Tree Hollow data analysis

Data type analysis:

Independent variable:

1. number of tree hollow increase

* The increase will be a non-negative number for each tree – no limitation to models like Poisson, Binomial (they only take positive input)
* The increase will be no limits on its numbers – the binomial requires the dependent variable to be zero or one, we cannot scale the dependent variable in this case, so binomial model is not suitable
* The number of tree hollows is a discrete variable, so the increase is also a discrete variables, so we’d better fit a discrete variable model, therefore, poisson might be suitable distribution to fit
* However, the discrete model might not always fit data well, we might use some continuous variable distribution to approximate a discrete variable, for example, normal(guassian), Gamma
* The detailed link function could be determined by plotting those data with each link function and see how they are performed

Thinning will increase the number of tree hollows across both treatments relative to the control

1. Construct a new variable, the difference between the tree hollows, that is diff =
2. Add a field to each tree to label which treatment the tree is applied
3. Use the glm() in r to run the regression: diff ~ treatment
4. Apply the different model family to see the diagnosis plot, choose the best performance model family, the supposed model family is
5. Poisson
6. Guassian
7. Gamma
8. And see the coefficient for each treatment is positive or not and the significance level for it

Note: the data might not be fitted the “best model” well, we will only choose the one with better performance than others

There will be a greater number of tree hollows develop in T2 as compared to T1

We could still use the model above, and check whether the coefficient for T2 is larger than T1, and we need to notice whether both coefficients are statistically significant

**CWD data**

**Problems to be noticed:**

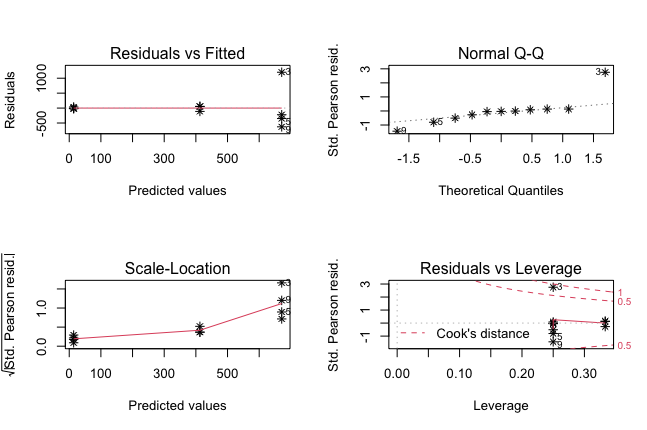
**Data need to be further corrected**

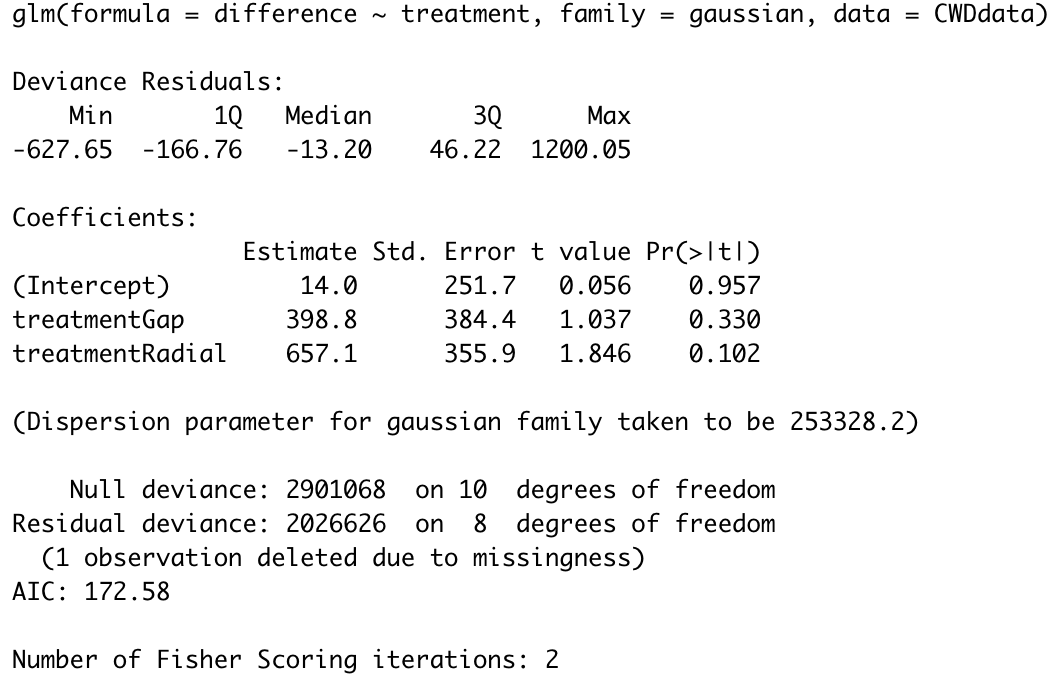
1. **Many outliers are appearing when model fitting, plot 3,5,9 are especially outstanding, which all come from treatment Radial**
2. **Violation of the model assumption is common, we might need more variables to be introduced & data modifications? Any suggestions?**

