ARTICLE PRODUCTION BY GRADUATE STUDENTS IN BIOCHEMISTRY PH.D. PROGRAMS

Lumia Giulia, Bongiovanni Sabrina Anna, D'Antoni Marco

ABSTRACT

The article production of 915 Ph.D. biochemistry students was analysed in order to detect which factors may have an influence on it. A comparison among different models was carried out, since problems related to overdispersion and zero-inflation were detected.
The optimal model turned out to be the Negative Binomial GLM.

OUTLINE

1. Introduction ..... pag. 1

2. Descriptive analysis ..... pag. 2

3. Poisson Generalized Linear Model ..... pag. 6

4. Negative Binomial Linear Model ..... pag. 9

5. Zero-Inflated Negative Binomial Model pag.12

6. Hurdle Negative Binomial Model pag.14

7. Conclusion pag.16

1. INTRODUCTION

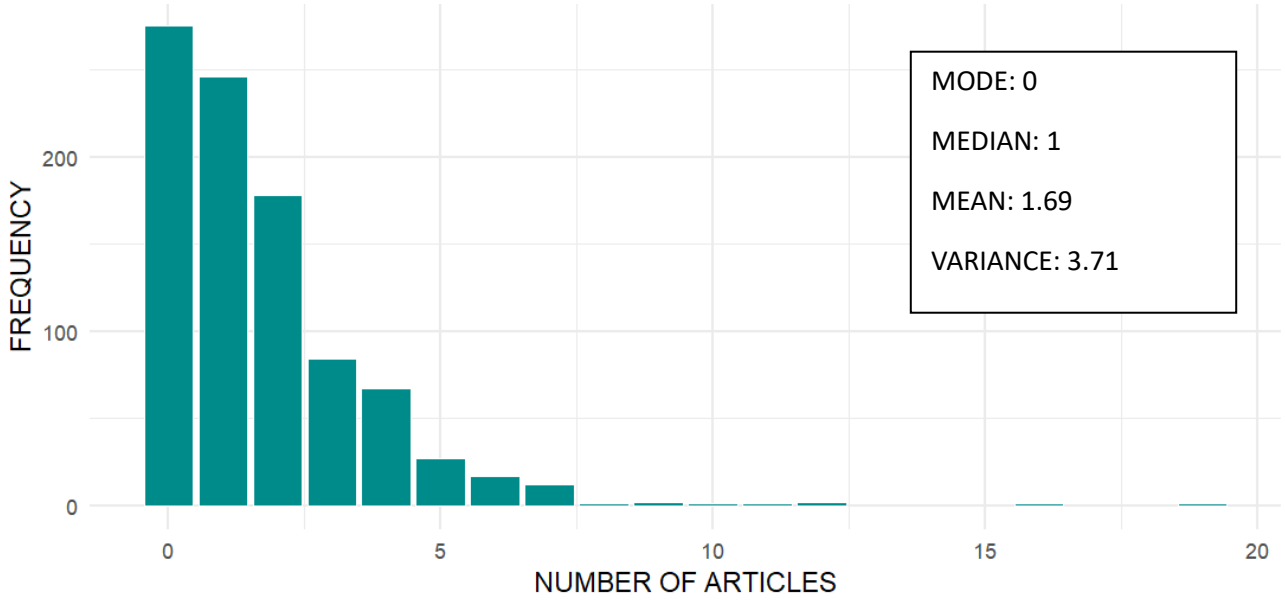
The number of articles that a Ph.D. student write and publish during its research period may be influenced by various factors. As an example, it is reasonable to assume that married students with young kids may be less active thus publishing less articles.
However, It worths noting that most Ph.D. students may have published their first article according to the degree thesis, so, most of them probably started the program having already published an article. In order to assess whether some identified factors influence the number of produced articles, some data were collected with respect to the following features:

- Number of articles published by the student during the last 3 years of Ph.D. ;
- Number of articles published by the student mentor during the last 3 years of Ph.D.;
- Students Sex;
- Students marital status (Married or not);
- Number of children aged 5 years old or younger;
- Prestige of the Ph.D. department.

2. DESCRIPTIVE ANALYSIS

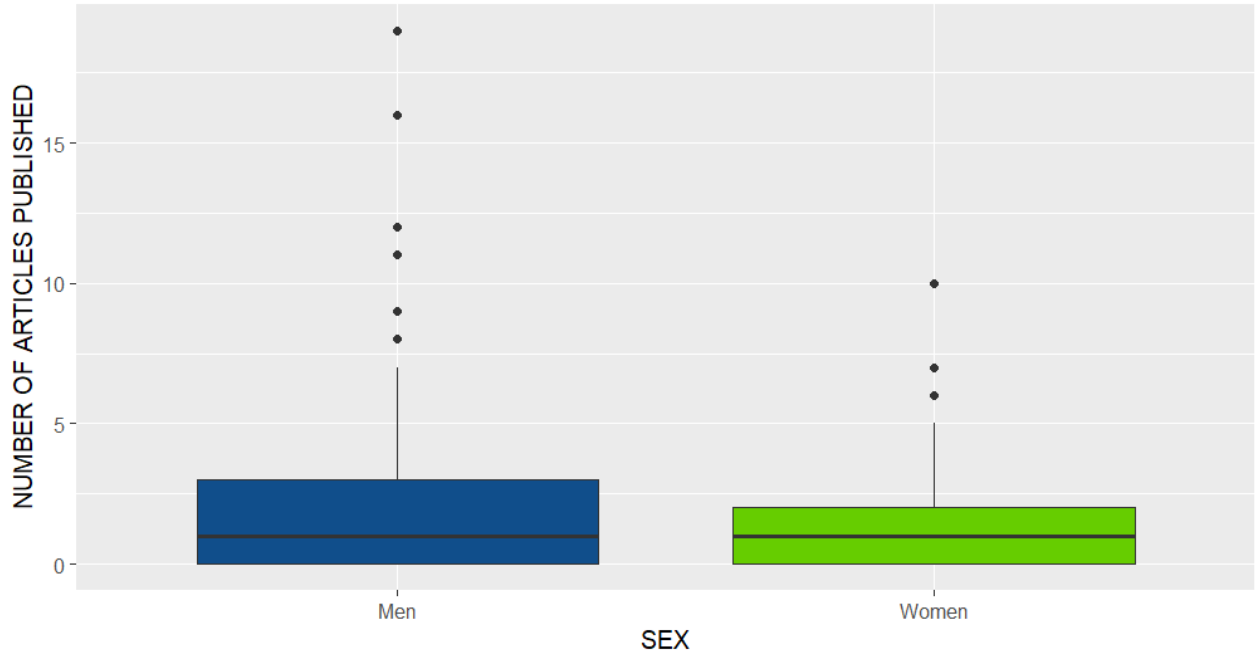
First of all, an overall description was carried out. The marginal and conditional distribution of the number of articles published by the students during the last three of the Ph.D. program, given the other features were graphically represented.

