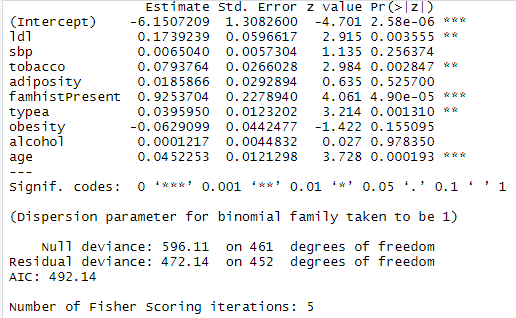
Heart Disease Assignment

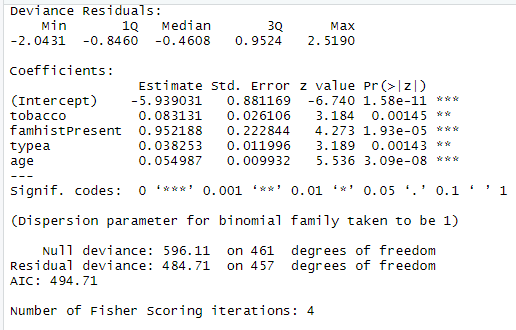
Cardiovascular Heart Disease is one of the leading causes of death among adults in the United States. Using the heart.csv dataset, it can be analyzed what factors contribute the most to cardiovascular heart disease. The factors analyzed are Systolic Blood Pressure, Tobacco use, LDL cholesterol, Adiposity, Family History of chd, stress level (type A), Obesity, Alcohol use, and Age. Using logistic regression, the factors will be analyzed to predict chd based on the significance of the factors.

**Full Model**



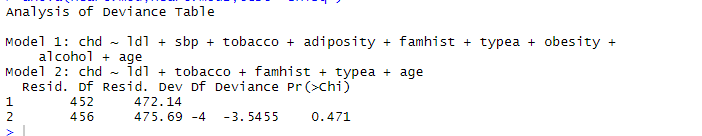
Using all the predictors present in the dataset, the full model is show with ldl, tobacco, family history, typea, and age to be statistically significant with family history and age being slightly more statistically significant than the other predictors.

**Step Reduced Model**

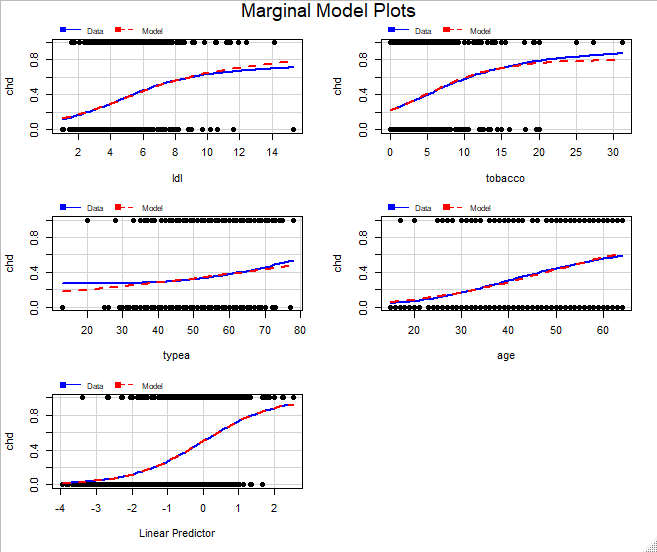


Using the step command, all the predictors that weren’t statistically significant are filtered out and the model is trimmed down to only the most significant ones. Based on this model we can use these predictors as the best method of predicting chd in men.

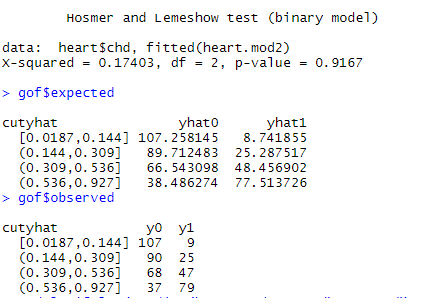
**Comparison Between Models**



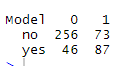
Comparing the two models using a chi squared test, we can see that the difference from the residual deviance between the models is 3.5455 and the P value is 0.471 meaning that the trimmed down model is less statistically significant. This could be due to the fact that the predictors in the stepped model are all so statistically significant that it views none of them as significant because without the predictors it would be hard to predict chd in men.



Plotting the Marginal model plots, it can be seen that the predictors when comparing the observed values from the model versus the actual values from the data that the model follows the expected predictors very closely and in some cases like in tobacco, age, and ldl are almost indistinguishable.



**Confusion Matrix**



Out of all the predictions 256+87 = 353/462 = 0.764 were correct so the predictions were right 76.4% of the time.

In conclusion, the study conducted had a wide range of possible predictors to cardiovascular heart disease. The most predictive variables that came out of the study were ldl, tobacco use, type a stress, and age, so it perhaps could be used as a real world model if the study was conducted with random samples although not specified. All things considered, both models indicated that a family history of cardiovascular heart disease and age were the most significant variables that contributed to actually having the disease, which are easy variables to test for.