless of illness. On the other hand, a second group of students might have any number of absences in a given year. A natural remedy to this is to use a two-part model such as the hurdle or zero-inflated model.

Hurdle model

The basic idea of a hurdle model is that a binomial probability determines the binary outcome of whether a count is a zero or a positive realization. If a realization is positive, the hurdle is crossed, and the conditional distribution of the positive count is governed by a zero-truncated count model (such as Poisson or negative binomial).

$$f(Y=y) = \begin{cases} f_1(0) & \text{if } y = 0 \\ \frac{1-f_1(0)}{1-f_2(0)} f_2(y) & \text{if } y \geq 1 \end{cases}$$

Similar to the hurdle model, zero-inflated models assume the observed counts are the result of a two-part process: one that generates "structural zeros", and one that generates random counts. The count process is often modeled as either a poisson distribution or a negative binomial distribution. On the other hand, the structural zeros are assumed to follow a Bernoulli process with parameter π.

Zero inflated Poisson (ZIP)

This model is very similar to the hurdle model in that it is composed of a mixture of distributions. However, in the ZIP model, zeros can be generated from both the count and structural processes. These models can use up to twice the number of parameters, so overfitting can be called into question.

P.m.f.

$$P(Y_j) = \begin{cases} P(Y_j = 0) = \pi + (1-\pi)e^{-\lambda} \\ P(Y_j = h_i) = (1-\pi)\frac{\lambda^{h_i} e^{-\lambda}}{h_i!} \end{cases}$$

Zero inflated Negative Binomial (ZINB)

This is a slight variation on the ZIP model, where we model the counts with a negative binomial distribution, in lieu of the Poisson. One drawback is that fitting this model can be difficult on smaller datasets due to convergence issues [2].

P.m.f.

$$P(Y) = \begin{cases} \pi + (1-\pi) f_2(0) & \text{if } y = 0 \\ (1-\pi) f_2(y) & \text{if } y \geq 1 \end{cases}$$

Where $f_2 =$ the negative binomial P.m.f.

[2] With our smaller dataset, we found that the ZINB log likelihood often fails to converge. After reducing the number of covariates in the model, we still experienced frequent convergence failures when estimating out-of-sample prediction error. This is also the case with the ZIP model, but to a lesser degree.

Other models

To consider alternate methods of predicting counts, we also considered treating our discrete outcome as a continuous variable in three different settings. We will only briefly summarize these methods, as they involve a bit more technical detail than the count models above.

Random Forest

We chose to use this nonparametric model in an attempt to improve prediction accuracy over the standard count models. In a random forest, no assumption is made about the distribution of the data. This method uses an aggregation of trees in order to lower the variance of the estimators.

We also used a Support Vector Regression (**SVR**) in an attempt to model possible nonlinearity in the data. We only list the Gaussian kernel SVR since it always led to better predictions than the linear kernel. We also considered a **Tobit** model, which is a simple censored version of linear regression.

[3]

Alternative methods of comparing count models

Although we are comparing these models based on prediction accuracy, in practice, other metrics relying on likelihood calculations are used to compare model performance[2]. Two commonly used measurements are the Akaiki information criterion (AIC) and Vuong's closeness test for comparing two models. For a given model, AIC measures the goodness of fit (based on the likelihood function), while penalizing for the number of parameters used. It does not give a measure of how good the model is, instead, it gives a metric to compare a set of candidate models. Usually, AIC is used in model selection within nested models, but it is not restricted to this class. In general, the model with the lowest AIC value is preferred. Vuong's closeness test, on the other hand, is a likelihood ratio test based on the Kullback-Leibler information criterion. This is used to compare two different models, but a series of tests can be used to compare multiple models. Similar to AIC, this test is not restricted to nested models.

[3]Tobit models do require a homoskedastic normality assumption on the error term. However, the error is in the latent part of the model, which we do not observe. Also, checking this assumption is more involved than checking a residual plot in OLS. Papers have been published suggesting different ways to check this [5] [6]. Since we are only interested in prediction accuracy, we bypass this detail.

