# 20250716_Abner

2025-07-16

## Project Summary

- We are using data from the American Society of Microbiology's (ASM) 12 published journals to investigate the relationship between the number of citations (variable 'is.referenced.by.count') a published scientific article receives and if the authors have included access to their raw sequencing data (variable 'da', data availability) in the manuscript.
- We are trying to understand if publishing raw data helps to improve citation metrics. We have data from 2000-2024, and will also adjust for time published (variable 'age.in.months'), as older papers have had the opportunity to accumulate more citations over time.

## Model Format

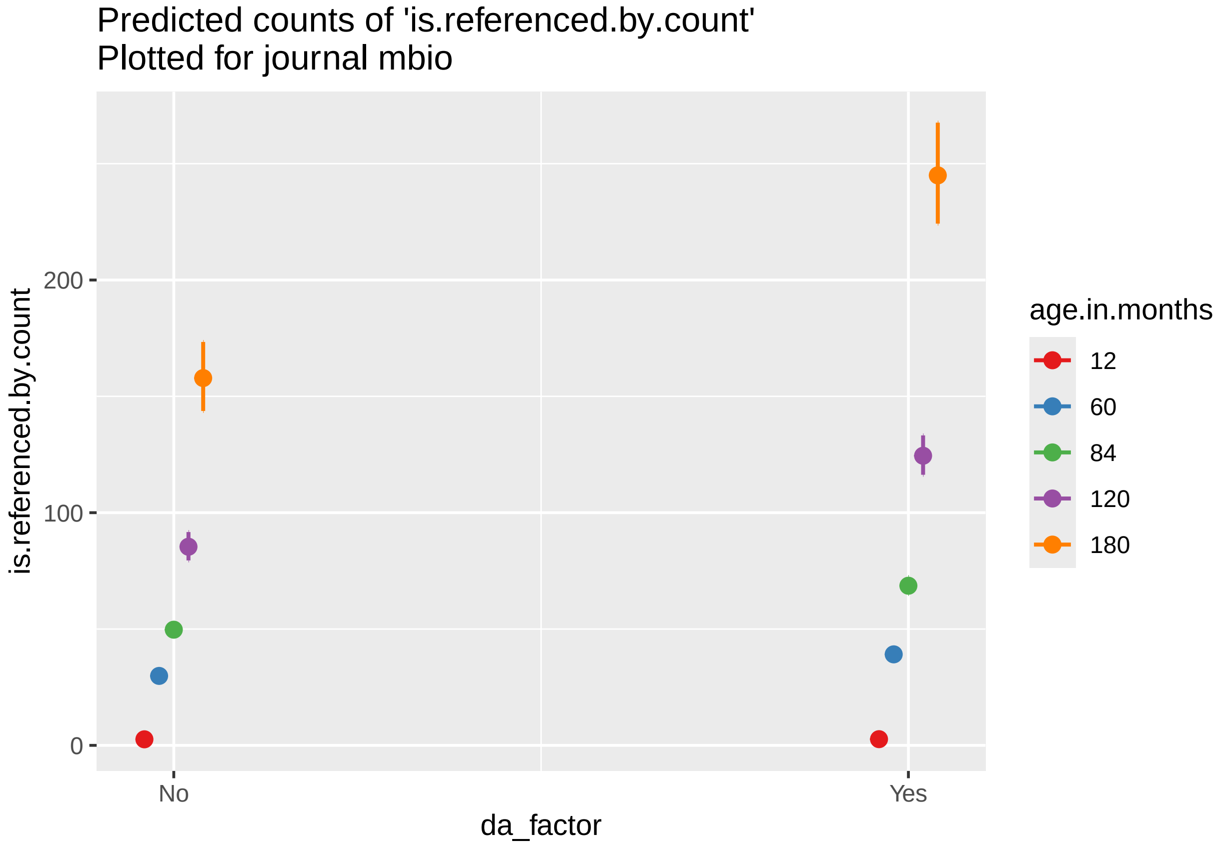
- Use model format for data from each journal:
    - MASS::glm.nb(is.referenced.by.count~ da_factor + log(age.in.months) + log(age.in.months)*da_factor, data = <each journal>, link = log)
    - Data with N > 10 to create model
    - is.referenced.by.count = Number of citations received as of 2/1/25
    - age.in.months = age of paper in months
    - da_factor = Is raw sequencing data available? Yes/No as a factor.

