Bayesian Analysis: Gender Effects on Success in Academia

Suoyi Yang

12/4/2019

1 Abstract

Gender biases in academia is a very well studied topic and was a much stronger issue back in the 1960s. This paper examines the factors that potentially affected the professional success of Biochemist Ph.D. students from the 1950s to the 1970s using Bayesian analysis, and focuses on gender, in particular, to determine if it has a larger impact than other factors. Gender was found to be a significant predictor of publication counts, with female Ph.D. students publishing fewer papers than their male counterparts. In addition, having more young children to raise was also found to negatively impact publications counts. Conversely, marriage was found to have a positive effect on publications, while the prestige of the Ph.D. department and publications by mentors did not have strong predictive power in the models. In addition, due to zero-inflation and overdispersion in the data, several different likelihoods were used to find the best fit Bayesian model. The Bayesian model with the Negative Binomial model ended up outperforming all the other models tried in this paper.

2 Introduction

While gender bias in academia has seemingly improved significantly over the years, it is still a popular point of contention due to the strong gender imbalance present in academia just decades ago, particularly in scientific fields. Therefore, to understand some of these biases still present today, it would be beneficial to take a closer look at the factors which have impacted the career and success of students in the past. In particular, we want to see if gender truly had a much larger impact compared to other potential factors.

A common measure used in academia as a proxy for a person's professional success is their publication output. In an effort to better understand the direct and indirect effects of gender on the careers and productivity of academics, Long (1990) gathered data on 915 students who obtained Ph.D.s in biochemistry between 1950 and 1970. In his paper, Long (1990) used a frequentist approach to measure the effects of gender and other factors, such as marital status or reputation of a mentor, on the early career outcomes of Ph.D. students.

In this paper, a similar study will be conducted but instead, we are using a Bayesian approach to the problem. The goal is to understand which factors impact the publication output of biochemistry students and if gender truly plays a large role. In addition, rather than using

the linear regression Long used in his paper, this paper will examine several different Bayesian models and attempt to find the best fit for the data.

3 Analysis and Results

3.1 Initial Analysis of Data

Before we start any bayesian analysis, we need to first take a close took the data provided by Long (1990).

There is a total of 915 student data collected by Long and each student is an observation. The predicted variable we are interested in is the number of articles, `art`, produced by the student during the last three years of his or her Ph.D. program. There are 5 additional covariates in the data:

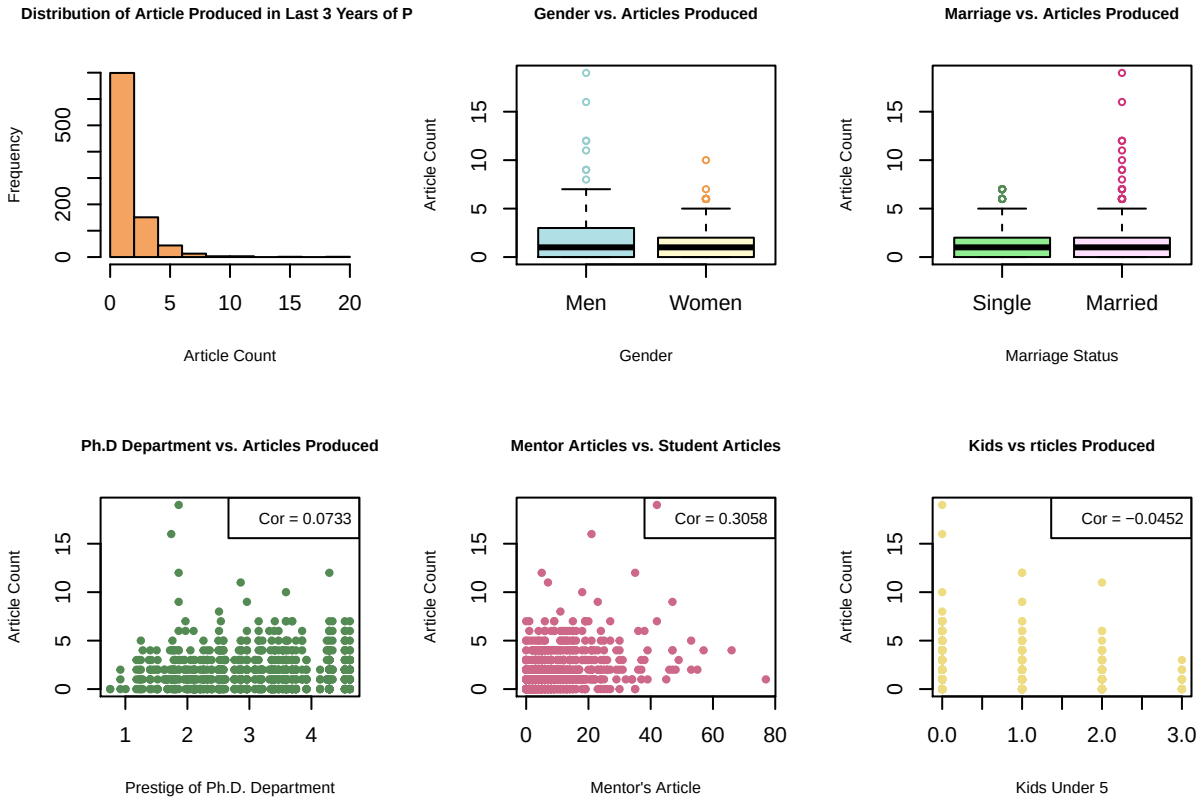
- `fem`: A binary indicator variable for gender (female=1)
- `mar`: A binary indicator variable for marital status (married=1)
- `kid5`: The number of children under the age of five
- `phd`: Scores for departmental prestige range from a low of 1 to a high of 5.
- `ment`: The number of articles produced by the student's mentor within the same three-year period that the student's productivity was measured

