How a Student's Number of Days of Absence is Indicated by the Score in a Standardized Mathematics Test

APSTA-GE 2123 Project

Zixuan Zhou

5/15/2020

I. Introduction

In this report, the topic discussed is how a student's number of days of absence is indicated by the score in a standardized mathematics test, and the data set used is form UCLA Institution for Digital Research & Education Statistical Consulting. The data is firstly accessed by delivering both numerical and graphical summaries. Two models- Negative Binomial Model and Poisson Model are then implemented to fit the data, and this is followed by model comparison. Before the final wrap-up of the report, the posterior predictive distribution is shown.

II. Description of the Data

The data set contains attendance data on 314 high school juniors from two urban high schools in the file nb_data. The response variable of interest is days absent, daysabs. The variable math gives the standardized math score for each student. The variable prog is a three-level nominal variable indicating the type of instructional program in which the student is enrolled.

Numerical Summary

```
dat <- read.dta("https://stats.idre.ucla.edu/stat/stata/dae/nb_data.dta")
dat <- within(dat, {
    prog <- factor(prog, levels = 1:3, labels = c("General", "Academic", "Vocational"))
    id <- factor(id)
})

summary(dat)
```

```
##        id         gender         math          daysabs              prog
##  1001   :  1   female:160   Min.   : 1.00   Min.   : 0.000   General   : 40
##  1002   :  1   male  :154   1st Qu.:28.00   1st Qu.: 1.000   Academic  :167
##  1003   :  1                Median :48.00   Median : 4.000   Vocational:107
##  1004   :  1                Mean   :48.27   Mean   : 5.955
##  1005   :  1                3rd Qu.:70.00   3rd Qu.: 8.000
##  1006   :  1                Max.   :99.00   Max.   :35.000
##  (Other):308
```

Each variable has 314 valid observations and their distributions seem quite reasonable. The unconditional mean of our outcome variable is much lower than its variance. This might imply that Poisson Model is not likely to be a suitable model to fit the data.

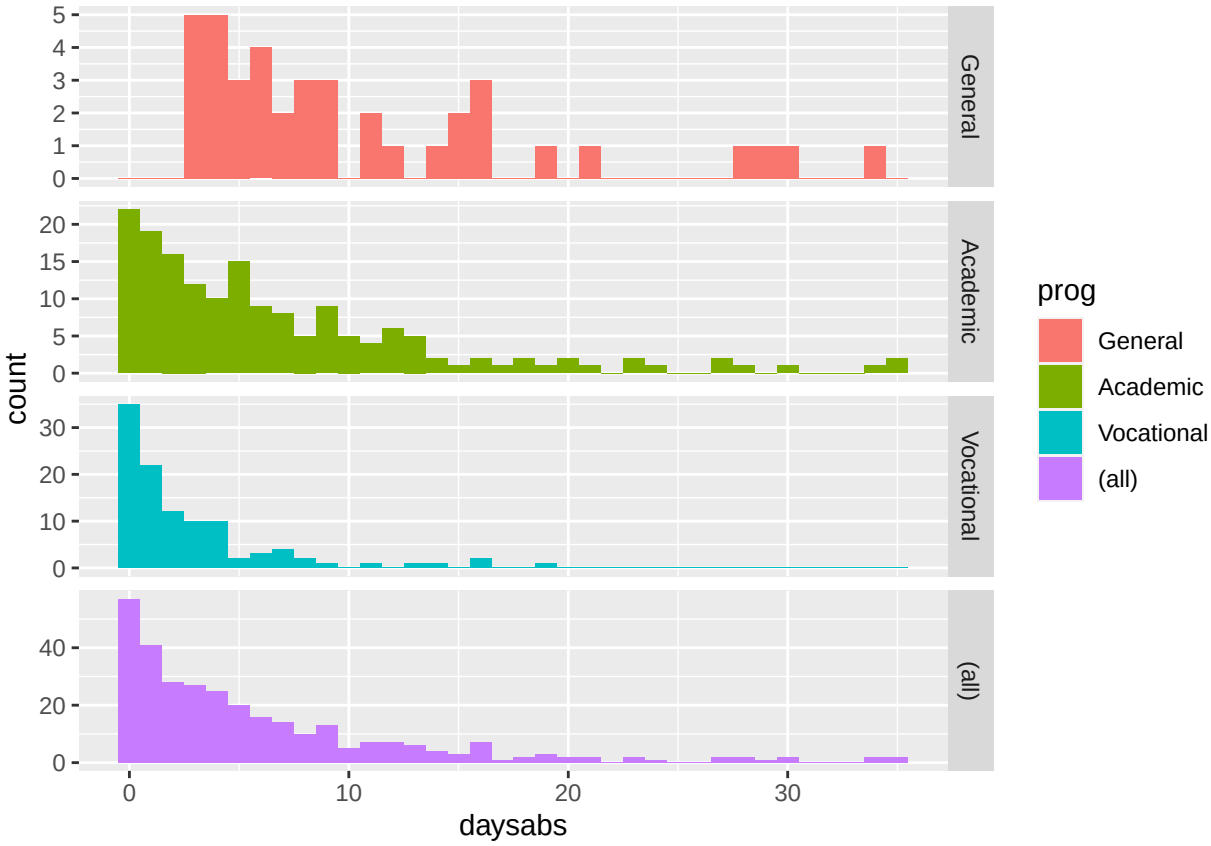
```
with(dat, tapply(daysabs, prog, function(x) {
    sprintf("M (SD) = %1.2f (%1.2f)", mean(x), sd(x))
}))
```

```
##                 General                Academic              Vocational
## "M (SD) = 10.65 (8.20)"  "M (SD) = 6.93 (7.45)"  "M (SD) = 2.67 (3.73)"
```