## Journal Case Study: mBio

- mBio is an ASM family journal that has been around ~15 years (2015-2025)
- See below for the plot of the model, the simulated residuals from DHARMa, and the residual plots.

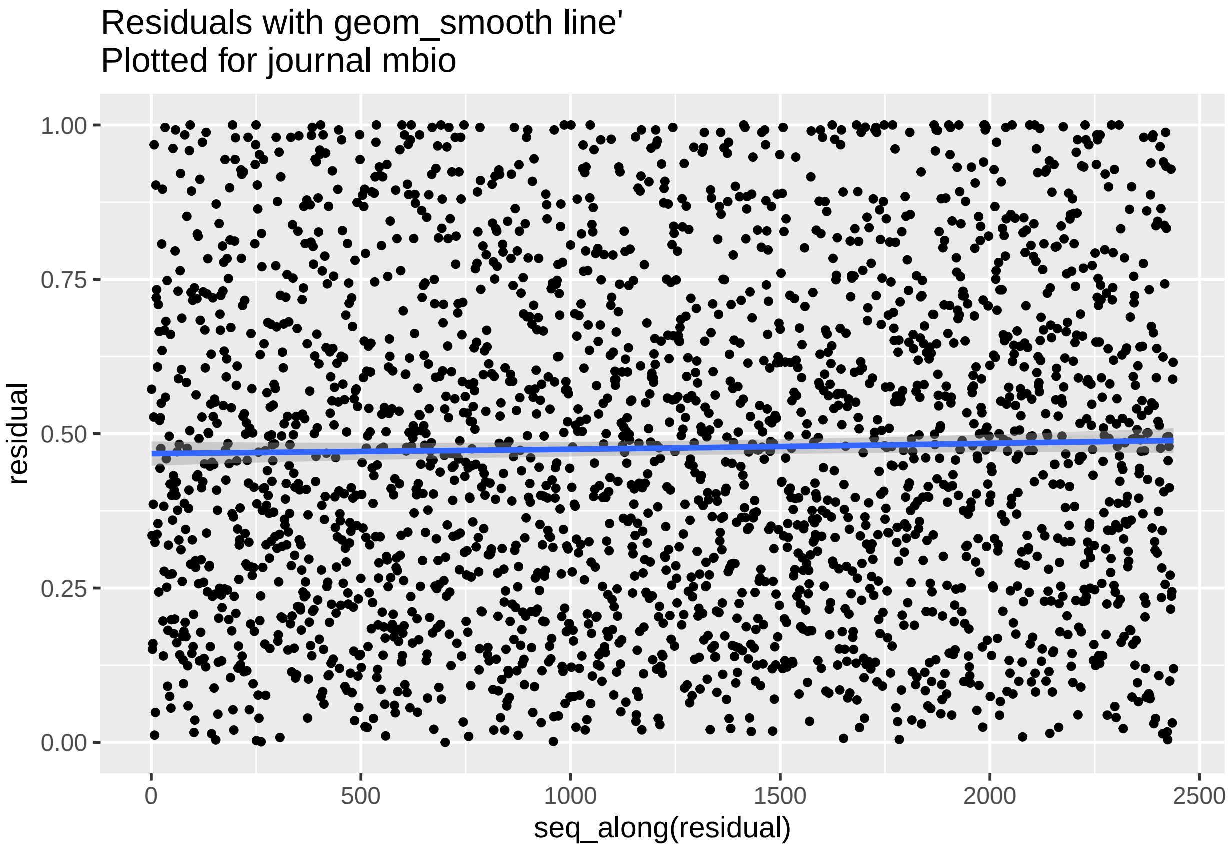
```r
#smaller model with 2 terms
two_term_glmnb <-function(model_data, model_name) {

  total_model <-MASS::glm.nb(is.referenced.by.count~ da_factor + log(age.in.months) +
       + log(age.in.months)*da_factor + log(age.in.months)*da_factor, data = model_data, link = log)

  return(total_model)

}

journals <-
  nsd_yes_metadata %>%
  count(journal_abrev) %>%
  filter(journal_abrev != "jmbe")
```

```
j <- 8 #mbio


  journal_data <-
  nsd_yes_metadata %>%
    filter(journal_abrev == journals[[j,1]]) %>%
    mutate(da_factor = factor(da))


  model <- two_term_glmnb(journal_data, journals[[j,1]])


  plot_model <- plot_model(model, type = "pred", terms = c("da_factor", "age.in.months[12,60,84,120,180]

 print(plot_model)
```



