BST 224 Homework 1

4/20/2020

Student Name: ***Bohao Zou***

Student ID: ***917796070***

Student Degree: ***Master***

Question 2:

(A)

In this project, the researchers used two different treatments (1. Receive Azathioprine alone (AZ, group 1) 2. Receive Azathioprine plus Methylprednisommne (AZ+MP, group 2)) for suffers and we want to investigate (i) whether both treatments lower AFCR over 18 month period and (ii) whether treatment with AZ+MP results in different immune system response than does AZ alone, and, (iii) if so how it is different in terms of response over time.

At the first step, we should explore the data structure. In this data set, the *id* variable represents the identity of suffers. The *time* is the longitudinal variable and it indicates how long has passed after the initiation of treatment. The *AFCR* is the response variable. Low value of AFCR are evidence that immunity is improving and may have a better treatment effect. The *group, prior treatment indicator and age* are x-covariates. Those variables may affect the response variable AFCR in some situations.

X-Covariates Explore

We must notice that the immunity is different between different age. if received the prior treatment is also important factor which effects the immunity. So, we could not ignore those essential factors.

There are 150 suffers participated in this project. The youngest suffer is 32 years old and the oldest is 73 years old. There are 92 suffers who has received prior treatments and 58 suffers who did not receive the prior treatments. The distribution of age and prior treatments are displayed below.

