Group Regularization for Zero-Inflated Count Regression Model with an Application to Female PhD Student's Scientific Productivity

Rommel Bartolome

Abstract

The systemic differences in the outcomes that men and women achieve in health, education, economy and politics, also known as the Gender Gap, has been one of the biggest problem the world is trying to solve. In particular, this is evident in the academic world where women are underrepresented especially in academic science programs. Several studies have been made regarding the Gender Gap and some of these studies use zero inflated data and does not use grouped covariates, which can provide a better method of analysis compared to regular EM LASSO in terms of model error and in computational overhead. In this paper, we demonstrate the use of an algorithm called Group regularization for zero-inflated model (GOOOGLE) developed by Chatterjee et Al. and Chowdhury et Al. (2018) using the productivity data of PhD students in terms of published articles, with gender as one of the response variables. The GOOOGLE method in its Poisson Regression Form using a Bridge variable selector showed that the quality of the graduate program and the mentor is the biggest indicator of a productive PhD student. However, it also showed that being a female and having young kids are indicators of a non-productive PhD student. To battle the Gender Gap, it is recommended that graduate programs and mentors of the highest quality to take more female students.

Introduction

The Gender Gap has existed throughout history (Kleinberg, Jay Kleinberg, and Goldin 1991). While we have seen a significant closing of the gap in recent years, it is estimated that the gap between men and women, measured in terms of political influence, economic gain and health and education will take 99.5 years to close according to the World Economic Forum (WEF). The lack of representation of women in emerging fields such as Cloud Computing and AI is contributing to this gap, even though we are seeing that more women are currently taking positions of power throughout the world (Han and Budig 2019).

One of the most factors that are checked in the Gender Gap issue is gender equality in terms of educational attainment. In the 2019 WEF report, only 35 countries has achieved this. However, compared to other Gender Gap areas, this is the most attainable in the near future where WEF expects that all countries will eliminate the Gender Gap in the next 12 years (Munene and Wambiya 2019) (Alfarhan and Dauletova 2019). It is important to tackle this at a doctoral level, as this is the period where many women are in the age of having kids and thus in the traditional society, are expected to leave their careers to pursue motherhood solely.

The nature of most data regarding post-doctoral studies on the scientific publications of PhD students are zero-inflated. This means that most of these students don't have publications, or are still working on creating one. Normal Negative Binomial and Poisson Regression model cannot handle this due to the number of zeros in the dataset.

In this paper, we implement two new methods of modelling for a zero-inflated dataset with group regularization using a Negative Binomial Regression Model (Chatterjee et al. 2018) and then using a Poisson Regression Model (Chowdhury et al. 2019). For the dataset, we used the data from the AER package in R called PhDPublications from (Long 1990). To our knowledge, this is the first implementation of such dataset to this newly created models.

Review of Related Literature

Gender Gap has been one of the most talked about topics in the past century. Most of the studies made are economic ones (Kleinberg, Jay Kleinberg, and Goldin 1991), usually tackling pay (Han and Budig 2019) as the main point. There are also studies about the political aspect of such (Mueller 1988). Interestingly, physiological effects have also been studied such as Gender Gap in heart failure (Lin and Greenberg 2019), life expectancy (Oecd and OECD 2018) and even on the genes itself (Barash and Lipton 2017).

Moreover, Gender Gap tackling specifically education has also been well-studied. In (Alfarhan and Dauletova 2019), they have revisited the gap in terms of academic achievement while in (Munene and Wambiya 2019) they revisited it in terms of patriarchy. Several authors have also given in-depth solutions in terms of academic participation (Stout et al. 2013), cognitive and social-emotional skills gap bridging (Nakajima et al. 2019), and most importantly relating this gap to earnings (Han and Budig 2019).

In the work of (Long 1990), he was able to pinpoint the origin of sex difference in science. In his work, he related the number of publications made by PhD students in relation to the graduate program, mentor, number of young kids, gender and if the student is married. The data he used is zero-inflated, with many of the students having no publications in the last 3 years of their PhD. During his time, he used the usual zero inflated models.

