Absenteeism at work

Nicole Salani

2022-05-13

Contents

Introduction

The goal of this analysis is to explore what individual and social-context attributes are strongly associated with absenteeism at work. Absenteeism is any failure to report for or remain at work as scheduled, regardless of the reason (Cascio & Boudreau, 2015). Absenteeism can have a severe impact on the workplace. According to the Centers for Disease Control and Prevention (CDC), "productivity losses from missed work cost employers $225.8 billion, or $1,685 per employee, each year." Therefore understanding why employees fail to report to work and identifying the patterns in absenteeism behavior is an invaluable tool for human resource management. Such insights can help guide and inform organizational decision-making to ensure greater employee engagement and with that higher productivity.

Background

Absences at work can arise due to an array of factors. Winkelmann (1999) found that absenteeism is dependent on factors such as wages and even the firm size. Organizational policies also impact the level of absenteeism (Ruhle and SilB, 2020). For example, Halpern and colleagues (2001) found that the smoking policy in the workplace affects absenteeism and productivity. Their research concludes that current smokers tend to have significantly higher absenteeism than former smokers and non-smokers. Absenteeism is repeatedly reported to be strongly correlated with employees' health status. For instance, Tunceli and colleagues (2005) found that for employees with diabetes, the absolute probability of reporting for work for male and female employees is 7.1 % and 4.4 % points less compared to the individuals without diabetes.

Gates and colleagues (2008) found that moderately or extremely obese workers experience a 4.2% loss in productivity, which is tantamount to 1.18% more than all other employees. Other health-related factors with well documented impact on absenteeism are social-context conditions like cold and influenza seasons during winter. For some time now, studies have shown health-related workplace absenteeism correlates well with the prevalence of influenza-like illiness and reaches seasonal peaks in conjunction with influenza activity.

Our goal is to understand the relationship between absenteeism and employee-level and social-context attributes. Using research evidence on employee absenteeism and what know about data at hand, we consider two count models to assess absenteeism and attributes that affect it: (1) Poisson regression model, and, (2) Negative binomial.

Data

The data used consists of records of employee absences collected over a three year period from July 2007 to July 2010 at a courier company in Brazil[1]. The dataset is a panel at the employee level of 36 employees, comprising of a total of 740 instances recorded over 21 attributes per instance.

Based on the outlined research evidence, we consider explanatory relationships underlying the reality of employee absenteeism using the following variables:

- Body mass index (BMI): is convenient rule of thumb used to broadly categorize a person as underweight, normal weight, overweight, or obese based on tissue mass (muscle, fat, and bone) and height. BMI is modelled as a continuous variable. *Seasons: Four seasons of the year* Social.drinker: Indicator that takes value "Yes", if an employee is a social smoker, and, "No" otherwise.
- Social.smoker: Indicator that takes value "Yes", if an employee is a social drinker, and, "No" otherwise.
- Disciplinary.failure: Indicator variale that takes value "Yes" for a past disciplinary failure, and, "No" otherwise

Model

Before building our model, we first consider the context of the data and then outline assumptions. The data is on employees at a courier company. A courier a service is a premium, all-inclusive service which collects and delivers shipments in the shortest possible time frame. In cities, there are often bicycle couriers or motorcycle couriers but for consignments requiring delivery over greater distance networks, this may often include trucks. A courier may be assigned routes but will also pick up and deliver individual orders that are placed same day. It seems reasonable that a courier company provides around the clock service and so employees are working varied hours in the day. We assume they can be assigned to work first, second, or third shifts based on an hours schedule that is fixed, rotating, or split, or on-call. We assume employees work independently of one another,and are assigned to different types of shifts of varying length based on the nature of consignments and their length of employment at the courier company. With that in mind, we can consider each sample data or instance i of absenteeism in hours, Y_i , as a random count that can take values 0 up to some large unknown number, $Y_i \in \{0, 1, 2, ..., n\}$, and so the poisson model makes a reasonable candidate for modelling our data.