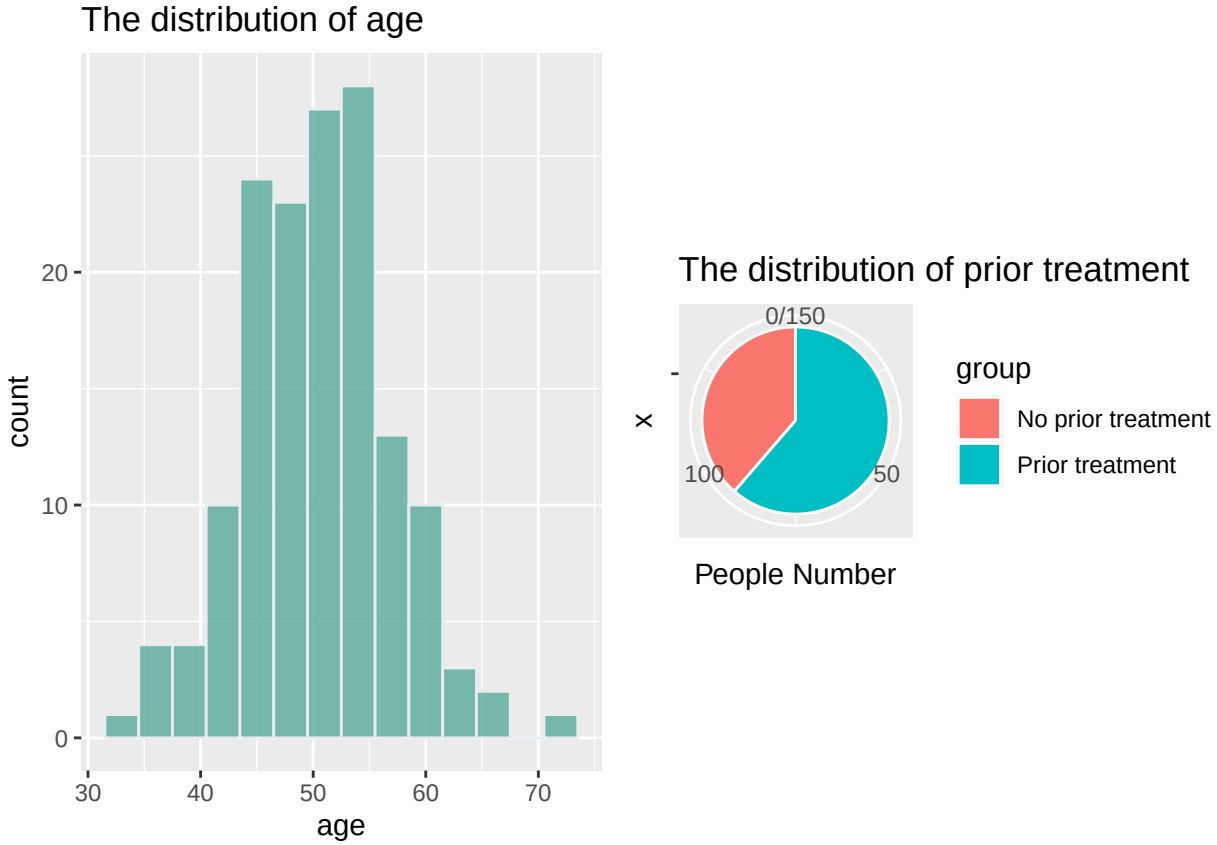

Figure 1 : The left side is the distribution of age variable. The right side is the distribution of Prior treatment indicator.

From those distributions, we can get that the age of most suffers are concentrated in the interval of 42 - 56 and the number of suffers who recieved prior treatment is more than the number of suffers who didn't receive the prior treatment.

Longitudinal Variable Explore

In the data set, the longitudinal variable time is an discrete variable. It contains 7 unique time points which are 0, 3, 6, 9, 12, 15 and 18 respectively. Those numbers indicate how many months have passed after the first treatment. The longitudinal variable is an discrete variable. It is difficult to use curve to display the tendency of AFCR. So, we should use box plot to show the relation between time t and response variable AFCR.