PLOT 2.1: DISTRIBUTION OF THE NUMBER OF ARTICLES PUBLISHED BY THE PH.D. STUDENT

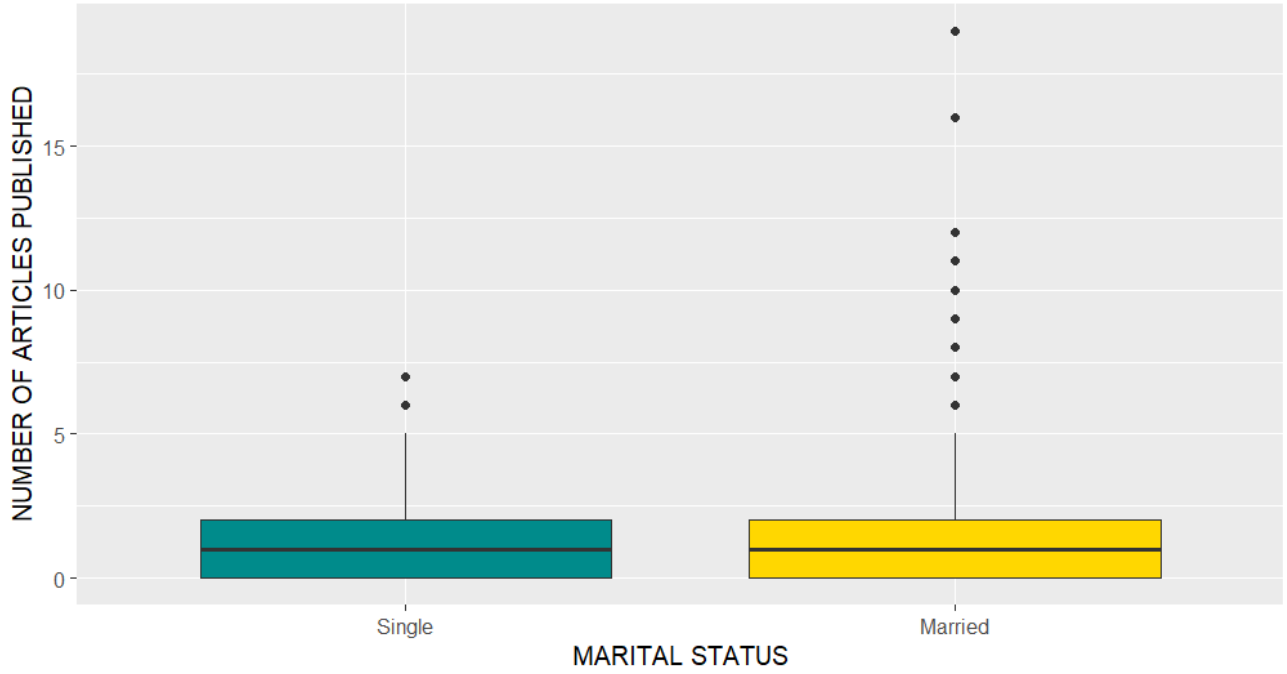


PLOT 2.1 shows on the x-axis the quantity of articles published in the last three years while on the y-axis there is the number of students who published that certain number of articles. It is possible to notice that most students published just few articles: the modal amount is zero. A very limited number of students was capable of publishing more than 7 articles during the last three years.

PLOT 2.2: CONDITIONAL NUMBER OF ARTICLES DISTRIBUTION GIVEN THE STUDENTS SEX

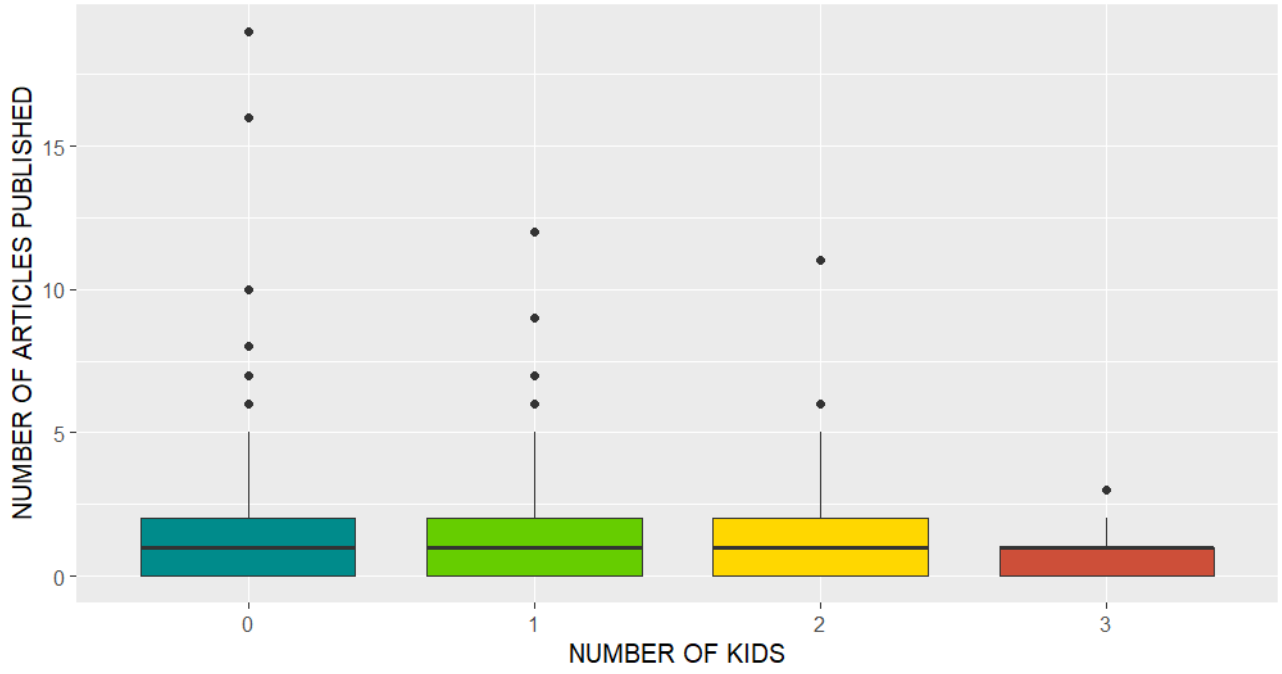


PLOT 2.3: CONDITIONAL NUMBER OF ARTICLES DISTRIBUTION GIVEN THE MARITAL STATUS



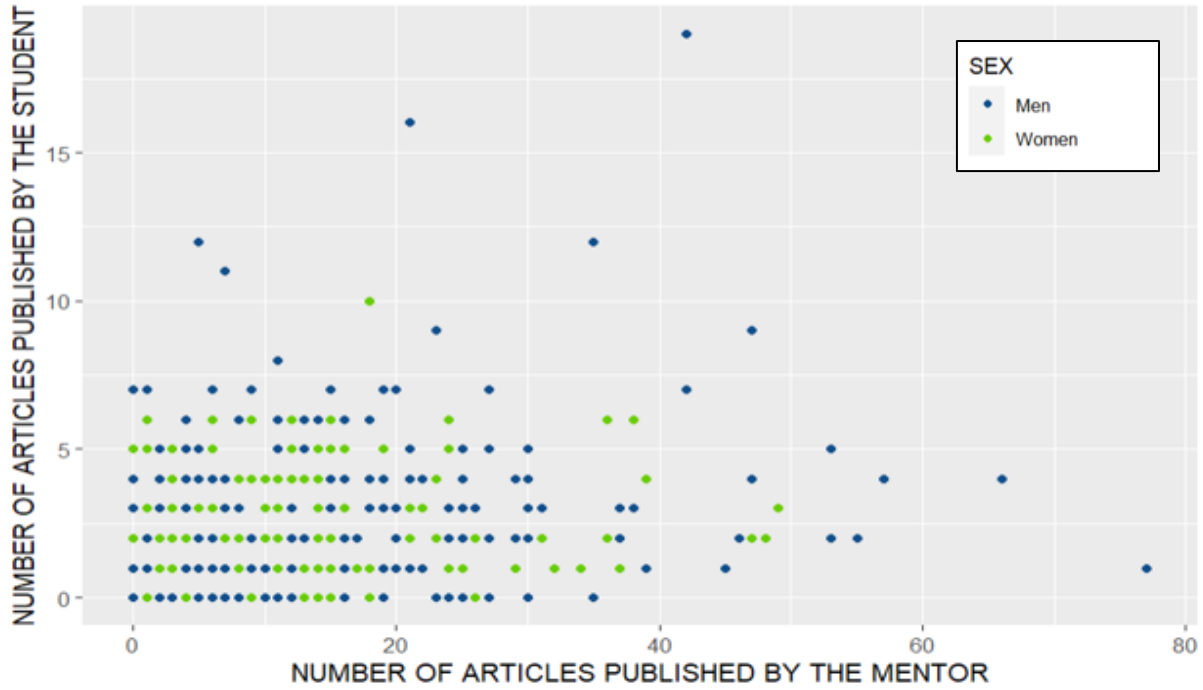
PLOTS 2.2 and 2.3 show how the distribution of the number of published articles changes according to the student sex and marital status. It is possible to notice that, regardless of these characteristics, half of the considered students published a most one paper. However, the variability changes: on the one hand, only male students published more than ten articles; on the other hand, while single students published at most seven articles, married students seem to publish more (at most 19 articles). The main reason may be that married Ph. D. students are also older, thus probably having more academic experience.