We assume a poisson data model for the number of absenteeism hours for each sample data or instance i (Y_i) in our data collection period, where the rate of absenteeism hours λ_i depends on the BMI, X_{i1}, Season, X_{i2}, Social.drinker, X_{i3}, Social.smoker, X_{i4}, and Disciplinary.failure, X_{i5}

Given that the counts of hours are aggregated over different amounts of time, or different exposures, we adjust for the length of employment using the variable, Service time in Months. This is represented as the offset variable in the model. Lastly, we use some flat priors to give weight to the data in what our models reveal.

[1]It is publicly available at the UCI machine learning repository: https://archive-beta.ics.uci.edu/ml/datasets/absenteeism+at+work)

Poisson Model

The first model we consider is the poisson model.

$\beta_0 \sim \text{Normal}\,(0, 10)$
$\beta_1 \sim \text{Normal}\,(0, 10)$
$\beta_2 \sim \text{Normal}\,(0, 10)$
$\beta_3 \sim \text{Normal}\,(0, 10)$
$\beta_4 \sim \text{Normal}\,(0, 10)$
$\beta_5 \sim \text{Normal}\,(0, 10)$
$\eta_i = \log(\text{months})_i + \beta_0 + \beta_1 \times \text{BMI}_i + \beta_2 \times \text{Season}_i + \beta_3 \times \text{Social.smoker}_i + \beta_4 \times \text{Social.drinker}_i + \beta_5 \times \text{Discplinary failure}_i$
$\lambda_i = e^{\eta_i}$
$Y_i \sim \mathcal{P}l\rangle\int\int l\backslash\,(\lambda_i)$

```
poisson_covs <- stan("poisson.stan",
                                  data = list(s = 10,
                                              N = nrow(X), p = ncol(X),
                                              offset = log(data$Months.in.Service),
                                              X = X,
                                              y = y))
```

```
## Trying to compile a simple C file
```

