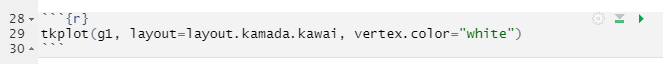
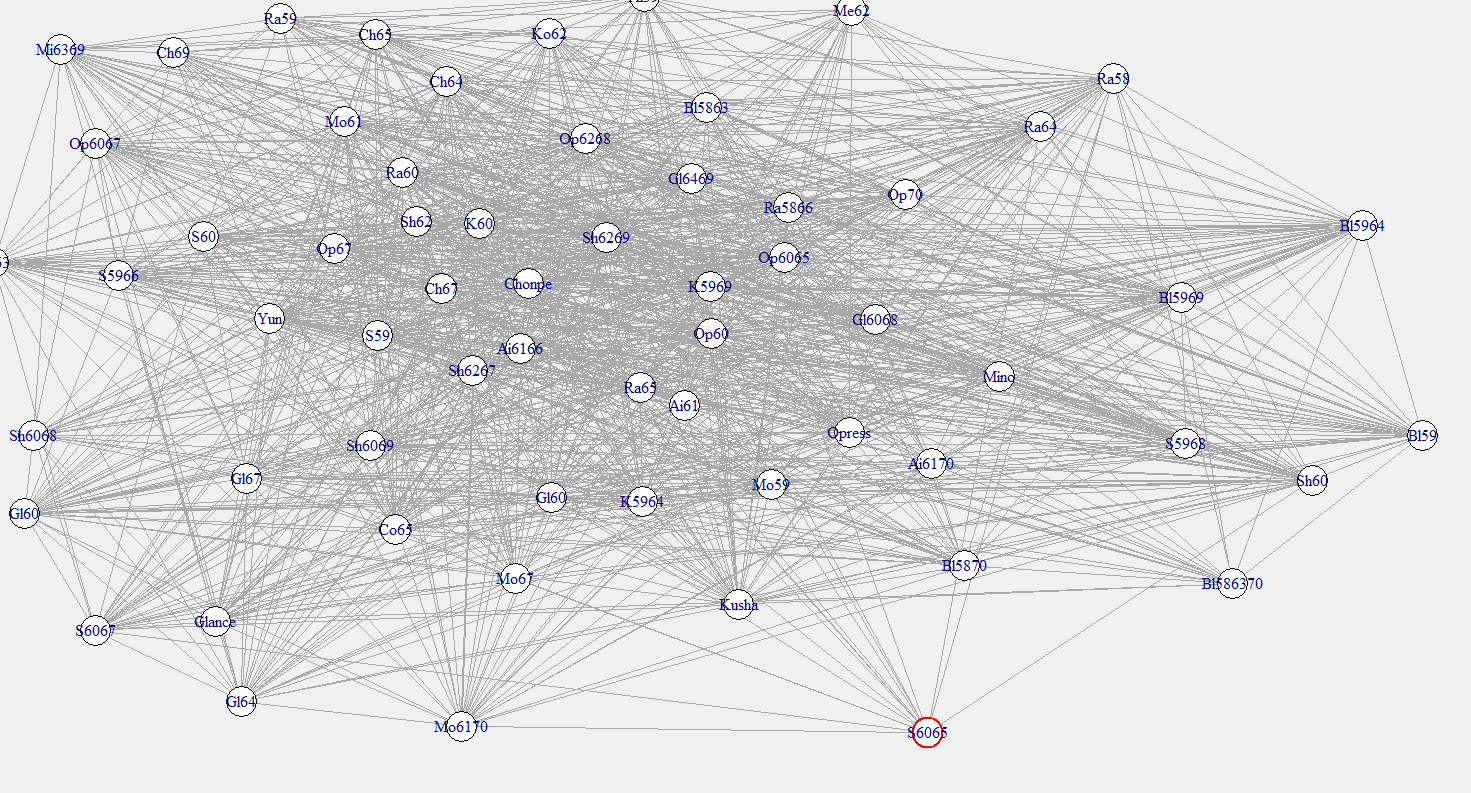
Forecasting and Prediction

