Chapter 12 Generalized Linear Models

Angelo LaCommare-Soto

2025-04-23

Rodent Diversity and Infected Tick Density

Research Question 1: Does rodent richness predict the density of *Borrelia*-infected ticks?

Section 1 - Importing Data

```r
# set working directory for all chunks in this file (default working directory is wherever Rmd file is)
getwd()
```

```
## [1] "C:/Users/Angelo L/Documents/GitHub/BIOL710/RCode710/RCode/working_directory"
```

```r
library(tidyverse)
```

```
## Warning: package 'purrr' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# importing the gene dataset
tick <- read.csv("tick.csv",header=TRUE,stringsAsFactors = TRUE)
str(tick)
```

```
## 'data.frame':    46 obs. of  58 variables:
##  $ Area_ha              : num  1193 1193 1193 1193 151123 ...
##  $ Log_Area             : num  3.08 3.08 3.08 3.08 5.18 5.18 5.18 2.27 2.27 2.27 ...
##  $ Perim_m              : num  47849 47849 47849 47849 339044 ...
##  $ Log_perim            : num  4.68 4.68 4.68 4.68 5.53 5.53 5.53 4.23 4.23 4.23 ...
##  $ PA_ratio             : num  40.1 40.1 40.1 40.1 2.2 2.2 2.2 91.2 91.2 91.2 ...
##  $ Euc_source_m         : num  4701 4701 4701 4701 0 ...
##  $ Log_Euc_source_m     : num  3.67 3.67 3.67 3.67 0 0 0 0 0 0 ...
##  $ Euc_source_km          : num  4.7 4.7 4.7 4.7 0 0 0 0 0 0 ...
##  $ Euc_forest_m           : num  7371 7371 7371 7371 0 ...
##  $ Log_Euc_forest_m       : num  3.87 3.87 3.87 3.87 0 0 0 2.89 2.89 2.89 ...
##  $ Euc_forest_km          : num  7.4 7.4 7.4 7.4 0 0 0 0.8 0.8 0.8 ...
##  $ CD_comm_source         : num  13919 13919 13919 13919 0 ...
##  $ Log_CD_comm_source     : num  4.14 4.14 4.14 4.14 0 0 0 0 0 0 ...
##  $ CD_comm_forest         : num  20769 20769 20769 20769 0 ...
##  $ Log_CD_comm_forest     : num  4.32 4.32 4.32 4.32 0 0 0 3.23 3.23 3.23 ...
##  $ CD_pred_source         : int  7932 7932 7932 7932 0 0 0 0 0 0 ...
##  $ Log_CD_pred_source     : num  3.9 3.9 3.9 3.9 0 0 0 0 0 0 ...
##  $ CD_pred_forest         : int  13389 13389 13389 13389 0 0 0 1311 1311 1311 ...
##  $ Log_CD_pred_forest     : num  4.13 4.13 4.13 4.13 0 0 0 3.12 3.12 3.12 ...
##  $ CD_deer_source         : int  17998 17998 17998 17998 0 0 0 0 0 0 ...
##  $ Log_CD_deer_source     : num  4.3 4.3 4.3 4.3 0 0 0 0 0 0 ...
##  $ CD_deer_forest         : int  27382 27382 27382 27382 0 0 0 2216 2216 2216 ...
##  $ Log_CD_deer_forest     : num  4.4 4.4 4.4 4.4 0 0 0 3.3 3.3 3.3 ...
##  $ Natural_1km            : num  3.14 3.14 3.14 3.14 3.03 3.03 3.03 2.13 2.13 2.13 ...
##  $ Patch_proximity        : logi  NA NA NA NA NA NA ...
##  $ Site                   : Factor w/ 14 levels "CCSP","FL","HOS",..: 1 1 1 1 2 2 2 3 3 3 ...
##  $ Year                   : int  2016 2018 2019 2021 2018 2019 2021 2016 2018 2019 ...
##  $ Rodents_Present        : Factor w/ 22 levels "MICA","MICA, NEFU, PECA, PEMA, PETR",..: 18 20 18 3
##  $ Rodent_Rich            : int  3 5 3 4 4 3 5 3 3 2 ...
##  $ Rodent_Shannon         : num  0.726 1.164 0.68 0.116 1.01 ...
##  $ NEFUwPERO_abund        : int  32 34 44 33 89 44 36 43 29 35 ...
##  $ Pero_abund             : int  9 33 32 25 86 42 28 32 25 13 ...
##  $ Nefu_abund             : int  23 1 12 8 3 2 8 11 4 22 ...
##  $ Wildlife_Richness      : int  4 6 5 7 NA 1 1 2 7 4 ...
##  $ Wildlife_Present       : Factor w/ 41 levels "","Cow, Skunk, Turkey, Squirrel",..: 38 26 15 16 1
##  $ Wildlife_Shannon       : num  0.695 1.381 0.811 1.324 NA ...
##  $ Predator_Richness      : int  0 2 1 3 NA 0 0 1 4 2 ...
##  $ Predators_Present      : Factor w/ 32 levels "","Bobcat","Bobcat, Coyote",..: 1 9 16 19 1 1 1 25
##  $ Predator_Shannon_NoCat : num  0 0.683 0 1.099 NA ...
##  $ Predator_Shannon_wCat  : num  0 0.683 0 1.099 NA ...
##  $ AllMammal_Richness     : int  7 11 8 11 NA 4 6 5 10 6 ...
##  $ DOT                    : int  253 747 478 114 183 103 202 96 249 276 ...
##  $ DOL                    : int  199 633 308 9 140 45 129 89 229 266 ...
##  $ DOA                    : int  12 6 3 1 8 0 8 0 1 2 ...
##  $ DON                    : int  42 108 167 104 35 58 65 7 19 8 ...
##  $ NIP_weight             : int  42 108 167 102 35 58 65 7 19 8 ...
##  $ Lagged_DON             : int  33 167 88 NA 58 71 NA 0 8 15 ...
##  $ DIN_Bbss               : int  2 25 37 10 5 1 1 0 4 0 ...
##  $ Lagged_DIN_Bbss        : int  NA 37 15 NA 1 4 NA NA 0 0 ...
##  $ DIN_Bbsl               : int  2 25 51 11 5 2 1 0 4 0 ...
##  $ Lagged_DIN_Bbsl        : int  NA 51 17 NA 2 9 NA NA 0 0 ...
##  $ Qnymph_Bbss_prev       : num  0.05 0.24 0.223 0.099 0.015 0.018 0.015 0 0.211 0 ...
##  $ Lagged_Qnymph_Bbss_prev: num  NA 0.223 0.171 NA 0.018 0.058 NA NA 0 0 ...
##  $ Qnymph_Bbsl_prev       : num  0.05 0.24 0.31 0.108 0.015 0.035 0.015 0 0.211 0 ...
##  $ Lagged_Qnymph_Bbsl_prev: num  NA 0.31 0.193 NA 0.035 0.13 NA NA 0 0 ...
##  $ Qnymph_Rt_prev         : num  0 0 0 0 0 0.1 0 0.2 0 0 ...
##  $ Qnymph_Bm_prev         : num  0.05 0.0619 NA 0 0.0294 NA 0.0615 0 0 NA ...
##  $ Min_OspC_Rich          : int  NA 4 5 NA NA 1 2 NA NA NA ...
```