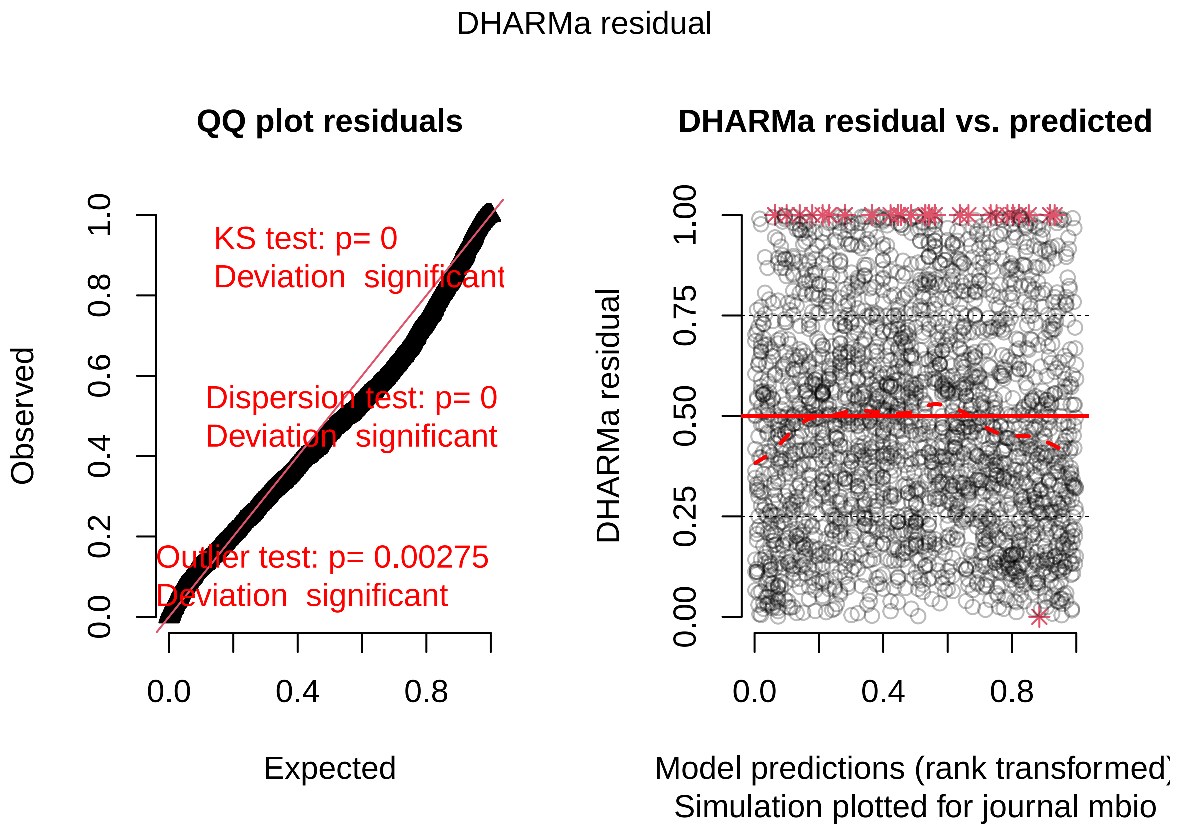
```
  simulation <- simulateResiduals(fittedModel = model, plot = F)

  # residuals(simulation)  %>%
  # plot(main = paste0("Residuals plotted for journal ", journals[[j,1]]), pch = ".")
  residuals <-
   residuals(simulation) %>%  tibble(residual = .) %>%
     ggplot(aes(y = residual, x = seq_along(residual) )) +
     geom_point(size = 1) +
     geom_smooth() +
    ggtitle(paste0("Residuals with geom_smooth line' \nPlotted for journal ", journals[[j,1]]))
  print(residuals)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
  plot(simulation, sub = paste0("Simulation plotted for journal ", journals[[j,1]]))
```