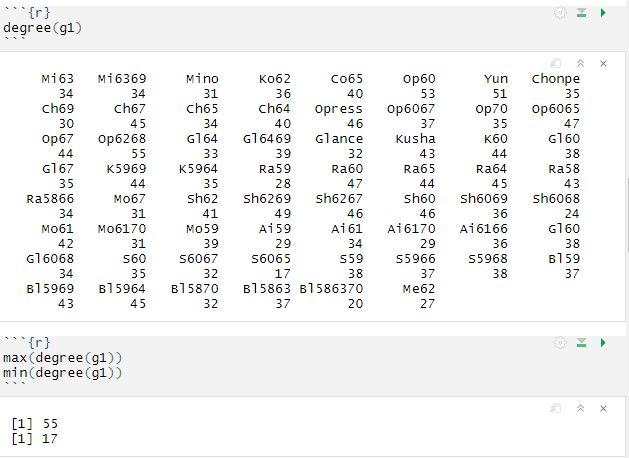
1. Social network analysis

For this analysis, I used the mac dataset to analyze the relationships between 62 Japanese macaques. The initial plot of this data shows clusters that are highly connected, and too close together to examine fully. Applying tkplot, we can interreact with each node, see the relationships a bit clearer, and identify the connectivity of each node. However, seeing as the data is largely connected, not all of the analysis can be done visually.

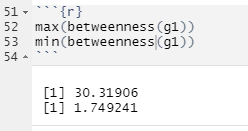




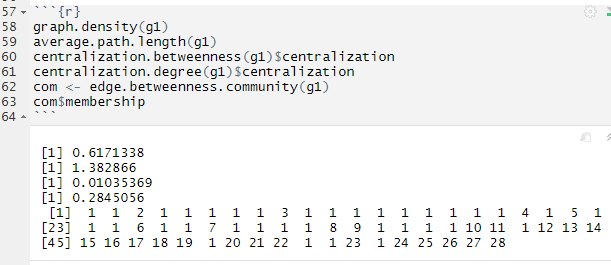
Using degree to quantify the relationships for each nod, I found the most connected as Op6268 with 55, and the least as S6065 with 17. These numbers make sense when looking back at the tkplot, the node in red is the least connected, and all other nodes have between 17 and 55 connections which explains the overlaps.

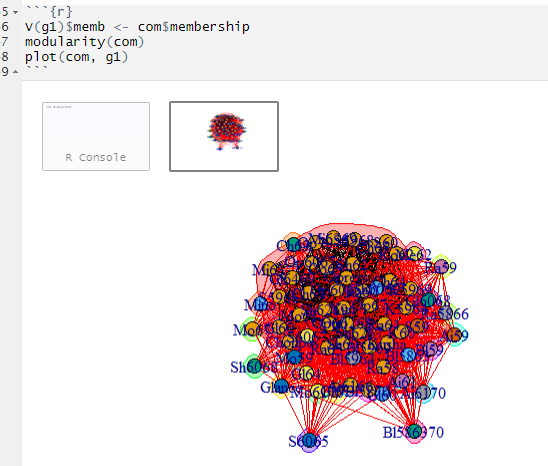


Applying betweenness showed a range of importance between 1.75 and 30.32, both being the same two nodes mentioned from the degree. Since that range is somewhat small, and there are very few nodes with an importance of 6 or lower, this is also a good example that shows the nodes are highly connected.



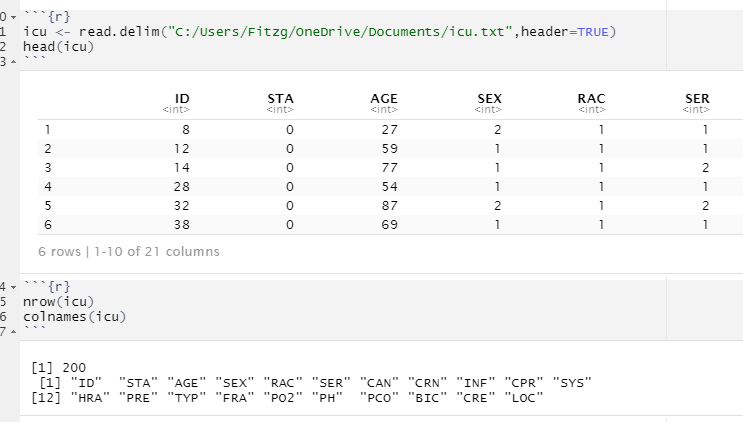
Using the cliques function was not very useful in identifying specific groups that appear to associate more often because the least results I got with a min of 13 was 50. Anything more than 13 didn’t show any results, and anything less than that gave too many results to analyze, which again points to the data being highly connected. As expected, graph density shows a fairly high value at .62, and average path length of 1.38 shows the nodes are closely related. A very low centralization shows the nodes are relatively evenly distributed. Community shows that there are many clusters that include 10+ other nodes, again showing high connectivity.



