Predicting the Number of Publications by Biochemistry Ph.D. Students (1950-1970)

Shuchi Goyal, Suoyi Yang, Heather Zhou

STATS 202B
Instructor: Dr. Chad Hazlett
University of California, Los Angeles

March 23, 2019

1 Introduction

Researchers must often choose between a wide array of statistical methods, including minor variations of standard models, when deciding the best way to analyze a set of data. On the one hand, the high-tech capabilities of the numerous machine learning algorithms proposed in recent years are tempting when accurate estimates are required from high-dimensional or otherwise complex data. On the other hand, some basic regression methods have the benefit of interpretability and often, with minor adjustments, can be made to suit data which violate traditional assumptions.

The "best" model for analysis clearly depends on the context of the problem that researchers are trying to address. However, the decision about the best model can be difficult when so many methods are available, particularly for those who are not familiar with the different motivations and assumptions associated with the various statistical methods. In this paper, we address this challenge by analyzing a single data set using various models from three broader categories of statistical analysis. In doing so, we reflect on the types of inferences we can make using different methods and when each might be most useful.

We will begin by introducing our data set, which is a subset of data used by Long (1990). We split our data into training and test sets, and perform analysis on the training data using generalized linear regression, regularized regression, and tree-based methods. We then apply the models produced using the training data onto the test data. Finally, we conclude by comparing the performances of the various models and identifying their advantages and disadvantages, as well as recommending areas for further study.

2 Data

Among the population of academics in the United States, one measure sometimes used as a proxy for a person's professional success is their publication output. To understand the direct and indirect effects of gender on the careers of academics, Long (1990) gathered data on 915 students who obtained Ph.D.s in biochemistry between 1950 and 1970. Each student is considered an observation.

The variable we are interested in estimating is the number of articles produced by the student during the last three years of his or her program. Additionally, five covariates were considered:

- A binary indicator variable for gender (female=1)
- A binary indicator variable for marital status (married=1)
- The number of children below the age of five
- The prestige of graduate program (rated between 0.75 and 4.62 from an external source)
- The number of articles produced by the student's mentor within the same three-year period that the student's productivity was measured

While the aim of Long's paper (1990) was to measure the interaction effects of gender with other factors, such as marital status or reputation of mentor, on the early career outcomes of Ph.D. students, our goal is to understand which variables are the most useful predictors for publication output of biochemistry students according to different statistical methods.

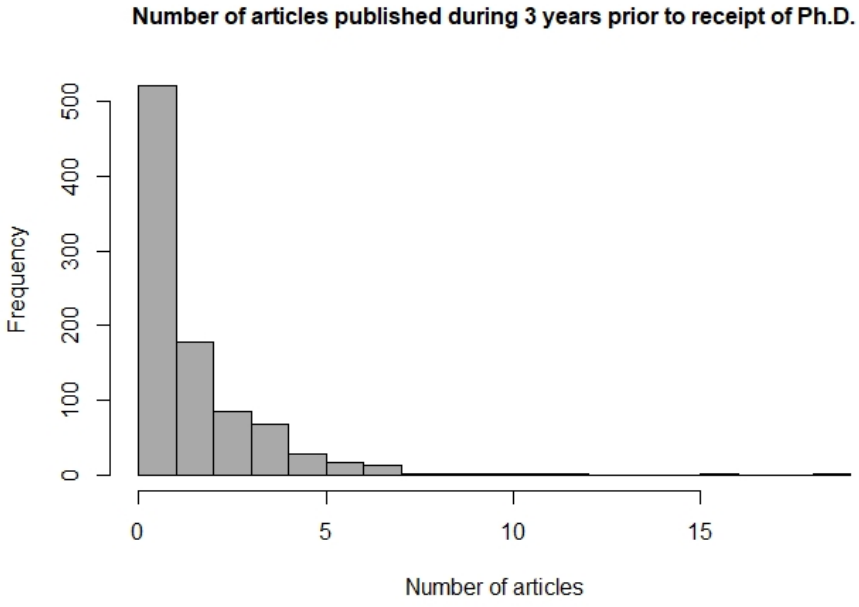


Figure 1: Histogram for number of articles published during final 3 years of Ph.D. program

A histogram of the number of publications across all students in the data set is shown in Figure 1. The first thing we note is that a large number of students in the data (over 500 observations) had not published any articles by the time they received their Ph.D. We therefore consider methods for our analysis which can accommodate the large count of 0s and judge if we need to account for this potential issue to produce more accurate predictions. We also observe a large variance in the response outcome. For example, one student was found to have been included as author on 20 publications.