3 Data

Our dataset contains information on 316 students at two Los Angeles high schools. [4] The number of days absent during the school year (`daysabs`) is the outcome of interest. Covariates include `gender`, student's ethnicity as a categorical variable (`ethnic`) [5], which of the two high schools was attended (`school`), student's standardized math and language test scores (`mathnce` and `langnce`), and student's bilingual status (`biling`). [6] Looking at the mean and variance of *daysabs* in the table below, one can see that the data demonstrates over-dispersion ($\text{Var}(Y_i \mid \mathbf{X}) = 47 >> \mathbf{E}[Y_i \mid \mathbf{X}] = 5.4$).

Variable	Min	Mean	Median	Max	StDev	Freq
`daysabs`	0	5.38	3	45	6.87	
`mathnce`	1	48.75	48.94	98.99	17.88	
`langnce`	1	50.06	50	98.99	17.94	
Only English						37%
`school (=2)`						49.7%

Table 1: Descriptive statistics on covariates.

White	Asian	African American	Hispanic	Other	Total
98	25	39	144	10	316

Table 2: Count of ethnicities.

[4] http://www.ats.ucla.edu/stat/stata/notes/lahigh.dta

[5] Because there is only one Native American, nine Filipinos, and zero Pacific Islanders, those three categories are combined as "other". This is to ensure that our categorical variable has the same number of levels in both training and testing data.

[6] The Bilingual variable comes with 4 levels: 0=English only, 1=Proficient in native language and tested proficient in English, 2=Good English, recently transitioned from English-as-Second-Language (ESL) program, 3=In an ESL program. We simplified this into a binary variable of Native English Only={0} and Native Language + English={1,2, and3}.

4 Results

We do not give much attention to the coefficients and standard errors due to the fact that we are specifically interested in prediction accuracy.

Table 3: **Test MSE**

Linear	ZIP	Poisson	SVR	Tobit	Hurdle	NB2	ZINB	R.Forest
67.06	67.83	67.91	68.15	70.70	71.40	71.78	73.58	75.17

Table 4: **Test Median Absolute Error**

Poisson	ZIP	NB2	Tobit	Linear	Hurdle	SVR	ZINB	R.Forest
3.04	3.09	3.18	3.19	3.26	3.36	3.94	4.05	4.11

Table 3 and Table 4 present the testing mean squared error (MSE) and median absolute error (MAE). These are based on 100 observations left aside for testing (out of 316 total). In computing the MSE and MAE, we use the fitted value $\mathbf{E}[y_i|x_i]$. It is important to note that the regular Poisson model performs as well as the ZIP model. To our surprise, OLS outperformed ZIP with respect to MSE, even though the distributional assumptions are not appropriate. In general, this illustrates how poor our count models are at prediction. See Appendix Figure 8, in which predicted values deviate from the true value with a significant margin(45 degree line).

We believed that $\mathbf{E}[y_i|x_i]$ is not a good concept for fitted values in ZInf models. Thus, we tried a different definition which we loosely [7] refer to as a "Bayes" estimate. Here, we define a predicted value as

$$\hat{y}_i = \underset{j}{argmax}\Big\{P(Y_i = j|X_i)\Big\}.$$

These are not available directly from R, but can be computed. Given a set of covariates, the model gives probabilities of different counts, and we take the fitted value to be the outcome with the highest probability.

[7] In the sense that this is not generative classification.

This alternative definition of a fitted value confirms the conclusion that the Poisson model performs as well as ZIP in out-of-sample error [8] [9]. This was puzzling because the ZInf models, ZIP in particular, predicted total counts (both zeros and non zeros) that closely matched actual counts observed in the data set. [10]

Table 5: **Test Mean Absolute Error (Bayes)**

Poisson	ZIP	Hurdle	ZINB	NB2
5.28	5.86	6.41	6.75	6.75

Table 6: **Test Mean Square Error (Bayes)**

Poisson	ZIP	Hurdle	ZINB	NB2
70.02	80.94	109.51	117.63	117.63

We concluded that total predicted counts obtained by commands like `sum(predict(zip.out, testdata ,type="prob")[,1])` say nothing about actual prediction accuracy. This statistic is frequently mentioned in other papers, giving the impression that ZInf models are correctly predicting zeros counts. In reality, it is just a weak statement that indicates the overall expected count of zeros given the whole data set. This is not a statement indicating a zero will be predicted when the true value is zero.