The recent work of (Chatterjee et al. 2018) and (Chowdhury et al. 2019) has showed a lot of promise in modelling zero-inflated data. In their work, they utilized group regularization to improve the overall performance of the Negative Binomial and Poisson Zero Inflated Model. Zero-inflated models have been used in many modelling applications in number of falls in the elderly (Ob and Yusuf 2017), aphid population study (Carvalho, Santana, and Sampaio 2019), HIV mortality (Musal and Aktekin 2013), disease mapping (Torabi 2017), dental caries (Gilthorpe et al. 2009) and cardiac surgery in children (Z. Wang et al. 2014), among others. The data and the work of (Long 1990) can surely benefit in this newly minted technique.

Methodology

Zero-inflated models assume that a significant population will have zero as their response variable. From (Greene 1994), Zero-inflated negative binomial (ZINB) mixture distribution can be written as:

$$P\left(y_i = k\right) = \begin{cases} p_i + (1-p_i)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}} & \text{if } k = 0 \\ (1-p_i)\frac{\Gamma\left(\alpha^{-1}+k\right)}{\Gamma(k+1)\Gamma(\alpha^{-1})}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^k\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha'} & \text{if } k = 1, 2, \cdots \end{cases}$$

For n observations, the ZINB negative log-likelihood is as follows:

$$\begin{aligned} L(\theta) = & -\sum_{y=0}\log\left[p_t + (1-p_t)\left(\frac{\alpha^{-1}}{\mu_t+\alpha^{-1}}\right)^{\alpha^{-1}}\right] \\ & -\sum_{y\leq 0}\log\left[(1-p_t)\frac{\Gamma\left(\alpha^{-1}+y_1\right)}{\Gamma\left(y_t+1\right)\Gamma\left(\alpha^{-1}\right)}\left(\frac{\mu_t}{\mu_t+\alpha^{-1}}\right)^n\left(\frac{\alpha^{-1}}{\mu_t+\alpha^{-1}}\right)^{\alpha^{-1}}\right] \end{aligned}$$

For the Poisson model, we assume that the variance should at least be well-approximated by the mean. The Negative Binomial model is used to account for overdispersion, only a slightly less parsimonious.

The groupings the authors created are least absolute shrinkage and selection operator (LASSO), bridge, smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP) which carries different penalty terms. The equations below shows that penalty terms of various group methods under GOOOGLE. The full explanation of the methodology can be seen in (Chatterjee et al. 2018) and (Chowdhury et al. 2019).

Group LASSO Penalty Term:

$$\lambda_n \sum_{k=1}^{K} \sqrt{m_k}\, \|\theta_{A_k}\|_2$$

Group bridge Penalty Term:

$$\lambda_n \sum_{k=1}^{K} \sqrt{m_k}\, \|\theta_{A_k}\|_1^v, 0 < v < 1$$

Group SCAD Penalty Term:

$$\sum_{k=1}^{k} \rho\left(\|\theta_{A_k}\|_2; jm_k\lambda_n, v\right), \rho\left(x_i, \lambda_n, v\right) = \lambda_n \int_0^k \min\left\{1, (v - t/\lambda_n)_+ / (v-1)\right\} dt, v > 2$$

Group MCP Penalty Term:

$$\sum_{k=1}^{k} p\left(\|\theta_A\|_2; v_m\lambda_n, v\right), \rho\left(x; \lambda_n, v\right) = \lambda_n \int_0^\infty \left(1 - t/\left(v\lambda_n\right)\right)_+ dt, v > 1$$

All codes and the complete methodology can be seen in this Github link:

- https://github.com/rlbartolome/stat269_finalproject

Results

In this paper, we implement the group regularization for zero-inflated regression model in both the Negative Binomial and the Poisson form, using PhD Publications data from the R AER package. The dataset contains 915 observations pertaining to the cross-sectional data on the scientific productivity of PhD students in biochemistry. This was from the study of Long J.S. where they tackled the origin of sex difference in science. There were six variables in the study, and all were used in the paper. Table 1 shows the description of the predictor variables, stratified by groups, in the PhD Publications data. Take note that similar to how the original authors of the GOOOGLE method, we express each continuous predictor as a group of three cubic splines variables (*prestige, mentor and kids*), resulting to a total of 11 candidate predictors, three of which are triplets and two are singleton groups.