There will be a greater increase in coarse woody debris in years post thinning in T1 & T2 compared to the control

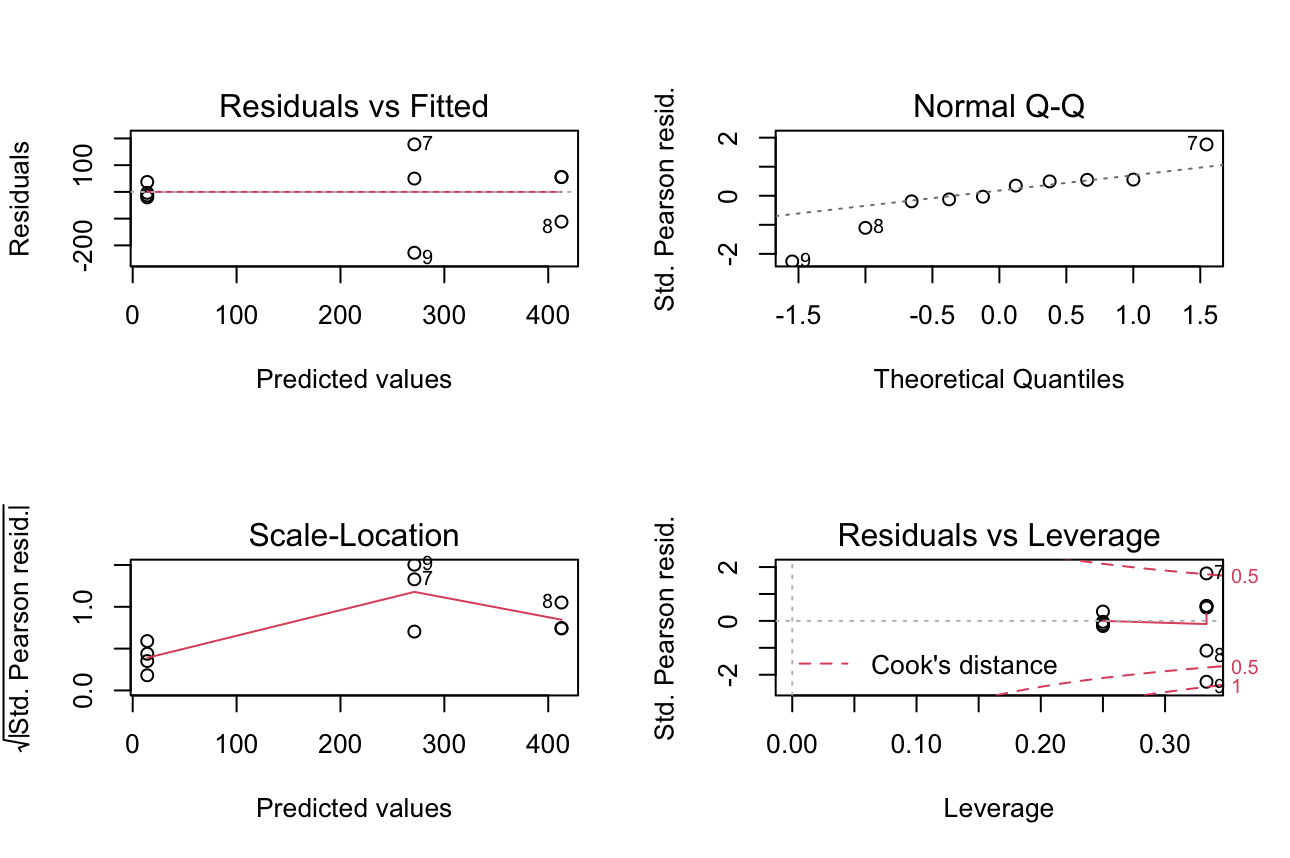
1. Construct a new variable, the difference between the CWD total, that is difference
2. Add a field to each plot to label which treatment the plot is applied
3. Use the glm() in r to run the regression: difference ~ treatment
4. Apply the different model family to see the diagnosis plot, choose the best performance model family, the supposed model family is
5. Poisson
6. Quasipoisson
7. Guassian
8. Gamma