Based upon Figure 1 and 2 , we see that, on average, the Poisson model is not assigning a high probability of zero when the true outcome is zero. However, for the low non-zero counts, the Poisson model is assigning a higher probability to the true outcome compared to other competing models. This is where the Poisson model is making up ground in the overall measure of error. This also explains why the Poisson and ZInf models have similar

[8] This concept of a fitted value does not allow us to evaluate the machine learning methods, hence the tables only feature our count models.

[9] This Bayes estimate turned out to be poor for the negative binomial models because, given the dispersion paramter of 1.003, the distribution of any negative binomial pmf with any mean always has the most mass at zero. Hence, the bayes estimate is always zero.

[10] Our test set had 21 zeroes and ZIP predicted 19. `sum(predict(zip.out, testdata ,type="prob")[,1])=19.2`

MSE and MAE. [11]

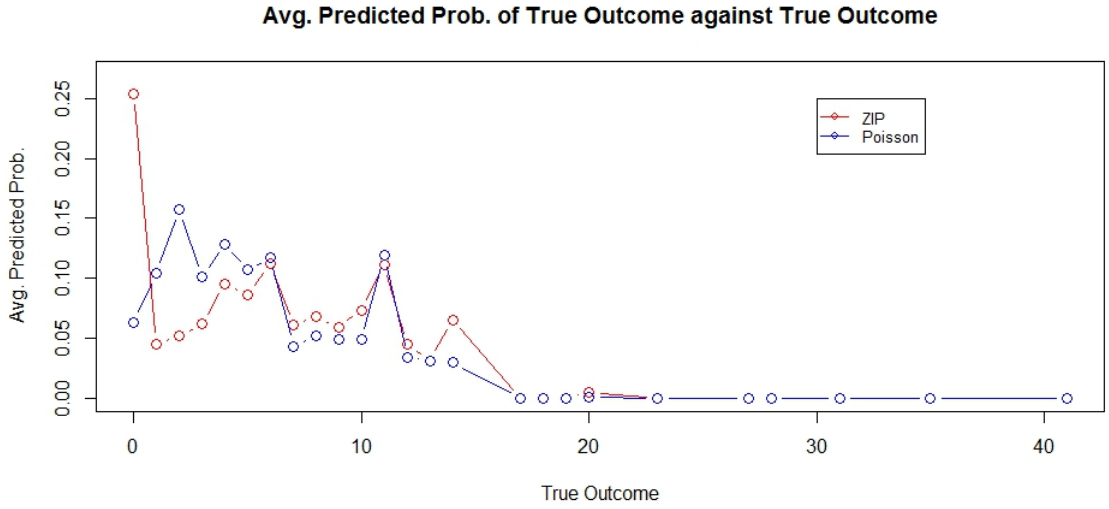


Figure 1

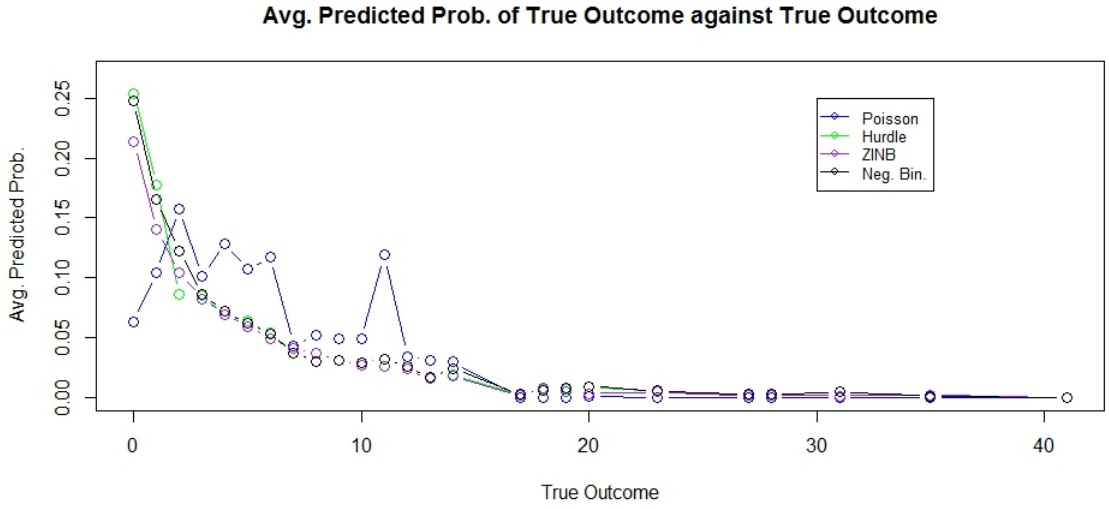


Figure 2

In Figure 3 and 4 we inspect the individual Bayes predicted values against the true value, and sort the index by size of error. The striking feature is that the ZIP correctly predicts many true zeros, but also returns zero-count predictions almost indiscriminately.

[11] Although the NB2 is not a zero-inflated model, the nature of its predictions closely follow the zero-inflated models due to its over dispersion parameter.

This gives us insight into why the average predicted probability of non-zero counts was so low for ZInf models. This can be seen in Figure 1 and 2.

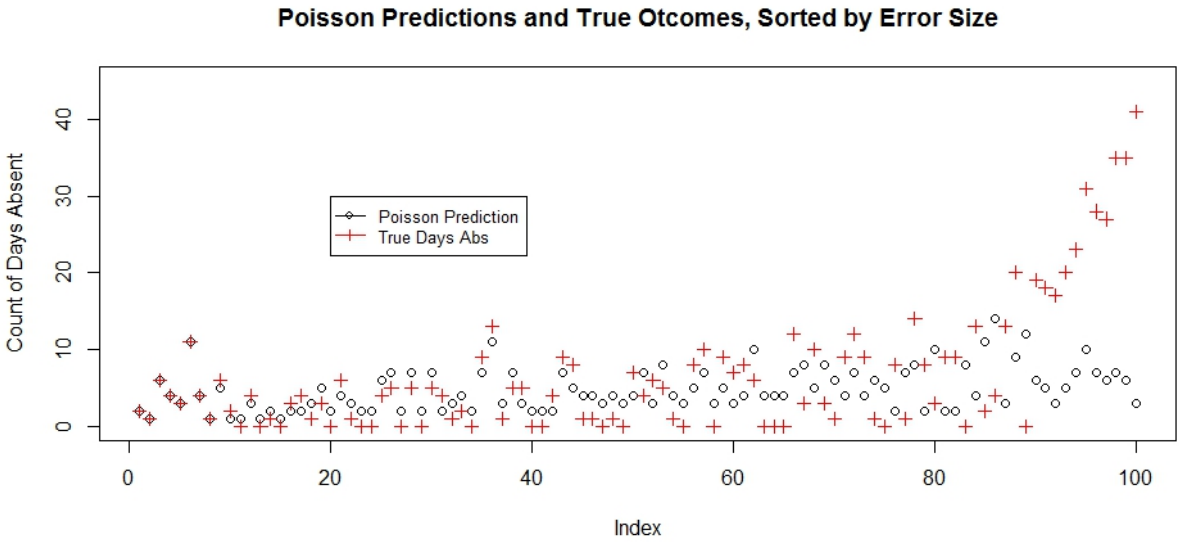


Figure 3: Bayes predictions and true value sorted by error.