The table above demonstrates the average numbers of days absent by program type and seems to suggest that program type is a good candidate for predicting the number of days absent, our outcome variable, because the mean value of the outcome appears to vary by prog. The variances within each level of prog are higher than the means within each level. These are the conditional means and variances. These differences suggest that over-dispersion is present and that a Negative Binomial model would be appropriate.

Graphical Summary

```
ggplot(dat, aes(daysabs, fill = prog)) + geom_histogram(binwidth = 1) + facet_grid(prog ~
    ., margins = TRUE, scales = "free")
```



III. Fit the Models

In this report, the Negative Binomial Model and the Poisson Model are implemented due to following reasons.

Poisson regression is often used for modeling count data. Poisson regression has a number of extensions useful for count models.

Negative binomial regression can be used for over-dispersed count data, that is when the conditional variance exceeds the conditional mean. It can be considered as a generalization of Poisson regression since it has the

same mean structure as Poisson regression and it has an extra parameter to model the over-dispersion. If the conditional distribution of the outcome variable is over-dispersed, the confidence intervals for the Negative binomial regression are likely to be narrower as compared to those from a Poisson regression model.

Negative Binomial Model

```
get_prior(formula = daysabs ~ math + prog, data = dat, family = negbinomial)
```

```
##                 prior     class           coef group resp dpar nlpar bound
## 1                             b
## 2                             b           math
## 3                             b   progAcademic
## 4                             b progVocational
## 5 student_t(3, 1, 10) Intercept
## 6   gamma(0.01, 0.01)     shape
```

```
priors <- prior(normal(0, 1), class = "b", coef = "math") +
  prior(normal(0, 0.1), class = "b", coef = "progAcademic") +
  prior(normal(0, 0.1), class = "b", coef = "progVocational") +
  prior(normal(0, 1.5), class = "Intercept") +
  prior(exponential(1), class = "shape")
```

```
nb <- brm(daysabs ~ math + prog, data = dat, family = negbinomial,
          prior = priors, verbose = TRUE)
```