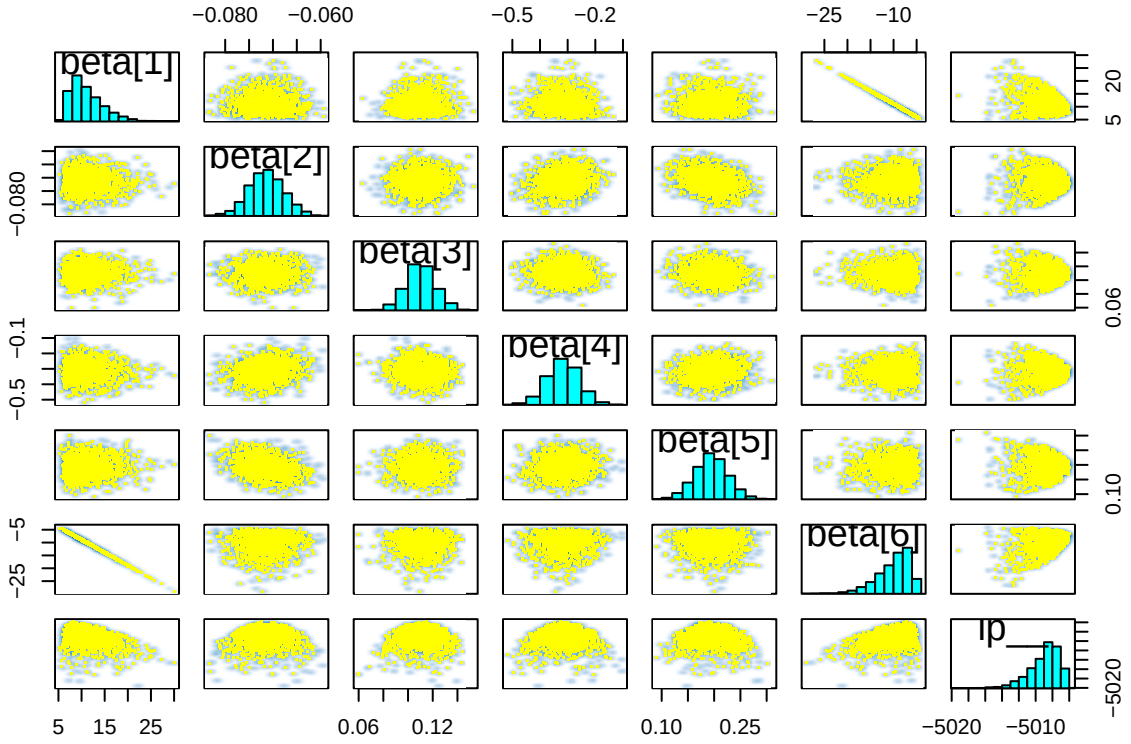
```
## Warning: There were 2853 transitions after warmup that exceeded the maximum treedepth. Increase max_
## https://mc-stan.org/misc/warnings.html#maximum-treedepth-exceeded
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

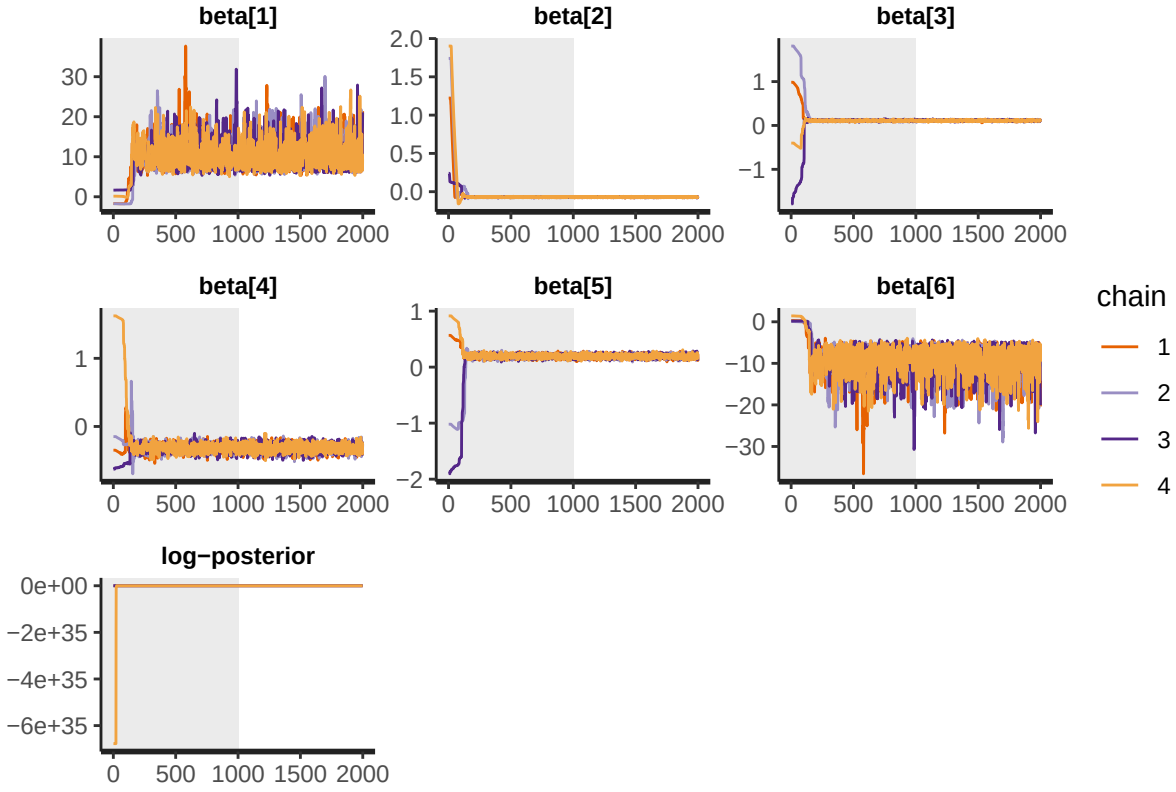
```
#check_hmc_diagnostics(poisson_covs)
```

```
pars = c("beta","lp__")
pairs(poisson_covs, pars = pars)
```

Poisson Model Checks