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```



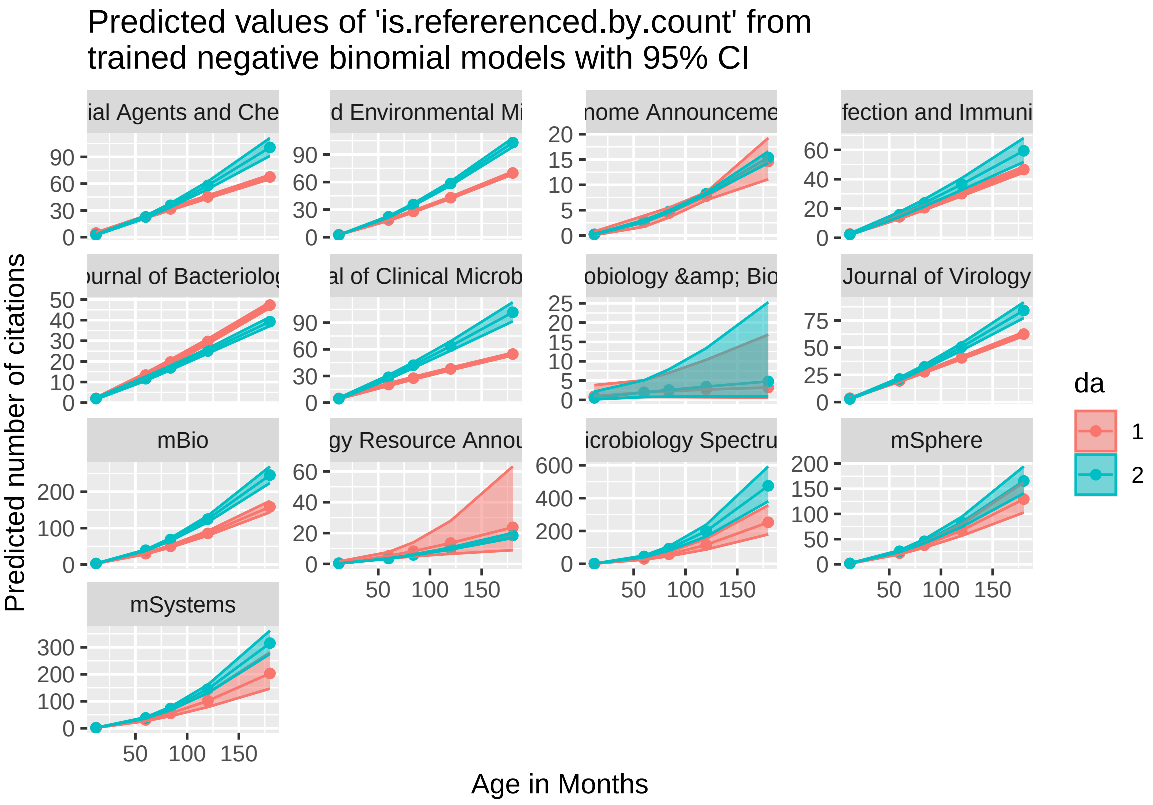
## Questions

- This is one example from our 12 journal set, if all of the plots look roughly like this, we are assuming the models fit well and can proceed with using them.
- We've looked at the residuals 3 ways using DHARMa.
  - When using the QQ plot, should we be worried about the significance of the three tests?
  - Does this occur because there are a large number of observations in the dataset (N = 2438)?
  - Should we be interpreting this differently?
- In the plotted model data, we see that the slope increases with time, and there are confidence intervals on these estimates.
  - Is this enough to say that these data are different?