A numerical summary of these variables is shown in table 1.

art	fem	mar	kid5	phd	ment
Min. : 0.000	Men :494	Single :309	Min. :0.0000	Min. :0.755	Min. : 0.000
1st Qu.: 0.000	Women:421	Married:606	1st Qu.:0.0000	1st Qu.:2.260	1st Qu.: 3.000
Median : 1.000			Median :0.0000	Median :3.150	Median : 6.000
Mean : 1.693			Mean :0.4951	Mean :3.103	Mean : 8.767
3rd Qu.: 2.000			3rd Qu.:1.0000	3rd Qu.:3.920	3rd Qu.:12.000
Max. :19.000			Max. :3.0000	Max. :4.620	Max. :77.000

Table 1: Summary of Data

Visual summaries of these variables are shown in the following graphs.



One of the graphs shows a histogram distribution of the number of publications for all students in the data. Note that there seems to be a zero-inflation issue for the predicted `art` as over 500 of the 915 students has a count of 0 for the number of publications. In addition, there seems to also be an overdispersion problem as, despite the majority of students producing between 0 to 5 publications, one student seems to have produced around 20 publications, heavily skewing the data. Therefore, when we are fitting the model in the next section, we need to consider methods that can accommodate these issues.

The correlation between `art` and other numerical covariates are indicated on the graph, and the numeric summaries of `art` for different categories of categorical variables are shown in table 2. While there does seem to be some difference in mean of publications between men and women (women publishes less) and between married and single (married students public less), we will have to look at models in the next section to determine whether these differences are significant.

Table 2: Article Summaries for Categorical Variables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Men	0	0	1	1.883	3	19
Women	0	0	1	1.470	2	10
Single	0	0	1	1.592	2	7
Married	0	0	1	1.744	2	19

3.2 Bayesian Analysis of Data

The goal of frequentist inference is to extract information from only the data while making minimal assumptions about it. In contrast, the Bayesian approach begins with making assumptions using prior beliefs about the data and specifying prior distributions for the parameters. Bayesian inference allows us to combine prior information about parameters such as regression coefficients with observed outcomes (y) and covariate data (X) based on Bayes theorem:

$$\underbrace{p(\theta|X,y)}_{\text{posterior}} \propto \underbrace{p(\theta)}_{\text{prior}} \underbrace{p(y|X,\theta)}_{\text{likelihood}}$$

We will first specify the probability model by determining a likelihood (section 3.2.1) for the observed data and a prior distribution for the covariates (3.2.2).

After model specification, we need to get an updated distribution of the covariates conditional on the observed data. To do this, we draw samples from the posterior distribution by using Markov Chain Monte Carlo (MCMC) methods in R (3.2.3). We will then check the diagnostics of all the models (3.2.3) and finally determine the best fit model for the data (3.2.4).

3.2.1 Likelihood

The specification of the likelihood is typically based on the outcome type and is often the same as a frequentist analysis. We will try to model the data with different likelihoods in this paper and attempt to find the best fit.

We will first use the Poisson likelihood as it is the most common likelihood to use for count data. However, recall in section 3.1 that we noted there is overdispersion in the predictor `art` variable. Since the Negative Binomial distribution is good for handling overdispersion, we will also try that as one of our likelihood distributions.

Besides overdispersion, we also need to take into account the zero-inflation problem found in the data. Since hurdle Poisson, hurdle Negative Binomial, zero-inflated Poisson, and zero-inflated Negative Binomial are designed to accommodate excess 0 counts in data, we will also try these four distributions as out likelihoods.

3.2.2 Priors

In Bayesian analysis, prior information about covariates and the data should not come from analyzing the data itself, but from either expertise or outside information. Thus in order to get informed priors for the covariates, I researched and read several papers discussing gender and academic success from around the 1950s to the 1970s.

- `fem`: Around the 1950s to 1970s, it has been found that on average, female scientist publishes fewer articles than male scientists (Cole & Zuckerman 1984). A study done by Cole and Zuckerman in 1984 found that the odds ratio of female to male article publications is around 0.57. Thus I chose the prior for `fem` to be $N(\log(.57) = -0.5621189, 1)$.
- `pdh`: Some studies have found that the prestige of the Ph.D. department of graduate education of a student had a large impact on his or her success and career (Crane 1965). However, Long and McGinnis (1985) found that when variables about mentor prestige are included in the model, the prestige of the Ph.D. department had a much smaller effect on the success of students. So I chose the prior for `phd` to be $N(0,1)$ since information about mentors will also be added into the model.
- `ment`: Long and McGinnis (1985) found that mentor prestige (which includes mentor publications) had a significant impact on the publication count of students. As a result, I placed an informed prior of $N(0.5621189, 1)$ so that it also indicates a strong effect on publication count like the `fem` covariate, but in the positive direction instead.
- `kid5`: In Long's paper (1990), he states that after looking through several different research paper, the impact of kids on publications seem to be mixed from having no impact, to somewhat negative impact. To reflect this "weak" negative effect, I made its prior normal with a negative location value, but relatively smaller in magnitude compared to `ment` and `fem`: $N(-0.22,1)$.
- `mar`: Long's paper (1990) once again indicates varied effects of marriage on publications, ranging from positive effect to no effects to negative effects. To indicate this, I gave this a weakly informative prior of $N(0,1)$.