3 Model Training

We mainly considered 3 categories of statistical methods:

- GLM methods
  - Poisson, negative binomial
  - Hurdle Poisson, hurdle NB
  - Zero-inflated Poisson, zero-inflated NB
- Regularized methods: Ridge, Lasso, KRLS
- Tree-based methods: CART, random forest, XG Boost

3.1 GLM Methods

For GLM methods, we begin with the Poisson regression as it is the most basic GLM model for count data. To account for over-dispersion in our data, we can use quasi-Poisson regression or negative binomial regression. Since negative binomial regression is usually favored over quasi-Poisson nowadays, we chose negative binomial regression.

To account for the large number of PhD students with 0 publications, we considered 4 GLM models that are designed to accomodate excess 0 counts in the data: hurdle Poisson, hurdle NB, zero-inflated Poisson, and zero-inflated NB. The results of our GLM models are summarized in Table 1. We will discuss the interpretation of the coefficient estimates in the next section.

The conventional way to compare different GLM methods is to look at their log likelihood along with the complexity of the model (measured by number of parameters). We prefer models with a high log likelihood and a small number of the parameters. From this regard, negative binomial regression achieves a happy medium.

Hurdle negative binomial regression is a natural extension of negative binomial regression. It also improves upon the negative binomial regression in terms of RMSE without sacrificing the log likelihood too much.

Table 1: Results from GLM Methods

Motivation		**Baseline**	**Overdispersion**	**Excess 0's**			
		Poisson	NegBin	Hurdle Poisson	Hurdle NB	ZI Poisson	ZI NB
Coef. Estimate	int	0.114	0.049	0.497 **	0.139	0.354 *	0.049
	female	-0.219 **	-0.208 *	-0.252 **	-0.264 *	-0.237 **	-0.208
	married	0.249 **	0.243 *	0.172 *	0.189	0.218 **	0.243
	numKid$\leq$ 5	-0.250 ***	-0.240 ***	-0.250 ***	-0.272 **	0.251 ***	-0.240 ***
	phdRating	0.050	0.058	0.026	0.049	0.042	0.058
	mentRating	0.025 ***	0.029 ***	0.020	0.025 ***	0.022 ***	0.029 ***
Model Eval.	numParameters	6	7	7	8	7	8
	Loglikeli	-1080.441	-1025.808	-1077.468	-1043.847	-1063.258	-1025.808
	RMSE (Training)	2.184	2.182	1.821	1.820	1.801	1.827

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Therefore, moving forward, we will focus on these two methods (negative binomial regression and hurdle negative binomial regression).

3.2 Regularized Methods

Although we do not expect regularization to be necessary since our data set has only five covariates, we will still try three regularized methods: ridge, LASSO, and KRLS. In addition to being regularized, KRLS is also a kernelized method and may reveal new information in conjunction with the other methods.

The results of these methods are summarized in Table 2, along with the results of the two GLM models we chose to focus on.

Table 2: Comparison of GLM and Regularized Models

Motivation				**Regularization**		**Kernelization**
		NegBin	Hurdle NB	Ridge	Lasso	KRLS
Coef. Estimate	int	0.049	0.139	1.077	1.099	na
	female	-0.208	-0.264	-0.333	-0.340	-0.278
	married	0.243	0.189	0.358	0.372	0.181
	numKid$\leq$ 5	-0.240	-0.272	-0.321	-0.341	-0.149
	phdRating	0.058	0.049	0.051	0.034	0.059
	mentRating	0.029	0.025	0.054	0.058	0.045
Model Eval.	RMSE (Training)	2.182	1.820	1.787	1.787	1.746