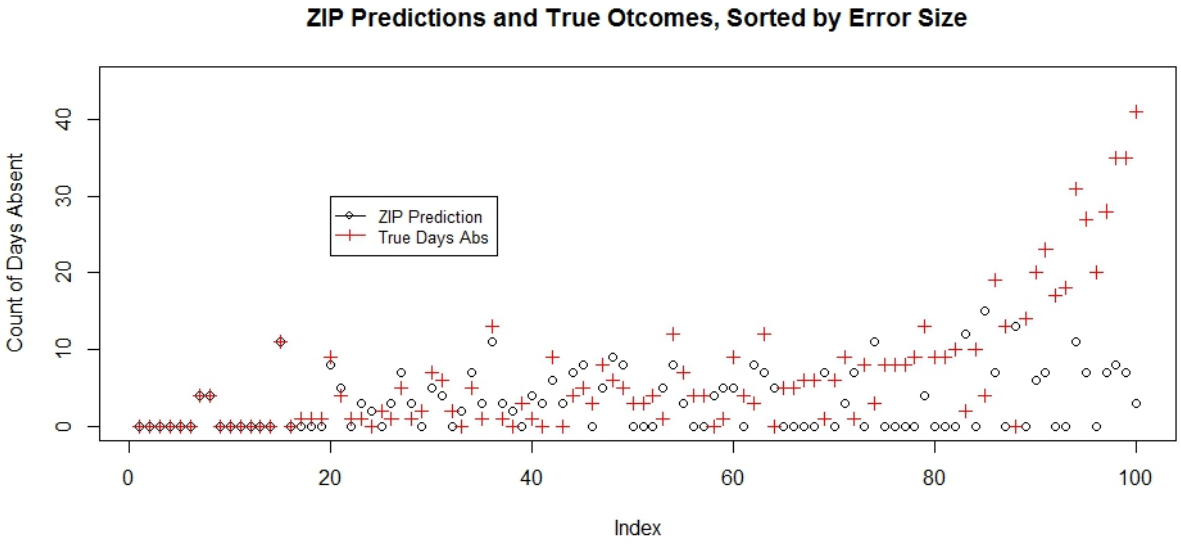


Figure 4: Bayes predictions and true value sorted by error.

5 Confirming on a Larger Data set

One major weakness in our study is that the test dataset has only 100 observations. In light of this, we found a larger dataset featuring 915 observations. We re-ran our analysis on this dataset, leaving aside 300 observations for testing. [12]

The outcome variable, `art`, is the number of articles published by biochemistry graduate students during the last three years of their Ph.D. Explanatory variables include: an indicator for females, an indicator for being married, (`fem` and `mar` respectively), number of children they have (under six years old) (`kid5`), prestige of the Ph.D program (ranging from 0.75 to 4.62) (`phd`), and number of articles published by the advisor in the last three years of the student's Ph.D (`ment`). [13] There is less overdispersion in this dataset compared to our smaller dataset ($\text{Var}(Y_i \mid \mathbf{X}) = 3.7 > \mathbf{E}[Y_i \mid \mathbf{X}] = 1.7$), but there are more zeros (30% are zero-counts, compared to 20% for our smaller dataset).

Table 7: **Test MSE**

SVR	Linear	Random Forest	ZIP	Tobit	Hurdle	ZINB	Poisson	NB2
3.18	3.39	3.54	3.73	3.78	3.93	4.04	4.20	4.54

Table 8: **Test Median Absolute Error**

Tobit	Random Forest	Linear	ZIP	Hurdle	ZINB	Poisson	NB2	SVR
0.97	1.07	1.12	1.12	1.13	1.14	1.17	1.18	1.32

Unlike our smaller dataset, the non-ZInf models did not do as well as the ZInf models when using the standard definition of a fitted value. However, this margin is extremely small. The discrepancy in Table 7 is driven by large prediction errors in higher counts. The plot in Appendix Figure 8 looks almost identical to our larger dataset. Median Absolute Error between our ZInf and non-ZInf is very close due to the median being more robust to large error. Our alternative, non-count methods, outperform ZIP in either definition of

[12] http://www.stata-press.com/data/lf2/couart2.dta

[13] For the curious: All variables turn out to be significant in all models except program prestige. Being married helps. Having children and being female both had negative coefficients.

error. We believe this is because of the increased size of the dataset.

When looking at the Bayes MSE and MAE [14], we see that errors are quite similar. The reason why ZInf models do not do exceedingly better is the same as in the previous section. This is clarified by the average predicted probability charts (Figure 5 and 6).