Poisson Model

```
po <- update(nb, family = poisson)
```

IV. Model Comparison

As mentioned before, the Poisson model is a special case of the negative binomial model as the overdispersion (shape) parameter goes to infinity. Clearly, its posterior distribution is small, indicating considerable overdispersion relative to a Poisson model.

```
nb
```

```
##  Family: negbinomial
##   Links: mu = log; shape = identity
## Formula: daysabs ~ math + prog
##    Data: dat (Number of observations: 314)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          2.21      0.15     1.92     2.50 1.00     5253     2906
## math              -0.01      0.00    -0.01    -0.00 1.00     5173     3036
## progAcademic       0.05      0.08    -0.11     0.21 1.00     4241     2783
## progVocational    -0.28      0.09    -0.45    -0.10 1.00     4533     3032
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     0.92      0.09     0.76     1.11 1.00     4403     3002
```

```
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
po
```

```
##  Family: poisson
##   Links: mu = log
## Formula: daysabs ~ math + prog
##    Data: dat (Number of observations: 314)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          2.44      0.06     2.32     2.56 1.00     3384     3321
## math              -0.01      0.00    -0.01    -0.01 1.00     4485     3379
## progAcademic      -0.17      0.05    -0.27    -0.08 1.00     2504     2049
## progVocational    -0.77      0.06    -0.89    -0.66 1.00     2361     2338
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

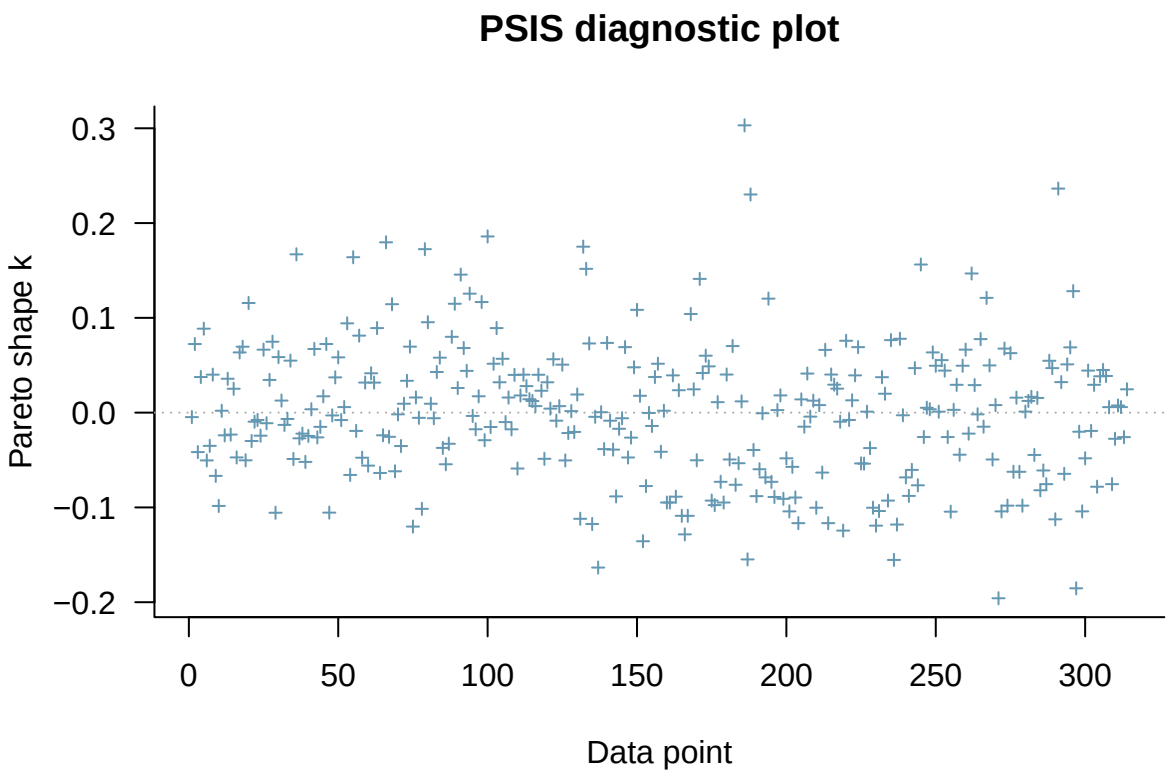
The Pareto k estimates for the negative binomial model are all fine.

```
loo(nb)
```

```
##
## Computed from 4000 by 314 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -882.3 19.6
## p_loo         3.5  0.4
## looic      1764.5 39.1
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