Table 1: Description of predictor variables (stratified by groups) in the PhD Studies data

Group	Variable	Description
1	prestige1, prestige2, prestige3	First, second, and third degree polynomial of prestige of the graduate program.
2	mentor1, mentor2, mentor3	First, second, and third degree polynomial of number of articles published by student's mentor.
3	kids1, kids2, kids3	First, second, and third degree polynomial of number of children less than 6 years old.
4	gender	1 if female, 0 otherwise
5	married	1 if married, 0 otherwise

The variable that we want to predict is the number of publications a PhD student makes. We look at the distribution of the said data in Figure 1. Here, we can see that clearly, a huge chunk of the data are zeroes which tells us that traditional models such as regular Poisson and Negative Binomial regression cannot fully model the said dataset.

In the paper of (Chatterjee et al. 2018) and (Chowdhury et al. 2019), they already showed that GOOOGLE

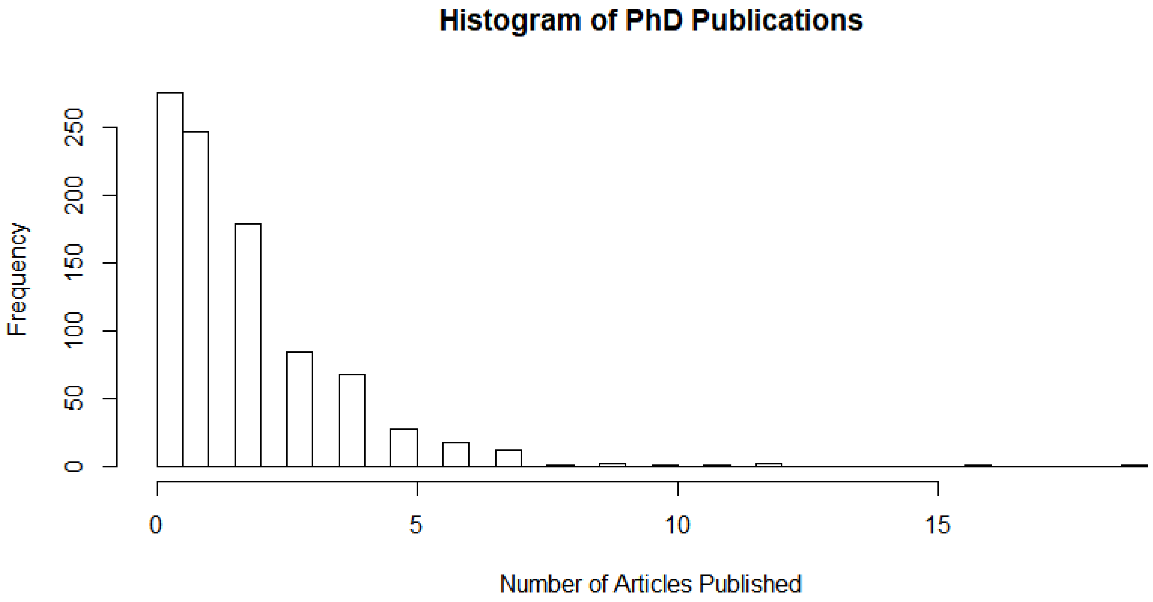

Figure 1: Number of articles published during last 3 years of PhD.

significantly outperforms published non-grouped EM LASSO method in out-of-sample prediction. They also showed that GOOOGLE was significantly faster in terms of computational overhead compared to EM LASSO which was 20 to 25 times slower. In this paper, we show the performance of GOOOGLE in the PhD Studies dataset in Poisson and Negative Binomial Regression form. We also show the performance per group variable selection method (LASSO, MCP, SCAD and Bridge).