Question Answers

a. The ‘tick’ dataset contains 46 observations of 58 variables.
b. The variable ‘DIN_Bbsl’ is an integer response variable.

Section 2 - Fitting a Poisson GLM

We are using a Poisson model here because our response variable(tick infection density)is a count. Poisson models are designed for count data that are non-negative integers and often right-skewed. In Poisson GLMs, the variance increases with the mean, which often fits ecological count data better than assuming constant variance and normality.

```r
# Fit GLM with Poisson family
tick_glm <- glm(DIN_Bbsl ~ Rodent_Rich, family = "poisson", data = tick)
summary(tick_glm)
```

```
##
## Call:
## glm(formula = DIN_Bbsl ~ Rodent_Rich, family = "poisson", data = tick)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.31505    0.27030  -1.166    0.244
## Rodent_Rich  0.48401    0.07654   6.324 2.55e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 458.20  on 45  degrees of freedom
## Residual deviance: 418.47  on 44  degrees of freedom
## AIC: 489.83
##
## Number of Fisher Scoring iterations: 7
```

Stop, Think, Do:

The coefficient for rodent richness is 0.484 and it represents the predicted increase in the logged count of infected ticks with each additional rodent species. In this model, the coefficient for rodent richness is statistically significant, with a p-value of 2.55*10^-10.