PLOT 2.4: CONDITIONAL NUMBER OF ARTICLES DISTRIBUTION GIVEN THE NUMBER OF KIDS

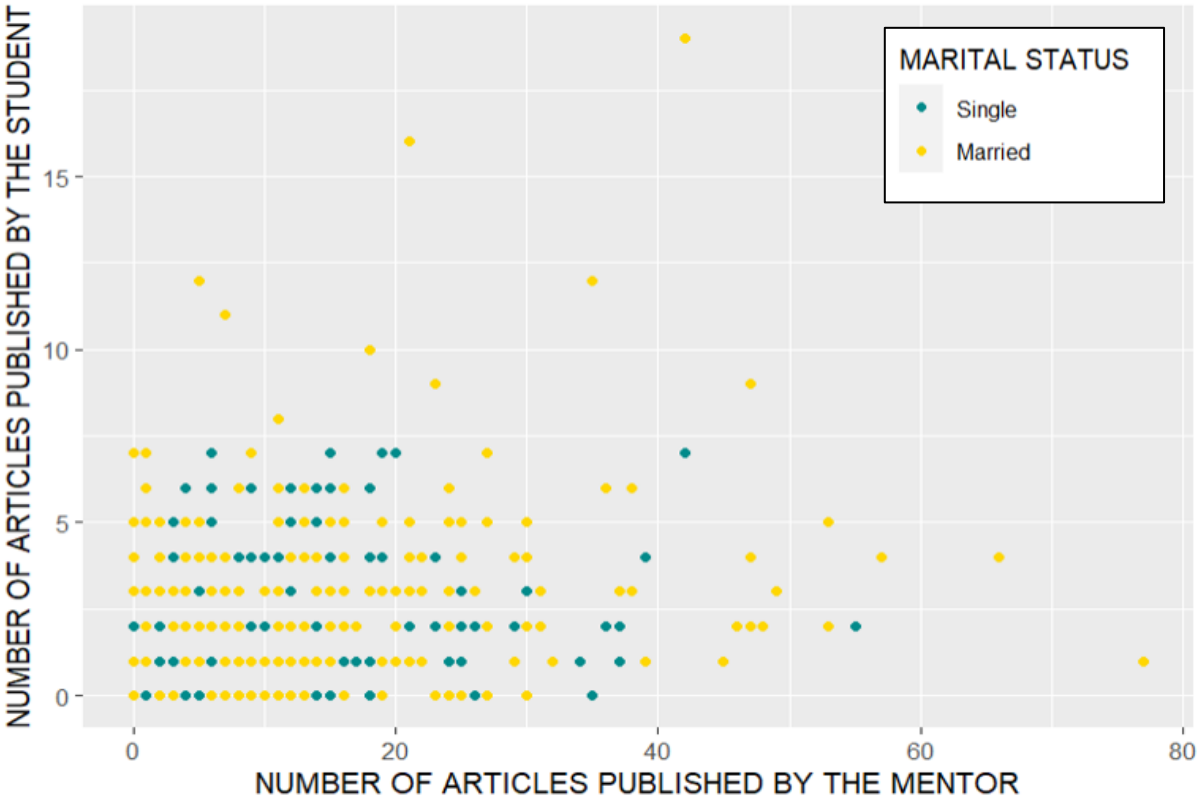


In PLOT 2.4 it is possible to notice that the variability of number of articles published decreases as the number of kids increases. In fact, students with 3 kids have published a maximum of 3 articles; instead, students without kids have published at most 19 articles. However, regardless of number of kids, the median value of number of articles is 1.

PLOT 2.5: NUMBER OF ARTICLES PUBLISHED BY THE STUDENT AND ITS MENTOR GIVEN THE SEX

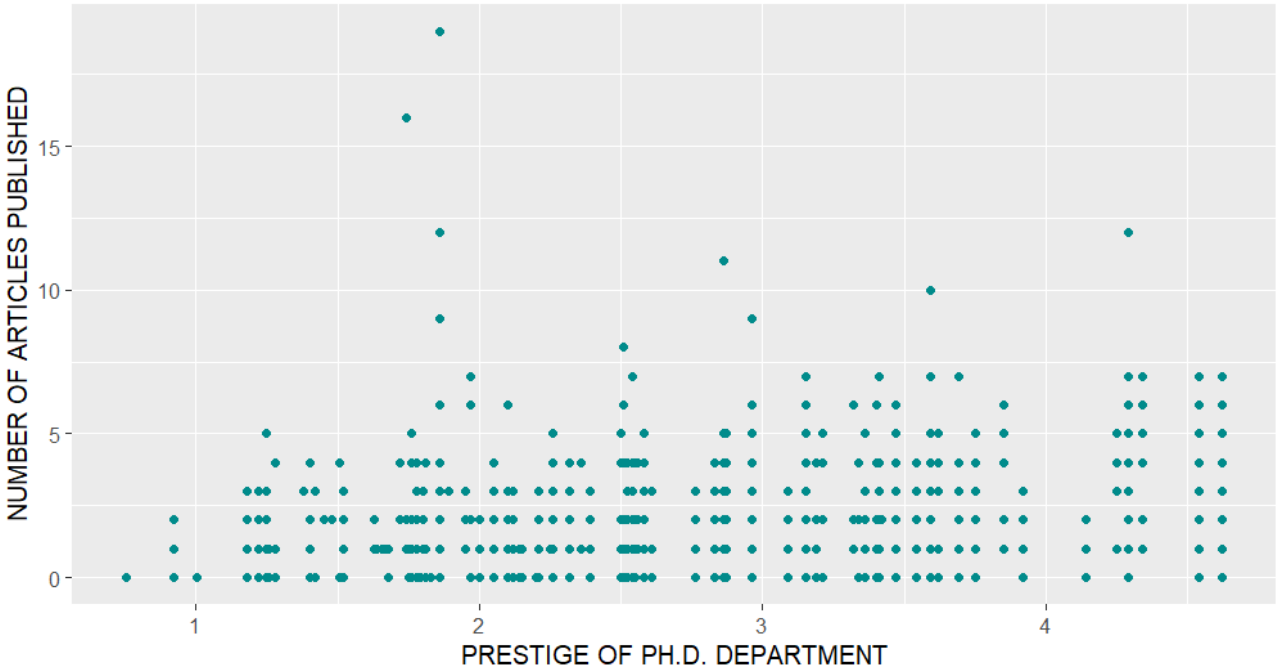


PLOT 2.6: NUMBER OF ARTICLES PUBLISHED BY THE STUDENT AND ITS MENTOR GIVEN THE MARITAL STATUS



PLOTS 2.5 and 2.6 highlight how the number of articles published by the Ph.D. student varies according to the amount published by his/her mentor, given the student sex and marital status. It is possible to notice that as the number of articles published by mentor increases, the number of articles published by the student increases too (the Pearson linear correlation coefficient is about 0.31). Moreover, it is possible to notice that some students seem to be not influenced by the mentor activity: some married men publish a great number of articles even when the mentor publish just few articles and vice versa.

PLOT 2.7: NUMBER OF ARTICLES PUBLISHED AND PRESTIGE OF PH.D. DEPARTMENT



The scatterplot 2.7 shows that the Ph.D. department prestige seems to be linearly independent with respect to the number of articles. In fact, the correlation between them is approximately zero (0.07).

3. POISSON GENERALIZED LINEAR MODEL

Poisson generalized linear models are a class of models able to predict how the expected number of articles published by a Ph.D. student having certain features, i.e. the expectation of the related conditional Poisson distribution, changes according to the considered features.