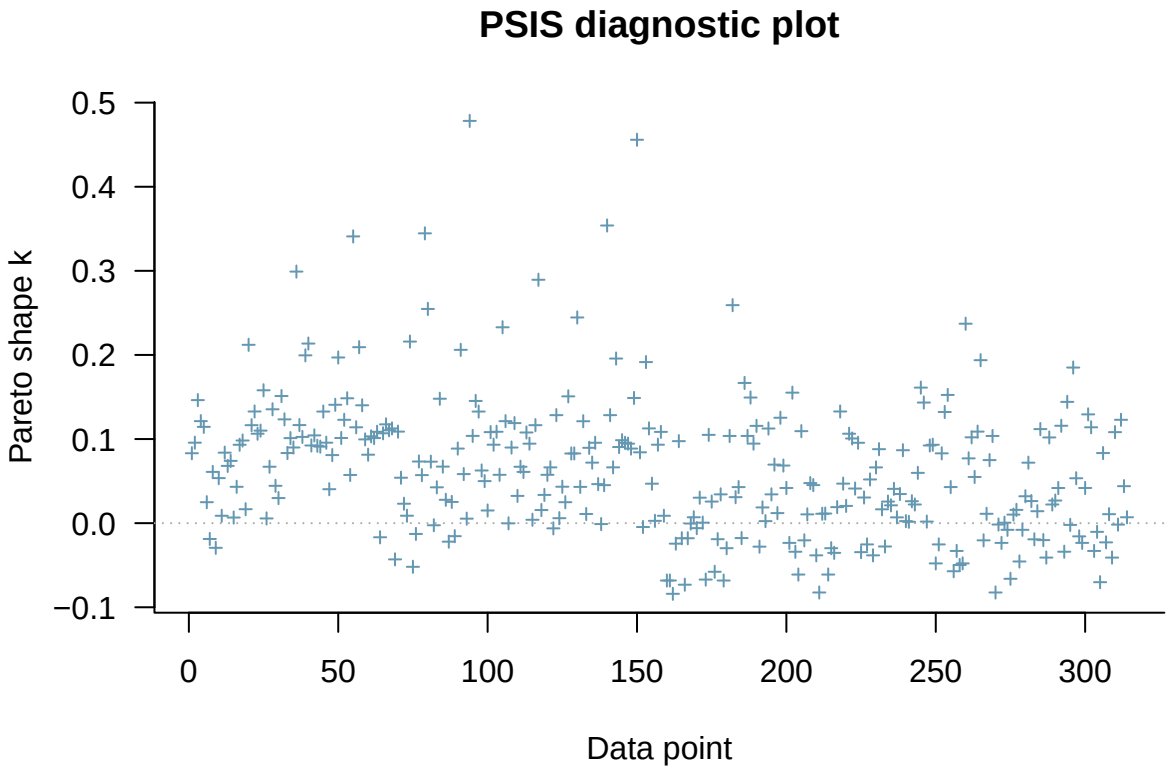
The PSIS diagnostic plot of the Negative Binomial Model is fine as well as shown below.

```
plot(loo(nb), label_points = TRUE)
```



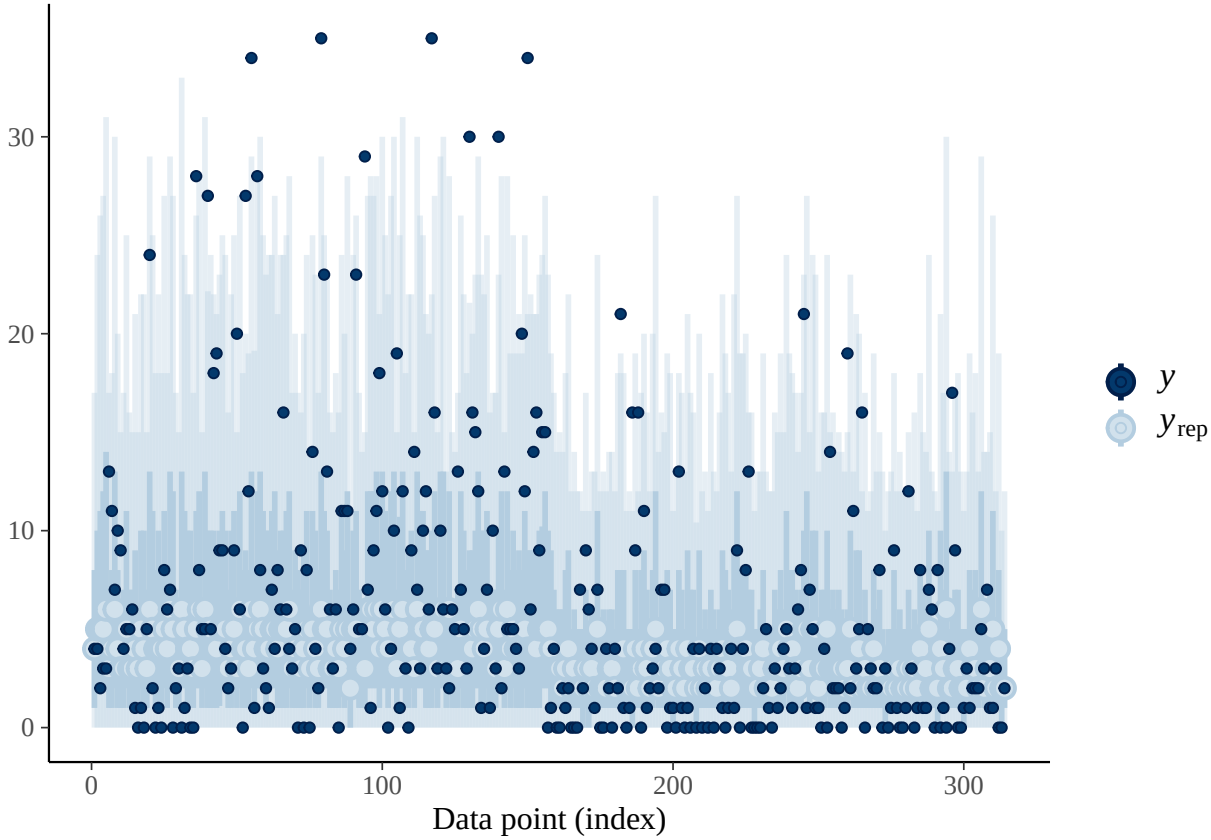
Whereas that is not true for the Poisson model, indicating that its posterior distribution is sensitive to particular observations(the one above the line in the graph below).

```
plot(loo(po), label_points = TRUE)
```



