Week 8 Lab Report

Nguyen Tien Anh Quach

2024-03-29

Histogram of the moons

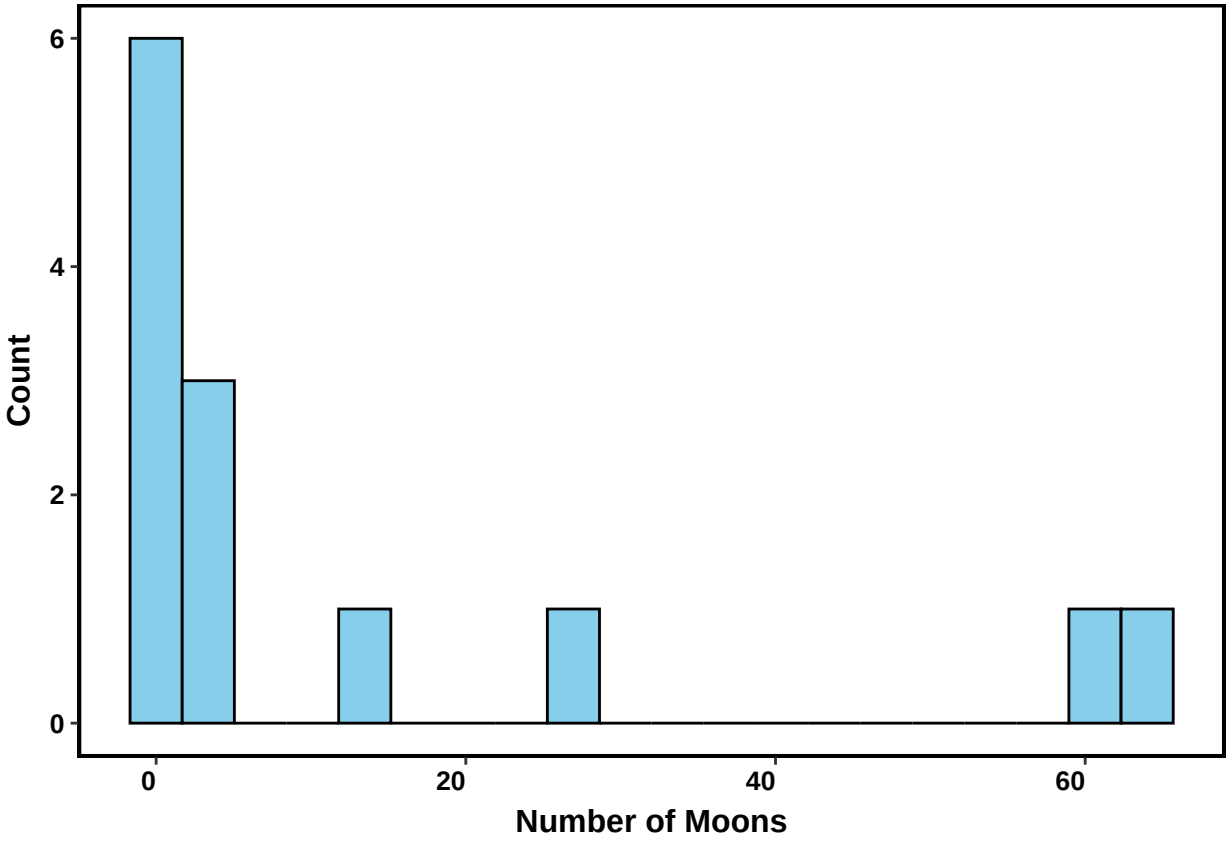

Figure 1. Histogram of the number of moons among different planets.

Linear model

I ran the linear model with Moons as the response variable and Distance, Diameter, and Mass as the predictor variables. The model is significant ($F(3, 9) = 101.8$, $p < 0.001$) with only Diameter being the significant predictor ($p < 0.001$).