3.1 DISTRIBUTIONAL ASSUMPTION VALIDITY

In order to adapt a Poisson GLM, it is necessary to assess whether the observed published articles counts can be described by a Poisson distribution. A simple graphical method, called Poissonness Plot, can be used for checking if the observed published articles counts are Poisson distributed. It displays how the observed counts differ from a straight line whose equation is obtained by equating the observed with the expected frequencies.
If the number of published articles follows a Poisson distribution and the constant rate assumption is suitable then data will follow the following line:

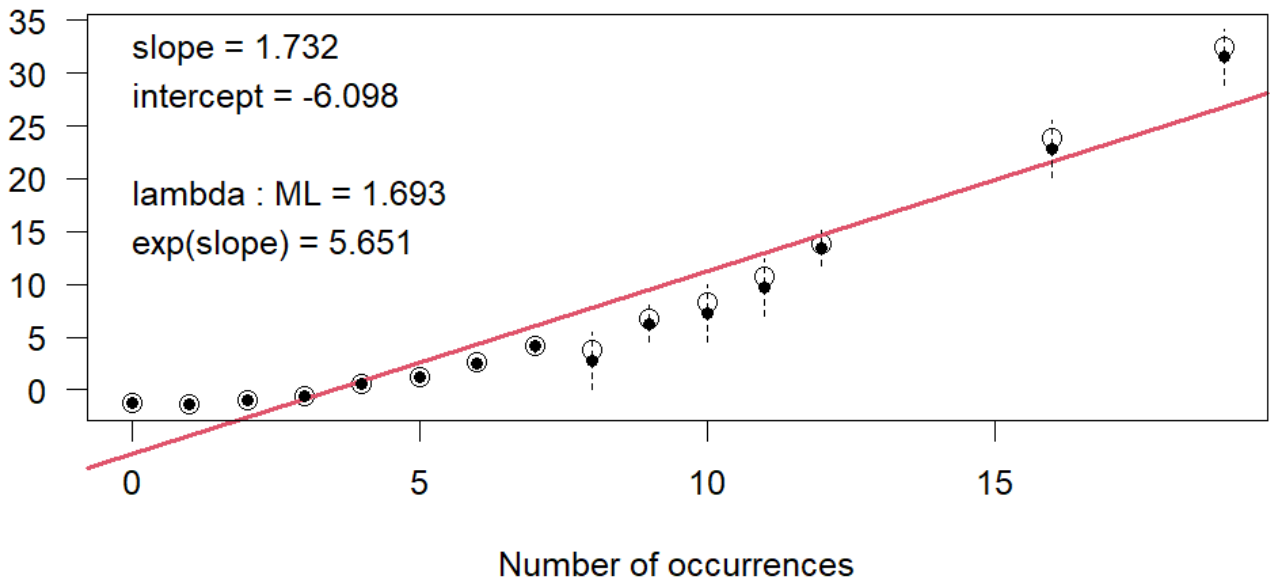
$\mathbf{E[Y] = \lambda, \quad Var[Y] = \lambda}$ ← FIRST TWO DISTRIBUTION MOMENTS

$P_\lambda\{Y = k\} = e^{-k} \frac{\lambda^k}{k!}$ ← POISSON PROBABILITY DISTRIBUTION

$N = n_{k=1} + n_{k=2} \ldots$ ← NUMBER OF OBSERVATIONS

$\mathbf{log}(\frac{y_k k!}{N}) = -\boldsymbol{\lambda} + \mathbf{log}\,(\boldsymbol{\lambda})\boldsymbol{k}$ ← POISSONESS PLOT EQUATION

PLOT 3.1.1: POISSONESS PLOT



The Poisonness plot shows a slightly nonlinear pattern. The averaged lambda value (Maximum Likelihood estimate) seems to differ from the one corresponding to the line slope. A Poisson distribution may be not the most appropriate distribution. Moreover, PLOT 2.1 showed that mean and variance are not equal. For reasons of simplicity and model interpretability and since the situation is unclear, a Poisson model can be adapted and then eventually improved.

3.2 MODEL ADAPTATION

In order to adapt the Poisson Generalized Linear Model it was necessary to choose a link function to connect the discrete response variable with a linear predictor, ranging in the real numbers set. The choice fell on the canonical link, that is the logarithmic one. The resulting model is highlighted below (formula 3.1).

$$log(E[Yi|Xi]) = log(\mu_i) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} Sex_W + \boldsymbol{\beta_2} Mar_M + \boldsymbol{\beta_3} Kid5 + \boldsymbol{\beta_4} PhD + \boldsymbol{\beta_5} Ment + \boldsymbol{\beta_6} Sex_W PhD \quad (3.1)$$

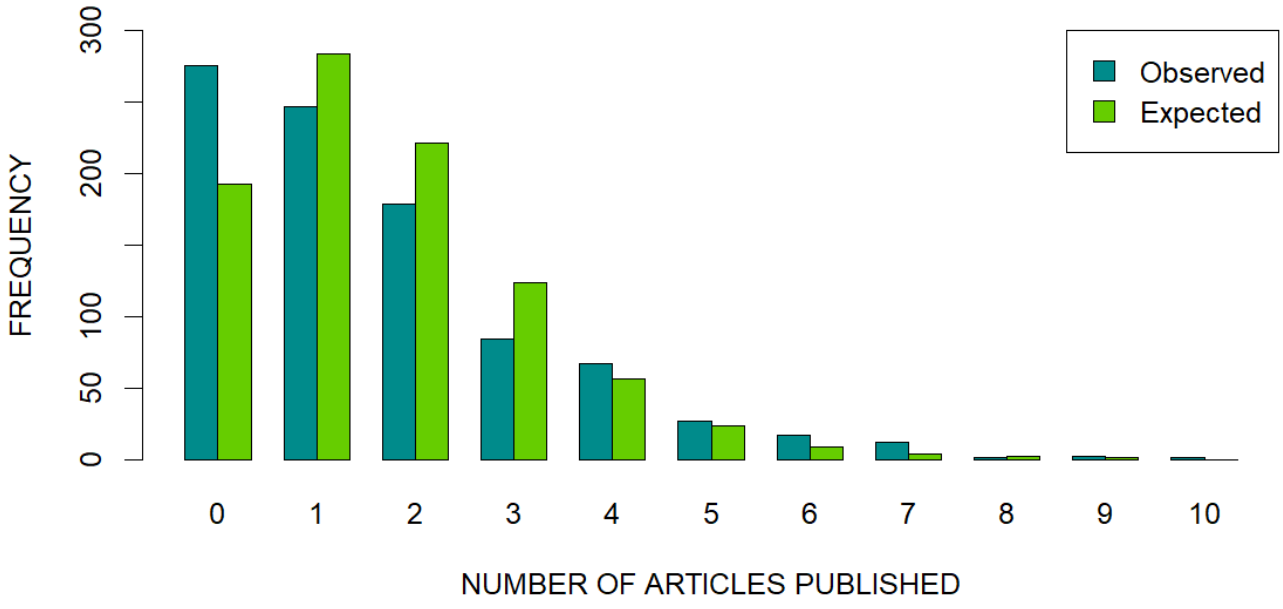
$$\mu_i = \exp^{(\mathbf{0.47} - \mathbf{0.63}\, Sex_W + \mathbf{0.14}\, Mar_M - \mathbf{0.19}\, Kid5 - \mathbf{0.03}\, PhD + \mathbf{0.02}\, Ment + \mathbf{0.12}\, Sex_W PhD)} \quad (3.2)$$

The chosen optimal Poisson Generalized Linear model was selected on the basis of the Akaike Information Criterion, a metric that quantifies the divergence between the model and the ground truth function, taking into account the complexity of the evaluated model.
For each individual, the probability of publishing a certain number of articles is obtained according to the formulae 3.1 and 3.2. The expected number of publications is lower for the female students and for who has a bigger number of kids (<5 years old).
In contrast, the number of articles published by the student's mentor and being married have a positive impact on the expected publication activity.