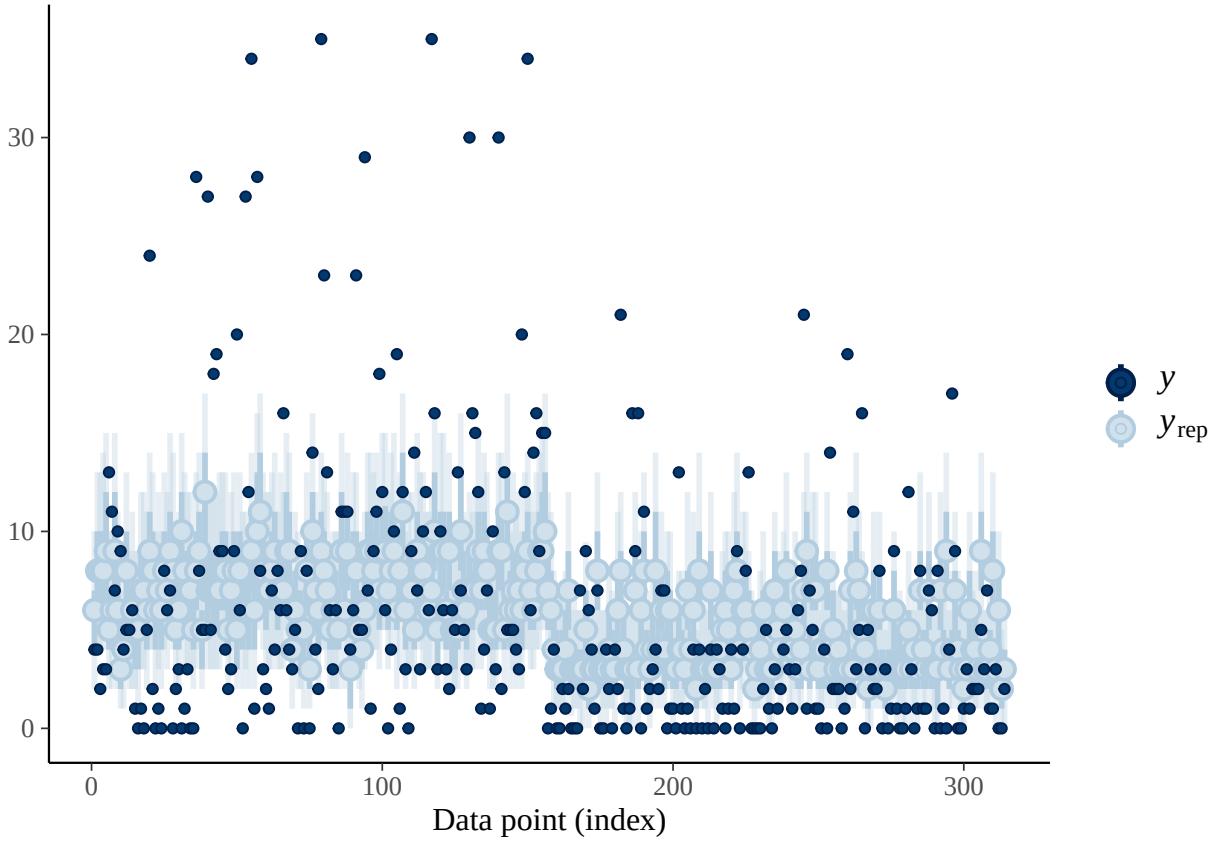
```
pp_check(nb, type = "loo_intervals") + ggplot2::scale_y_continuous()
```

```
## Using all posterior samples for ppc type 'loo_intervals' by default.
```




```
pp_check(po, type = "loo_intervals") + ggplot2::scale_y_continuous()
```

```
## Using all posterior samples for ppc type 'loo_intervals' by default.
```

