Can you predict the number of Academy Awards Nominations just from IMDb Ratings???

Cedric Lam
Probability and Bayesian Statistics

INTRODUCTION

As a child I used to get really excited at watching the Academy Awards on TV. Around November each year I usually make guesses on which films will be nominated and get them wrong most of the time because the Academy usually nominates films that most people haven't seen before…

I figured this project would be a good chance for me to see if I could get better at doing this. IMDb (Internet Movie Database) is a very accessible website so I downloaded a dataset of almost 100,000 film ratings in order to make the (very rough) connection between **audience reception** and **critical acclaim**. I will cover both the Frequentist and Bayesian approaches in doing so.

METHODOLOGY

It took some time to get the dataset into a decent shape (a lot of merging, text cleaning, and filtering). After doing some exploration, I realized that the nominations are not normally distributed. It is heavily right-skewed (Fig. 1), as with most count-based distributions. Thus I had to use Poisson Regression instead. This is a slightly modified version of Regular Linear Regression, taking the log of the mean as a linear combination of the predictors (log-link function). The Bayesian approach also requires a prior belief about the likelihood of an event. I used R to simulate this belief 100 times (Fig. 3).

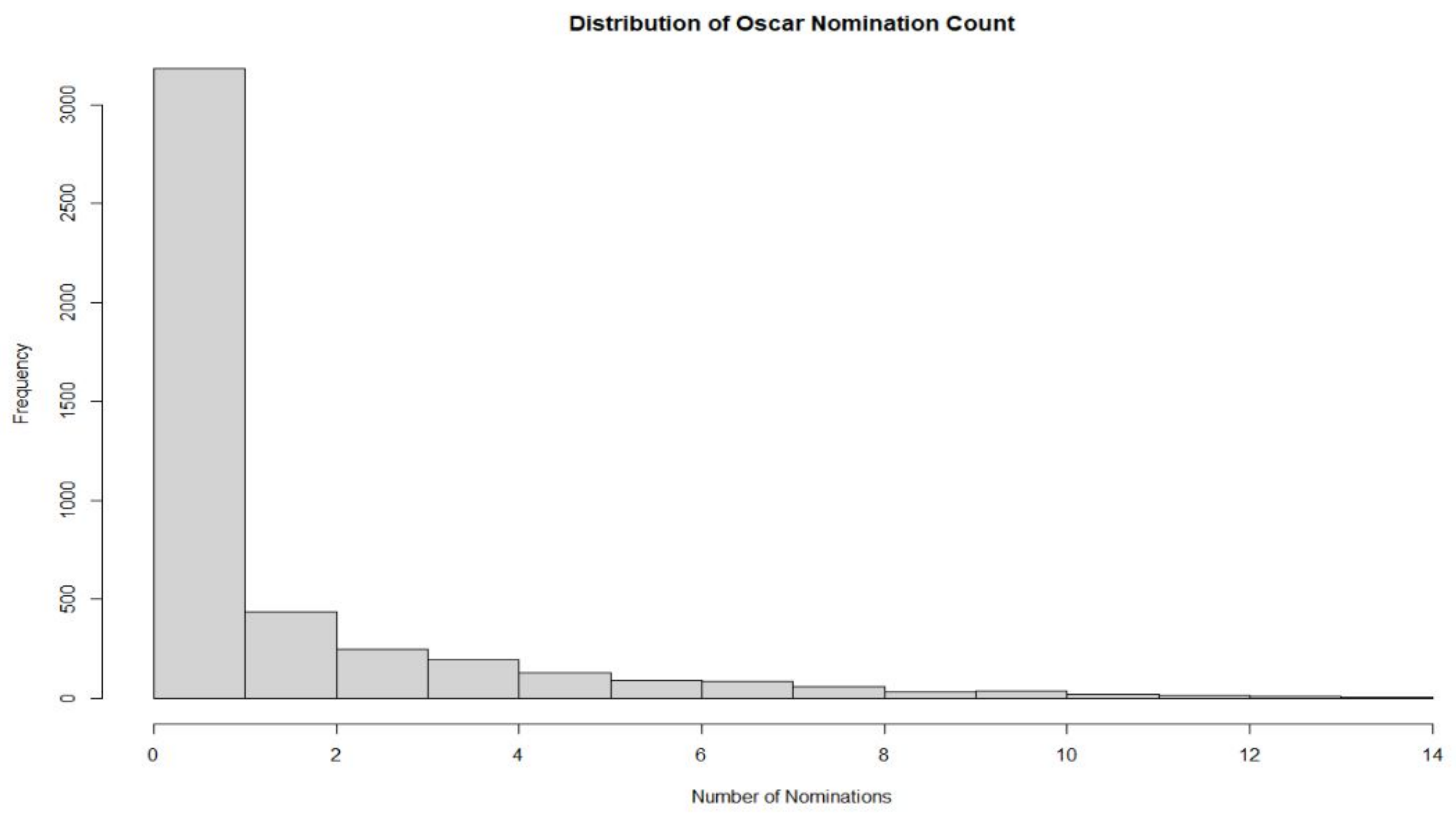
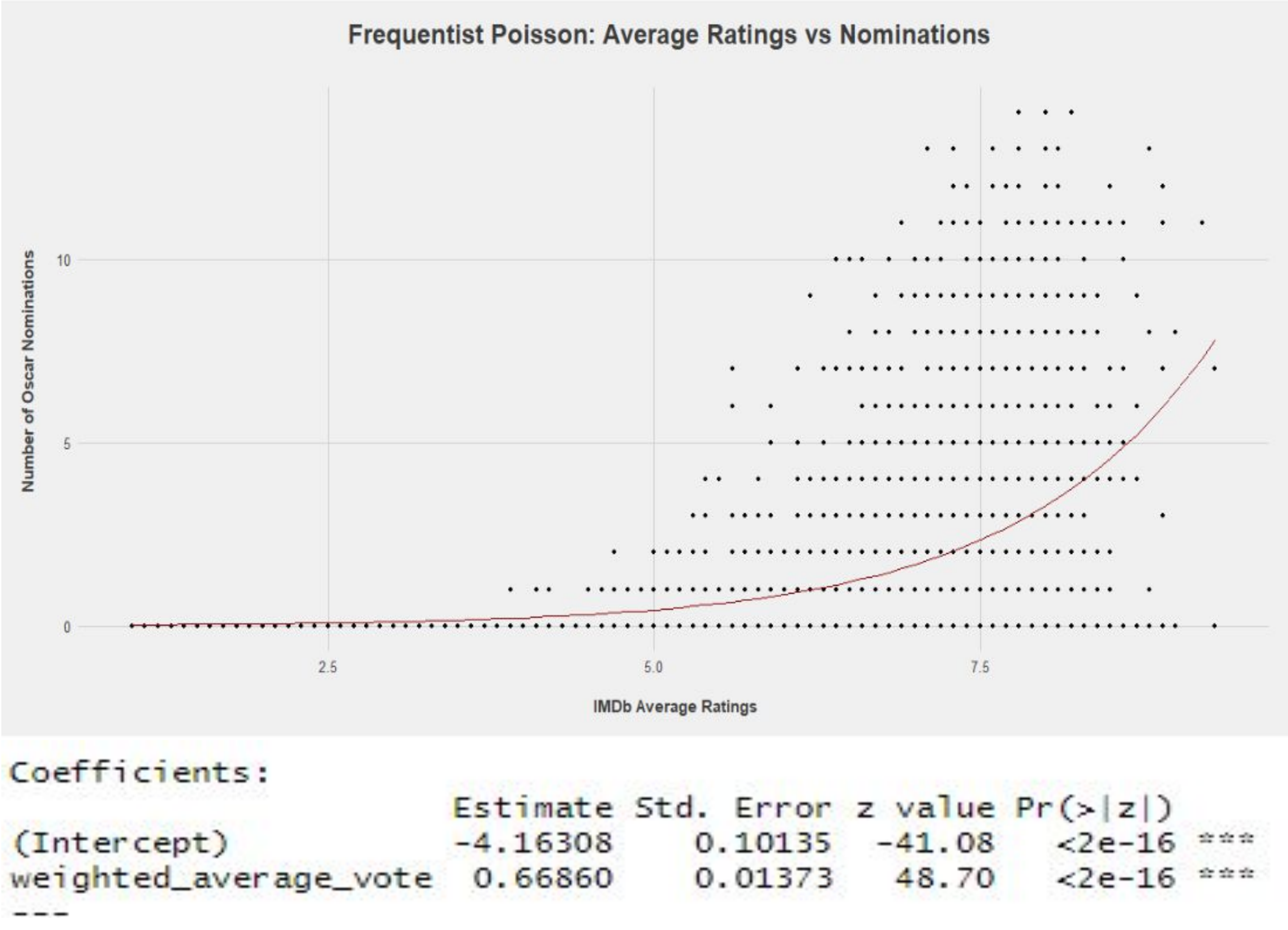