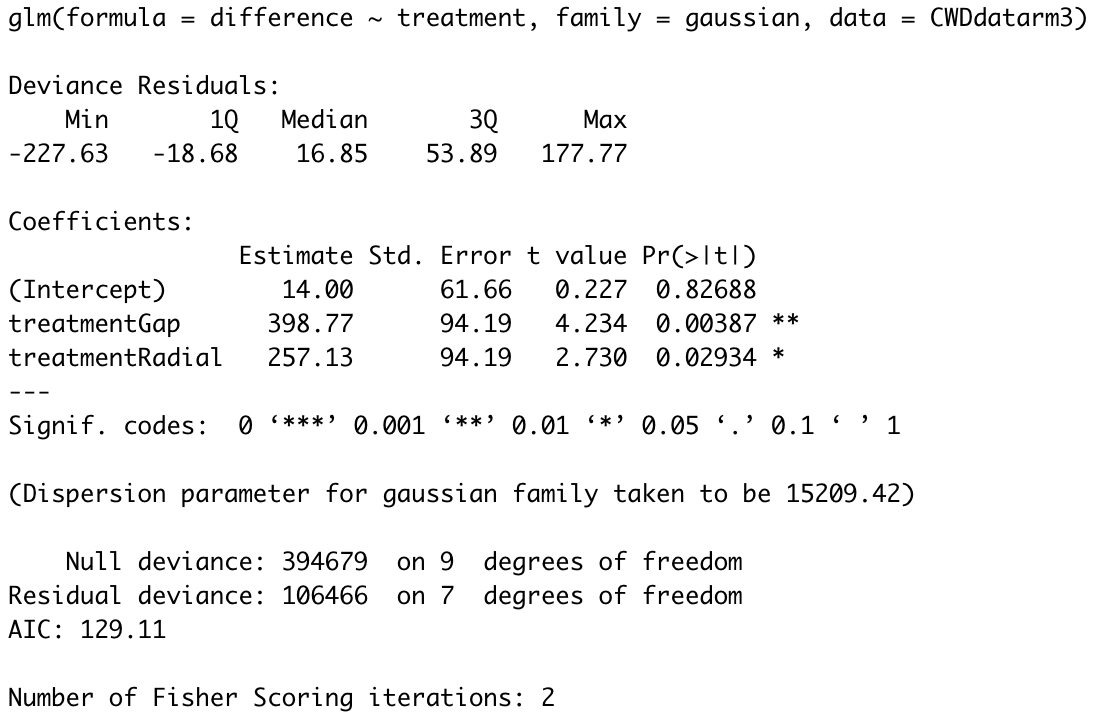
The guassian family might not fit the model well. The scale-location plot shows an increasing trend, it states the heteroskedasticity, that is the spread of the residuals is not equal, and the cook’s distance shows that the point 3 are a significant influential point. All 3, 5, 9 are outliers, and all of them are from treatment Radial. All the coefficient is not significant in this case, so even though all of them are positive, we cannot draw conclusion from it.