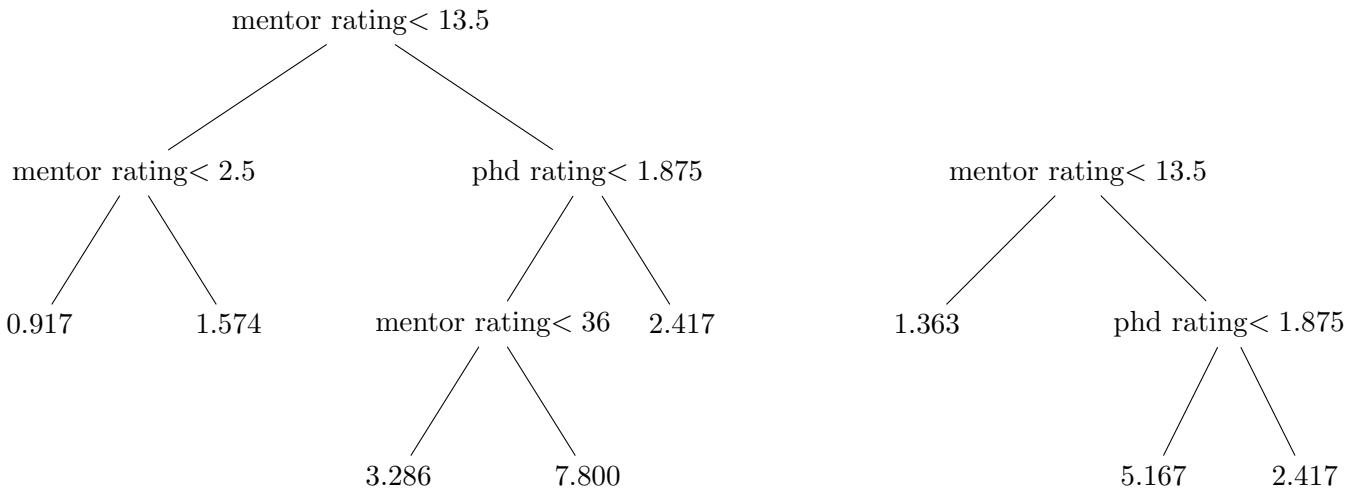
3.3 Tree-based Methods

For tree-based methods, we ran CART, random forest, and XG boost.

For CART, the full tree and the pruned tree (number of terminal nodes selected via cross validation) are shown in Figure 2 below. We can see that the only two covariates selected by the model, in both the full and pruned versions, are mentor and PhD rating.

For random forest, we tuned the parameter mtry based on cross validation.

For XG boost, we tuned the parameter nround using the default values for the other parameters. Unfortunately, due to time restriction, we were not able to tune all the parameters.



(a) Full CART model

(b) Pruned CART model (best=3)

Figure 2: Trees produced by CART method

3.4 Training RMSE Comparison

We compare the training root mean squared error (RMSE) for our models in the table below. We can see that the XG Boost model performed the best with a RMSE of 1.562. It is followed closely behind by KRLS, ridge, and LASSO.

Table 3: Training RMSEs (Unscaled)

Method	**GLM**		**Regularization/KRLS**			**Tree-based Methods**		
	NegBin	Hurdle NB	Ridge	Lasso	KRLS	CART (n=3)	Random Forest	XG Boost
RMSE	2.182	1.820	1.787	1.787	1.746	1.796	1.829	1.562

4 Interpretation

Predictive power of the models aside, we are also interested in the the effects that variables such as gender and children have on the number of articles produced by a student in the last 5 years of

their PHD.

Our GLM and regularized methods all produced consistently negative coefficients for the variables "female" and "number of kids under the age of 5." Recall that the interpretation of the coefficient for negative binomial and Poisson GLM are not the same as those of linear regression. If we increase a predictor x in our model by one unit, our predicted value will be multiplied by $\exp(\beta_x)$. Thus since the coefficients of female and children are negative, we are multiplying the predicted variable by a value less than 1, thus decreasing the predicted value (which is the number of articles produced). Therefore, all our models seems to indicate that both being a woman and having young children negatively impacts the amount of article a student produces, and consequently, the productivity of a student's graduate career.