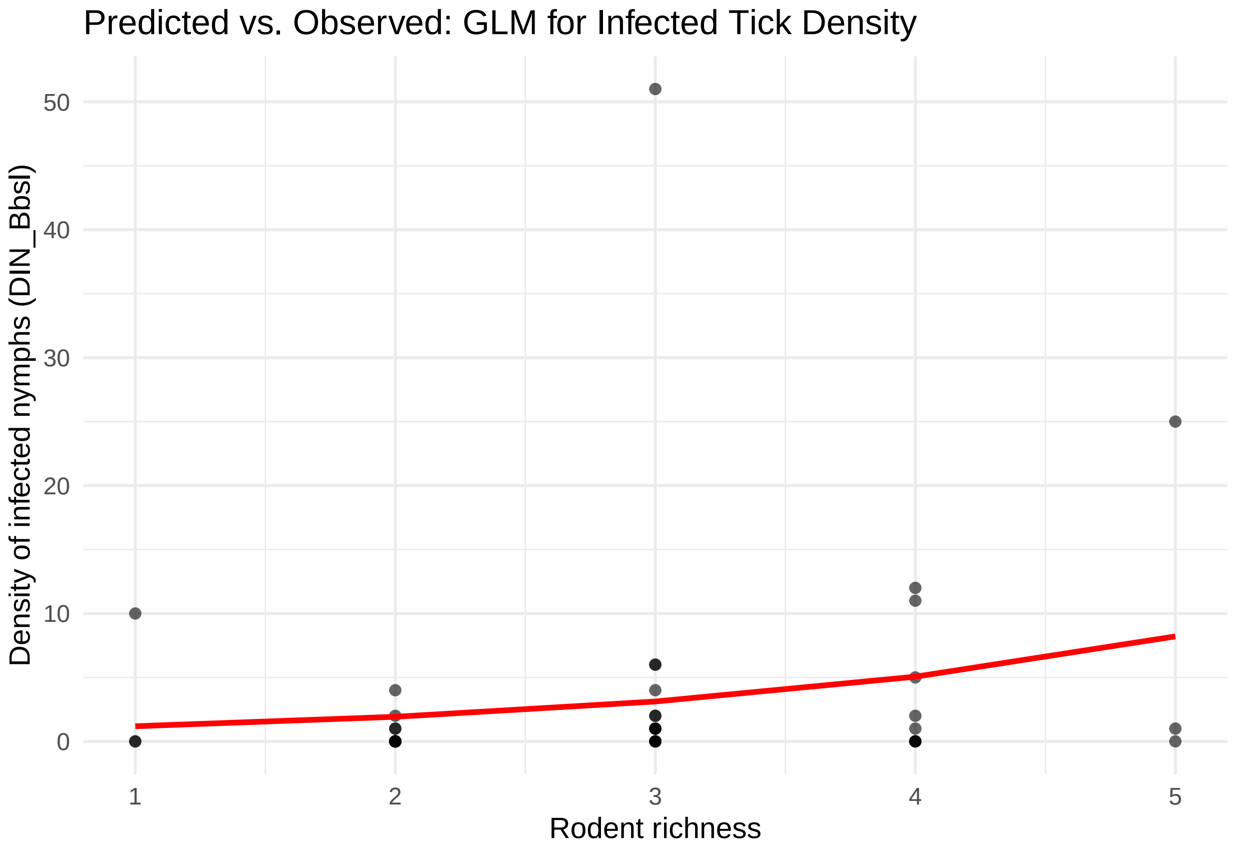
Section 3 - Plotting the GLM Predictions

After fitting a GLM, it’s useful to visualize the model predictions alongside your raw data. To do this, we can use the predict() function, which takes a fitted model and returns predicted values. These predictions can then be added as a new column in our dataset and plotted against the observed data to assess model fit and shape.

```r
# Add model predictions to the dataset
tick$predicted <- predict(tick_glm, type = "response")
```