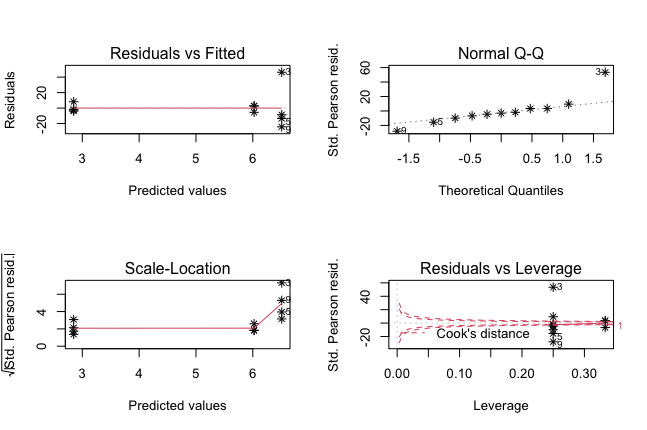


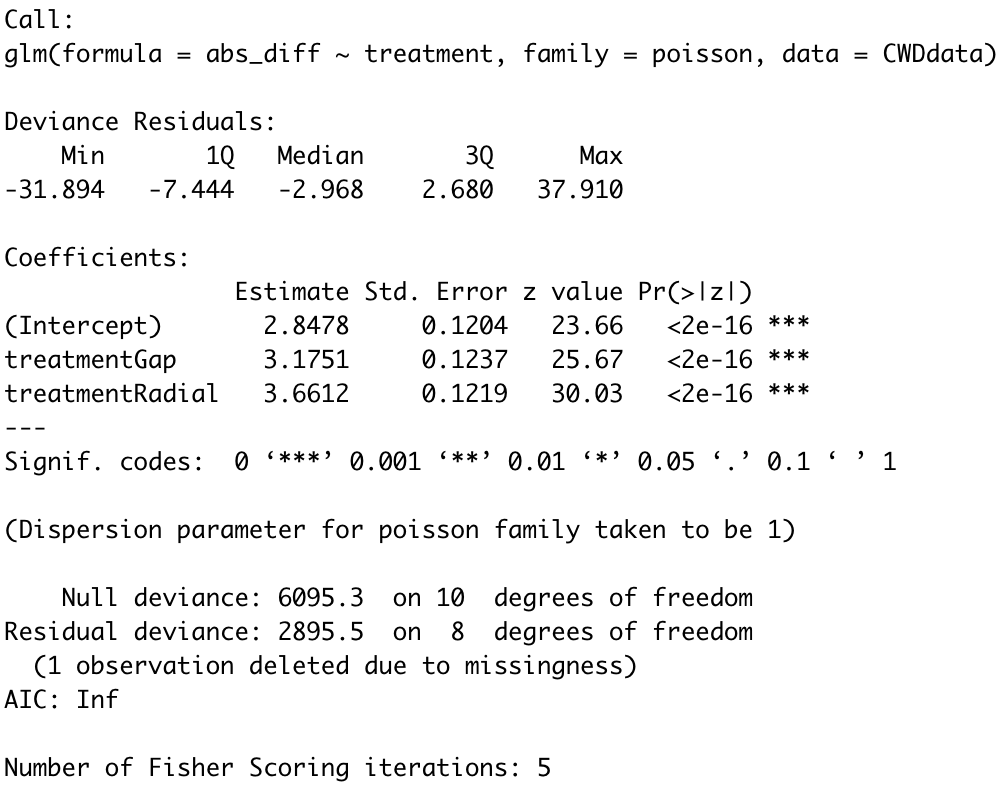
Then we tried to remove the plot 3(most significant outlier) and plot 12, the outstanding outliers become 7,8,9, both 7 and 9 are influential points, besides, QQplot and scale location plots state that violation of the model assumption, even though the Gap treatment and radial treatment become significant now, and AIC is smaller than before.