It should be noted that the LASSO, MCP and SCAD group variable selection offers all variables in the model while the Bridge group variable selection offers a selection procedure where only the most significant coefficients are added. The selection of the "best" model in terms of statistical measures is not part of the scope of this paper.

Table 2 and 3 shows the coefficient estimates for the count and zero submodel in the PhD Studies data analysis for the Negative Binomial regression form of GOOOGLE. From Table 4, the zero submodel appears to be inflated in many of the group variable selection method and thus, does not appear to be the best selection for a GOOOGLE model. Overall, the Negative Binomial count submodel shows that prestige and mentor positively affects the number of articles published by a PhD student while having kids and being females affects it negatively.

Table 4 and 5 on the other hand shows the coefficient estimates for the count and zero submodel in the PhD Studies data analysis for the Poisson regression form of GOOOGLE. Similar to the zero count model previously seen, the coefficients appears to be inflated with confusing signs, depending on the group value number. Again, it appears that the zero count (logistic) model is not appropriate given the dataset. The Poisson count model shows a better level of coefficient factors. Similar to the Negative Binomial count submodel, it also indicates that prestige and mentor is a positive factor for number of articles published by a PhD student while kids and being a female are negative factors.

Table 2: Coefficient estimates for the count submodel in the PhD Studies data analysis (Negative Binomial)

	Lasso	MCP	SCAD	Bridge
intercept	0.0414	0.0356	0.0357	0.0319

	Lasso	MCP	SCAD	Bridge
prestige1	0.0089	0.0016	0.0017	0.1577
prestige2	0.1526	0.1486	0.1487	0.0396
prestige3	-0.1246	-0.1283	-0.1282	0.0000
mentor1	0.8552	0.8539	0.8540	0.7808
mentor2	1.8460	1.8402	1.8403	1.7645
mentor3	0.1045	0.0862	0.0866	0.0000
kids1	-0.1464	-0.1532	-0.1530	0.0000
kids2	0.3276	0.3173	0.3175	0.0000
kids3	-0.8629	-0.8755	-0.8753	-0.5700
gender	-0.2060	-0.2082	-0.2081	-0.1653
married	0.0493	0.0537	0.0536	0.0000

Table 3: Coefficient estimates for the zero (logistic) submodel in the PhD Studies data analysis (Negative Binomial)

	Lasso	MCP	SCAD	Bridge
intercept	-1.6706	-2.1861	-2.2169	0.4903
prestige1	64.4056	64.5828	64.5777	56.2164
prestige2	1.2253	-0.3574	-0.4257	13.8734
prestige3	35.1612	35.6163	35.6246	29.5592
mentor1	-49.2659	-43.8046	-43.4197	-11.2431
mentor2	-1334.8851	-1782.7423	-1807.8653	0.0000
mentor3	-1592.8203	-2113.5617	-2142.7963	0.0000
kids1	-1307.6120	-1718.9893	-1743.3815	0.0000
kids2	3275.9583	4303.5054	4364.0379	0.0000
kids3	-3944.5979	-5179.9834	-5252.8859	0.0000
gender	-0.8741	-0.9102	-0.9111	0.0000
married	-3.4983	-3.4753	-3.3998	0.0000

Table 4: Coefficient estimates for the count submodel in the PhD Studies data analysis (Poisson)

	Lasso	MCP	SCAD	Bridge
intercept	0.0414	0.0356	0.0357	0.0319
prestige1	0.0089	0.0016	0.0017	0.1577
prestige2	0.1526	0.1486	0.1487	0.0396
prestige3	-0.1246	-0.1283	-0.1282	0.0000
mentor1	0.8552	0.8539	0.8540	0.7808
mentor2	1.8460	1.8402	1.8403	1.7645
mentor3	0.1045	0.0862	0.0866	0.0000
kids1	-0.1464	-0.1532	-0.1530	0.0000
kids2	0.3276	0.3173	0.3175	0.0000
kids3	-0.8629	-0.8755	-0.8753	-0.5700
gender	-0.2060	-0.2082	-0.2081	-0.1653
married	0.0493	0.0537	0.0536	0.0000