3.2.3 Posteriors

Now that we have specified the prior and likelihood, it is time to move onto drawing from the posterior distribution using Markov Chain Monte Carlo (MCMC) by combining our priors with our likelihood in order to update our beliefs about the predictor variables.

This is done with `stan_glm` in package `rstanarm` for the Poisson and Negative Binomial model, and `brm` in package `brms` for hurdle Poisson, hurdle Negative Binomial, zero-inflated Poisson, and zero-inflated Negative Binomial. The estimates for the mean of each of the covariate predictors are shown below for all 6 models.

	Intercept	femWomen	marMarried	kid5	phd	ment
Poisson	0.3070	-0.2268	0.1533	-0.1852	0.0128	0.0256
NB	0.2545	-0.2175	0.1498	-0.1775	0.0157	0.0293
ZI Poisson	0.5507	-0.2323	0.1321	-0.1709	0.0034	0.0214
ZI NB	0.2822	-0.2208	0.1484	-0.1776	0.0134	0.0288
Hurdle Poisson	0.6679	-0.2292	0.0964	-0.1426	-0.0125	0.0187
Hurdle NB	0.3215	-0.2498	0.1044	-0.1545	-0.0018	0.0242

Table 3: Covariate Estimates

The GLM methods above seemed to have all consistently produced negative mean estimates for the variables `fem` (around -0.23) and `kid5` (around -0.1678). Recall that this means that with all other covariates help constant, the expected number publications for a female is around $e^{-0.23} = 0.7945336$ times the number publications for a male. Similarly, with all variables held constant, as you increase the number of kids under 5 by 1 unit, the expected number of publications for a student decreases by a factor of 0.8455229. Thus following these interpretations, we see that being a woman and having more kids under the age of 5 both negatively impact publication count, while marriage seems to have a positive effect on publication counts. Department prestige and mentor article count are also positive, indicating a potentially positive effect on publication count, but their magnitudes are relatively small compared to the other variables, indicating their effect on publication count is not great.

3.2.4 GML Diagnostics

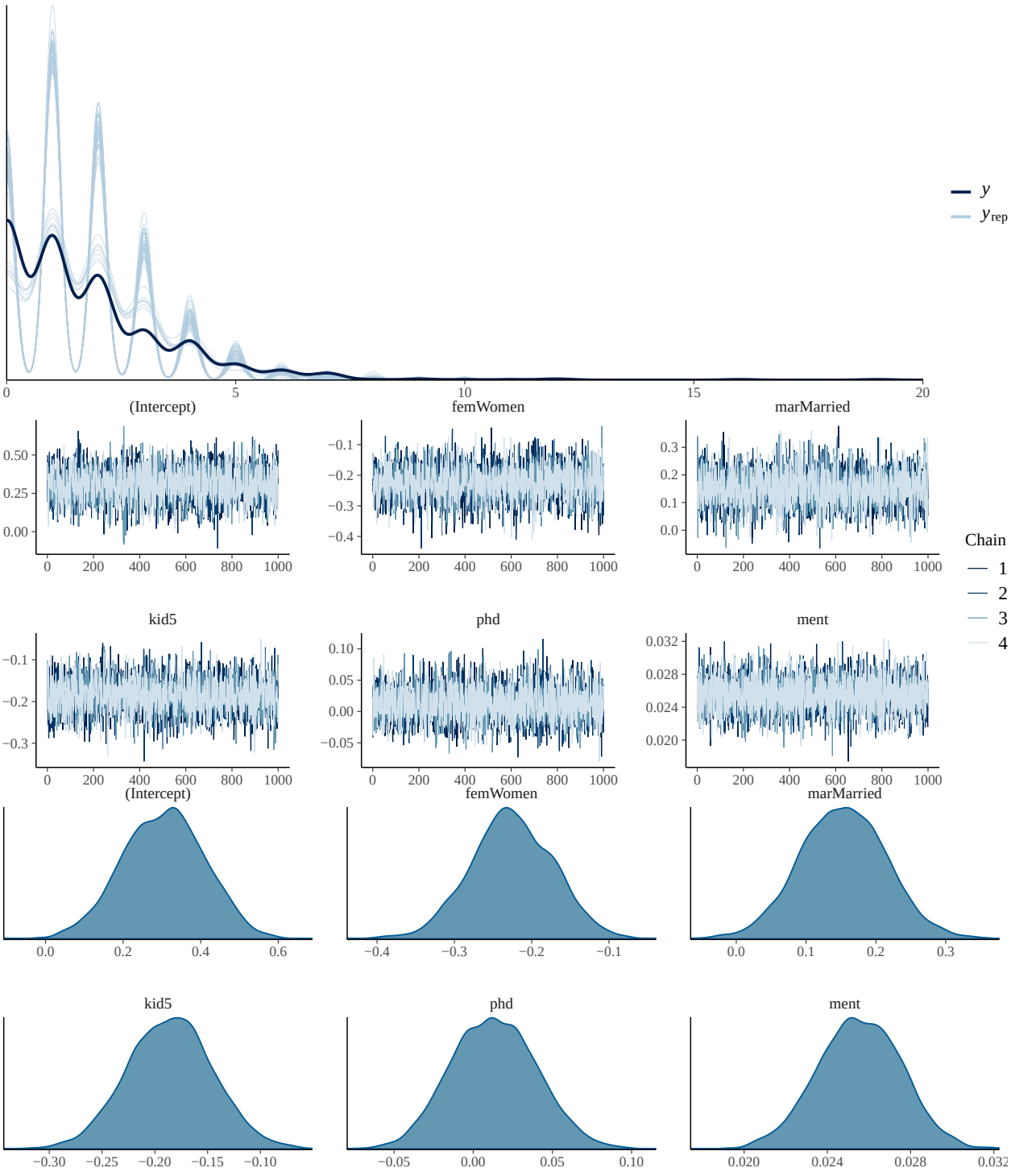
Besides simply looking at the significance of certain predictors in the model, we must also check the diagnostics of our model to check if our models are even valid. In particular, since MCMC was used in creating our models, we need to makes sure that convergence occurred in our models. This can be done using trace plots. The diagnostics for each of the model is shown below.