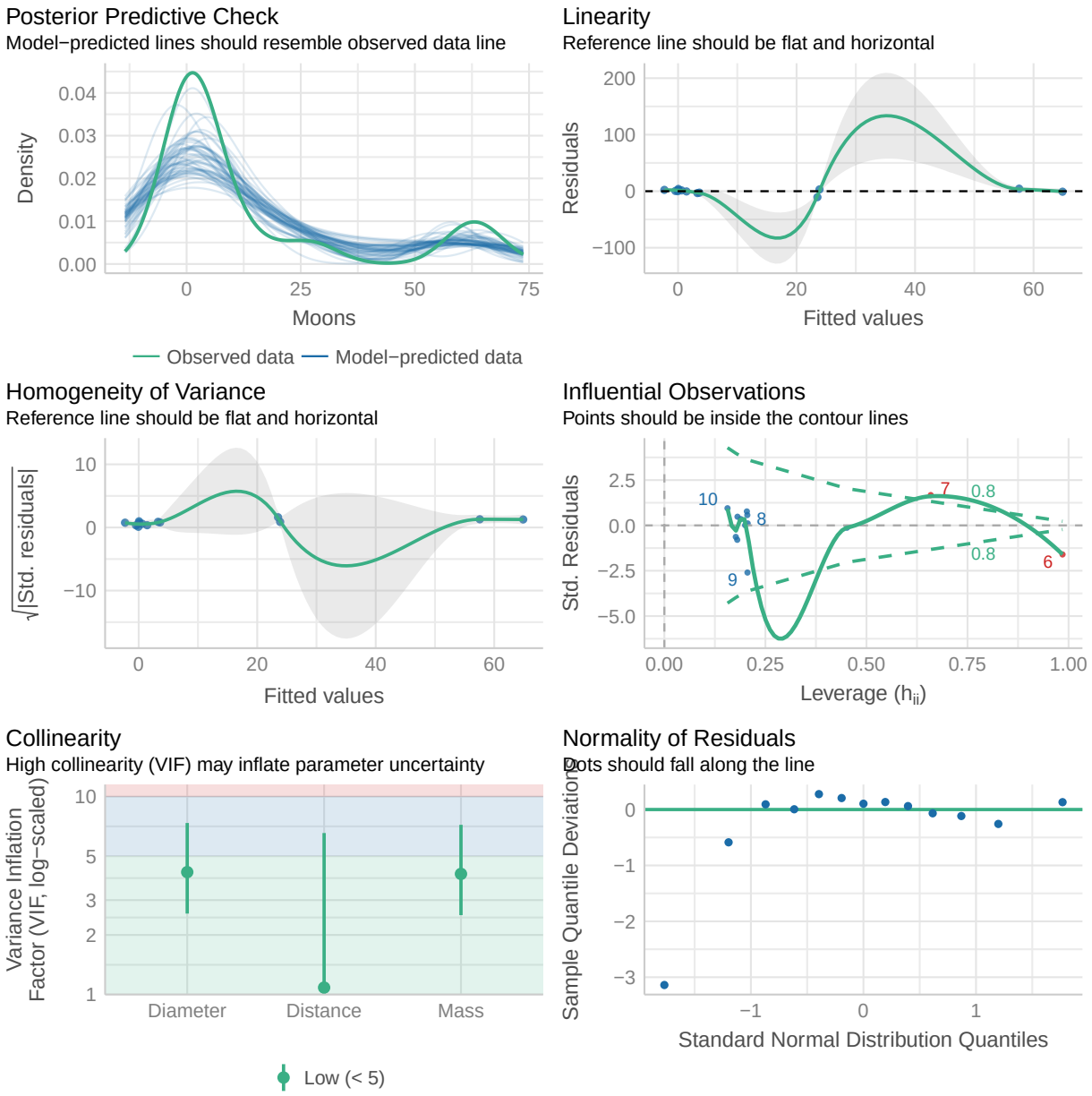

Figure 2. Model assumptions check.

As shown in Figure 2, VIFs are all below 5. Therefore, the multicollinearity assumption is met. In addition, normality of residuals is met, as points fall approximately along the line. However, the linearity and homogeneity of variance assumptions are not met as the lines are not flat and horizontal. In conclusion, with several assumptions violated, linear model may not be the best option.