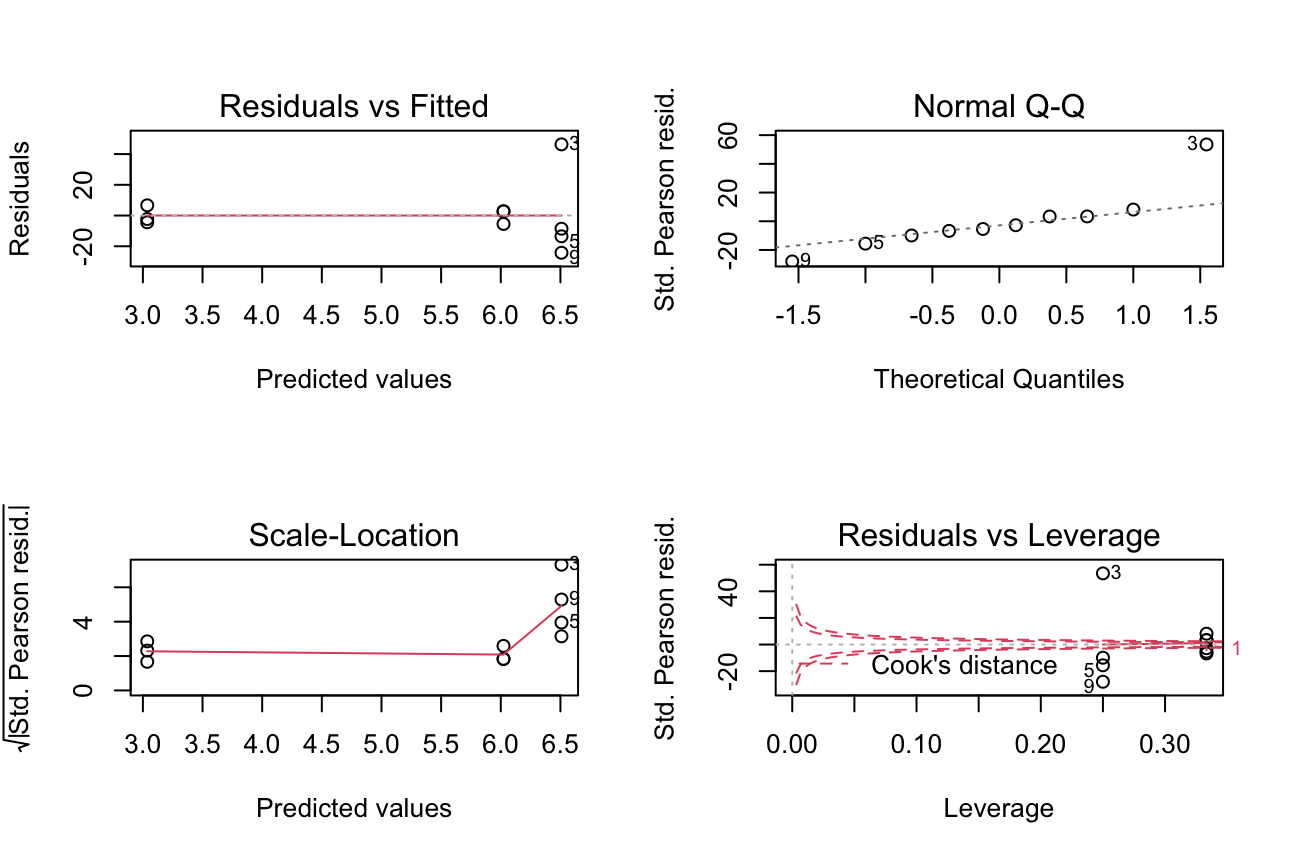


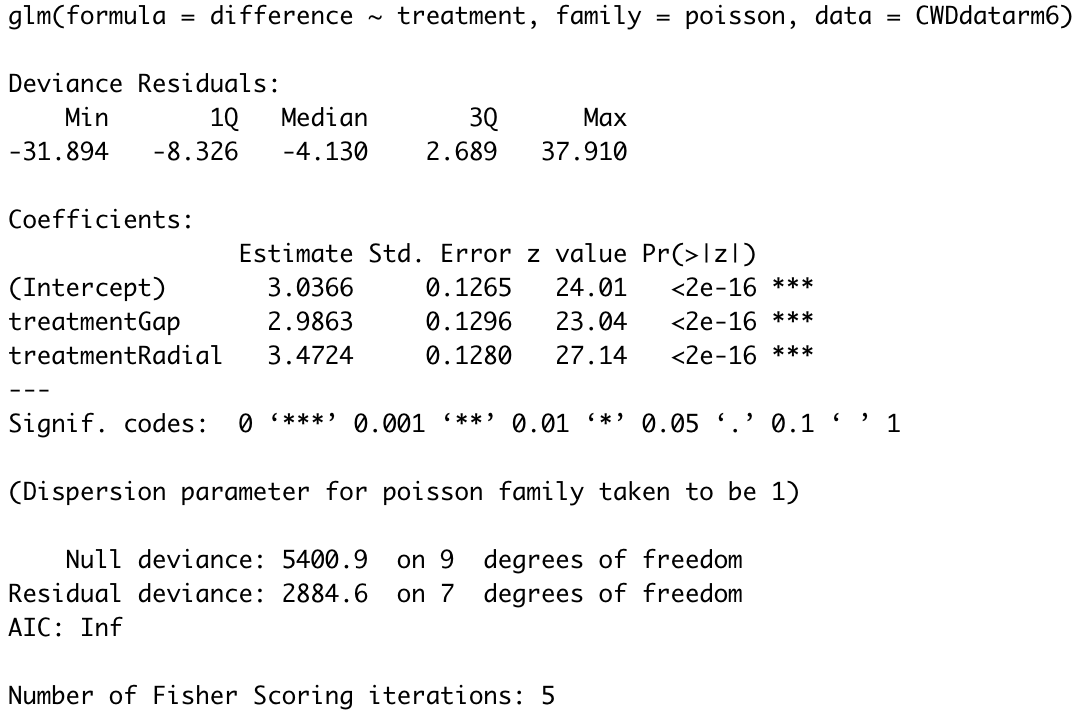
The poisson model cannot be used directly, since the one of the difference in CWD is -6.5, so if we want to fit the poisson model, we have to convert all different to positive using the abs() in r.(pros: no information is lost + we only care the direction rather than the magnitude; cons: might not be reasonable to do the transformation like this) this time the scale-location model seems better, but it still shows an increasing trend, all 3,5, 9 becomes significant influential this time, almost all points become influential. So, this fitting is worse than the original guassian model fitting. All the coefficient is statistically significant and positive.





Then we tried to remove the only negative value, it does not make much difference compared to the absolute value transformation. (magnitude is smaller than before for both treatment)