We are unable to determine the effects that children and gender has on the number of articles produced for tree-based methods. While they are good predictive models, they do not produce easily interpretable coefficients like the other models do. However, it can be seen from the CART models that the two variables with the most significant predictive power are "mentor rating" and "phd rating," while variables like "female" and "children" do not play a significant role in the model.

5 Test Results

We now look at the performance of our models on the test data set.

The plots in Figure 3 show the actual density of the number of publications in the test data set (yellow bars) against the density of the predicted number of publications (red curves).

5.1 Model Comparison based on RMSE

We calculated two measures of test accuracy: RMSE and scaled RMSE. For RMSE, we simply calculated

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Based on RMSE, ridge and LASSO performed the best, closely followed by KRLS and random forest. Negative binomial regression performed the worst in terms of RMSE (RMSE=2.343).

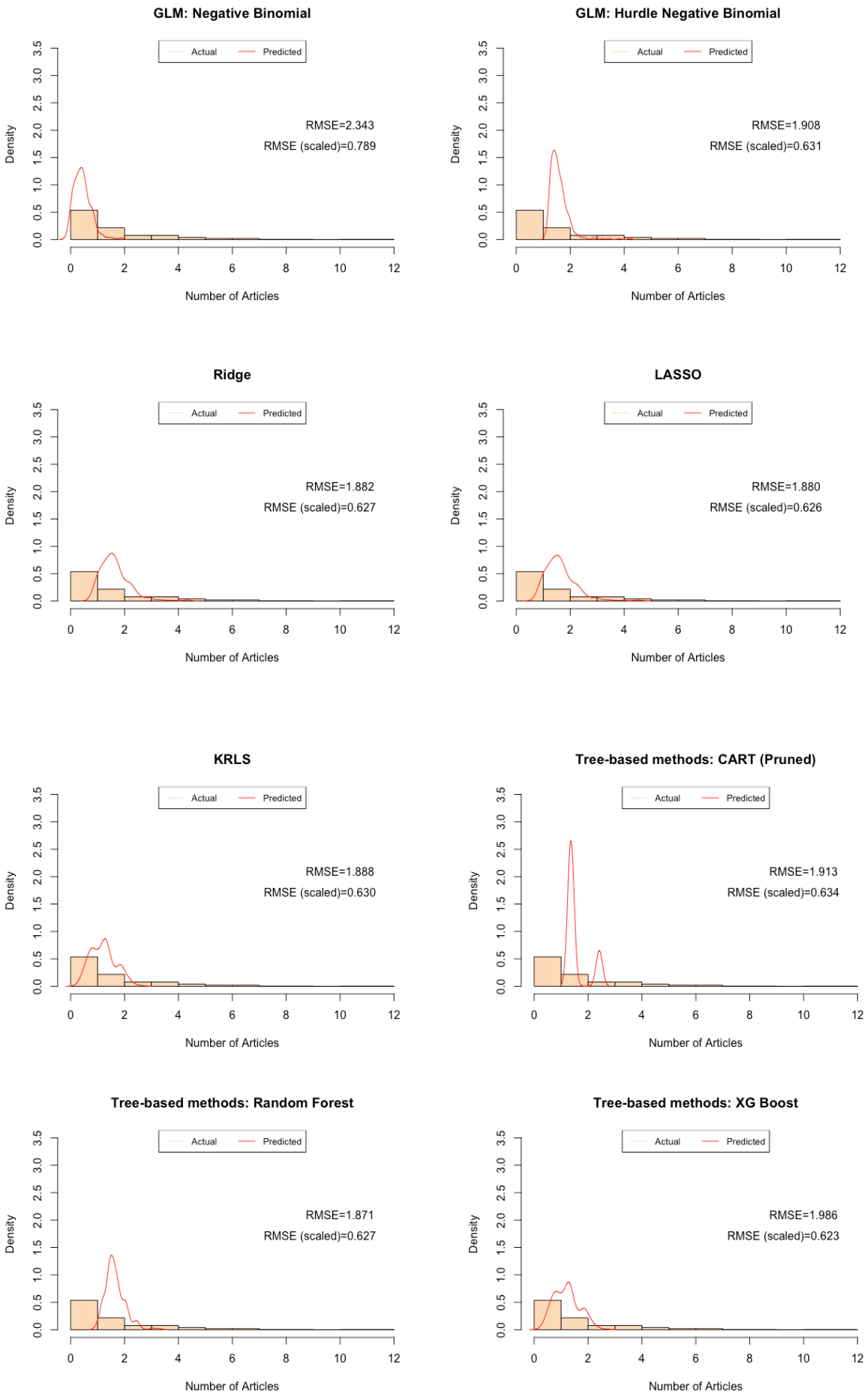


Figure 3: Test Results

5.2 Model Comparison based on Scaled RMSE

As Professor Hazlett and our fellow classmates pointed out during our presentation, when you look at the actual versus predicted densities, negative binomial regression seems to capture the shape of the density quite well. However, it has the highest RMSE among the models considered. This is because the negative binomial model does not predict high numbers of publications well, and that is heavily penalized in terms of RMSE. This is counterintuitive, however, because generally, the difference between a student with one publication and a student with no publications is likely more meaningful that the difference between a student with 14 publications and a student with 15 publications.

To accommodate this issue, we also calculate the RMSE when the response variable has been log-scaled. That is,

$$\text{RMSE (scaled)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big[\log(y_i+1) - \log(\hat{y}_i+1)\Big]^2}$$

To further understand the reasoning behind this, we consider two cases:

- *Case 1*: the actual number of publication of a student is 2, but we predicted 0
- *Case 2*: the actual number of publication of a student is 16, but we predicted 14

We want to penalize the error in Case 1 more heavily than the error in Case 2. But RMSE penalizes both situations equally. With scaled RMSE, we penalize Case 1 more, since

$$\log(2+1) - \log(0+1) = 1.099$$

whereas

$$\log(16+1) - \log(14+1) = 0.125$$