```
traceplot(poisson_covs, inc_warmup = TRUE, pars = pars)
```



Negative Binomial Model

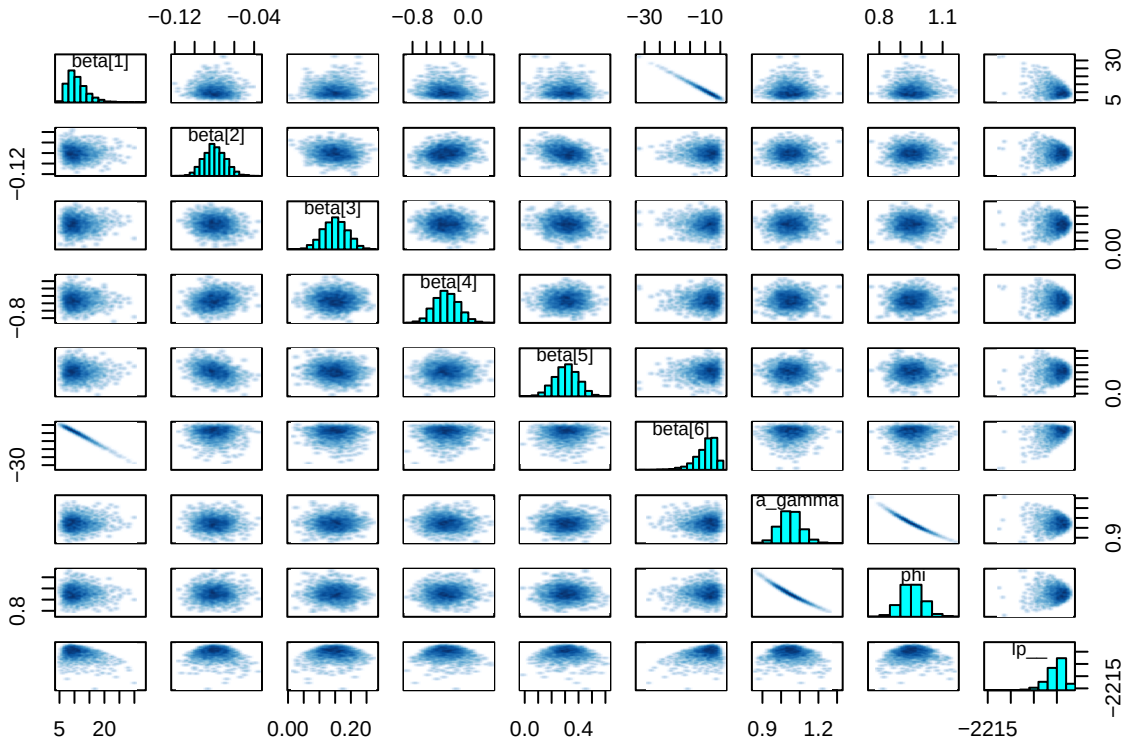
There seems to be a lot of unexplained variation in absenteeism rate in our poisson model which, presumably arises from unobserved influences that vary from case to case, generating variation in the true λ_i's.So we consider an extension of Poisson GLMs by swapping the Poisson distribution for something the NEGATIVE BINOMIAL distribution, or also sometimes called the GAMMA-POISSON distribution. This model assumes that each poisson count observation, count of absenteeism in hours for our case, has its own rate. It estimates the shape of a gamma distribution to describe the Poisson rates across cases.

$$
\begin{aligned}
\beta_0 &\sim \text{Normal}\,(0, 10) \\
\beta_1 &\sim \text{Normal}\,(0, 10) \\
\beta_2 &\sim \text{Normal}\,(0, 10) \\
\beta_3 &\sim \text{Normal}\,(0, 10) \\
\beta_4 &\sim \text{Normal}\,(0, 10) \\
\beta_5 &\sim \text{Normal}\,(0, 10) \\
\eta_i &= \log(\text{months})_i + \beta_0 + \beta_1 \times \text{BMI}_i + \beta_2 \times \text{Season}_i + \beta_3 \times \text{Social.smoker}_i + \beta_4 \times \text{Social.drinker}_i + \beta_5 \times \text{Discplinary f} \\