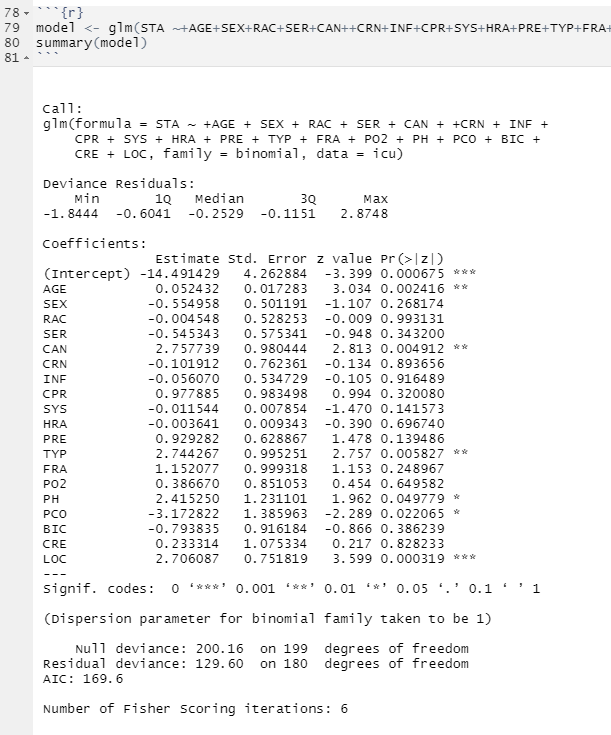


2. Supervised Prediction

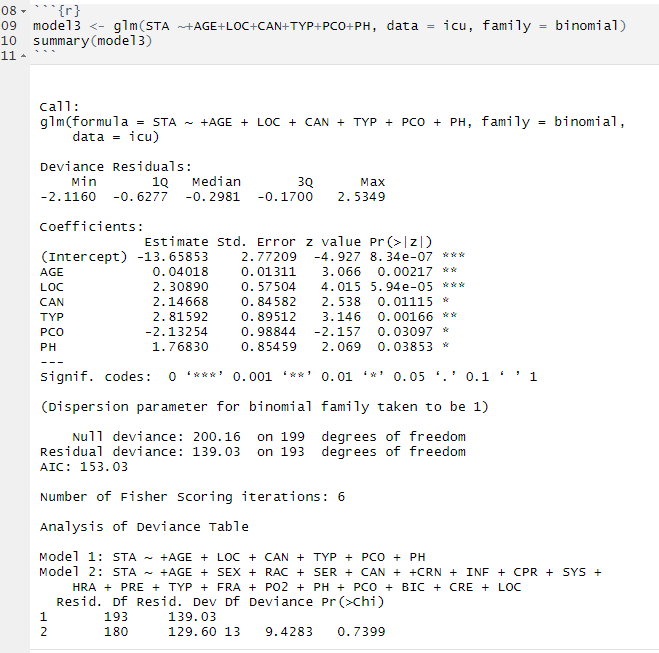
Using the ICU data set to predict survival rates on patients admitted to the ICU after discharge, I applied logistic regression in R. Importing the data, and finding the number of rows I ran the following code.



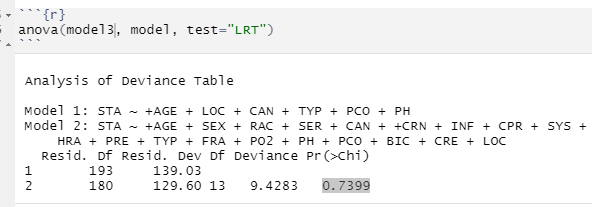
Using the glm function in R returned the following regression results:



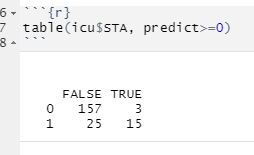
I decided to identify the variables with higher P values, and decided to run a new model without those variables. Those variables were PRE, SYS, FRA, SEX, CPR, SER, BIC, PO2, HRA, CRE, CRN, INF, and RAC. Running the new model in glm calculated the below results. As you can see, the deviance residuals are more closely mirrored than the original model. The residual deviance and the AIC value are also both smaller, indicating a better fit.