Poisson model

The Poisson model showed that diameter ($p < 0.001$) and mass ($p < 0.001$) of a planet are significant predictors of the number of moons.
Holding other variables constant:
An increase of one unit in diameter is associated with an increase in the expected number of moons by 56%.
An increase of one unit in mass results in a decrease in the expected number of moons by 0.4%.

Poisson model assumptions

Multicollinearity

All VIFs are below 5, indicating that the multicollinearity assumption is met!

Dispersion

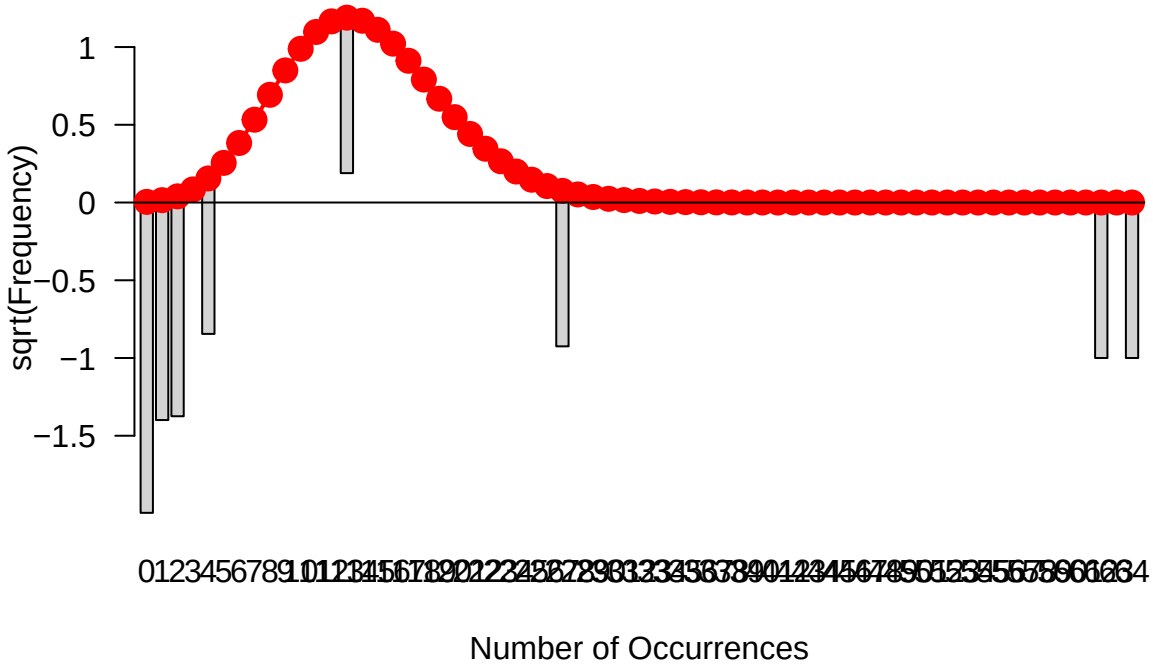

Figure 3. Poisson model rootogram.

As shown in the rootogram (Fig. 3), most bars hang below 0, indicating underfitting. I also calculated the dispersion statistic and observed that the value is 4.66, which means that a different model may be needed.