The first diagnostic we see for each of the models is the posterior predictive check. It generates several simulations from the posterior predictive distribution (light lines), using the covariate data used to fit the models, and compares them to the observed outcome (dark line). If the models are a good fit, then simulations from posterior predictive distribution should look very similar to the data we observed. Indeed we see that the posterior predictive check for all the model seems to show the simulations looks fairly similar to the observed data (although the model using Poisson and hurdle Negative Binomial likelihood performed a little worse than the others).

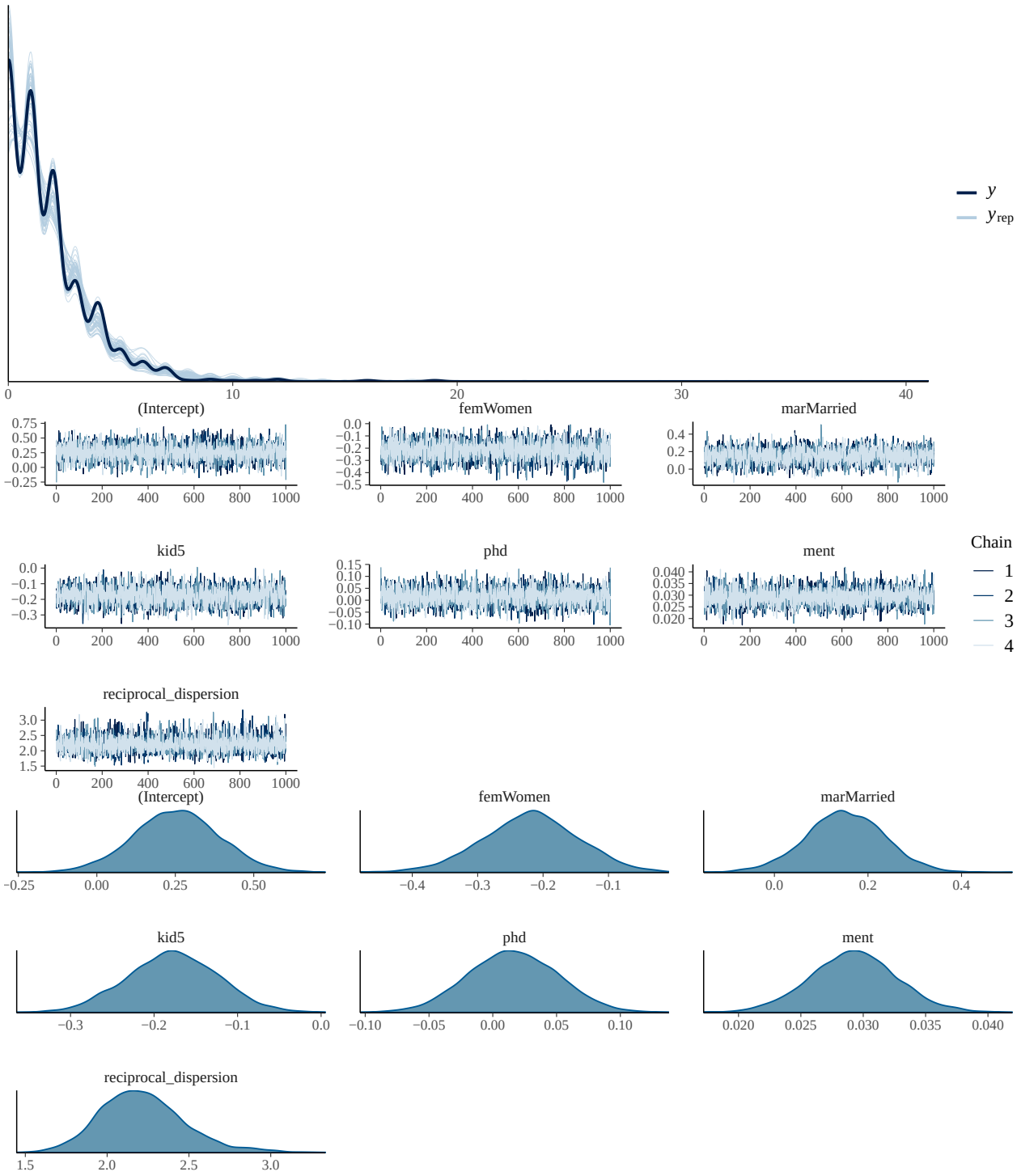
The trace plots are also shown for each model. We see that in all the model diagnostics, chains in each trace plot are center around one value and varying randomly around it. In addition, all the chains seem to be overlapping one another in the trace plot This indicates that the chains are stable and well mixed, meaning convergence has occurred in MCMC for all the models.