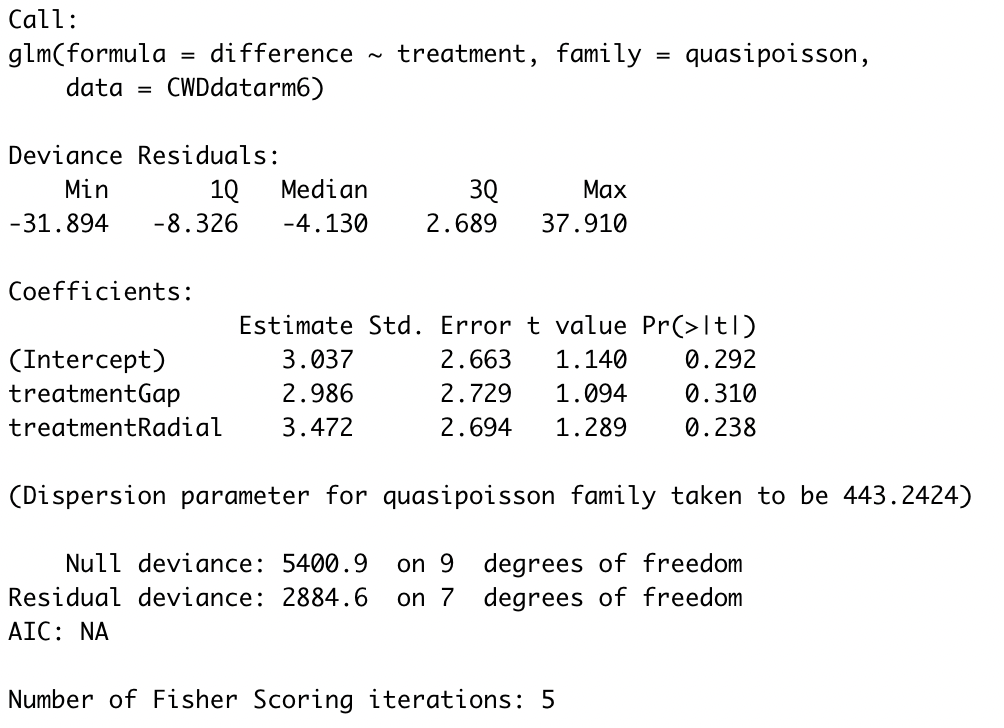




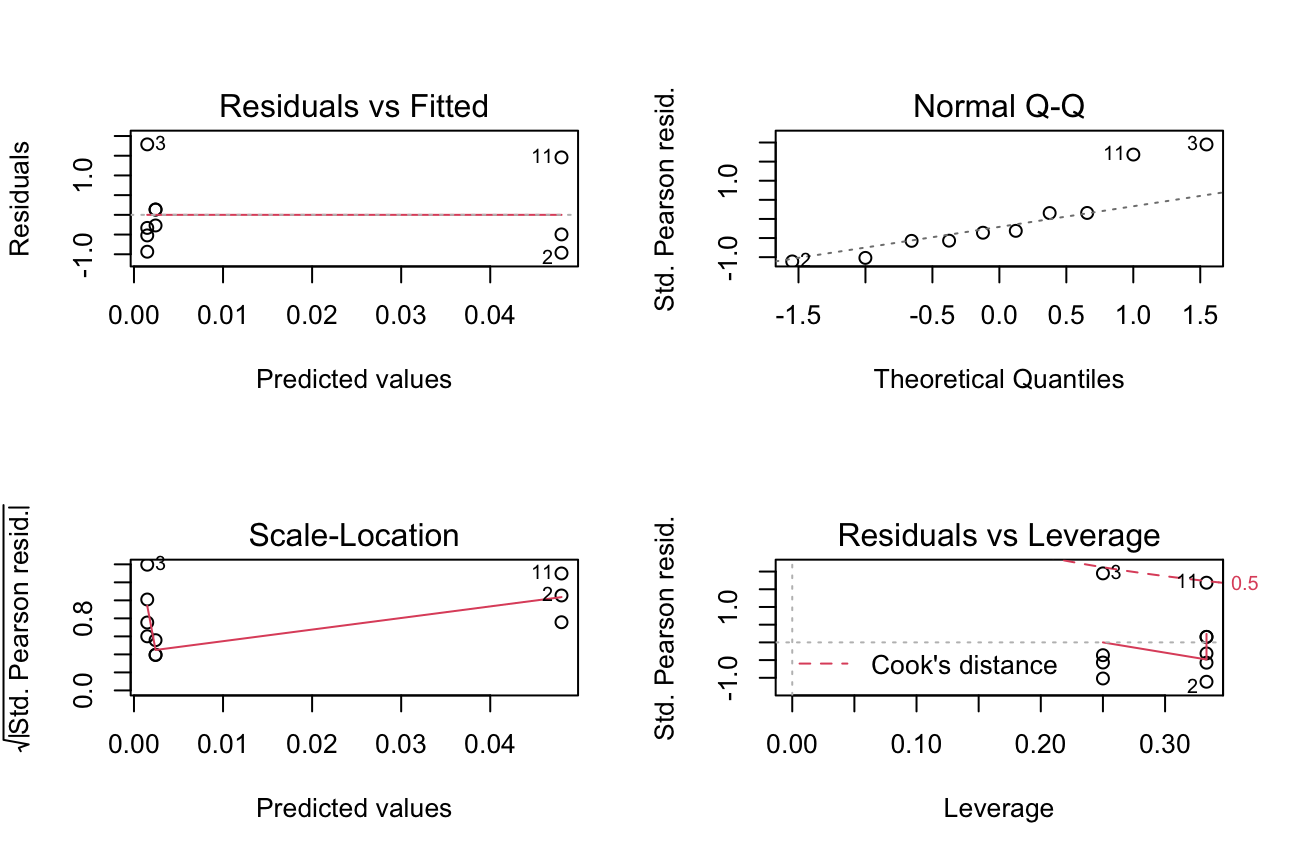
The quansi-poisson cannot be used directly, we use the plot 6 removel to further fit into the quasi-poisson model to resolve the overdispersion problem, the scale-location still shows an increasing trend, the outlier this time is point 3,5,9, and all of them are from the treatment Radial. Plot 3 is the only influential point. But all the coefficients are not significant but are positive.