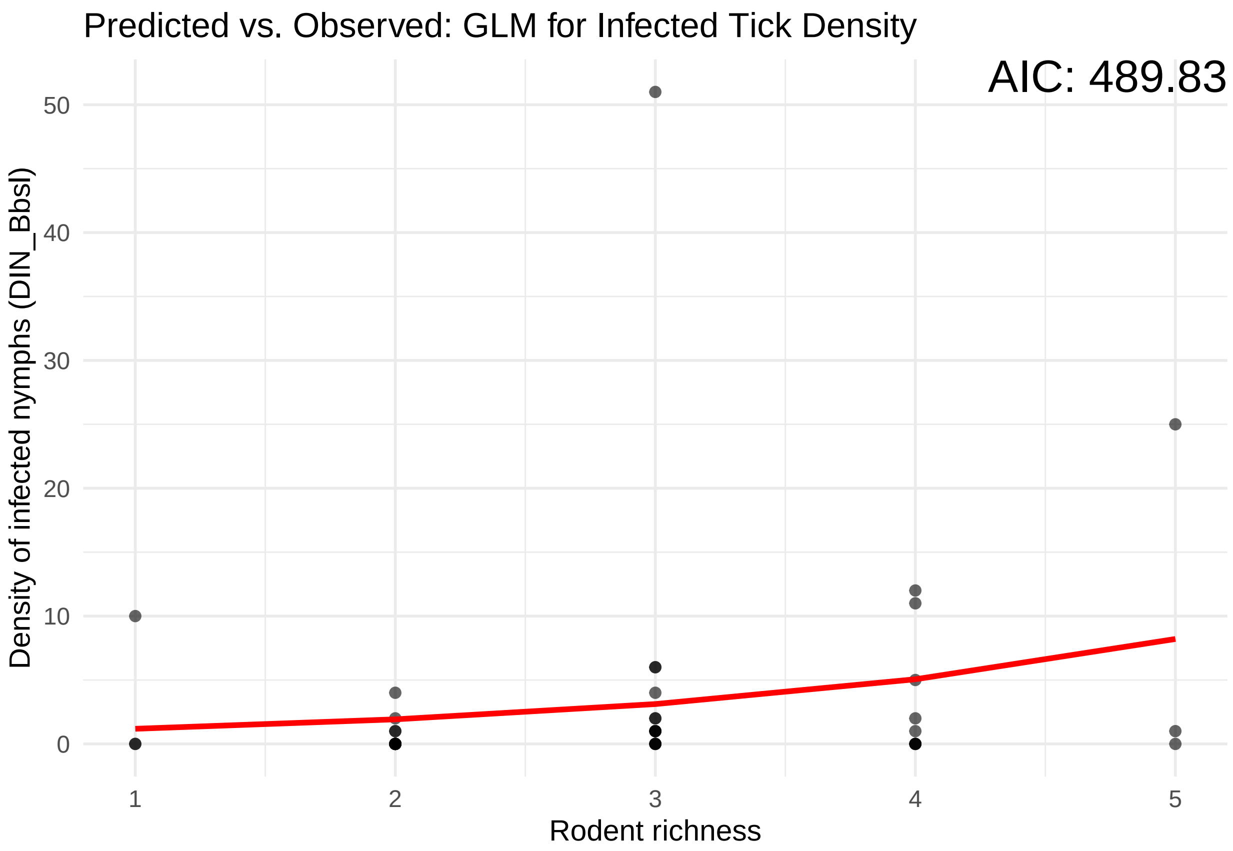
```r
# Plot observed and predicted values
ggplot(tick, aes(x = Rodent_Rich, y = DIN_Bbsl)) +
  geom_point(alpha = 0.6) +
  geom_line(aes(y = predicted), color = "red", size = 1) +
  labs(
    x = "Rodent richness",
    y = "Density of infected nymphs (DIN_Bbsl)",
    title = "Predicted vs. Observed: GLM for Infected Tick Density") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