Table 5: Coefficient estimates for the zero (logistic) submodel in the PhD Studies data analysis (Poisson)

	Lasso	MCP	SCAD	Bridge
intercept	0.1621	0.1393	0.1398	0.1971
prestige1	-0.5197	-0.5417	-0.5412	-0.7075
prestige2	0.5050	0.4958	0.4960	0.0000
prestige3	-0.5725	-0.5886	-0.5882	-0.4069
mentor1	-8.0249	-8.0737	-8.0729	-4.7199
mentor2	17.5229	17.7458	17.7418	1.7371
mentor3	-42.5254	-43.9041	-43.8773	0.0000
kids1	323.0554	277.2466	278.1835	0.0000
kids2	-804.0099	-689.5879	-691.9278	0.0000
kids3	966.9319	829.5317	832.3416	0.0000
gender	0.1114	0.1072	0.1073	0.0000
married	-0.6743	-0.6587	-0.6591	0.0000

Overall, we will choose the GOOOGLE method in its Poisson Regression form with Bridge as the group variable selection method as our model. Here, we can see that the top contributor on the number of articles a PhD will publish in their last 3 years of PhD is the quality of the mentor, specifically the number of articles published by their mentor has. Next will be the prestige of the graduate program, with the higher prestige of the program the better chances of having more published articles. This is intuitive as we would expect that the quality of mentorship and the graduate program itself would be the main factor for productivity.

It has also been found that having kids is the most significant factor in publishing less articles, followed by being a female. This is also intuitive as we would usually have parents take more leaves and would need more personal time with their young kids compared to those without one. This takes a toll more on females given that they are usually tasked to take care of the kids.

Conclusions and Recommendations

We have successfully demonstrated the effectiveness of the Group regularization for zero-inflated model (GOOOGLE) method created by (Chatterjee et al. 2018) and (Chowdhury et al. 2019) using PhD Studies data by (Long 1990). Among all GOOOGLE methods, the Poisson Regression Form using a Bridge variable selector was selected to model the said data. It has been found that the quality of the graduate program and the mentor is the biggest indicator of a productive PhD student. This is expected as quality education (higher prestige score) and educators (mentors that are also productive) are very likely the reasons why a PhD student is able to produce more published research papers.

However, the same model also showed that being a female and having young kids are indicators of a non-productive PhD student. Interestly, being married is not considered to be a significant factor, either as a good or bad influencer of PhD productivity. It appears that bearing a child and taking care of it burdens the female PhD student in being more productive. This is explainable as female PhD students will likely be in majority, if not solely, the caretaker of their young child.

To battle the Gender Gap, it is recommended that graduate programs and mentors of the highest quality to take more female students. This will enable more productive female PhDs. In addition, the fact the having young kids and being female are both indicators of a less productive PhD student, there should be more opportunities for female PhDs to focus on their careers while having a kids. This can be of the form of day care centers, nannies or just the simple sharing of responsibility of taking care of their young children with their husbands. It should be noted though that further studies much be taken before implementing the these recommendations.

Aside from that, other variations other than the Poisson and Negative Binomial count models can be explored. One model is the Poisson Inverse Gaussian Model, which provide for adjustment of greater amounts of overdispersion than the negative binomial model. Another path that can be taken is a time-series path, where we account for the temporal changes of the data. We can also do a Bayesian approach on this one, so we can add priors to our predictions.

Furthermore, improvements in the data can also been done. The data that was used in this study is relatively old. It would be beneficial to know the most recent status of the productivity of PhD students, as the realities might have changed already in the past few years.