\lambda_i &= e^{\eta_i} \\
\alpha_\gamma &\sim \text{Gamma}(a, b) \\
\phi_i &\sim \frac{1}{\alpha_\gamma} \\
Y_i \mid \lambda_i, \phi_i &\sim \text{NegBinomial}\,(\lambda_i, \phi_i)
\end{aligned}
$$

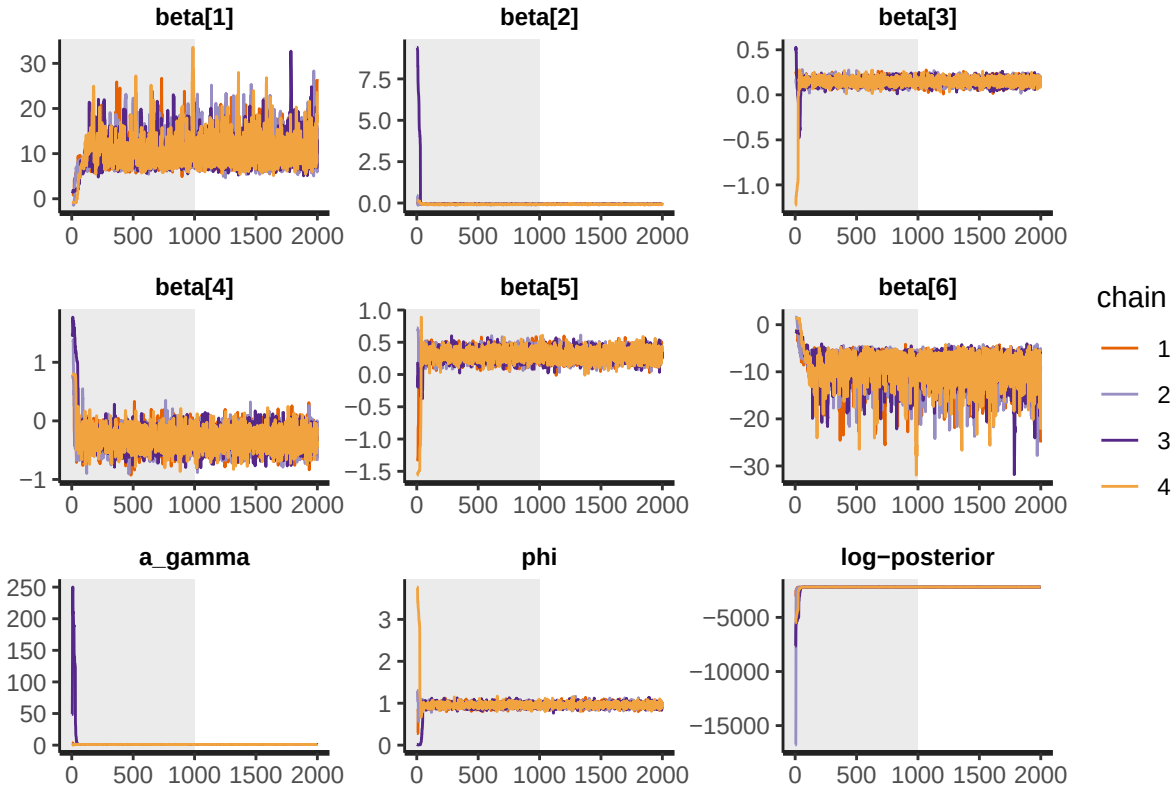
```
negbinomial_covs <- stan("negBinomial.stan",
                                data = list(a = 10^(-3), b = 10^(-3),
                                            s = 10,
                                            N = nrow(X), p = ncol(X),
                                            offset = log(data$Months.in.Service),
                                            X = X,
                                            y = y))
#check_hmc_diagnostics(negbinomial_covs)
```

Negative Binomial Model Checks We can assess the pairs plot to see if parameters all look reasonable centered with no strange patterns. Interestingly, the intercept and slope for Disciplinary.failure are skewed. This is also consistent with the traceplots which show some slight non-convergence for both the intercept and Disciplinary.failure slope. With that said, this model seems sufficiently converged compared to the poisson.