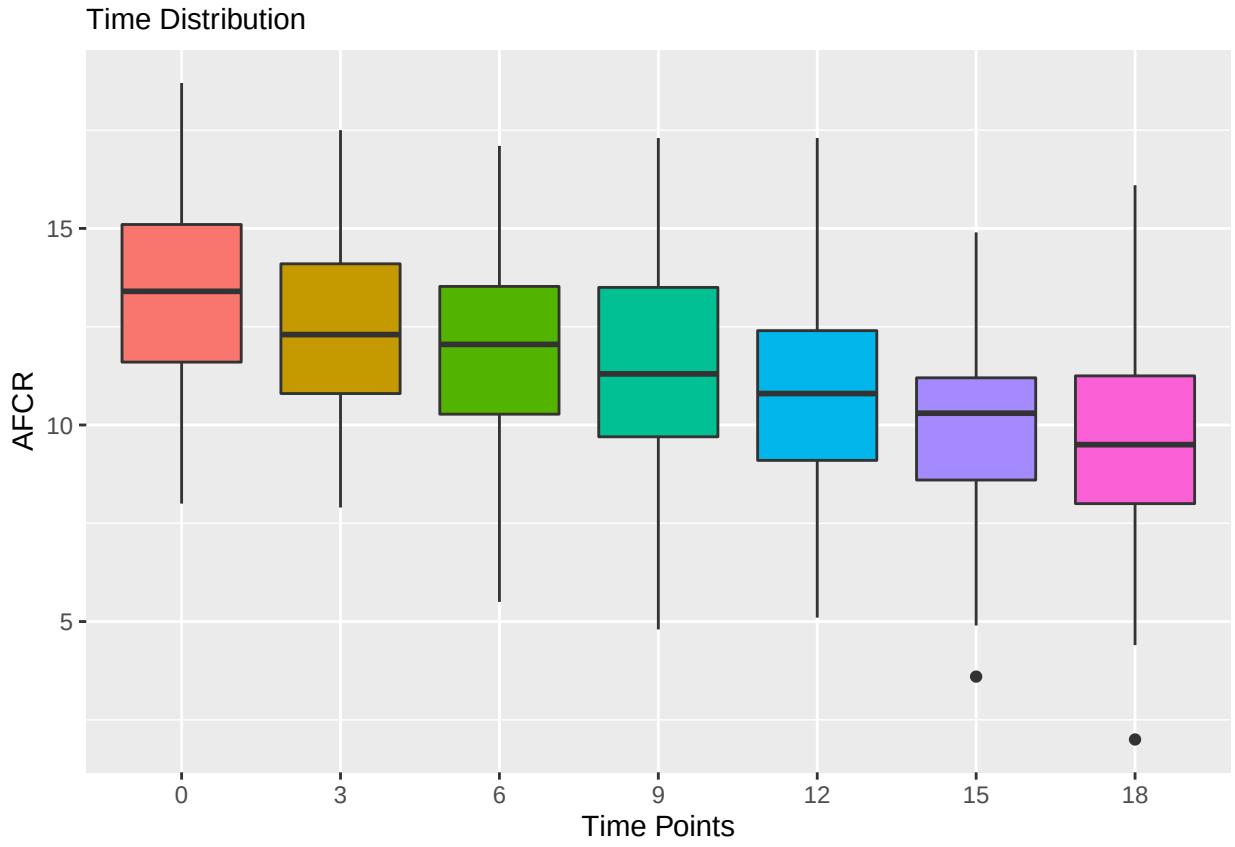

Figure 2 : The box plot of Time and response variable AFCR

From this box polt, we can know that with the time past, the tendency of AFCR is decreased significantly. We can roughly say that the immunity of those suffers are improving. We can make a initial judgement that those treatments are useful.

Data Details Explore

There are 150 suffers so that the subjectes in this data set is 150. For most of subjectes, the observations are 7 because those researcher planned clinic visits at 0, 3 , 6 , 9 , 12 ,15 and 18 months after accetpted the treatment. However, because of some reasons, a small mount of data are missing. So in some subjects, the observations are not 7. Because of those missing data, We need to explore the distribution of observations in each treatment group.

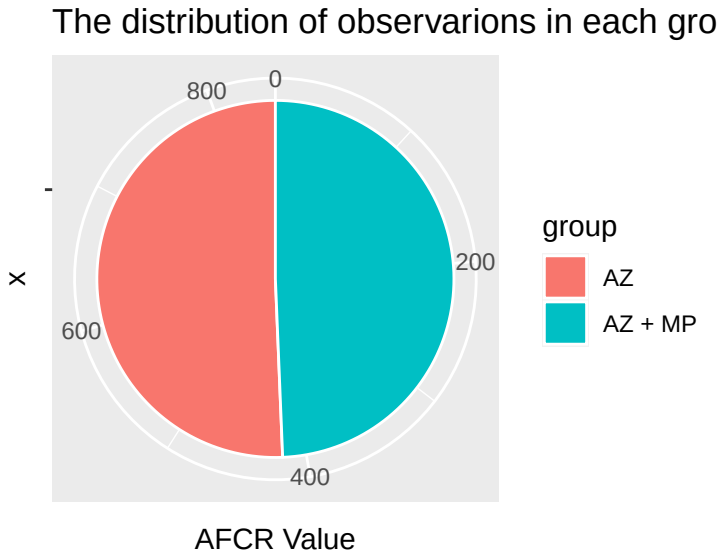

Figure 3 : The pie plot for the distribution of observations in each treatment group.

From the pie plot we can know that the numbers of observations in each treatment group are roughly same. The number of observations in group "AZ" is a little bit more than the number of observations in group "AZ+MP".

For those X-covariates *group*, *prior treatment* and *age*, the variables *group* and *prior treatment* are baseline variables because its will not change over time. The variable *age* changes over time because the clinic visits last more than a year. So, the *age* variable is a time-varying variable.

(B)

At this section, we will explore the tendency of the value of *AFCR* over time and give the first initial evidence that both treatments (AZ or AZ + MP) will make *AFCR* lower over the 18 month period.

From the *Figure 2* we can initial get the conclusion that with the time passing, the value of AFCR will be decreased. However, using box plot to show the tendency of a continuous response variable is not as accurate as using scatter points plot. But in this data set, the time variable is a discrete variable. It can not draw scatter plot because the response variables will in a vertical line. For solving this problem, I made a transformation to time varibale and used "KernSmooth", "linear regression" to draw the line of tendency. We can use those line to judge the tendency of the AFCR over the time. It is more clearly than box plot.