3.3 MODEL GOODNESS OF FIT

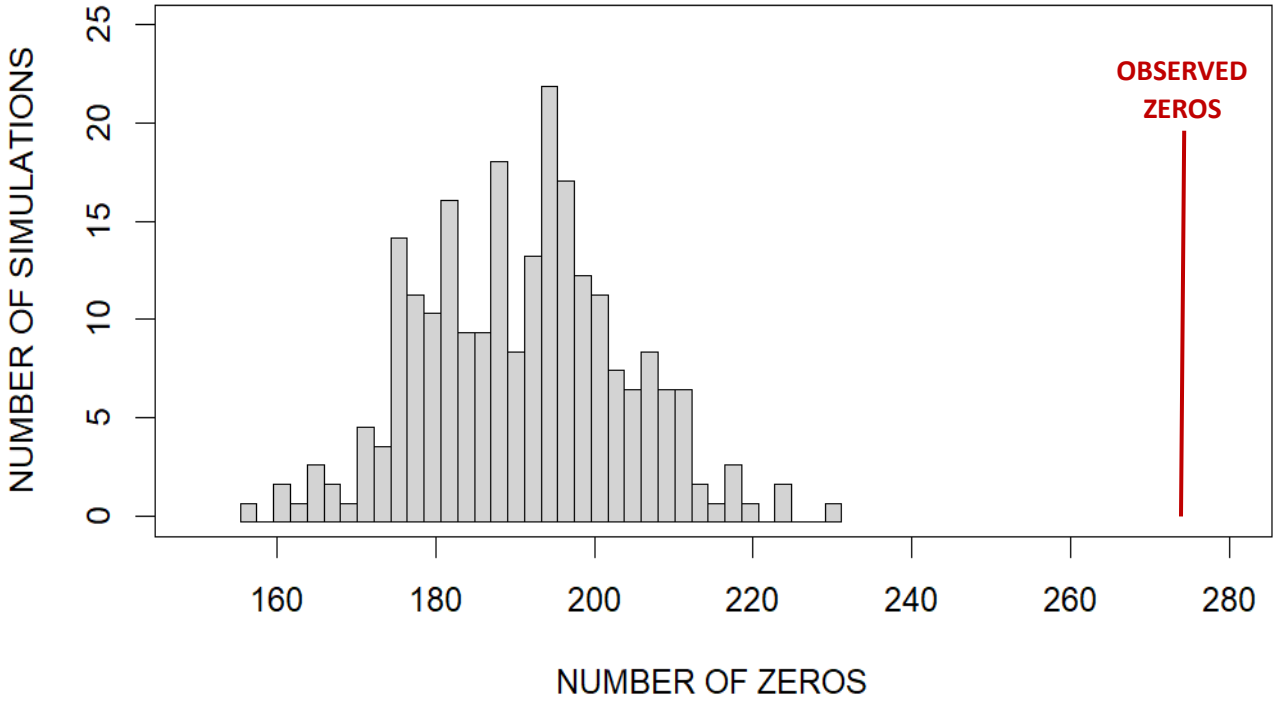
In order to assess the model goodness of fit, the first ten observed counts values were plotted and compared with the ones predicted by the chosen model (PLOT 3.3.1). It is possible to notice that the model was not able to catch the relevant presence of null and unitary response values previously highlighted by the exploratory analysis (PLOT 2.1).
In fact, the number of zeros was highly underpredicted, whereas the number of ones was overpredicted.

PLOT 3.3.1: COMPARISON BETWEEN OBSERVED AND PREDICTED NUMBER OF PUBLISHED ARTICLES (POISSON GLM)



The fact that the observed number of zeros is larger than the expected one under the standard Poisson GLM, in addition to the fact that the modal value is equal to zero and differs from the integer part of the mean value (1.69) suggest that the data may present a zero-inflation problem to take into account while modelling.
In order to test for the presence of the aforementioned problem 250 simulations were conducted, starting from the conditional expectations predicted by the selected Poisson Generalized Linear Model. The number of zeros in each simulation was computed and the corresponding distribution was graphically visualized and compared with the observed number of zeros (PLOT 3.3.2).
A hypothesis testing was then conducted: the ratio between the observed number of zeros and the averaged simulated one was equal to 1.43 and turned out to be statistically different from 1.

PLOT 3.3.2: COMPARISON BETWEEN OBSERVED AND SIMULATED NUMBER OF ZEROS (POISSON GLM)



The presence of a zero-inflation problem often leads to another problem, that requires ad hoc models: overdispersion. The overdispersion is a phenomenon that occurs when the count observations exhibit variability exceeding that predicted by the Poisson: this distribution assumes that the mean equal the variance, while in the current situation the mean is equal to 1.69 and the variance equals 3.71. In order to assess the presence of overdispersion the dispersion statistic was computed, starting from the Pearson residuals. This statistic turned out to be statistically greater than one, reporting a warning related to the presence of an overdispersion problem. This kind of problems can be easily overcome by the mean of extensions of the Poisson Generalized Linear Model.

4. NEGATIVE BINOMIAL GENERALIZED LINEAR MODEL

An extension of the Poisson GLM that has an extra parameter to account better for overdispersion is the negative binomial generalized linear model, in which the conditional expected value μ_i is linked to the linear predictor in the same way of the Poisson GLM. In the two-parameter negative binomial family, in contrast to the Poisson distribution, the variance is no longer mandatory equal to the expectation, but is a function of it and of a dispersion parameter φ, estimated to handle overdispersion.

4.1 DISTRIBUTIONAL ASSUMPTION VALIDITY

In order to adapt a Negative Binomial GLM, it is necessary to establish if it is reasonable to assume that the count of observed published articles come from a Negative Binomial distribution. In analogy to the Poisson distributional assumption validity (section 3.1) a plot like the Poissonness Plot can be displayed: the negative binomialness plot.

$\mathbf{E[Y] = \mu,\ \ Var[Y] = \mu + \frac{1}{\alpha}\mu^2}$ ← FIRST TWO DISTRIBUTION MOMENTS

$\frac{1}{\alpha} = \varphi$ ← DISPERSION PARAMETER

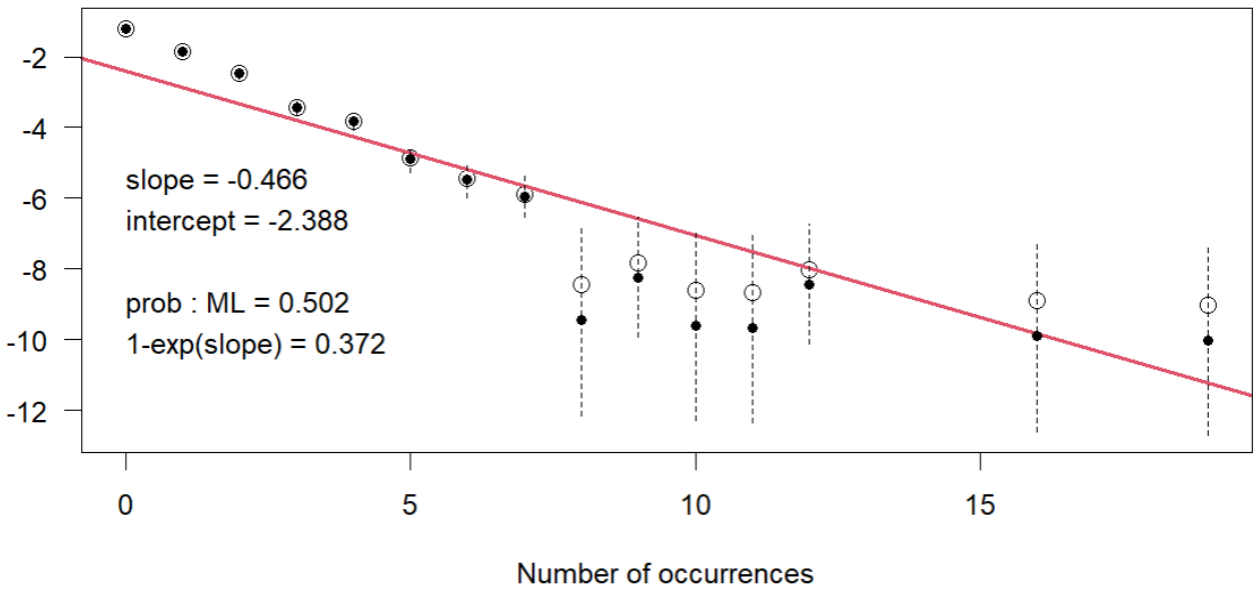
$\Pr(Y = k) = \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)\Gamma(k+1)}\left(\frac{\mu}{\mu+\alpha}\right)^k\left(\frac{\alpha}{\mu+\alpha}\right)^\alpha$ ← PROBABILITY DISTRIBUTION

$p = \frac{\alpha}{\mu+\alpha}$ ← ALTERNATIVE REPARAMETRIZATION

$N = n_{k=1} + n_{k=2} \ldots$ ← NUMBER OF OBSERVATIONS

$$\boldsymbol{log}\left(\frac{\boldsymbol{n_k}}{\boldsymbol{N}\binom{\boldsymbol{n_k+k-1}}{\boldsymbol{k}}}\right) = \boldsymbol{n_k}\mathbf{log(p)} + \boldsymbol{log}\,(\mathbf{1}-\boldsymbol{p})\boldsymbol{k}$$ ← NEGATIVE BINOMIALNESS PLOT (4.1.1)

PLOT 4.1.1: NEGATIVE BINOMIALNESS PLOT



PLOT 4.1.1 shows that the negative binomial distributional assumption appears to be adequate: the maximum likelihood estimate of the probability parameter is similar to the one obtained by adapting the equation (4.1.1).