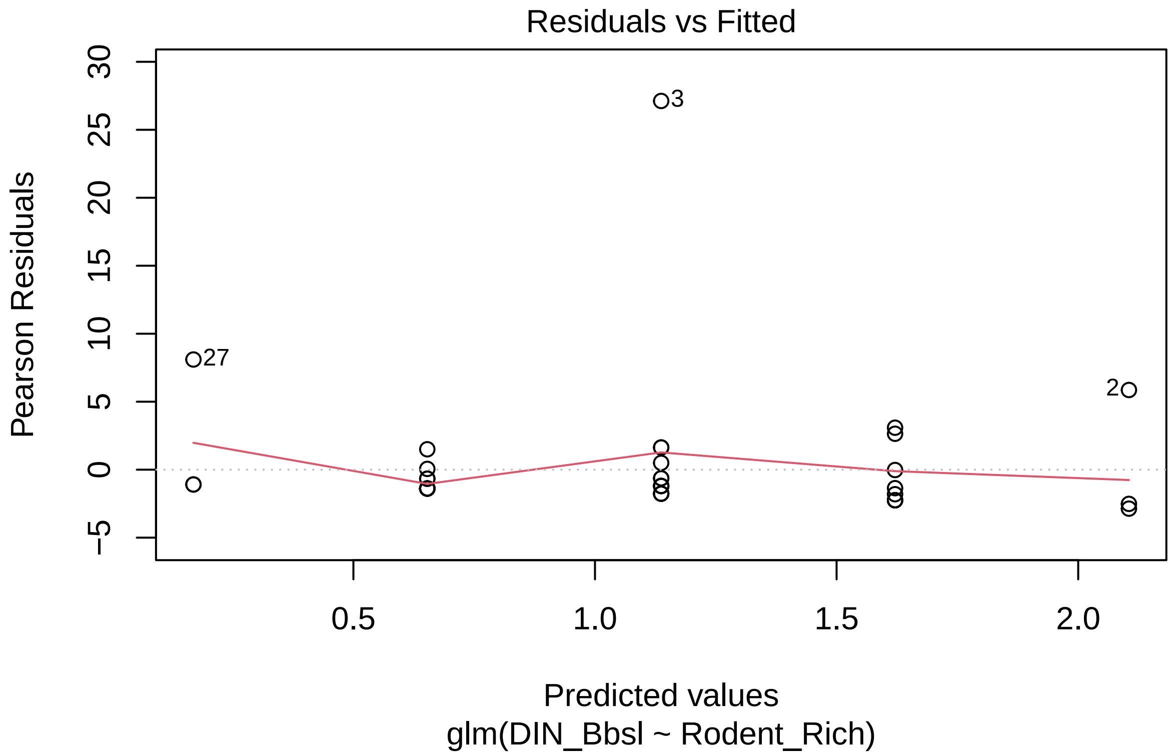
Challenge 1: Improve the above plot using an in-figure label.

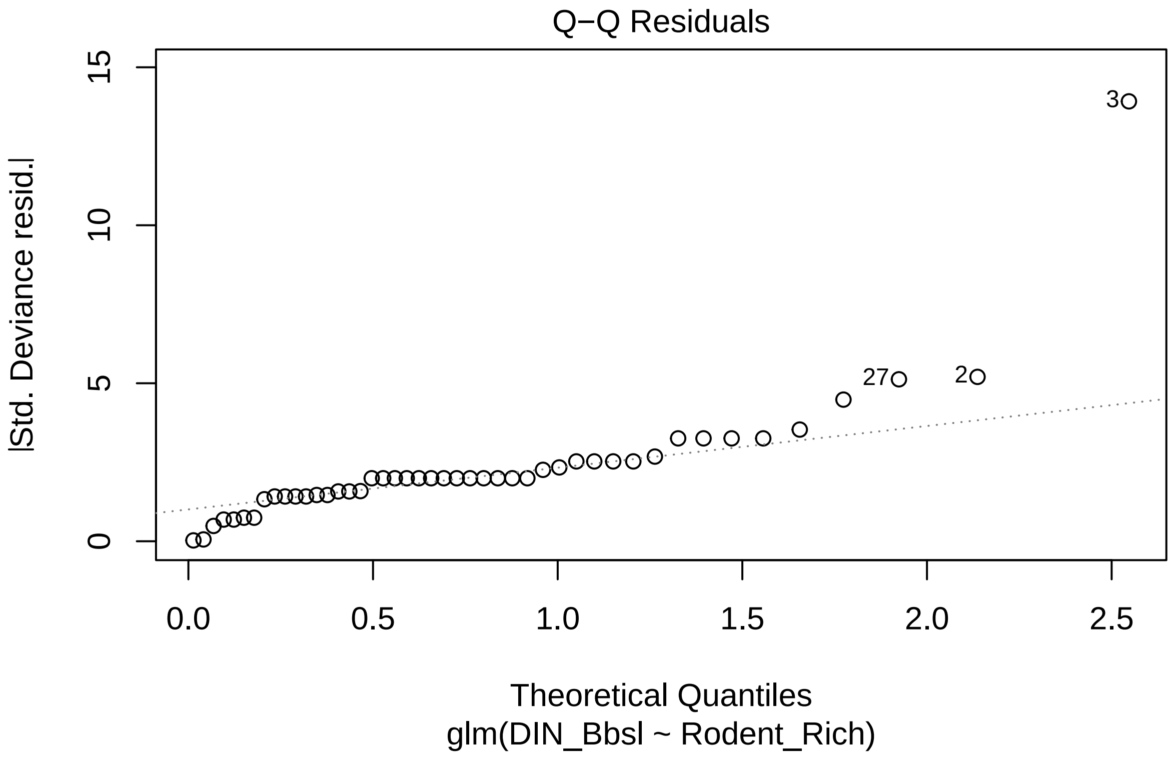
```r
# Plot observed and predicted values
ggplot(tick, aes(x = Rodent_Rich, y = DIN_Bbsl)) +
  geom_point(alpha = 0.6) +
  geom_line(aes(y = predicted), color = "red", size = 1) +