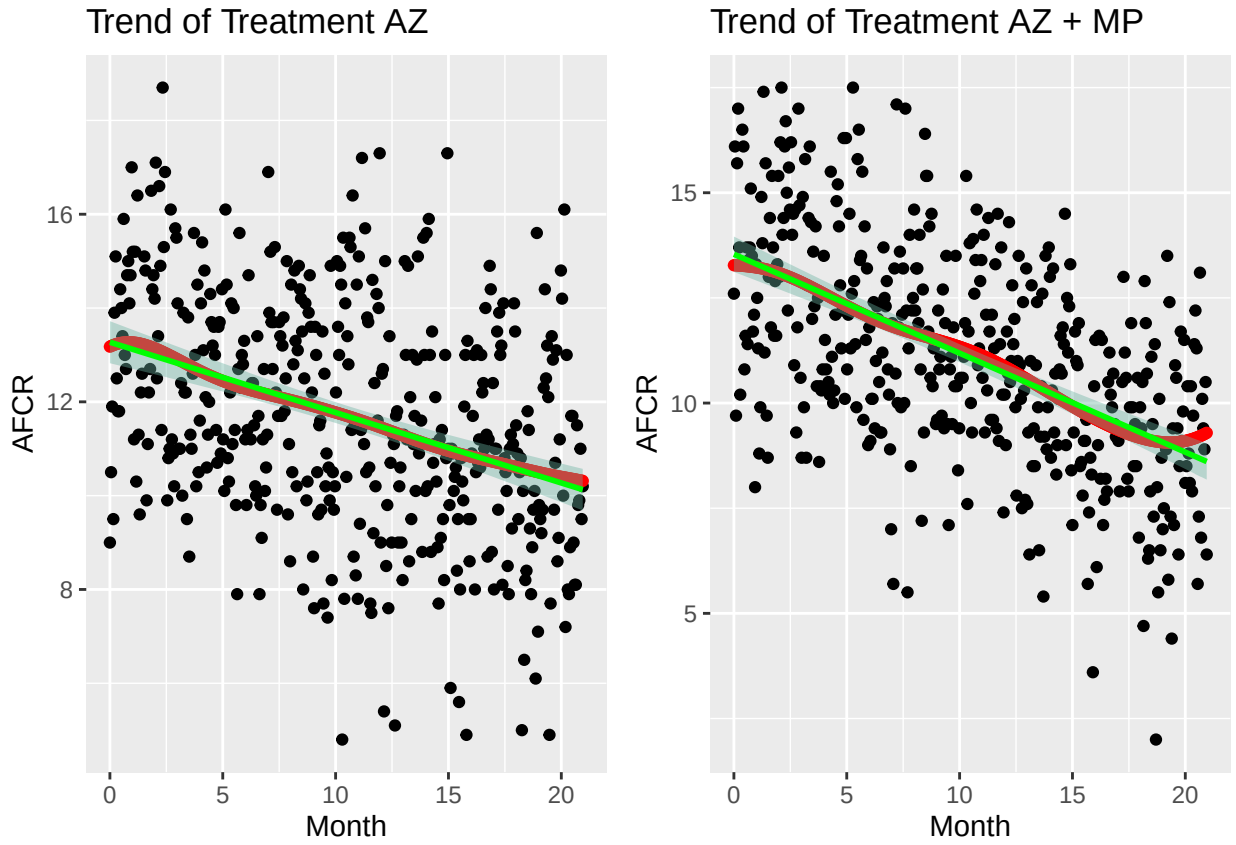

Figure 4 : The left is the AFCR tendency which treatment is AZ. The right is the AFCR tendency which treatment is AZ + MP. The red line is constructed by KernSmooth function and the green line is builded by linear regression function.

From above plot, we can know that when the time passing, the tendency of AFCR is decreasing in the both group which treatments are AZ and AZ + MP.

From this plot and *Figure 2*, we have sufficient evidences that both treatments (AZ or AZ + MP) will have a tendency that those treatments will lower AFCR over the 18 month period. For some aspects, using azathioprine (AZ) or using azathioprine plus methylprednisommne (AZ+MP) will improve the immunity over 18 month period.

(C)

Explore the variance of the response variable as a function of time

At this section, we will explore the variance of response variable AFCR as a function of time. We need to calculate the variance of response variable AFCR firstly. Then draw a line plot to initially explore the relationship between the variance of AFCR and time. Finally, we will build a linear model for AFCR as response variable and time as covariate.