4.2 MODEL ADAPTATION

The chosen link function, relating the expected number of published articles with the linear predictor, is the canonical one, i.e. the logarithmic function. The resulting model is:

$$log(\mu_i) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} Sex_W + \boldsymbol{\beta_2} Mar_M + \boldsymbol{\beta_3} Kid5 + \boldsymbol{\beta_4} PhD + \boldsymbol{\beta_5} Ment + \boldsymbol{\beta_6} Kid5\ PhD \qquad (4.2.1)$$

$$\mu_i = \exp^{(\mathbf{0.08} - \mathbf{0.21}\ Sex_W + \mathbf{0.16}\ Mar_M + \mathbf{0.16}\ Kid5 + \mathbf{0.06}\ PhD + \mathbf{0.03}\ Ment - \mathbf{0.11}\ Kid5\ PhD)} \qquad (4.2.2)$$

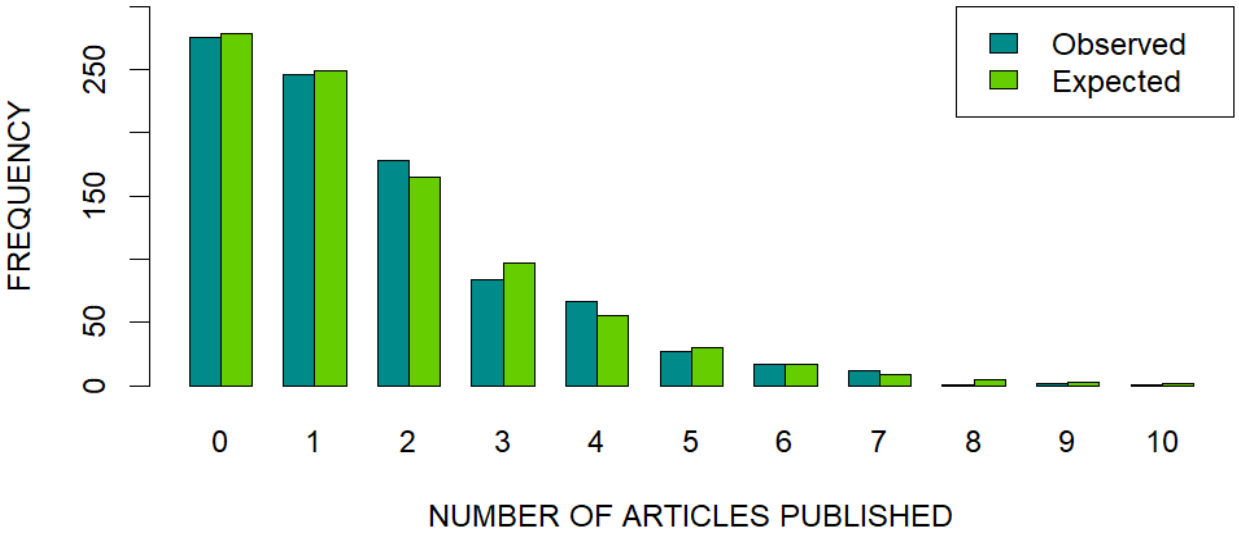
$$\varphi = 2.29 \qquad (4.2.3)$$

By looking at the coefficients signs it is possible to notice that, in analogy with the Poisson Generalized Linear model, the expected number of articles published decreases for women and increases for married students and as the number of articles published by the student mentor increases too. Furthermore, the interaction between the number of kids and the department's prestige score suggests that the expected number of published articles decreases when there is a simultaneous increase in the number of kids and prestige score.
The overdispersion parameter φ is a positive quantity referred to the overdispersion phenomenon. The greater the value of φ, the greater the overdispersion relative to the Poisson. As it tends to zero, the variance tend to the mean value and the negative binomial distribution converges to the Poisson. 4.2.3 result highlights again the presence of overdispersion.

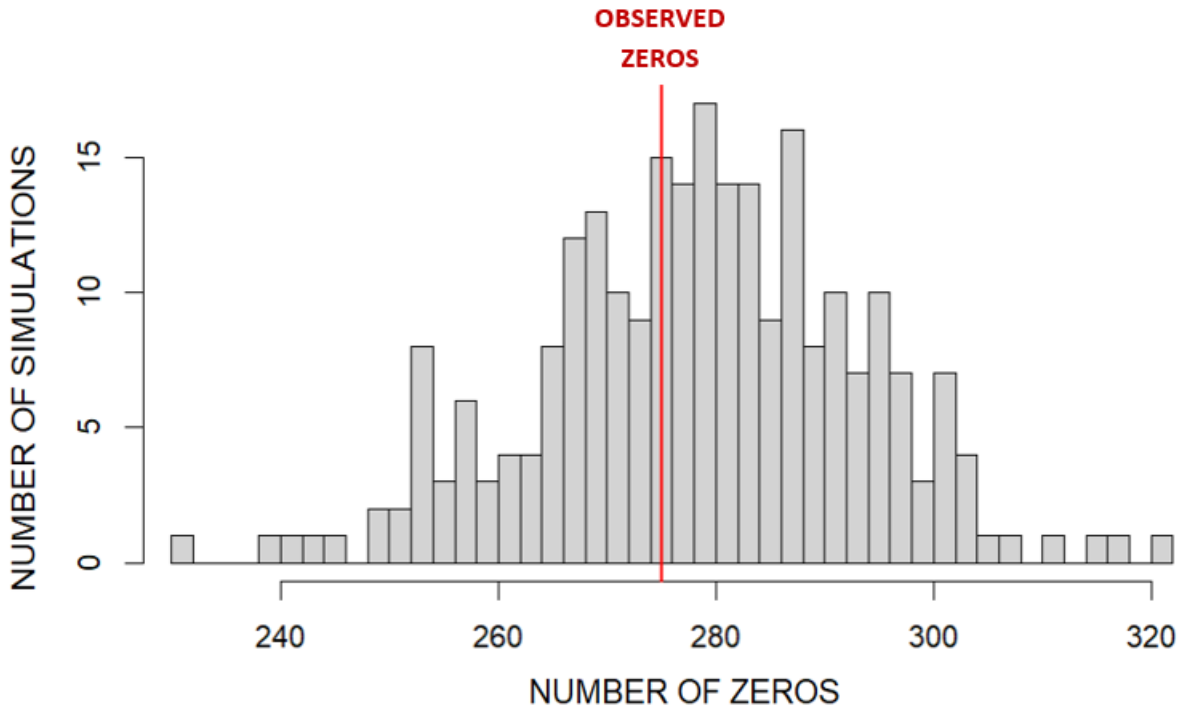
4.3 MODEL GOODNESS OF FIT

As in the previous model, in order to assess the model goodness of fit, the first ten observed counts values were plotted and compared with the ones predicted by the chosen model. In this case the model can catch far better the observed zeros and ones, since it just slightly overestimates both of them.

PLOT 4.3.1: COMPARISON BETWEEN OBSERVED AND PREDICTED NUMBER OF PUBLISHED ARTICLES (NEGATIVE BINOMIAL GLM)



PLOT 4.3.2: COMPARISON BETWEEN OBSERVED AND SIMULATED NUMBER OF ZEROS (NEGATIVE BINOMIAL GLM)



From PLOT 4.3.2 it is possible to notice that the observed number of zeros, equal to 275, is in line with the simulated counts, like in the Poisson case (PLOT 3.3.2), starting from the predicted conditional expectation values.

