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

```
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.1507209  1.3082600  -4.701 2.58e-06 ***
ldl             0.1739239  0.0596617   2.915 0.003555 **
sbp             0.0065040  0.0057304   1.135 0.256374
tobacco         0.0793764  0.0266028   2.984 0.002847 **
adiposity       0.0185866  0.0292894   0.635 0.525700
famhistPresent  0.9253704  0.2278940   4.061 4.90e-05 ***
typea           0.0395950  0.0123202   3.214 0.001310 **
obesity        -0.0629099  0.0442477  -1.422 0.155095
alcohol         0.0001217  0.0044832   0.027 0.978350
age             0.0452253  0.0121298   3.728 0.000193 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.14  on 452  degrees of freedom
AIC: 492.14

Number of Fisher Scoring iterations: 5
```

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

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0431  -0.8460  -0.4608   0.9524   2.5190

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.939031   0.881169  -6.740 1.58e-11 ***
tobacco         0.083131   0.026106   3.184  0.00145 **
famhistPresent  0.952188   0.222844   4.273 1.93e-05 ***
typea           0.038253   0.011996   3.189  0.00143 **
age             0.054987   0.009932   5.536 3.09e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 484.71  on 457  degrees of freedom
AIC: 494.71

Number of Fisher Scoring iterations: 4
```

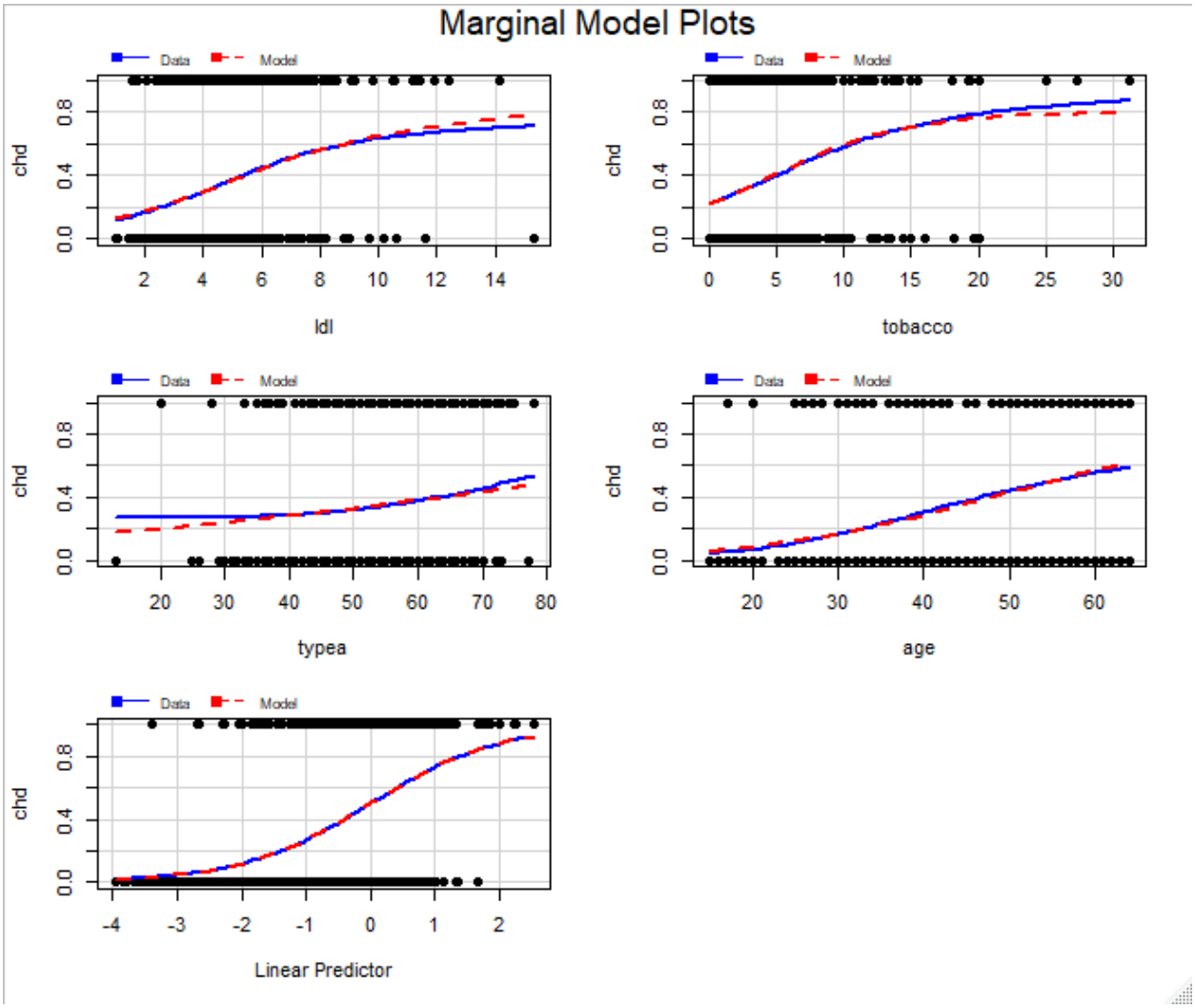
Using the step command, all the predictors that weren’t statistically significant are filtered out and the model is trimmed down to only the most significant ones. Based on this model we can use these predictors as the best method of predicting chd in men.

Comparison Between Models

```
Analysis of Deviance Table

Model 1: chd ~ ldl + sbp + tobacco + adiposity + famhist + typea + obesity +
    alcohol + age
Model 2: chd ~ ldl + tobacco + famhist + typea + age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       452     472.14
2       456     475.69 -4  -3.5455    0.471
> |
```

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.

```
        Hosmer and Lemeshow test (binary model)

data:  heart$chd, fitted(heart.mod2)
X-squared = 0.17403, df = 2, p-value = 0.9167

> gof$expected

cutyhat               yhat0      yhat1
  [0.0187,0.144] 107.258145   8.741855
  (0.144,0.309]   89.712483  25.287517
  (0.309,0.536]   66.543098  48.456902
  (0.536,0.927]   38.486274  77.513726
> gof$observed

cutyhat           y0  y1
  [0.0187,0.144] 107   9
  (0.144,0.309]   90  25
  (0.309,0.536]   68  47
  (0.536,0.927]   37  79
```

Confusion Matrix

```
Model   0   1
  no  256  73
  yes  46  87
```

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.