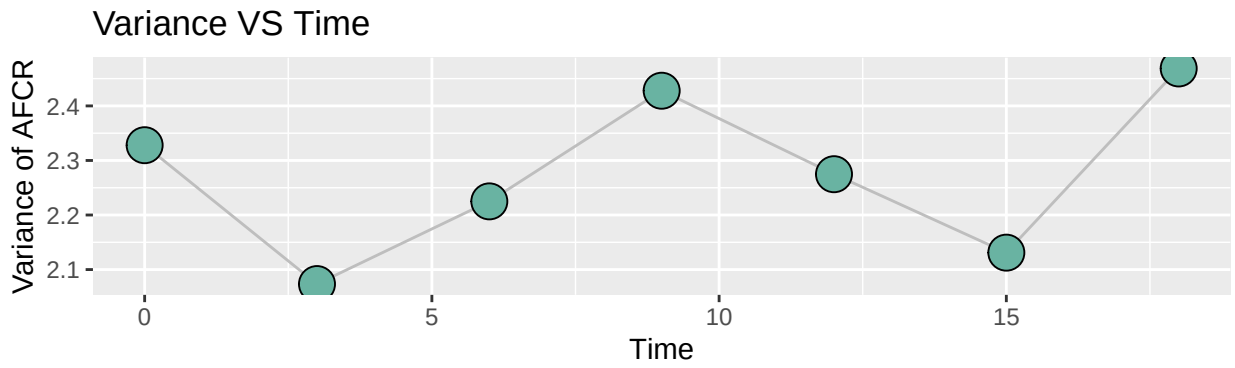

Figure 5 : The line plot of the variance of AFCR vs. Time

This plot implies that the variance may not have a linear relationship with time. It may contain some factors in its residuals. Next we should build the linear model and draw the plot of Residuals vs. Fitted Value to prove our guess.

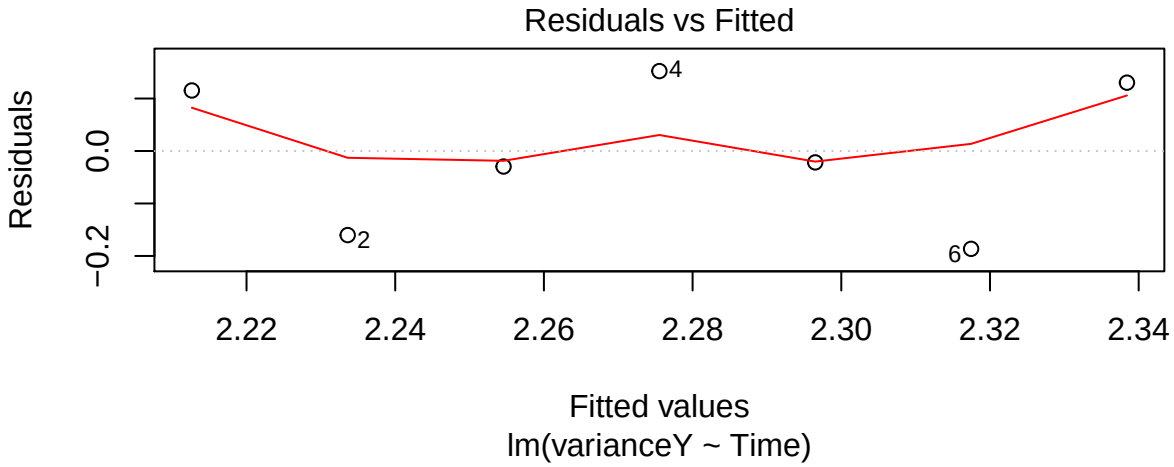

Figure 6 : The plot of Residuals Vs. Fitted Value

From *Figure 6* we can know that there are some factors in the residuals of this linear model. This indicates that we can't only consider the effects of time on the variance of response variable Y. We should consider more complex model.

Explore the correlation structure of the response variable using correlation matrices and the sample autocorrelation function.

For exploring the correlation structure among AFCR, we first build a linear model which contains all variables in this data set, *time, treatment group, prior treatment, age and all interaction terms between those variables.* Those variables are all treated as dummy variables except *age*. The response variable is *AFCR*. Then, get *the residuals of AFCR*.