The final diagnostics shown are the marginal posterior distributions (shown as densities) for the predictors we used in the models. We note that in all the models, all the variables seem to be very close to a univariate normal distribution, which is the desired marginal posterior distributions. The normal distribution is centered around the mean estimate for each covariate of the model. These mean estimates were shown and explained in section 3.2.3 and table 3.

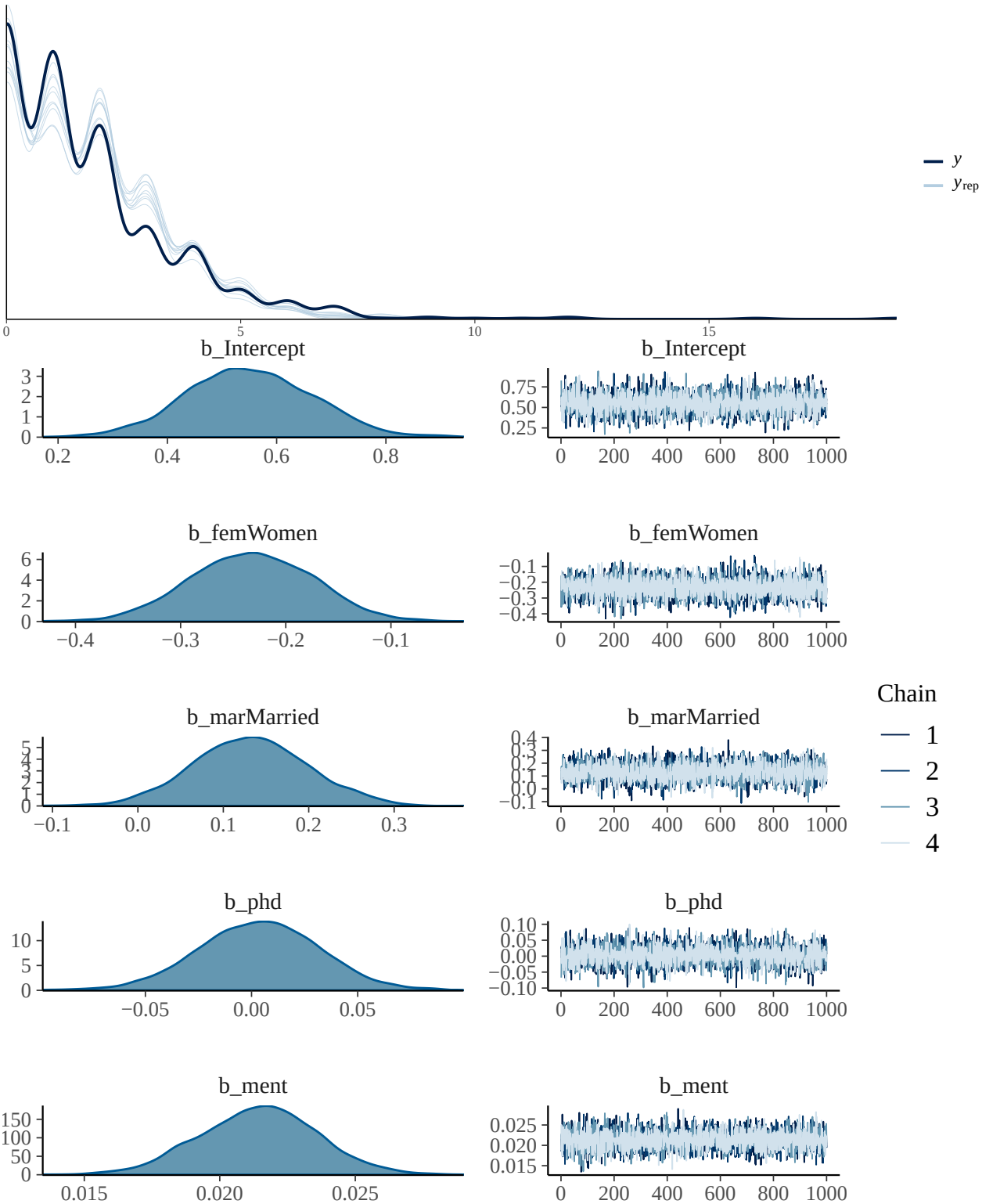
GLM Diagnostics: Poisson



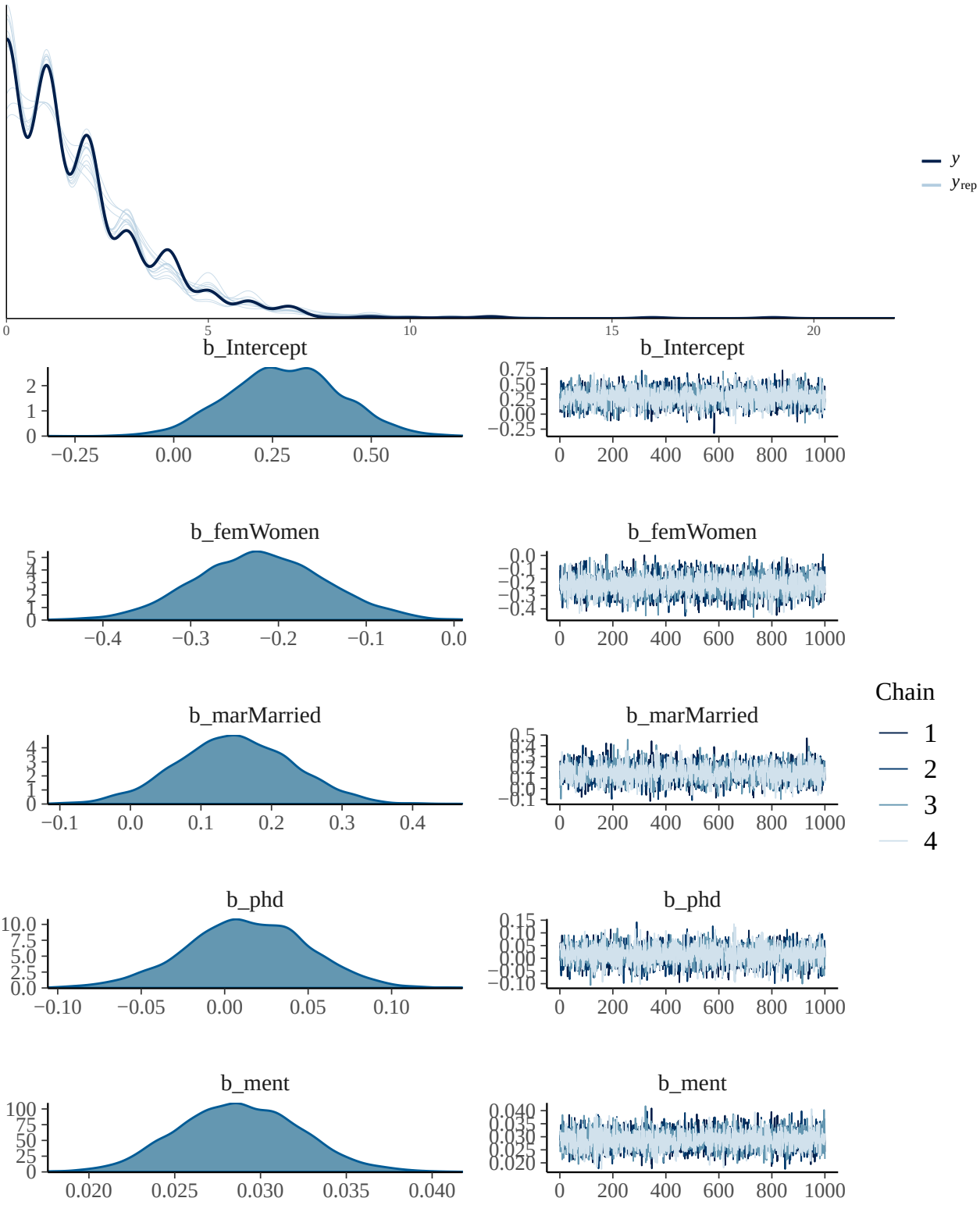
Generalized Linear Model: Negative Binomial