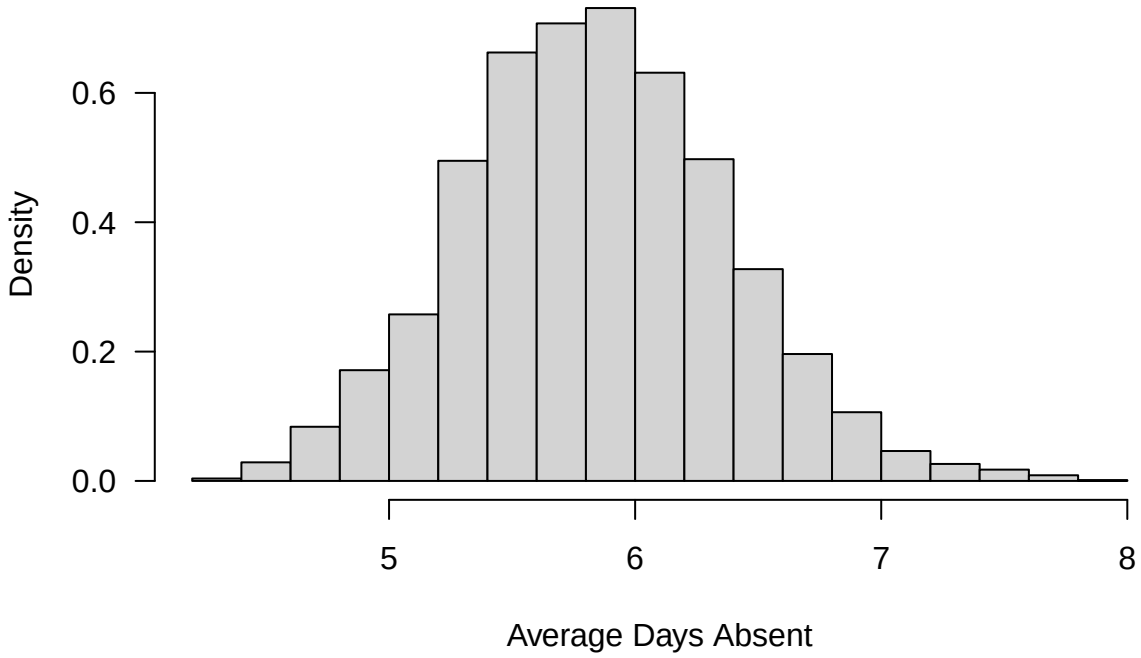


It can be seen from two graphs above, most of estimations of y lie in the confidence intervals in the Negative Binomial Model, however, the Poisson Model is way overconfident in its predictions.

V. Posterior Prediction

To describe the posterior beliefs about the students' average days absent, the graph below is derived.

```
PPD <- posterior_predict(nb, draws = 100, fun = exp)
hist(rowMeans(PPD), prob = TRUE, main = "", las = 1, xlab = "Average Days Absent")
```



VI. Conclusion

As discussed in IV section Model Comparison, the Negative Binomial Model is preferred as a model to be used to predict how a student's number of days of absence is indicated by the score in a standardized mathematics test. For further study, I am interested in implementing zero-inflated regression model and OLS regression for this data set.