In terms of scaled RMSE, XG Boost performs the best (scaled RMSE=0.623), closely followed by ridge, LASSO, KRLS, and random forest. While the negative binomial regression still gives the highest scaled RMSE, its performance relative to other models considered is better in the scaled RMSE analysis than in the unscaled RMSE analysis.

5.3 Remarks

In terms of how well the density of the predicted number of publications captures the shape of the density of the actual number of publications, we think that negative binomial regression, KRLS, and XG Boost performed the best. Negative binomial regression closely replicates the peak at 0, while KRLS and XG Boost replicate the overall shape the best.

During the course of this project, we did not have sufficient time to fully tune the parameters in XG Boost. But even with our limited parameter tuning, XG Boost is already showing a lot of potential. Therefore, one of the things we can continue to do with this project is to fully explore the potential of XG Boost.

6 Conclusion

Our analysis reinforced the overarching concept that while GLM methods are generally the most interpretable options for model fitting and tell us the most about *how* different variables effect our response, tree-based methods tend to be the most accurate. The choice between these categories of models is usually dependent on the researcher's goals. For example, if we wanted to follow Long's (1990) example, we may be interested in understanding the interaction between marital status and gender and its effect on article output, and we would probably prefer GLM methods. If, however, a student was simply interested in predicting her own success in publishing papers based on her mentor, she may choose tree-based methods in her analysis instead.

We also found that the ridge and LASSO regularization methods struck a good balance between interpretability and accuracy. However, it is unclear whether these models are appropriate for our data. Regularization methods are traditionally used to handle high-dimensional data, which was not the case with our data set.

We also note the lackluster performance of zero-accommodating GLM methods. This was surprising to us, as we had expected the hurdle and zero-inflation methods to improve the performance of the basic Poisson and negative binomial models. We suspect that the reason these methods were ineffective is that the large number of zeros does not necessarily distort the shape of the distribution. That is, even with the large number of zeros, the distribution is still unimodal.

References

[1] Long, J. *The Origins of Sex Differences in Science.* Social Forces, 68(4), 1297-1316. doi:10.2307/2579146