```
pars = c("beta","a_gamma","phi","lp__")
pairs(negbinomial_covs, pars = pars)
```




```
traceplot(negbinomial_covs, inc_warmup = TRUE, pars = pars)
```



Parameter Estimates

The confidence intervals of the partial slopes lie above zero, with two showing great magnitude. This implies that the attributes considered do have some significant effect on the variation of average or rate of monthly absenteeism. Of particular note are the slopes on the intercept, β_0 and on indicator for having a record of Disciplinary failure, β_5.

```
cat("Poisson\n")
```

```
## Poisson
```

```
print(poisson_covs, pars = c("beta"))
```

```
## Inference for Stan model: poisson.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##           mean se_mean   sd   2.5%    25%   50%   75% 97.5% n_eff Rhat
## beta[1]  11.13    0.13 3.60   6.30   8.44 10.33 13.17 19.89   734 1.01
## beta[2]  -0.07    0.00 0.00  -0.08  -0.07 -0.07 -0.07 -0.06  3429 1.00
## beta[3]   0.11    0.00 0.01   0.09   0.10  0.11  0.12  0.13  2453 1.00
## beta[4]  -0.32    0.00 0.06  -0.43  -0.35 -0.32 -0.28 -0.21  2464 1.00
## beta[5]   0.19    0.00 0.03   0.13   0.17  0.19  0.22  0.26  2198 1.00
## beta[6] -10.01    0.13 3.60 -18.69 -12.04 -9.23 -7.33 -5.23   734 1.01
##
## Samples were drawn using NUTS(diag_e) at Wed May 25 05:48:50 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

```
cat("Negative binomial\n")
```

```
## Negative binomial
```

```
print(negbinomial_covs, pars = c("beta","a_gamma"))
```

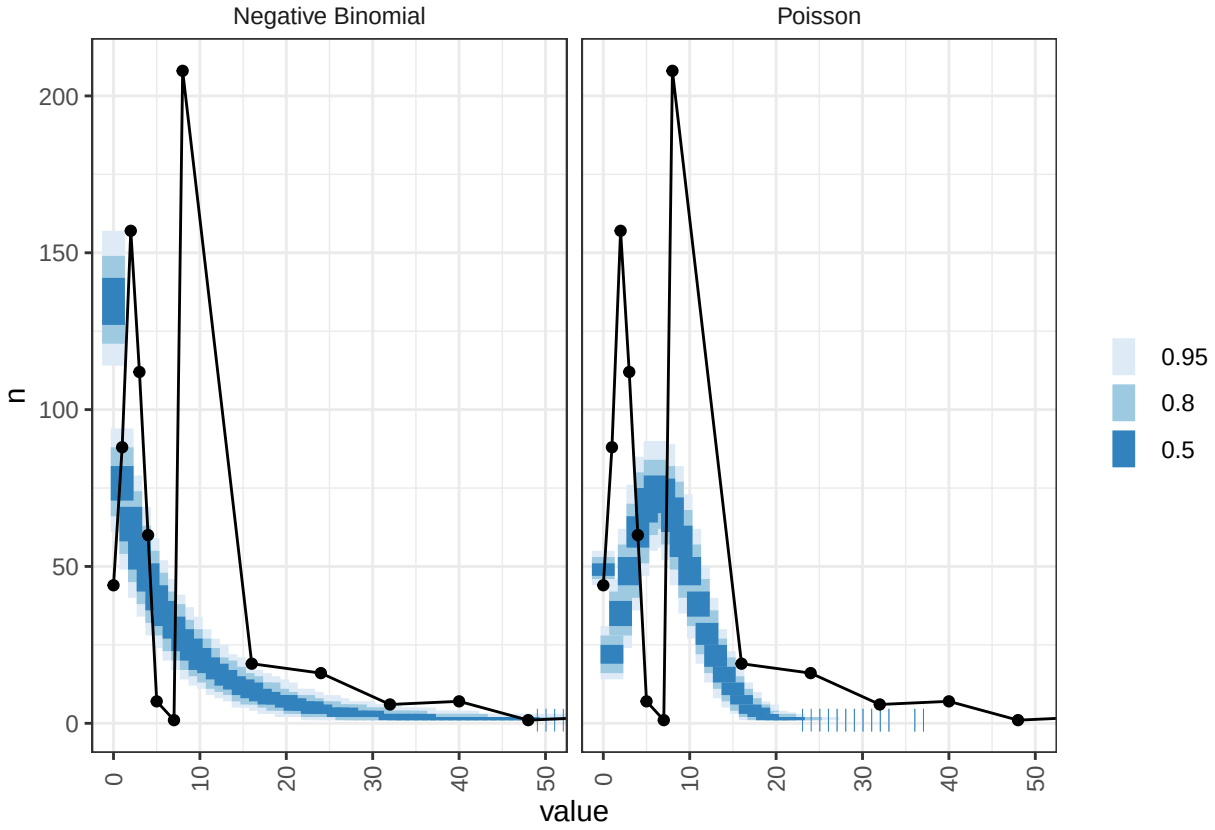
```
## Inference for Stan model: negBinomial.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##          mean se_mean   sd   2.5%    25%   50%   75% 97.5% n_eff Rhat
## beta[1] 11.02    0.11 3.49   6.41   8.53 10.28 12.82 19.66  1088    1
## beta[2] -0.08    0.00 0.01  -0.10  -0.09 -0.08 -0.07 -0.06  3393    1
## beta[3]  0.15    0.00 0.04   0.07   0.12  0.15  0.18  0.22  3568    1
## beta[4] -0.31    0.00 0.17  -0.64  -0.43 -0.32 -0.20  0.04  3641    1
## beta[5]  0.31    0.00 0.09   0.13   0.25  0.31  0.37  0.48  3137    1
## beta[6] -9.85    0.11 3.48 -18.49 -11.58 -9.08 -7.36 -5.30  1084    1
## a_gamma  1.05    0.00 0.06   0.94   1.01  1.05  1.09  1.17  3281    1
##
## Samples were drawn using NUTS(diag_e) at Wed May 25 05:54:04 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Model Comparison

The expected log pointwise predictive density (elpd) is higher for the negative binomial model, suggesting a better fit to our data.