Distribution

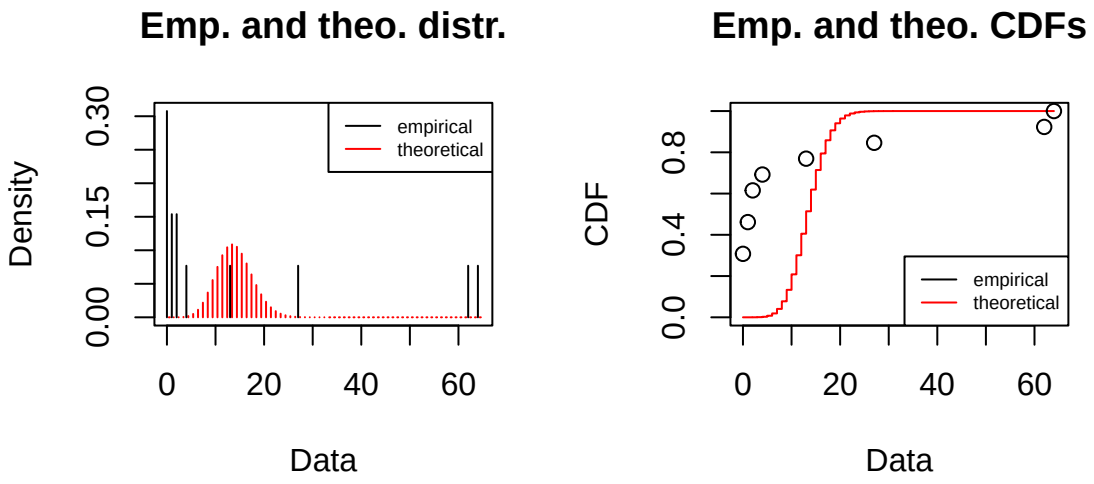

Figure 4. Empirical vs. Theoretical distribution.

From Figure 4, it is quite clear that the Poisson model is not fitting the distribution of the data very well.

Linearity

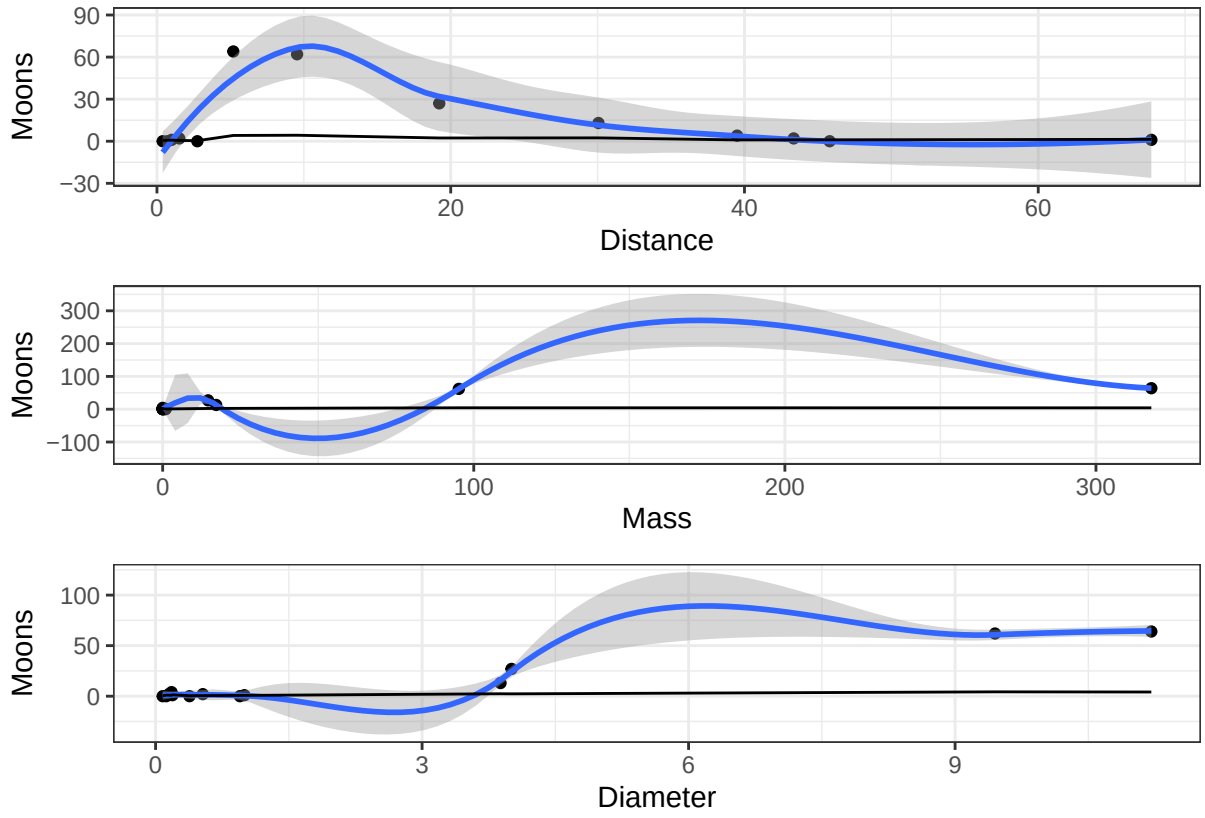

Figure 5. Linear assumption check.