  - Should we perform more hypothesis testing such as a T test?
  - Could we get a p-value to satisfy reviewers? How might we do that?

## Plotting Residual Histogram

- 20250716 - Abner advises:
  - AHB "In your plot"Predicted counts of 'is.referenced.by.count' " it is hard to see the differences between papers that share data and papers that don't. I suggest that you make a line plot with "age" in the horizontal axis, "predicted citations" in the vertical axis, color the lines by "da_factor", and put each journal in a separate box (i.e., facet by journal). This plot would emphasize the differences between sharing and not sharing data for each age value."

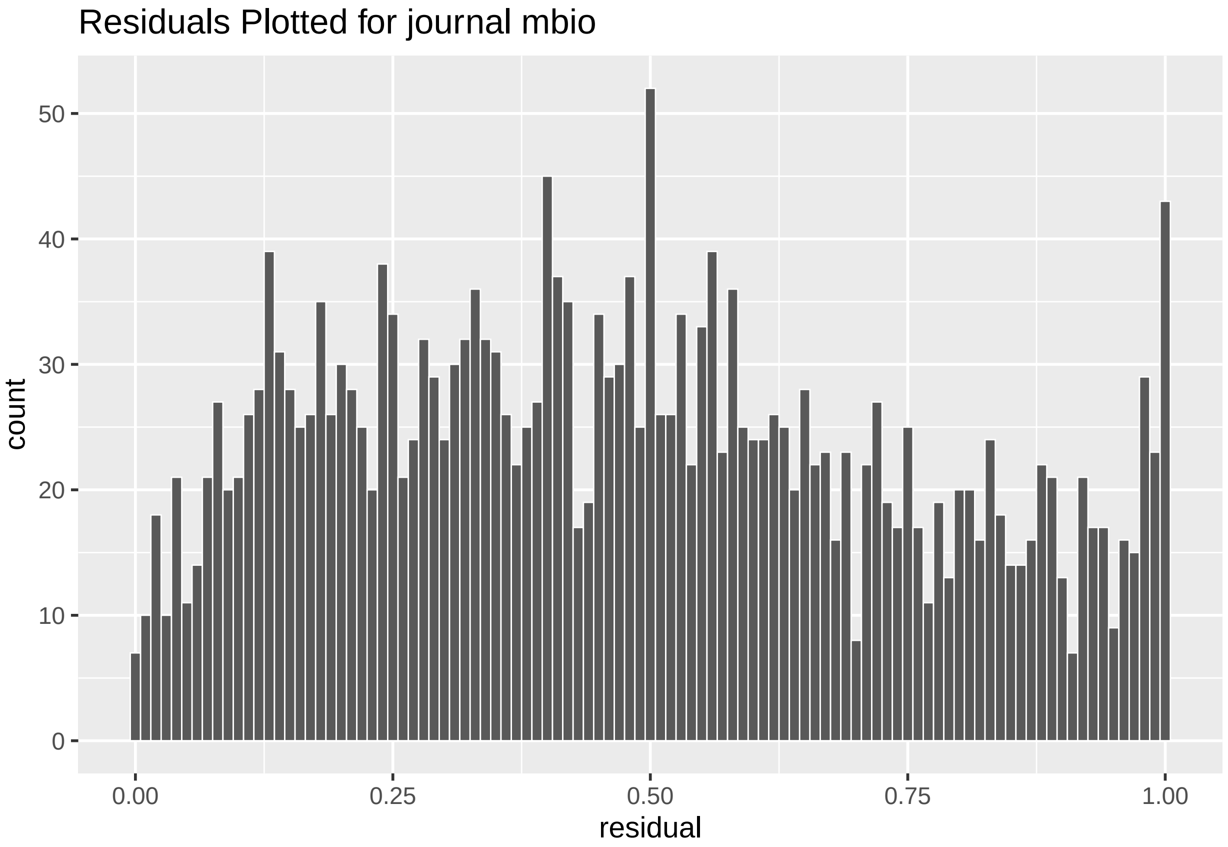
- This plot is for the negative binomial model from all journals.