  annotate('text', x=Inf, y=Inf, label = paste0("AIC: 489.83"), hjust=1, vjust=1, , size=6)+
  labs(
    x = "Rodent richness",
    y = "Density of infected nymphs (DIN_Bbsl)",
    title = "Predicted vs. Observed: GLM for Infected Tick Density") +
  theme_minimal()
```

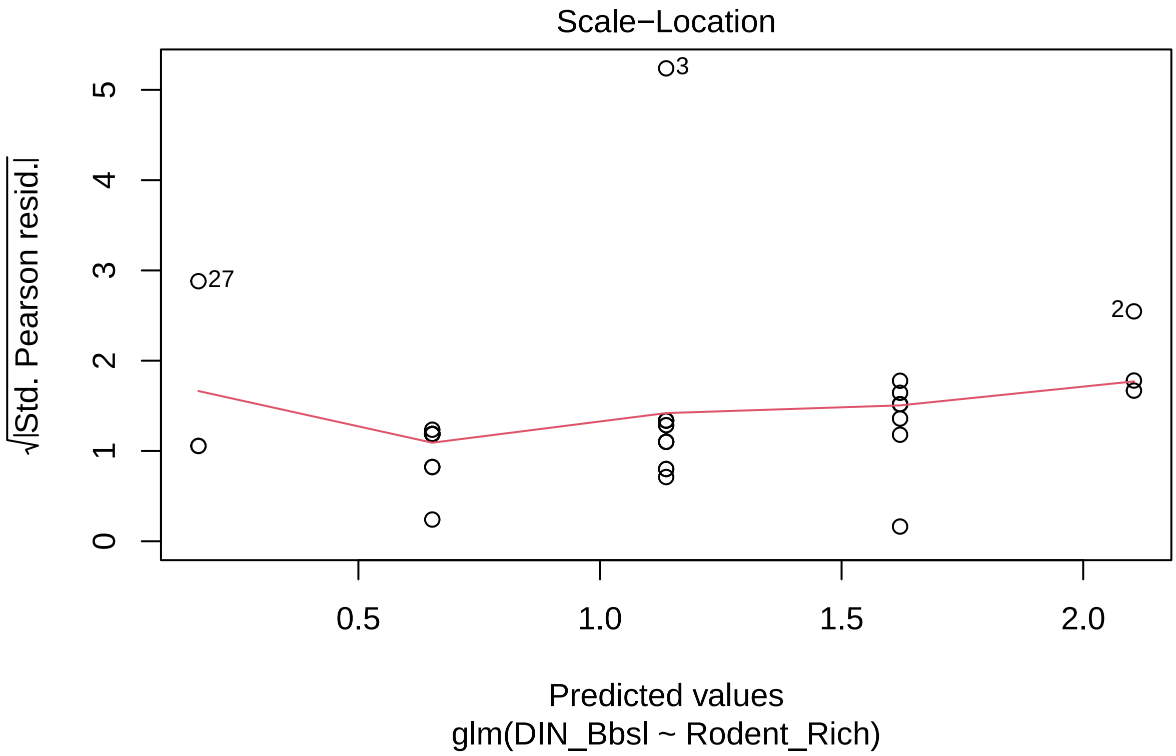


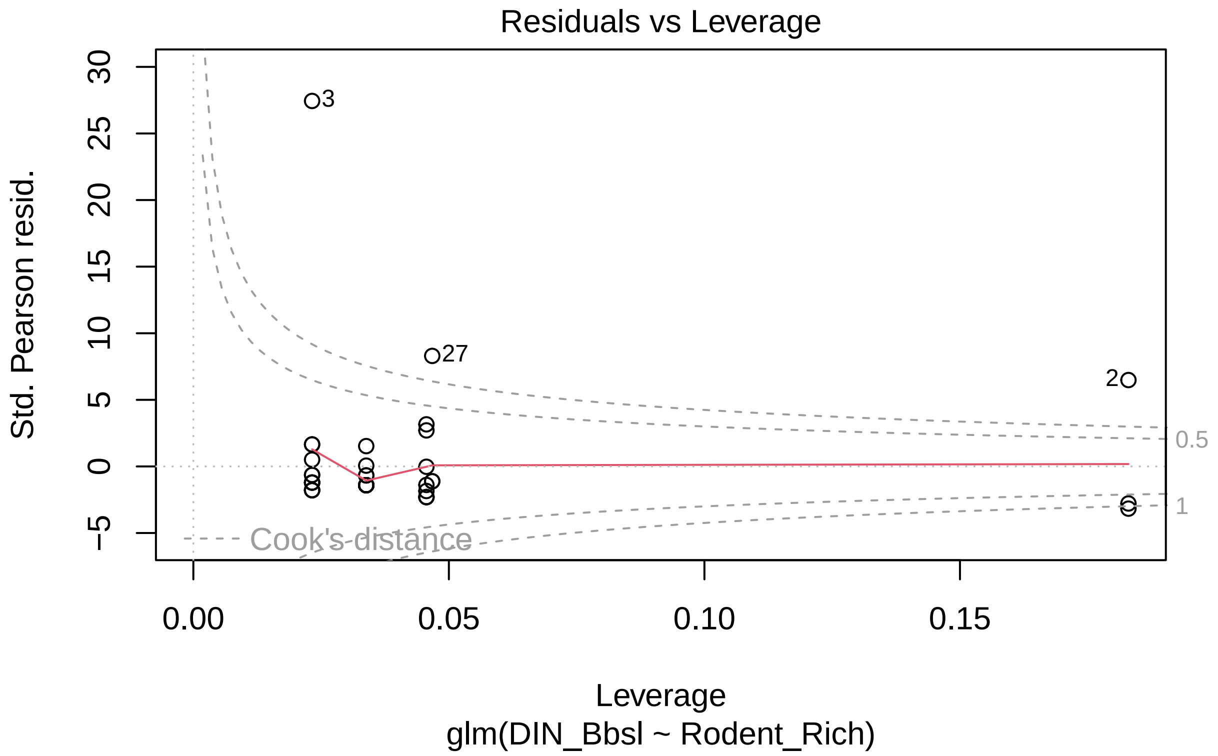
Section 4 - Plotting Model Diagnostics