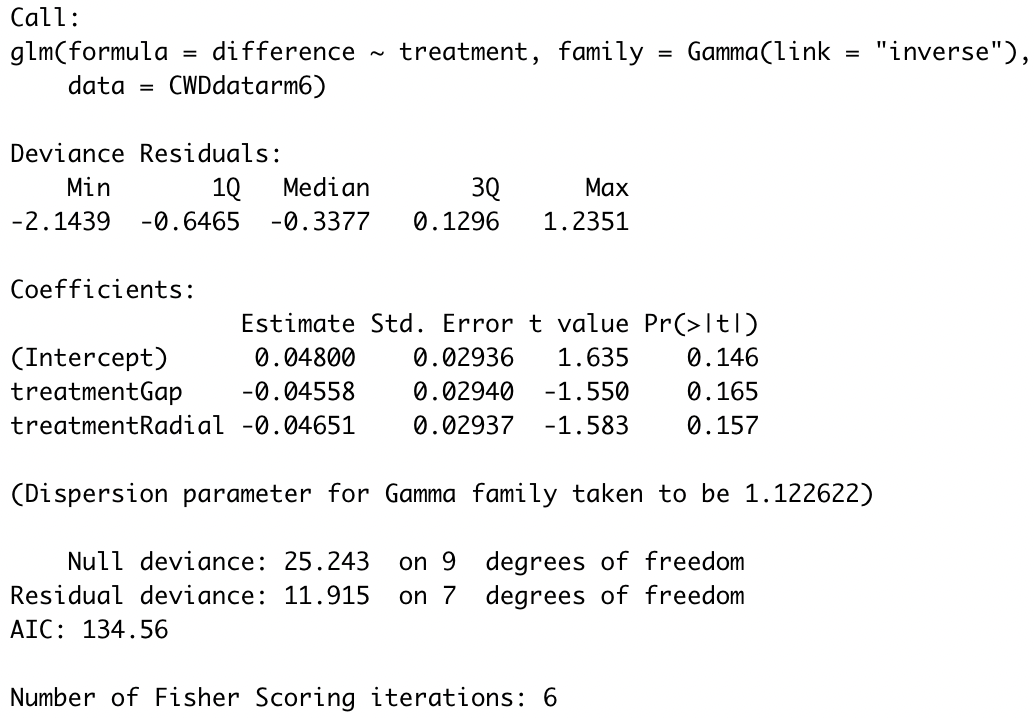
Diagram

Description automatically generated



The Gamma also takes no negative values, so the outlier becomes 2,11(Control), 3(Radial), and no significant points are presented, but the normal QQ plot is left-skewed, and the scale-location is showed a decreasing trend first and then increasing trend. All the coefficients from the model are positive but not statistically significant. But the AIC is the lowest among all the models.





Since the model has lowest AIC with the maximum information we can remained, so we further fit the log link function for it. This time the outlier become 3,5,9 again, and 3 is an influential point, all three coefs are statistically significant at 90% significance level but some violation about the model assumption.

Diagram

Description automatically generated