Through the above processing, we remove the time trends for each treatment groups. We draw the box plot of residuals of AFCR vs. time and the line plot of the median of the residuals of AFCR vs. time to show the variance and tendency of those residuals.

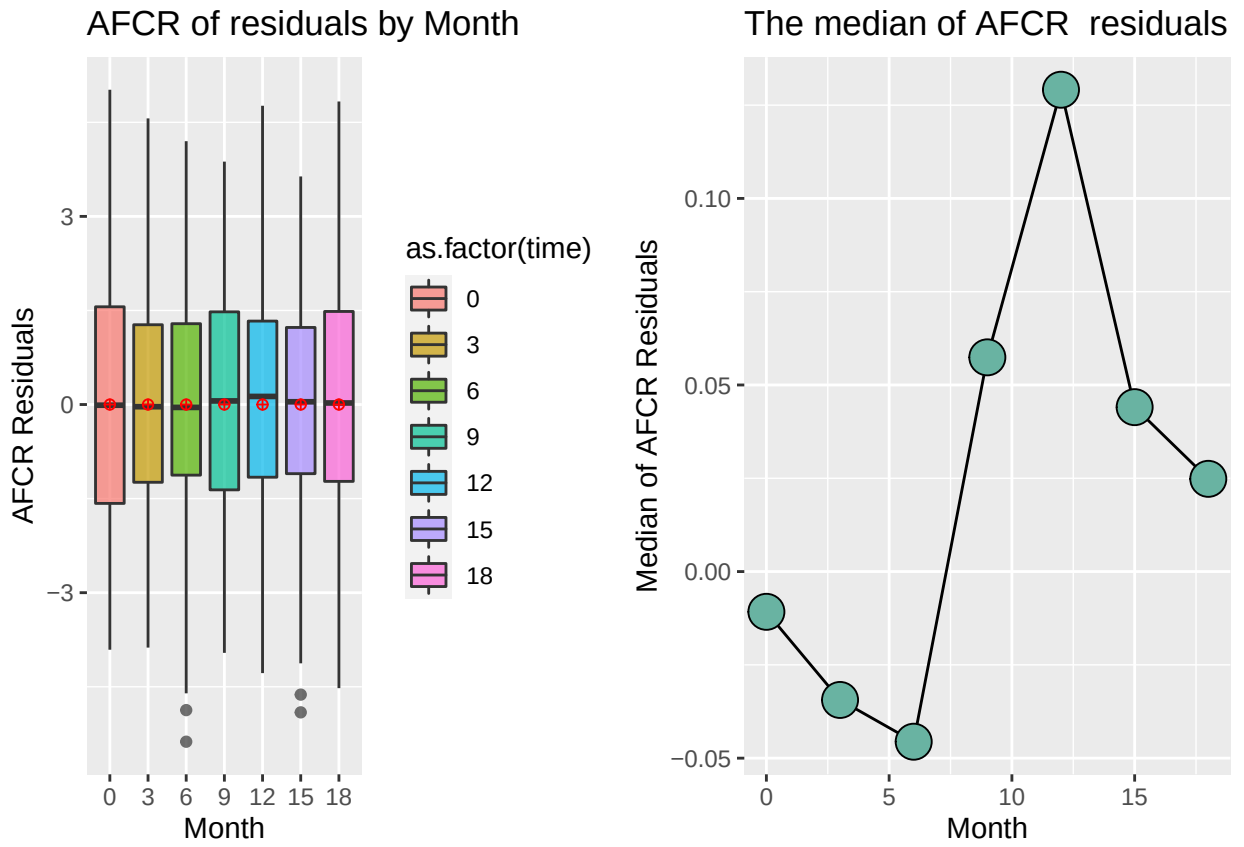

Figure 7 : Left: The plot of Residuals Vs. Time. The red point in each box represents the mean value of each box. Right: The plot of Median of Residuals vs. Time.

From the plot we can see that the variance of those residuals are fairly constant across time. Then line plot shows that the tendency of the median of those residuals are not stable but the distences between its are not far away and its all fluctuate around zero. Those phenomenons indicate that we have removed the covariate effects of *age, prior treatment, time, treatment group, and interaction between those covariates.*

Because we have removed all effects of covariates. At present, we can explore the correlation structure of response variable AFCR. At first, we will use correlation matrix to show the correlation structure of AFCR.