```
# Diagnostic plots for GLM
plot(tick_glm)
```


Residuals vs Fitted
Pearson Residuals
−5
0
5
10
15
20
25
30
0.5
1.0
1.5
2.0
Predicted values
glm(DIN_Bbsl ~ Rodent_Rich)
3
27
2


Q−Q Residuals
|Std. Deviance resid.|
0
5
10
15
0.0
0.5
1.0
1.5
2.0
2.5
3
27
2
Theoretical Quantiles
glm(DIN_Bbsl ~ Rodent_Rich)


Scale−Location
√|Std. Pearson resid.|
0
1
2
3
4
5
3
27
2
0.5
1.0
1.5
2.0
Predicted values
glm(DIN_Bbsl ~ Rodent_Rich)



Question Answers

a. In the residuals vs predicted plot, there are a few outliers throughout, with a slight increase in variation as predicted values increase. In the Q-Q plot, it is evident that the data does not employ a normal distribution.

Section 5 - Adding Another Predictor

In Chapter 10, we explored how using two explanatory variables together can help reveal interactions or explain more variation in a response. Here, we'll test whether adding another biologically relevant predictor - predator diversity — improves our model of tick infection density. This step mirrors the idea of exploring multiple factors at once, just as we did with two-way ANOVA, but it is not the same since it does not explicitly explore interaction.