Table 9: **Test Mean Absolute Error (Bayes)**

Poisson	ZINB	Hurdle	NB2	ZIP
1.35	1.37	1.40	1.43	1.49

Table 10: **Test Mean Square Error (Bayes)**

Poisson	ZINB	Hurdle	ZIP	NB2
4.30	4.33	4.54	4.85	4.92

We see in Figure 5 and 6 that the regular Poisson does not assign the same probability of zero when the true value is zero. However, it does not do much worse than all the other models. This is because there is much less overdispersion in the data. The Poisson model assigns higher mass when the true count is low (1 and 2), which is where most of our data takes its values. Because there is less overdispersion, the pattern observed in Figure 5 and 6 is not as dramatic as in Figure 1 and 2. [15]

[14] For the Bayes concept of a fitted value, we look at MSE over MAE. This is because MAE returns a whole number and will give ties.

[15] As said before, the larger discrepancy for MSE in Table 7 was driven by outliers.

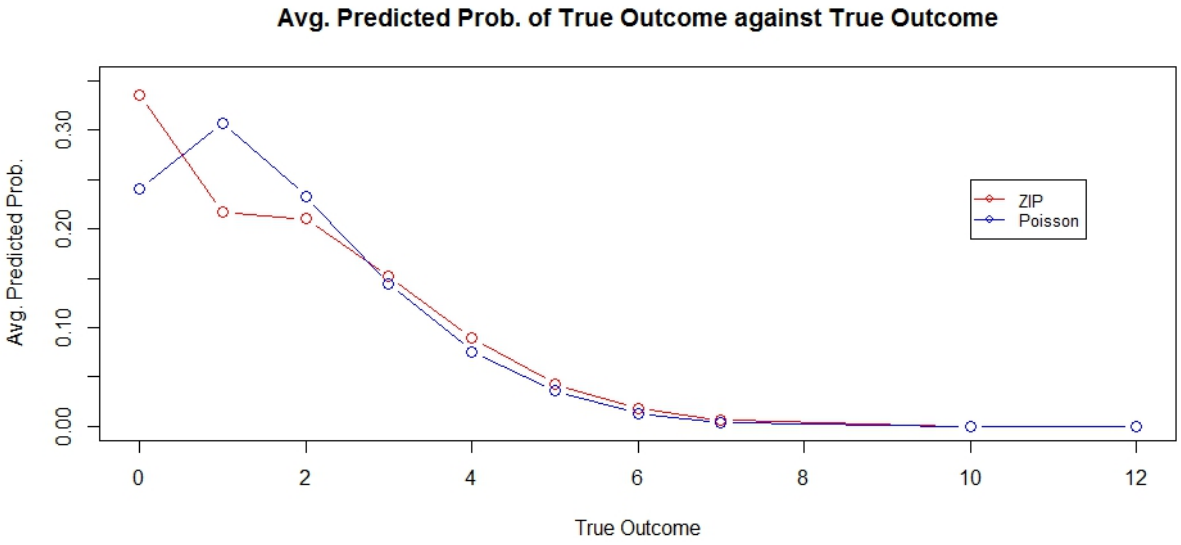


Figure 5: (Larger Dataset)