Month	0	3	6	9	12	15	18
0	1.0	0.17	0.21	0.35	0.22	0.27	0.22
3	0.17	1.0	0.14	0.27	0.22	0.17	0.27
6	0.21	0.14	1.0	0.25	0.28	0.13	0.22
9	0.35	0.27	0.25	1.0	0.27	0.28	0.31
12	0.22	0.22	0.28	0.27	1.0	0.22	0.32
15	0.27	0.17	0.13	0.28	0.22	1.0	0.15
18	0.22	0.27	0.22	0.31	0.32	0.15	1.0

Table 1 : The correlation matrix of response variable AFCR. (This matrix is builded by Pearson's correlation coefficient)

From the correlation matrix we can roughly get the conclusion that with the time passing, the correlation between different months trend increasing firstly and decreasing after the summit. This may indicate that those treatments would play a role as time passing and get the peak around 9 month passed. Then, the treatments effect decrese.

Besides using the correlation matrix to explore the correlation structure, we can also use the autocorrelation function to display the correlation structure of AFCR vaiable. We can draw the correlogram with 95% Tolerance limits to give a directive showing.

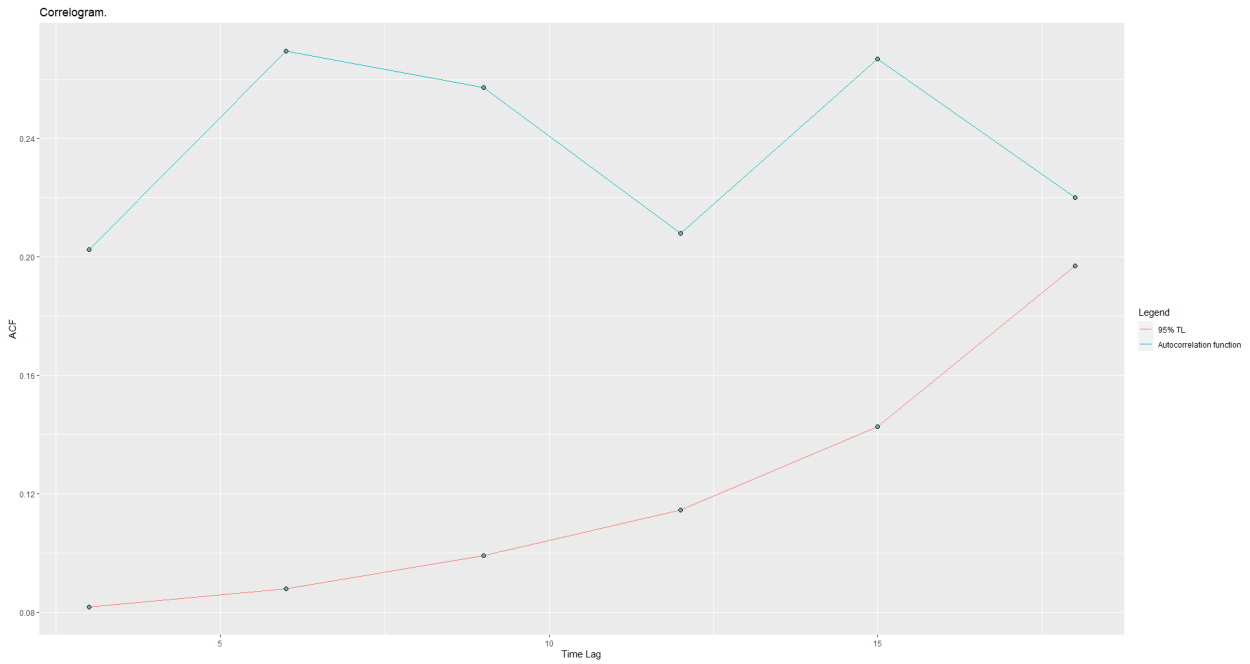

Figure 8 : The Correlogram plot with 95% Tolerance Limits

This plot shows that all correlation coefficients are all above the 95% Tolerance limits. This indicates that we have 95% confidence that the true correlation is not zero.