References

Alfarhan, Usamah F, and Victoria Dauletova. 2019. "Revisiting the Gender Academic Achievement Gap: Evidence from a Unique Environment." *Gender and Education*.

Barash, David P, and Judith Eve Lipton. 2017. *Gender Gap: How Genes and Gender Influence Our Relationships*. Routledge.

Carvalho, F J, D G de Santana, and M V Sampaio. 2019. "Modeling Overdispersion, Autocorrelation, and Zero-Inflated Count Data via Generalized Additive Models and Bayesian Statistics in an Aphid Population Study." *Neotrop. Entomol.*, November.

Chatterjee, Saptarshi, Shrabanti Chowdhury, Himel Mallick, Prithish Banerjee, and Broti Garai. 2018. "Group Regularization for Zero-Inflated Negative Binomial Regression Models with an Application to Health Care Demand in Germany." *Stat. Med.* 37 (20): 3012–26.

Chowdhury, Shrabanti, Saptarshi Chatterjee, Himel Mallick, Prithish Banerjee, and Broti Garai. 2019. "Group Regularization for Zero-Inflated Poisson Regression Models with an Application to Insurance Ratemaking." *Journal of Applied Statistics*.

Gilthorpe, Mark S, Morten Frydenberg, Yaping Cheng, and Vibeke Baelum. 2009. "Modelling Count Data with Excessive Zeros: The Need for Class Prediction in Zero-Inflated Models and the Issue of Data Generation in Choosing Between Zero-Inflated and Generic Mixture Models for Dental Caries Data." *Statistics in Medicine*.

Greene, William H. 1994. *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*.

Han, Joohee, and Michelle Budig. 2019. "Gender Pay Gap." *Sociology*.

Kleinberg, S Jay, S Jay Kleinberg, and Claudia Goldin. 1991. "Understanding the Gender Gap: An Economic History of American Women." *The Economic Journal*.

Lin, Felice, and Barry Greenberg. 2019. "Considering the Gender Gap in Heart Failure." *Eur. J. Heart Fail.*, December.

Long, J S. 1990. "The Origins of Sex Differences in Science." *Social Forces*.

Mueller, Carol Mcclurg. 1988. *The Politics of the Gender Gap: The Social Construction of Political Influence*. Sage Publications, Inc.

Munene, Ishmael I, and Paschal Wambiya. 2019. "Bridging the Gender Gap Through Gender Difference: Aiding Patriarchy in South Sudan Education Reconstruction." *Africa Education Review*.

Musal, Muzaffer, and Tevfik Aktekin. 2013. "Bayesian Spatial Modeling of HIV Mortality via Zero-Inflated Poisson Models." *Statistics in Medicine*.

Nakajima, Nozomi, Haeil Jung, Menno Pradhan, Amer Hasan, Angela Kinnell, and Sally Brinkman. 2019. "Gender Gaps in Cognitive and Social-Emotional Skills in Early Primary Grades: Evidence from Rural

Indonesia.” *Dev. Sci.*, December, e12931.

Ob, Yusuf, and O B Yusuf. 2017. “Zero Inflated Poisson and Zero Inflated Negative Binomial Models with Application to Number of Falls in the Elderly.” *Biostatistics and Biometrics Open Access Journal*.

Oecd, and OECD. 2018. “Regional Gender Gap in Life Expectancy at Birth (Female-Male), 2016.”

Stout, Jane G, Tiffany A Ito, Noah D Finkelstein, and Steven J Pollock. 2013. “How a Gender Gap in Belonging Contributes to the Gender Gap in Physics Participation.”

Torabi, Mahmoud. 2017. “Zero-Inflated Spatio-Temporal Models for Disease Mapping.” *Biometrical Journal*.

Wang, Zhu, Shuangge Ma, Ching-Yun Wang, Michael Zappitelli, Prasad Devarajan, and Chirag Parikh. 2014. “EM for Regularized Zero-Inflated Regression Models with Applications to Postoperative Morbidity After Cardiac Surgery in Children.” *Statistics in Medicine*.