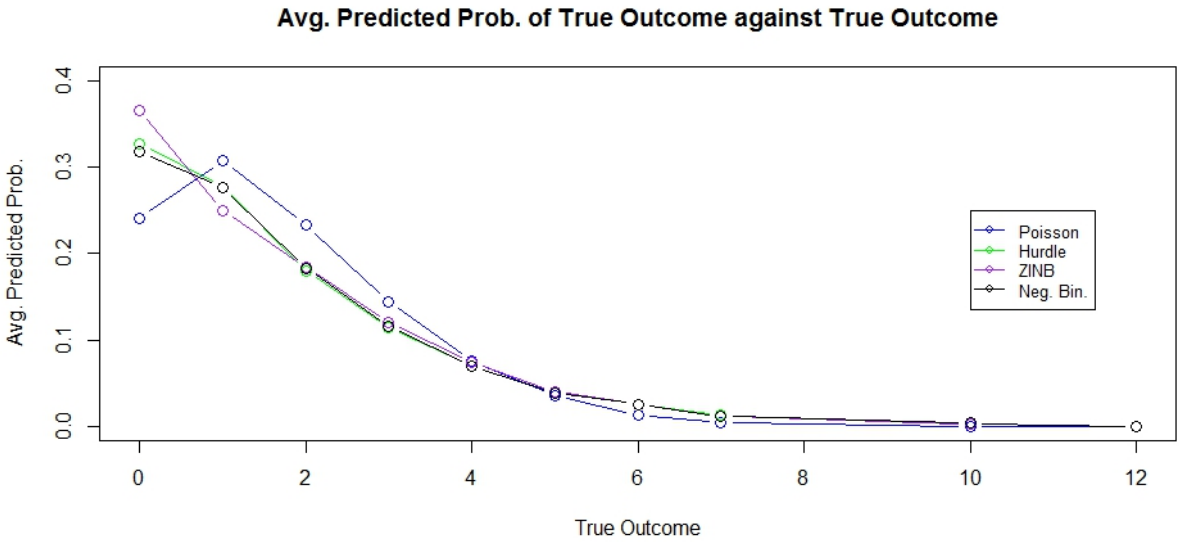


Figure 6: (Larger Dataset)

6 Conclusion

ZInf models claim to offer the best of both worlds: the ability to handle excess zeros, as well as the ability to model non-zero count outcomes. However, we found that ZInf models do not offer an increase in prediction accuracy over their non-Zinf counterparts. The total expected counts (of both zeros and non-zeros) from ZInf models match the data better

than non-ZInf models. This is often cited as one way to persuade the reader that ZInf models are superior. We have learned that this does not say anything meaningful about out-of-sample prediction accuracy.

We have seen that ZInf models do not provide better prediction accuracy compared to non-ZInf models. Generally, the prediction accuracies seem to be on par with each other, depending on the definition of fitted value used to compute errors. The main reason ZInf models did not do as well as expected is because they fail to assign a high probability to non-zero true outcomes. The regular Poisson model assigned a high probability to the true outcome in cases of low non-zero counts, but failed to do so for true zero counts.

Finally, if the data can support the methodology, we saw that alternative machine learning methods can provide better out- of-sample prediction error. This is not to say that ZInf models have no place in applied research. ZInf models still seem to be the most appropriately specified model from a probabilistic point of view [16], and their coefficient interpretations remain insightful for policy making.

[16] As it is confirmed by AIC, Vuong, and total expected count diagnostics.

7 Appendix

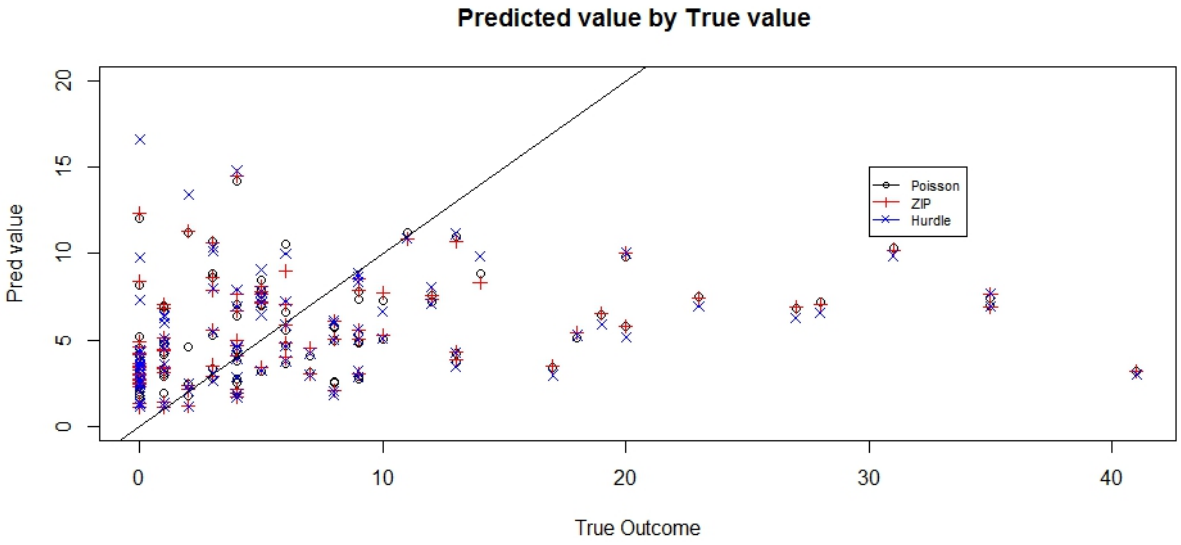


Figure 7: Predicted vs true on smaller dataset. (non-Bayes)