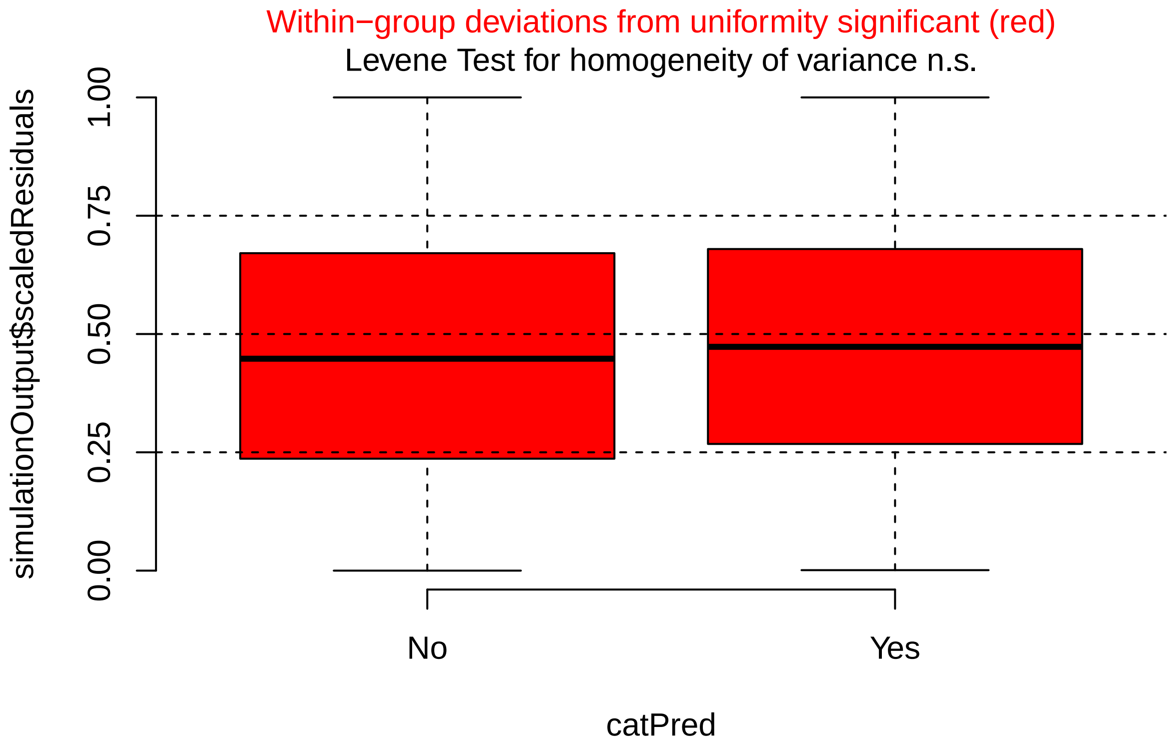
  - AHB “The QQ-plot of the residuals does show a pattern that suggests that your residuals have a distribution that leans to the right. I suggest you check this with a histogram of the residuals. The Residuals vs Predicted plot also hints at an inverted u-shape. In combination, these diagnostic plots make me suspect that there is a non-linear relationship between your current covariates and your outcome of interest. The easiest way to check is to make plots of Residuals vs Covariate for every covariate that you include in the model. DHARMa shows an example of these plots.”