Thus, the ratio between the observed zeros counts and the average of the simulated counts is not statistically different from one (p-value 0.78), meaning that the negative binomial GLM was able to handle the zero-inflation problem. Zero-inflation is less problematic for the negative binomial distribution with respect to the Poisson distribution.
In fact, while in the Poisson distribution the mode should be the integer part of the mean (in this case it should be 1 but it is equal to 0), the negative binomial distribution can have a mode of 0 regardless of the value of the mean. It means that, since both GLMs are aimed to model the conditional expectation, according to a Poisson distribution a mean of 1.69 should imply a mode of 1 whereas the true value is 0. This is also the reason because the Poisson GLM previously implemented overestimated the number of one values.

As in the previous case, to assess the presence of overdispersion the dispersion statistic was computed. This statistic turned out to be not statistically different from one (p-value 0.42), reporting that the negative was able to catch the overdispersion phenomenon.

5. ZERO-INFLATED NEGATIVE BINOMIAL MODEL

An alternative way to model the number of published articles, by taking into account presence of zeros in count data are the zero-inflated models:
mixture models that allow the presence of two different subpopulations, made up of "immune" and "susceptible" individuals.
The key assumption underlying such models is that data come from two separated distributions, one for each subpopulation. As an example, by considering the number of children for a sample of women, there will be two subpopulations:
The first "immune" population is given by infertile women that will always have a zero-response value (structural zeros), while the second "susceptible" subpopulation is given by fertile women that can either have children (non-zero response value) or not (zero response value).
The resulting mixture model is a two-part model, made up of:

- a binomial distribution to model the probability of belonging to each subpopulation;
- a count distribution, either Poisson or Negative Binomial, to model the counts.

Since the current data demonstrate not only a zero-inflation problem, but also an overdispersion problem, the previously adapted Generalized Linear Negative Binomial Model was compared with its zero-inflated version (ZINB model).

The probability of belonging to each subpopulation, i.e. the probability of have a structural zero or not, can be expressed as follows:

π ← PROBABILITY TO COME FROM THE "IMMUNE" SUBPOPULATION

$1 - \pi$ ← PROBABILITY TO COME FROM THE "SUSCEPTIBLE" SUBPOPULATION

$$\Pr(Y = 0) = \pi + (1 - \pi) * \left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha} \quad \leftarrow \text{Pr OF OBSERVING A ZERO VALUE} \quad (5.1)$$

$$\Pr(Y = k) = (1 - \pi)\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)\Gamma(k+1)}\left(\frac{\mu}{\mu+\alpha}\right)^{k}\left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha} \quad \leftarrow \text{Pr OF OBSERVING A NON-ZERO VALUE} \quad (5.2)$$

$$\text{E[Y]} = \theta = (1 - \pi)\,\mu \quad \leftarrow \text{ZINB EXPECTATION}$$

It is possible to notice that the probability of observing a zero value (5.1) is made up of a part related to the probability π of observing structural zero, and a part related to the zeros coming from a negative binomial distribution. 5.2 formula refers to the Power Mass Function related to the negative binomial counts.

$$logit\big(\pi(z)\big) = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_r z_r \quad \leftarrow \text{MODELLING THE Pr OF STRUCTURAL 0 VALUES}$$

$$log\big(\mu(x)\big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \quad \leftarrow \text{MODELLING COUNTS (STRUCTURAL 0 EXCLUDED)}$$

5.1 MODEL ADAPTATION

As highlighted in the previous section, the Zero-inflated negative binomial model is a two-part model, so the logit of the probability of observing a structural zero and the expected counts in each conditional distribution need to be simultaneously modelled. Thus, ZINB doesn't link directly the conditional expected value to the predictors, but two link functions are needed, one for the binomial part (logit link) and one for the negative binomial part (logarithmic link). The resulting overall optimal model, chosen on the basis of the AIC metric, is made up of the following models:

$$log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{\gamma_0} + \boldsymbol{\gamma_1} Sex_W + \boldsymbol{\gamma_2} Kid5 + \boldsymbol{\gamma_3} PhD + \boldsymbol{\gamma_4} Ment + \boldsymbol{\gamma} Kid5\ PhD \quad (5.1.1)$$

$$\pi = \frac{\exp^{(\mathbf{1.04-2.07}\ Sex_W + \mathbf{0.85}\ Kid5 - \mathbf{0.31} PhD - \mathbf{1.59}\ Ment + \mathbf{0.29}\ Kid5\ PhD)}}{1+\exp^{(\mathbf{1.04-2.07}\ Sex_W + \mathbf{0.85}\ Kid5 - \mathbf{0.31} PhD - \mathbf{1.59}\ Ment + \mathbf{0.29}\ Kid5\ PhD)}} \quad (5.1.2)$$

$$log(\mu_i) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} Sex_W + \boldsymbol{\beta_2} Kid5 + \boldsymbol{\beta_3} PhD + \boldsymbol{\beta_4} Ment + \boldsymbol{\beta_5} Kid5\ PhD \quad (5.1.3)$$

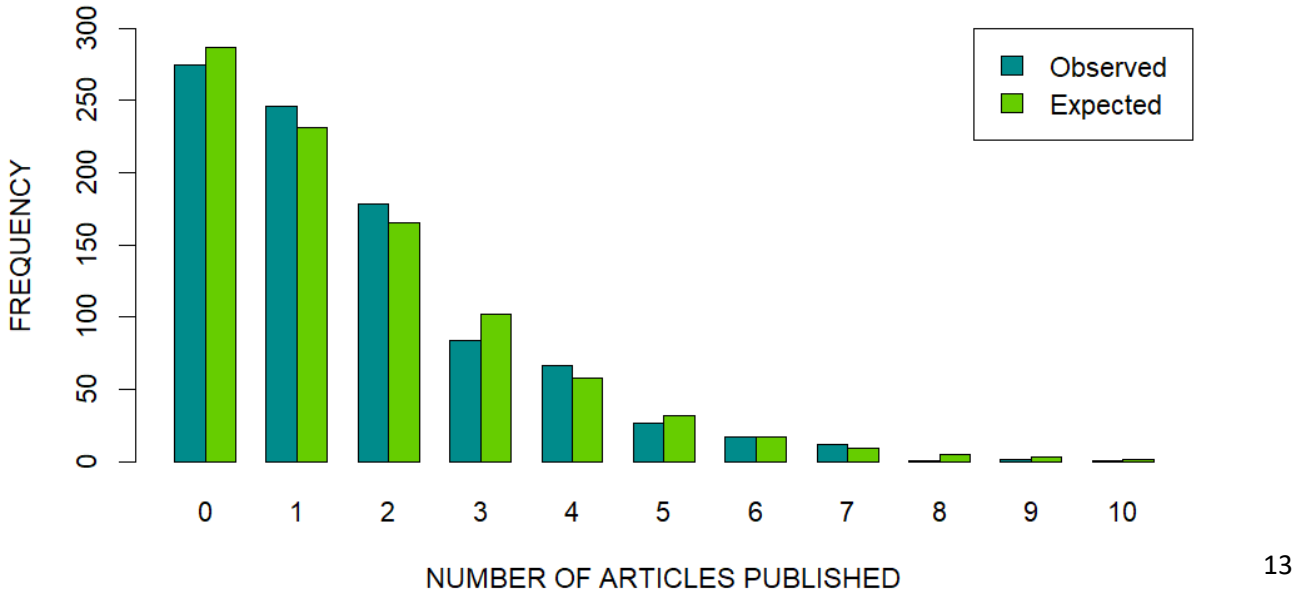
$$\mu_i = \exp^{(\mathbf{0.31-0.22}\ Sex_W + \mathbf{0.24}\ Kid5 + \mathbf{0.06}\ PhD + \mathbf{0.02}\ Ment - \mathbf{0.12}\ Kid5\ PhD)} \quad (5.1.4)$$

Formula 5.1.1 shows that the probability of observing structural zeros increases as the number of children under age 5 increases, while decreases for married students and as the prestige of the university or the number of mentor publications increases.