```r
# Compare two GLMs with AIC
tick_glm2 <- glm(DIN_Bbsl ~ Rodent_Rich + Predator_Shannon_wCat, family = "poisson", data = tick)

summary(tick_glm2)
```

```
##
## Call:
## glm(formula = DIN_Bbsl ~ Rodent_Rich + Predator_Shannon_wCat,
##     family = "poisson", data = tick)
##
```

```
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)             -0.1681     0.2902  -0.579    0.562    
## Rodent_Rich              0.4745     0.0759   6.251 4.07e-10 ***
## Predator_Shannon_wCat   -0.1732     0.1965  -0.882    0.378    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 450.87  on 43  degrees of freedom
## Residual deviance: 411.38  on 41  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 481.26
##
## Number of Fisher Scoring iterations: 7
```

```
AIC(tick_glm, tick_glm2)
```

```
## Warning in AIC.default(tick_glm, tick_glm2): models are not all fitted to the
## same number of observations
```

```
##           df      AIC
## tick_glm   2 489.8289
## tick_glm2  3 481.2592
```

Question Answer

a. The model that includes predator diversity with cats in addition to rodent species richness has the lower AIC when compared to the one with only rodent species richness.
b. Predator diversity only slightly added meaningful information to the model, as tick nymph infection density was not significantly impacted by predator diversity. AICs were very similar, with this model having a better balance of simplicity and fit.

Challenge 2: What if some unmeasured differences among sampling locations are influencing your results?

```
# Random effect of site
library(glmmTMB)
```

```
## Warning: package 'glmmTMB' was built under R version 4.4.3
```

```
tick_glm_site <- glmmTMB(DIN_Bbsl ~ Rodent_Rich + Predator_Shannon_wCat + (1|Site), family = 'poisson',
summary(tick_glm_site)
```

```
##  Family: poisson  ( log )
## Formula:          DIN_Bbsl ~ Rodent_Rich + Predator_Shannon_wCat + (1 | Site)
## Data: tick
##
```

```
##       AIC       BIC    logLik -2*log(L)  df.resid
##     219.5     226.6    -105.7     211.5        40
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  Site   (Intercept) 3.815    1.953
## Number of obs: 44, groups:  Site, 14
##
## Conditional model:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.555141   0.724772  -0.766    0.444
## Rodent_Rich            0.004049   0.136418   0.030    0.976
## Predator_Shannon_wCat -0.241316   0.267014  -0.904    0.366
```

```
# Compare the GLMs with AIC
AIC(tick_glm2, tick_glm_site)
```

```
##               df      AIC
## tick_glm2      3 481.2592
## tick_glm_site  4 219.4743
```

Accounting for site does greatly improve the model fit, but it does make the explanatory variable of rodent species richness lose statistical significance when combined with predator density. This suggests that the model using site as a random effect explains much more of the variation than just the fixed effects of rodents and predators.

Challenge 3: Fit a new model with both site and year as random effects.

```
# Random effect of site and year
tick_glm_site_year <- glmmTMB(DIN_Bbsl ~ Rodent_Rich + Predator_Shannon_wCat + (1|Site) + (1|Year), fami
summary(tick_glm_site_year)
```

```
##  Family: poisson  ( log )
## Formula:
## DIN_Bbsl ~ Rodent_Rich + Predator_Shannon_wCat + (1 | Site) +      (1 | Year)
## Data: tick
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##     157.5     166.4     -73.7     147.5        39
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  Site   (Intercept) 6.618    2.573
##  Year   (Intercept) 2.357    1.535
## Number of obs: 44, groups:  Site, 14; Year, 4
##
## Conditional model:
```

```
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)              0.7045     1.2139   0.580  0.56166
## Rodent_Rich             -0.7186     0.2233  -3.218  0.00129 **
## Predator_Shannon_wCat  -0.5254     0.4771  -1.101  0.27082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Compare the GLMs with AIC
AIC(tick_glm_site, tick_glm_site_year)
```

```
##                    df      AIC
## tick_glm_site       4 219.4743
## tick_glm_site_year  5 157.4743
```

Accounting for site and year does improve the model fit, and it highlights the explanatory variable of rodent species richness as statistically significant once again when combined with the fixed effects of predator density. A lower AIC suggests that the model using site and year as a random effect explains much more of the variation than just the random effect of site, combined with fixed effects of rodents and predators. This could be because the number of all organisms and species (ticks, rodents, and predators) vary greatly from year-to-year, so comparing one year to another might not be very fair in a model.

Discussion Question Answers

a. A GLM is different from a linear model in that it is able to analyze non-normal numerical data to generate an equation of best fit.
b. A Poisson GLM uses a log link function to create a best-fit model for positive interger values that have a right-skewed distribution (i.e. count data). The log link function works by exponentiating the relationship between the slope formula-adapted linear predictor variables and the response variables.
c. AIC tells us whether one GLM model is better than another at balancing simplicity with fit