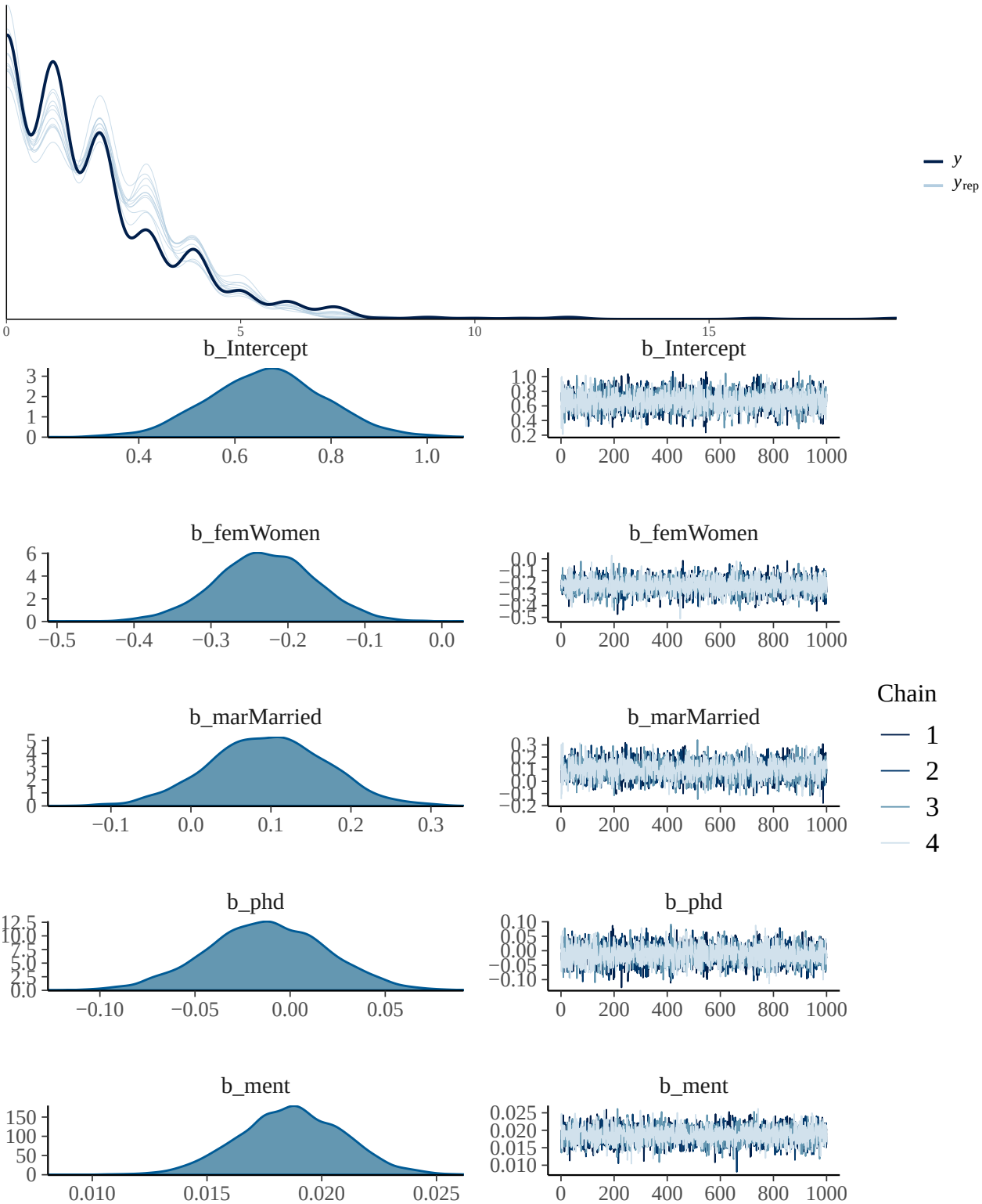
Generalized Linear Model: Zero-Inflated Poisson



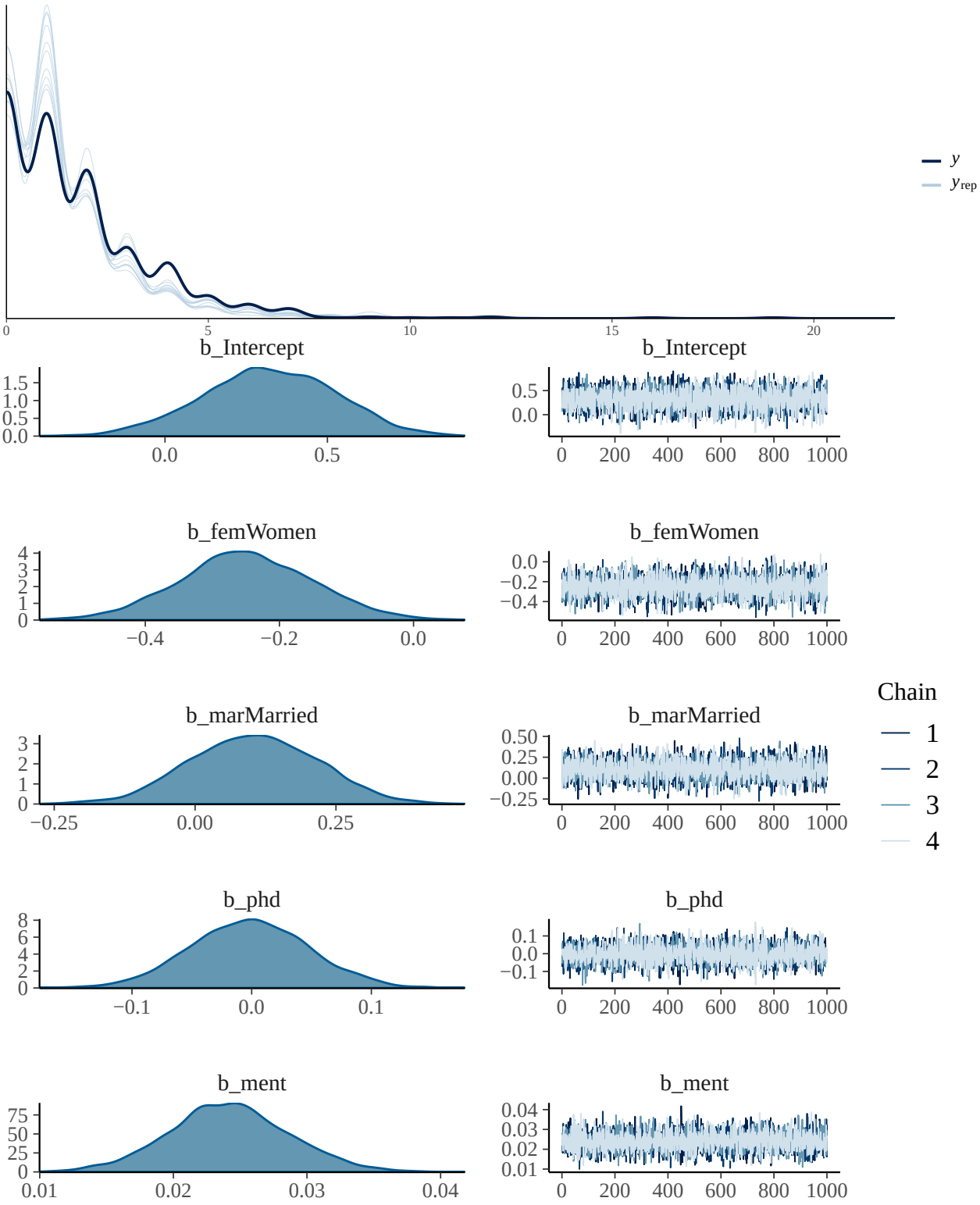
Generalized Linear Model: Zero-Inflated Negative Binomial