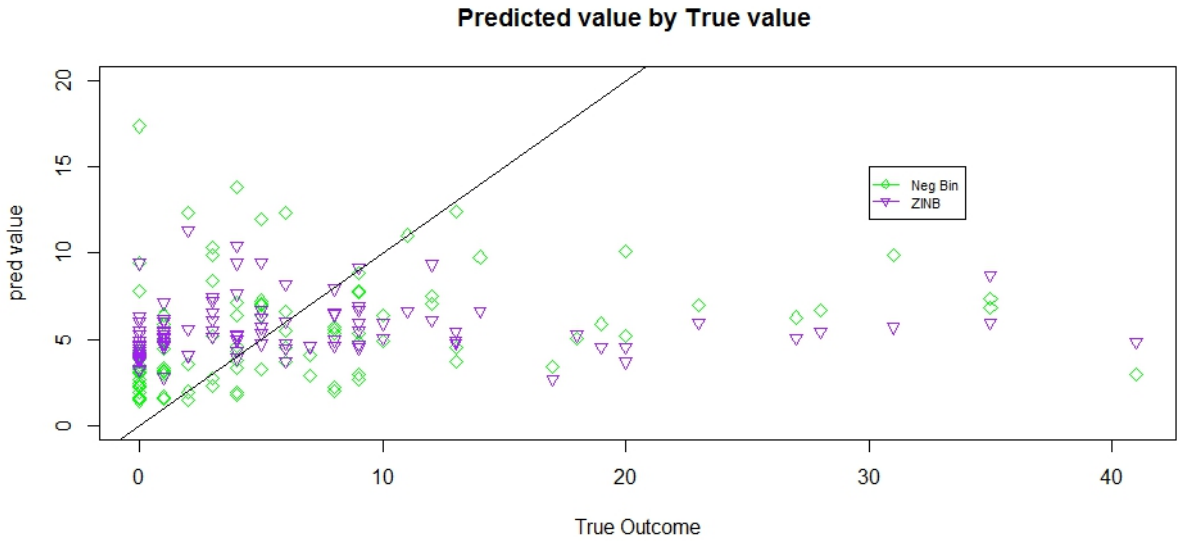


Figure 8: Predicted vs true on smaller dataset.(non-Bayes)

References

[1] Dietz E. Schlattmann P. Mendona L. Bhning D. and U. Kirchner. "The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology." In: Journal of the Royal Statistical Society: Series A (Statistics in Society) 162.4 (1999), 195209. DOI: 10.1111/1467-985X.00130.

[2] J. Cameron and P. Trivedi. Regression Analysis of Count Data. Cambridge University Press, Cambridge, 2013.

[3] Yin Bin Cheung. "Zero-inflated models for regression analysis of count data: a study of growth and development". In: Statistics in Medicine 21.10 (2002), pp. 1461–1469. ISSN: 1097-0258. DOI: 10.1002/sim.1088. URL: http://dx.doi.org/10.1002/sim.1088.

[4] Felix Famoye, John T. Wulu, and Karan P. Singh. "On the Generalized Poisson Regression Model with an application to Accident Data". In: Journal of Data Science (2004), pp. 287–295.

[5] D. Holden. "Testing for heteroskedasticity in the tobit and probit models ". In: Journal of Applied Statistics (2011), pp. 735–744.

[6] W.K. Newey. "Specification tests for distributional assumptions in the Tobit model." In: Journal of Econometrics (1987), pp. 125–145.

[7] C. E. Rose et al. "On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data". In: Journal of Biopharmaceutical Statistics 16.4 (2006), pp. 463–481.