```
library(loo)
#loo_compare(loo(poisson_covs), loo(negbinomial_covs))
loo_list <- list(loo(poisson_covs, moment_match = TRUE), loo(negbinomial_covs, moment_match = TRUE))
loo_compare(loo_list)
loo_model_weights(loo_list)
```

It seems even though the negative binomial better models the dispersion of the data, it still does not provide better fit for the two peaks at lower counts of absenteeism hours. We might consider the zero-inflated poisson model, a two-part model that captures two different types of zeros: structural zeros and poisson zeros. Structural zeros are zero-count observations that come from a subpopulation who can only have zeros while poisson zeros are those coming from subpopulation who can have non-zero values but by chance had zeros. In this context of hours of absenteeism at work, this does seem to be an appropriate model. And so for the next steps, we can consider finding attributes that better explain and identify the underlying grouping at lower counts of absenteeism hours and/or consider a more customized poisson model that factors in individual-level variations in absenteeism. We could also transform the categorical attributes to extract a clearer signal on their effects in terms of direction and magnitude, by defining a baseline or reference level. For example, defining "Winter" and "Yes" as reference levels for Season, and Disciplinary failure indicator, respectively.



Bibliography

1. TROTTER Y Jr, DUNN FL, DRACHMAN RH, HENDERSON DA, PIZZI M, LANGMUIR AD. Asian influenza in the United States, 1957-1958. Am J Hyg. 1959 Jul;70(1):34-50. doi: 10.1093/oxfordjournals.aje.a120063. PMID: 13670166.

2. McElreath, R. (2016). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman & Hall/CRC Press.

3. R. Winkelmann, "Wages, firm size and absenteeism," Applied Economics Letters, vol. 6, no. 6, pp. 337-341, 1999.

4. S. A. Ruhle and S. SilB, "Presenteeism and absenteeism at work-an analysis of archetypes of sickness attendance cultures," Journal of Business and Psychology, vol. 35, no. 2, pp. 241-255, 2020.

5. M. T. Halpern, R. Shikiar, A. M. Rentz, and Z. M. Khan, "Impact of smoking status on workplace absenteeism and productivity," Tobacco control, vol. 10, no. 3, pp. 233-238, 2001.

6. K. Tunceli, C. J. Bradley, D. Nerenz, L. K. Williams, M. Pladevall, and J. E. Lafata, "The impact of diabetes on employment and work productivity," Diabetes care, vol. 28, no. 11, pp. 2662-2667, 2005.

7. D. M. Gates, P. Succop, B. J. Brehm, G. L. Gillespie, and B. D. Sommers, "Obesity and presenteeism: the impact of body mass index on workplace productivity," Journal of Occupational and Environmental Medicine, vol. 50, no. 1, pp. 39-45, 2008.