Formula 5.1.4 highlight that, in analogy with the previously implemented models, the expected number of published articles, for those who are actually amenable to publication i.e. the "susceptible" subpopulation, decreases if the student is a woman, increases as university prestige or mentor's publications increases. Since the interaction between the number of kids and the department prestige has a negative sign, as these two features simultaneously increase the expected number of published articles decreases.

5.2 MODEL GOODNESS OF FIT

PLOT 5.2.1: COMPARISON BETWEEN OBSERVED AND SIMULATED NUMBER OF ZEROS (ZINB MODEL)



PLOT 5.2.1 shows that the model slightly overestimated the number of zeros while underestimated the number of unitary values. With respect to these two count values the negative binomial GLM appear to work better. In order to assess whether a more complex mixture model has brought a significant fitting improvement an appropriate test was conducted.
The Vuong test is a statistical test that can be used to test whether two non-nested models are equivalent. The Negative Binomial Generalized Linear Model and the Zero-Inflated Negative Binomial model turned out to be indistinguishable (p-value 0.12), thus the less complex version may be preferred.

6. HURDLE NEGATIVE BINOMIAL MODEL

An alternative approach to modelling zero-inflation uses a two-part model called a hurdle model. Hurdle models are a class of models that treat the zero-inflation problem in a slightly different way with respect to zero-inflated models. All the zero values are treated together by the mean of a binary model, e.g. a logistic model, while the non-zero values are treated by using a truncated count distribution, e.g. a truncated Poisson or a truncated Negati ve Binomial distribution. In contrast to the zero-inflated mixture models, Hurdle models do not distinguish between structural and non-structural zeros and are capable of handling not only zero-inflation problems but also zero-deflation problems, thus being more flexible.
According to the previous part of the analysis the adopted distribution is the negative binomial one. The P.M.F. of a Negative Binomial Hurdle is:

$\Pr(Y = 0) = (1 - \pi)$ ← Pr OF OBSERVING A ZERO VALUE

$\Pr(Y = k) = (\pi)\frac{Negative\ Binomial(k)}{1 - Negative\ Binomial(0)}$ ← Pr OF OBSERVING A NON-ZERO VALUE

$E[Y] = \theta = \frac{1 - f_1(0)}{1 - f_2(0)}\mu_2$ ← NB HURDLE EXPECTATION

The expectation of the negative binomial hurdle is obtained by multiplying the probability to obtain a number of published articles greater than zero and the expectation of the truncated negative binomial.

$logit(\pi(z)) = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_r z_r$ ← MODELLING THE Pr OF OBSERVING A ZERO

$log(\mu(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$ ← MODELLING NON-ZERO VALUES

6.1 MODEL ADAPTATION

In order to adapt the Hurdle Negative Binomial model, the AIC measure was adopted, leading to the following results:

$$log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{\gamma_0} + \boldsymbol{\gamma_1} Mar_M + \boldsymbol{\gamma_2} Kid5 + \boldsymbol{\gamma_3} Ment \quad (6.1.1)$$

$$\pi = \frac{\exp(\mathbf{0.13} + 0.36 Sex_W - 0.25\ Kid5 + 0.08\ Ment)}{1 + \exp(\mathbf{0.13} + 0.36 Sex_W - 0.25\ Kid5 + 0.08\ Ment)} \quad (6.1.2)$$

$$log(\mu_i) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} Sex_W + \boldsymbol{\beta_2} Kid5 + \boldsymbol{\beta_4} Ment \quad (6.1.3)$$

$$\mu_i = \exp^{(\mathbf{0.41} - \mathbf{0.25}\, Sex_W - 0.12\, Kid5 + \mathbf{0.02}\, Ment)} \quad (6.1.4)$$

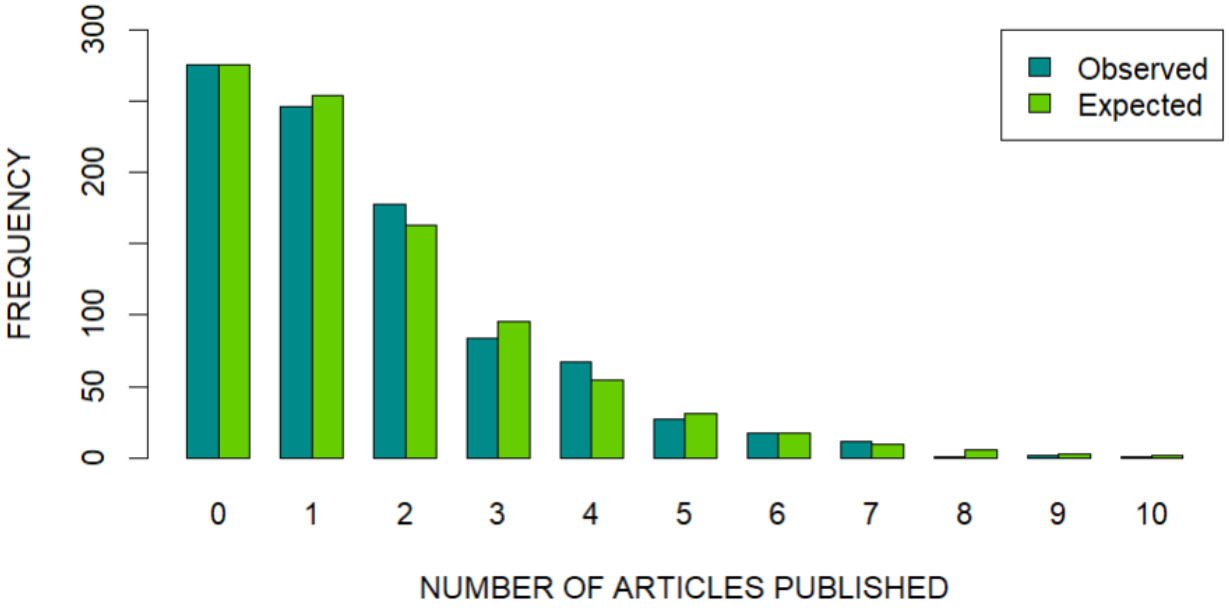
The first set of coefficients, related to the linear predictor 6.1.1 that models the logit of the probability of publishing at least an article, has a similar interpretation but an opposite direction with respect to the counter part in the ZINB model, since, in this case, the parametrization of the logit refers to the probability of not observing a zero value.
The probability of publishing at least an article during the last three years of the Ph.D. program increases for who is married and as the number of articles published by the mentor increases, whereas it decreases as the number of kids increases.

The second set of coefficients refers to the Truncated Negative Binomial part of the model (6.1.3 and 6.1.4), that is the linear predictor aimed to model the non-zero number of published articles. The considerations are coherent with those made on the basis of the other models previously seen: the expected number of published articles increases as the number of articles published by the mentor increases, while it decreases for women and as the number of kids under five years increases.

6.2 MODEL GOODNESS OF FIT

As expected and shown in PLOT 6.2.1 the Hurdle model was able to predict the exact observed number of zeros (it's a characteristic of such models).

PLOT 6.2.1: COMPARISON BETWEEN OBSERVED AND SIMULATED NUMBER OF ZEROS (HURDLE NB)



In order to assess which model best fits data, taking into account also the corresponding model complexity, the Vuong hypothesis testing was carried out.

The Negative Binomial Generalized Linear Model works equivalently with respect to the far more complex Hurdle version (p-value 0.28). The Vuong test also led to think the two implemented modelling alternatives, specific for zero-inflation problems, i.e. the Zero-inflation Negative Binomial and the Hurdle Negative Binomial, appears to be indistinguishable (p-value 0.26).

7. CONCLUSIONS

The carried-out analysis showed that the Zero-Inflated Negative Binomial and the Hurdle Negative Binomial models are not superior with respect to the standard and less complex Negative Binomial Generalized Linear Model.
In contrast to the Poisson GLM, including an additional parameter turns out to be enough to properly handle the excess of zero and the consequent overdispersion phenomenon.
Its performance is not significantly different from those of the more complex ZINB and Hurdle NB models, which require, however, to estimate more parameters because of their mixture nature. All things considered, it is better to choose a less complex and more parsimonious model such as the Negative Binomial GLM.