From Figure 5, it is clear that none of the predictor variables has a linear relationship with the response variable.

Evaluate Poisson model fit

Likelihood ratio test

The likelihood ratio test yielded significant results ($p < 0.001$). Therefore, including the predictor variables significantly increased the likelihood of the data.

Goodness of fit test

The goodness of fit test yielded significant p-value ($p < 0.001$). Therefore, the null hypothesis is rejected and the model did not fit the data well.

Negative binomial model

The negative binomial model showed similar results, that diameter ($p < 0.001$) and mass ($p = 0.027$) of a planet are significant predictors of the number of moons.
Holding other variables constant:
An increase of one unit in diameter is associated with an increase in the expected number of moons by 92%.
An increase of one unit in mass results in a decrease in the expected number of moons by 1%.

Dispersion

The dispersion statistic for the negative binomial model is now 1.32, which is much closer to 1 than the Poisson model. Several possible explanations why negative binomial is now a better model:
1. This dataset contains count data => Variance in this dataset is different from mean. 2. Negative binomial distribution introduces another parameter, k - dispersion parameter, that is in addition to lambda => this gives the model more flexibility.

Likelihood ratio test

Likelihood ratio test yielded significant p-value ($p < 0.001$). Therefore, there is a significant difference between the Poisson and negative binomial models.
The negative binomial model (-30.98) has higher log likelihood than the Poisson model (-37.95). Thus, it is safe to conclude that using negative binomial distribution improved the fit of the model (Fig. 6).

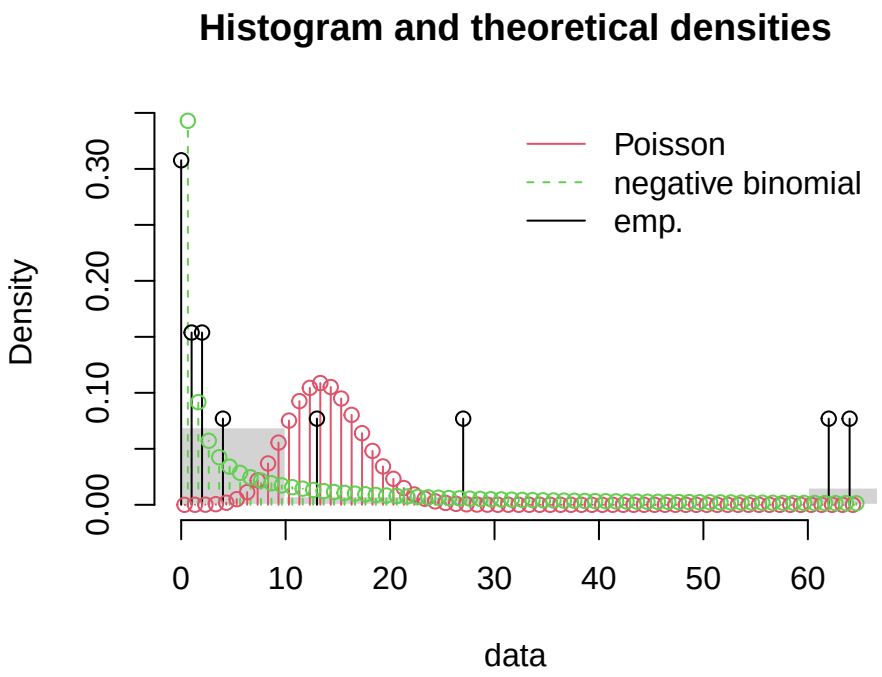

Figure 6. Negative binomial and Poisson distribution model comparison.