Fig. 1



```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.16308    0.10135  -41.08   <2e-16 ***
weighted_average_vote 0.66860    0.01373   48.70   <2e-16 ***
---
```

Fig. 2: Frequentist Poisson Regression

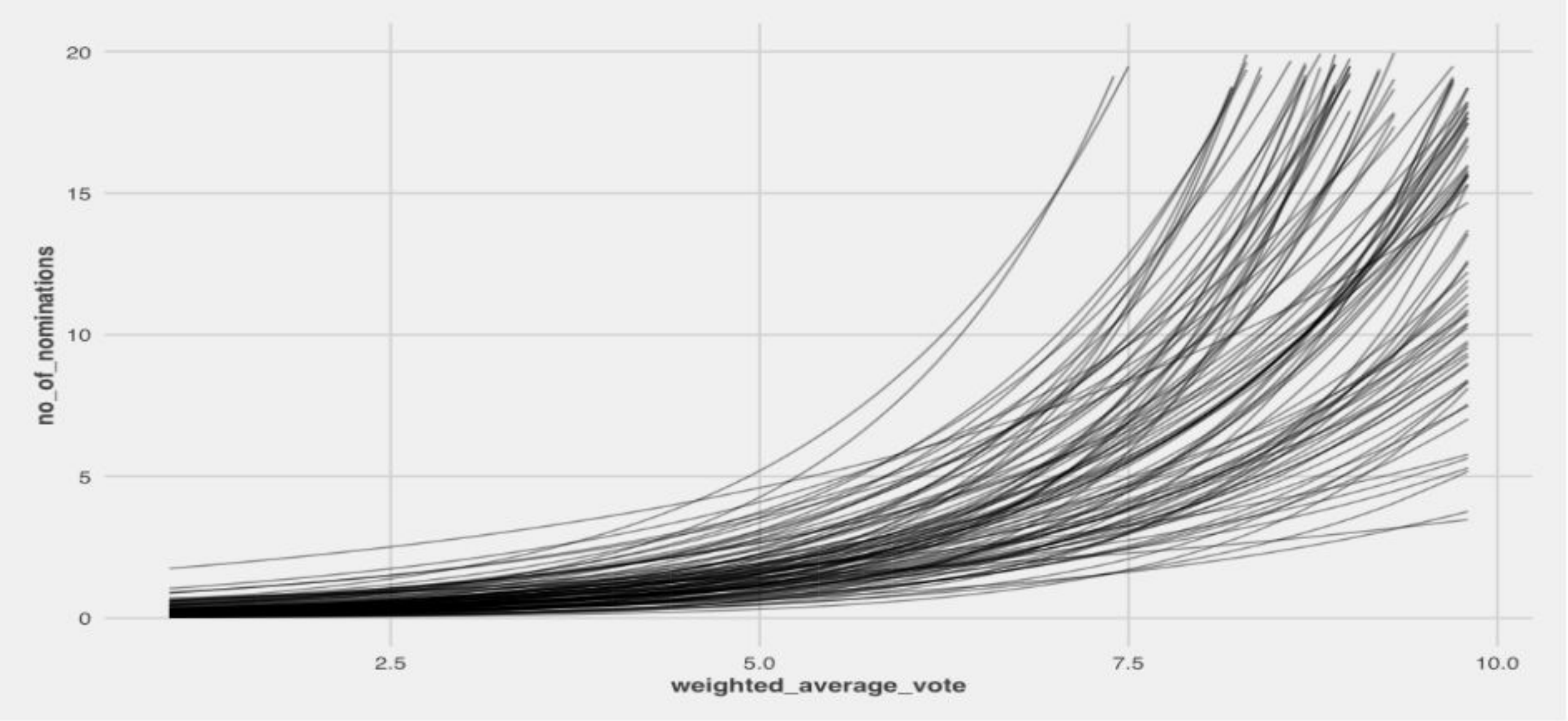


Fig. 3: 100 Simulated Priors for the Bayesian model

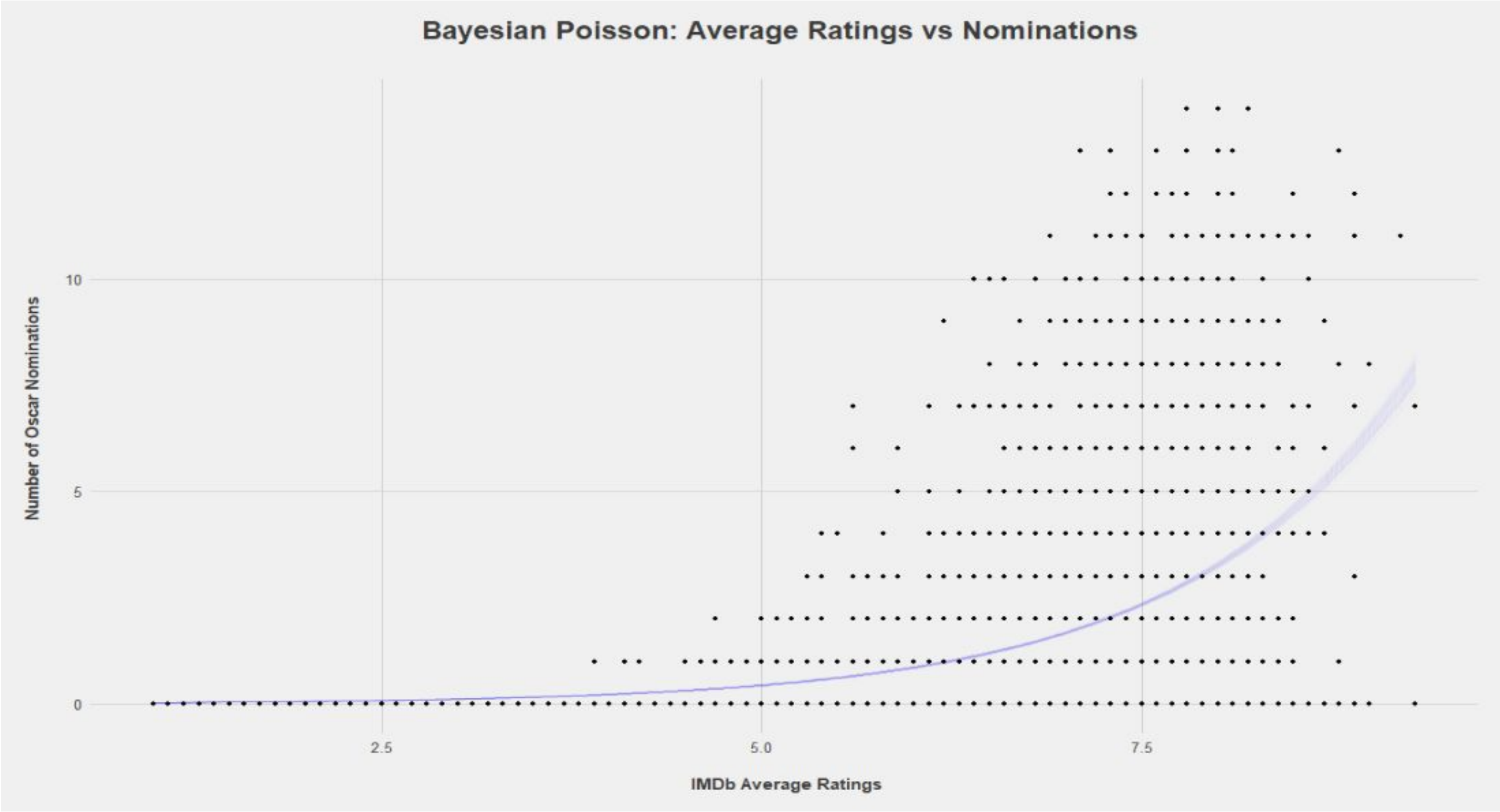


Fig. 4: Posterior Bayesian Poisson Regression using Priors from Fig. 3

RESULTS & CONCLUSION

- The frequentist test (Fig. 2) suggests that there's a statistically significant chance that the IMDb Average Rating is correlated with the number of Oscar Nominations. This is definitely no surprise, and the 95% credible interval (Fig. 4) also backs this up.

- The Bayesian approach is usually helpful in showing uncertainty and can adapt to new information quicker than the Frequentist approach that is taught in most Stats classes. In this case, the distinction between the two models isn't very clear. In fact, the fancy Bayesian simulation model took a minute and a half longer to get to a similar answer. With more complex models, the difference in computation time will be even more noticeable.

- Lastly, to answer the main question, **I think this model is a bit simplistic to be 100% useful, although it did give a good try.** I suspect previous nominations at other award shows would provide the best answer for this burning question, if such data is available online somewhere.

REFERENCES

1. This awesome book on Bayesian stats with loads of contemporary examples: https://www.bayesrulesbook.com/chapter-9.html

2. Poisson Regression tutorial: https://stats.idre.ucla.edu/r/dae/poisson-regression/

3. William Bolstad, James Curran Introduction to Bayesian Statistics (2016, Wiley)