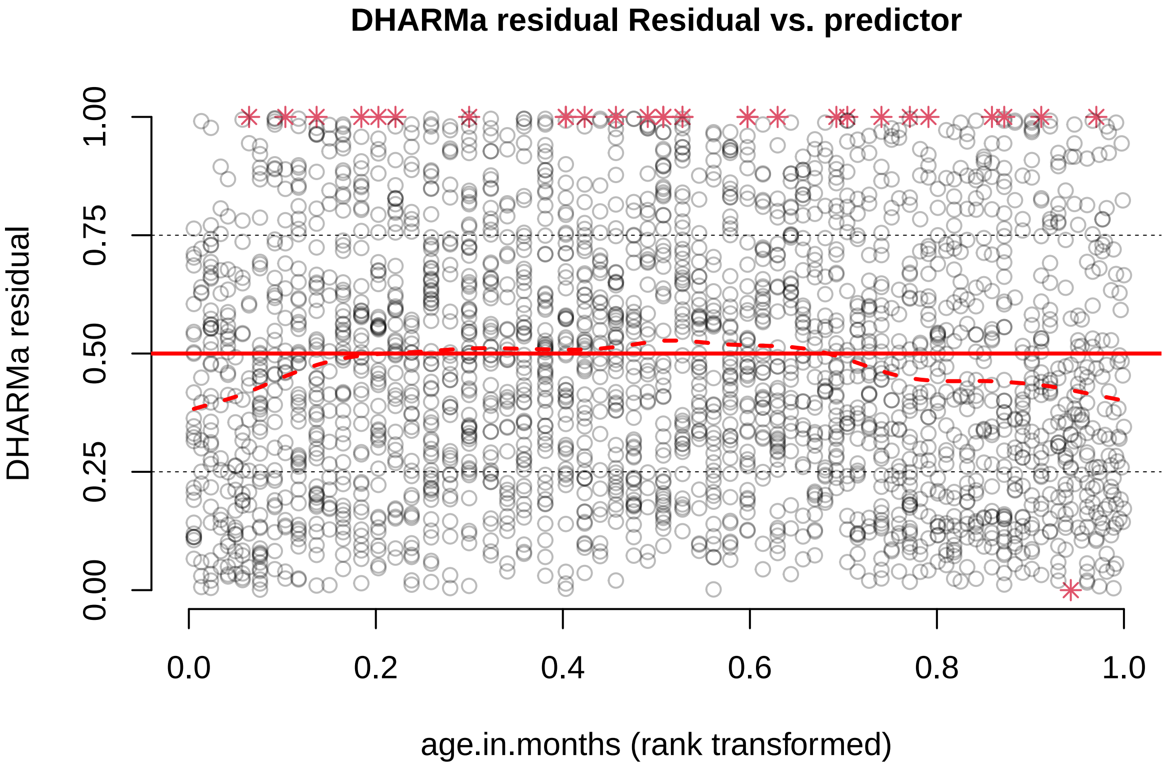
- See below for the histogram of the residuals.

- I have also plotted the residuals for each of the covariates for this model.



Residuals Plotted for journal mbio


Within−group deviations from uniformity significant (red)
Levene Test for homogeneity of variance n.s.
simulationOutput$scaledResiduals
0.00
0.25
0.50
0.75
1.00
No
Yes
catPred



- AHB “I am still not sure of the specific hypothesis you want to test. As I explained before, your current model does not have a unique parameter to compare”data vs no data”. This happens in part because you still have not decided how to represent the different journals into a single group of “data vs no data” “.
  - We would like to model the journals separately as we think it makes the differences between “data vs no data” clearer. We included only one journal model previously for simplicity as there are 12 journals and each journal had similar graphs available.
  - The parameter we would like to compare is the number of citations a paper has received. Please advise on how to continue with this residual distribution.
  - We have set a meeting for Tuesday 7/22 at 2pm.