Rootogram

The rootogram of the negative binomial distribution (Fig. 7) seems to fit the data so much better than the one of Poisson distribution (Fig. 3). Despite not being the best model, less numbers of bars are hanging below 0, indicating better prediction because most observations in Figure 3 (i.e., Poisson model) were shown to be underpredicted.

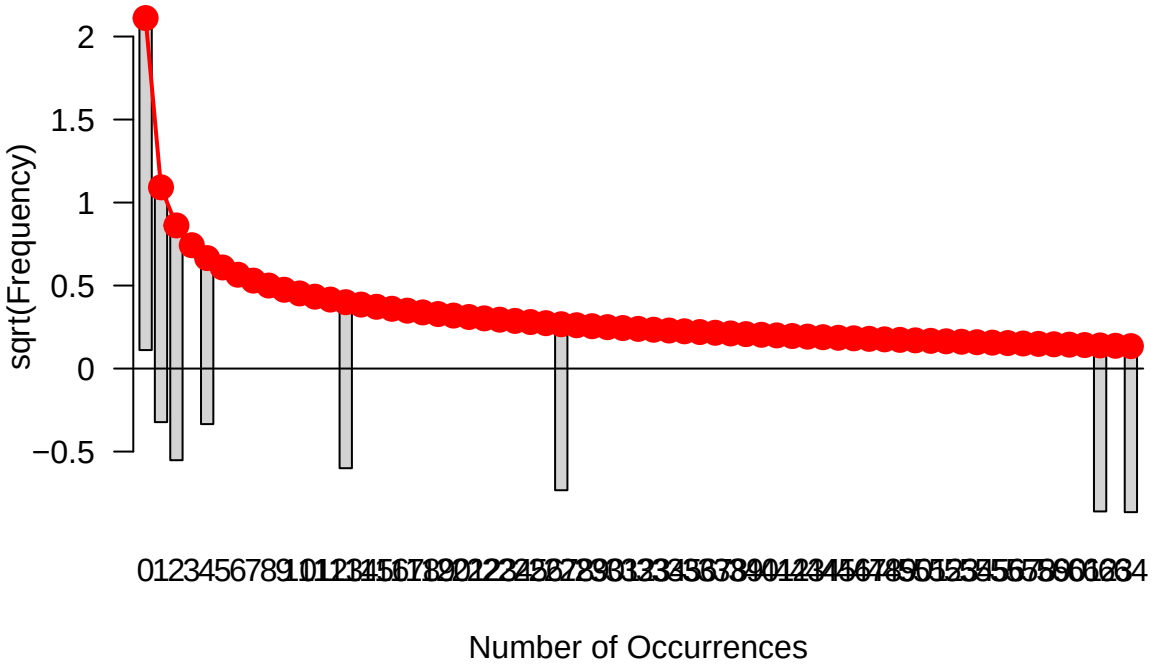

Figure 7. Negative binomial model rootogram.

Expected moons for Jupiter

The model equation is:

$$log(expectedmoons) = -0.426 + 0.020 \times Distance + 0.652 \times Diameter - 0.009 \times Mass$$

Therefore, the number of expected moons for Jupiter is: 62

Zero-inflated model

Because negative binomial distribution fit the data much better than Poisson distribution, I am using the zero-inflated negative binomial model for this part.

Zero-inflated negative binomial model

The zero-inflated negative binomial model showed similar results, that diameter ($p < 0.001$) and mass ($p < 0.001$) of a planet are significant predictors of the number of moons.
Holding other variables constant:
An increase of one unit in diameter is associated with an increase in the expected number of moons by 43%.
An increase of one unit in mass results in a decrease in the expected number of moons by 1%.

LR test and comparison of AIC

The negative binomial model is better because is has lower AIC (71.9), compared to the zero-inflated negative binomial model, with AIC of 80.7. Also, the log likehood of the NB model is -30.98, higher than the ZINB

model (log likehood = -32.35). Thus, it is safe to conclude that using negative binomial distribution fit the data better than the ZINB model.