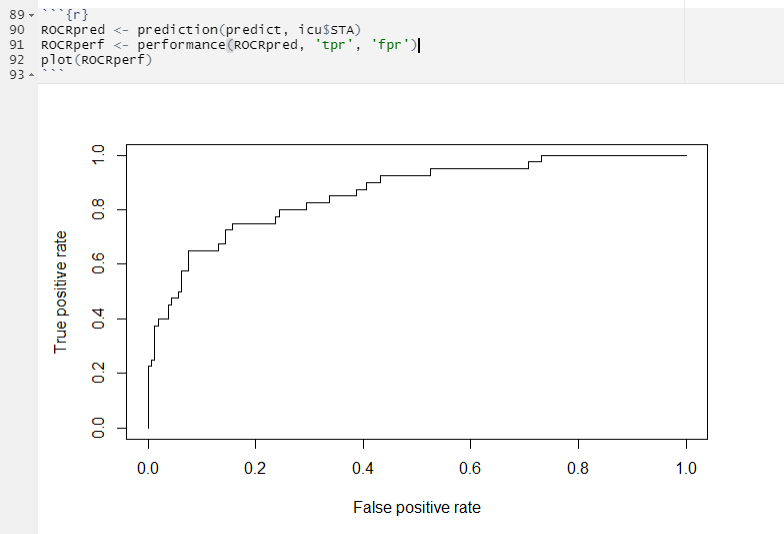


To test this further, I used the anova function with the likelihood ratio test in R to compare the two models. You can see the two models are displayed showing which variables are included, and you can see the difference in residual deviance is relatively low 9.4283, and we show a high P value of 0.7399 which means we can reject the null hypothesis. I did not find that the full model is much better than the model with those selected variables dropped, which means I can drop those variables.



This model made 157+15=172 correct predictions and 28 incorrect predictions. The accuracy rate for predicting vital status lived (0) is 157/160 or 98%, and the accuracy rate for predicting vital status died is 15/40 or 37.5%. This is only slightly worse than the full model with all variables, which shows the other variables have little to no influence on the accuracy of the predictions.

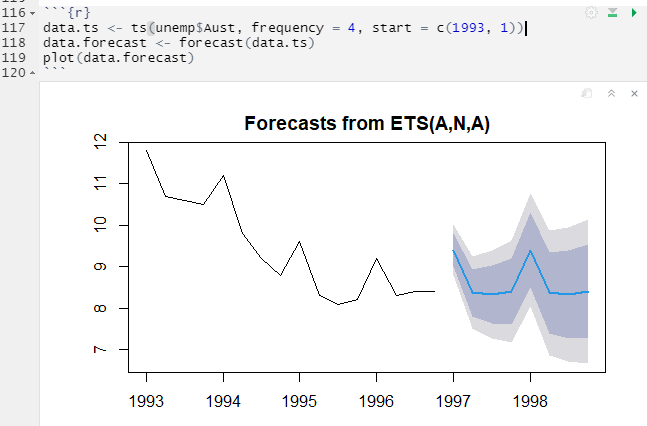




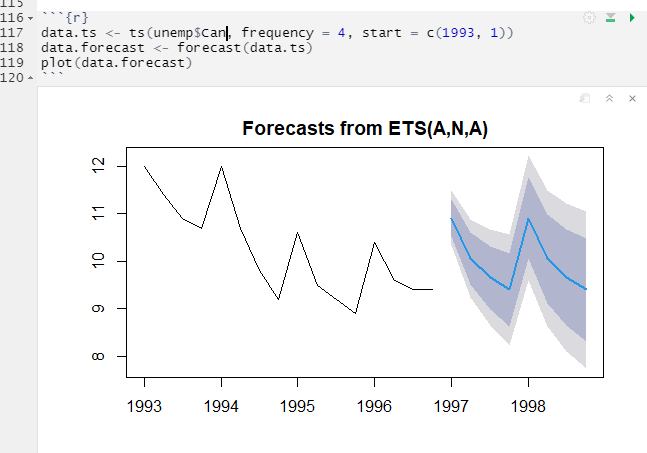
Our graph passes fairly close to the top left corner, which indicates that we have a good true positive rate and a low false positive rate.