Generalized Linear Model: Hurdle Poisson



Generalized Linear Model: Hurdle Negative Binomial



3.2.5 Model Selection

As the final part of our analysis, we wish to look at which of the models we built using the Bayesian approach was has the best fit. We can do this using Leave-One-Out (LOO) cross-validation which assesses the generalizability of a model. LOO estimates the prediction accuracy from a fitted Bayesian model by using the likelihood calculated at the posterior simulations of the covariate predictor values (Vehtari, Andrew Gelman, and Gabry 2016).

R can perform LOO on Bayesian models by using the `loo` and `compare` functions in package `loo`. The `compare` function takes in two Bayesian models and calculates an expected log predictive density (ELPD) difference. When the ELPD difference is positive, then the expected predictive accuracy for the second model is higher. **A negative ELPD difference favors the first model**.

The results of comparing the LOO of the models are shown in the matrix below. The **rows indicate the "first" model while the columns indicate the "second" model**. The values in the matrix are the ELPD differences for the first and second models.

	NB	ZI NB	Hurdle NB	ZI Poisson	Hurdle Poisson	Poisson
NB	0.0000	-1.0020	-26.8394	-63.1178	-81.8424	-92.9255
ZI NB	1.0020	0.0000	-25.8374	-62.1158	-80.8404	-91.9235
Hurdle NB	26.8394	25.8374	0.0000	-36.2784	-55.0030	-66.0861
ZI Poisson	63.1178	62.1158	36.2784	0.0000	-18.7245	-29.8077
Hurdle Poisson	81.8424	80.8404	55.0030	18.7245	0.0000	-11.0832
Poisson	92.9255	91.9235	66.0861	29.8077	11.0832	0.0000

Table 4: LOO Model Comparison

We see from the results above, the ELPD difference shows that the fit of the Bayesian models from *best* to *worst* is:

- Negative Binomial
- Zero-inflated Negative Binomial
- Hurdle Negative Binomial
- Zero-inflated Poisson
- Hurdle Poisson
- Poisson

This result is a little unexpected since I had previously thought the four models specialized to handle zero-inflation would perform best due to the heavy zero-inflation issue in the model.

4 Discussion

From our analysis, we see that our prior assumptions about the impact of gender and young children on publication count (and thus professional success in academia) are correct in that

they both have negative impacts on the publications. In addition, their coefficient estimates are relatively large compared to the other variables, indicating that they seem to be stronger predictors of publication count compared to other variables. These results do make sense. Having young children to take care of is very time consuming, which would likely have to divide their time between childcare and their education. It is also well known that the was a gender bias against women in the scientific field (particularly in higher graduate degrees), and the results of the model seem to indicate a similar result.

It is interesting to note that being married actually seems to have a relatively positive effect on publication counts in the models. A potential explanation for this could perhaps be that having spousal help with housework and responsibilities frees up some time for more academic pursuits.

Another interesting result is that both prestige of Ph.D. department and mentor publications (considered mentor's "prestige"), while having positive coefficient estimates, have small magnitudes. This indicates that they have less impact on publications in the models. A potential explanation could be the publication of a mentor in the last 3 years of a student's Ph.D. program is a very limited way of measuring prestige and does not take into account previous works of the mentor which can contribute to their "prestige".

Looking at the fits of the model, it was surprising that the bayesian model that used Negative Binomial as its likelihood performed best. While it made sense that it would perform better than the Poisson likelihood because of Negative Binomial accounts for overdispersion (which was present in the data), I thought zero-inflation would be a much larger issue and the hurdle and zero-inflation version of Poisson and Negative Binomial likelihood would perform better. A potential reason for this could be that the large zero counts for publication do not necessarily distort the shape of the unimodal distribution of the publication count. And in addition, perhaps the zero-inflation methods ended up being a little overzealous trying to adjust for the large number of zero counts and predicted a much higher number of publications than actually observed in the data.

5 References

Cole, Jonathan RX, and Harriet Zuckerem 1987. "Marriage and Motherhood and Research Performance in Science." Scientific American 256:119

Hamovitch, William, and Richard D. Morgansten 1977. "Children and the Productivity of Academic Women." Jounal of Higher Education 48:63345

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

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." Statistics and Computing 27.5 (2016): 1413–1432. Crossref. Web.