3. Forecasting

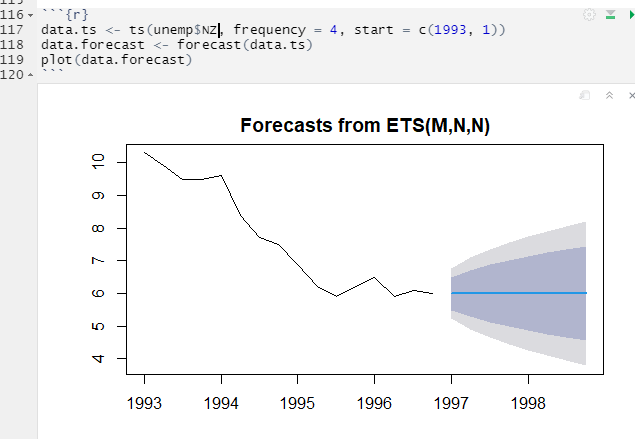
**Aust**



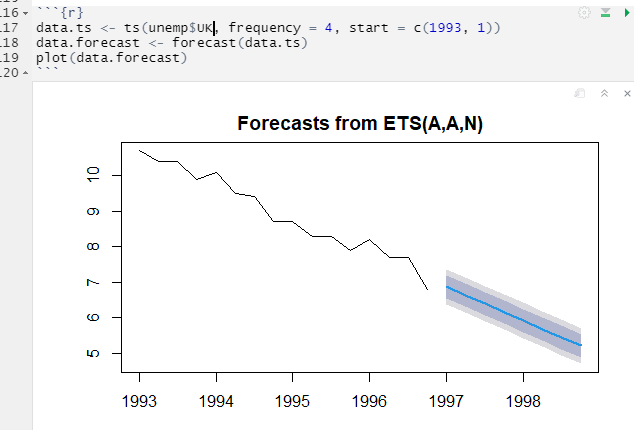
**Canada**



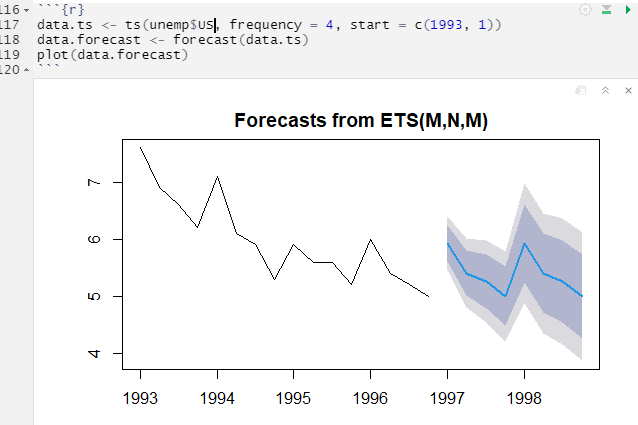
**New Zealand**



**UK**



**US**



Using the unemployment figures found in the “unemfive” dataset, I created a time series of the unemployment numbers for the countries listed, and plotted them on a quarterly basis. The model predicts Australia’s unemployment to spike up to roughly 9.5 plus or within range of the blue shading, and then fall back to the amount that we about saw in 1996 before starting to climb again. A similar pattern was predicted with Canada starting with a spike and falling back to around where the spike started, before going up again. The New Zealand forecast seemed to be a curveball since R calculated that it could do anything, and the blue shading looks wider than the other models, which supports the idea that R is unsure of the forecast. The UK model predicts the unemployment rate will continue to drop throughout the course of the next 4 quarters. This is likely due to the fact that it had been dropped relatively consistently since quarter 1 of 1993 with only small peaks towards getting higher. Finally, the US model shows that it is due for a spike in unemployment based on the past 4 years, but that spike will taper off throughout the next 4 quarters.