Final Project: Analysis

Julie Osborne

5/10/2019

Full Model

```
##
## Call:
## glm(formula = death ~ sex + education + alcoholfreq + exercise +
##     race, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2778  -0.6462  -0.5053  -0.3290   2.5090
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.27899    0.27480  -1.015 0.309971
## sex1         -0.63016    0.17163  -3.672 0.000241 ***
## education2   -0.63012    0.21252  -2.965 0.003027 **
## education3   -1.29079    0.21064  -6.128 8.90e-10 ***
## education4   -1.46143    0.39152  -3.733 0.000189 ***
## education5   -1.49688    0.32854  -4.556 5.21e-06 ***
## alcoholfreq1 -0.30000    0.26775  -1.120 0.262523
## alcoholfreq2 -0.79881    0.23645  -3.378 0.000729 ***
## alcoholfreq3 -0.05473    0.24094  -0.227 0.820304
## alcoholfreq4  0.06304    0.25719   0.245 0.806375
## exercise1     0.10939    0.23297   0.470 0.638671
## exercise2     0.44896    0.22988   1.953 0.050822 .
## race1        -0.10478    0.23291  -0.450 0.652801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1127.4  on 1140  degrees of freedom
## Residual deviance: 1024.0  on 1128  degrees of freedom
## AIC: 1050
##
## Number of Fisher Scoring iterations: 5
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: death
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
```

```
## NULL                         1140     1127.3
## sex          1   15.895      1139     1111.5 6.695e-05 ***
## education    4   63.144      1135     1048.3 6.329e-13 ***
## alcoholfreq  4   18.753      1131     1029.6 0.0008786 ***
## exercise     2    5.357      1129     1024.2 0.0686693 .
## race         1    0.205      1128     1024.0 0.6508918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the model summary, sex1, education2, education3, education4, education5, and alcoholfreq2 are significant predictors at the 5% level. Then, the ANOVA was run to determine which variables were significant to the model not at their individual levels. These variables are: sex, education, and alcoholfreq.

First Reduced Model

```
## Analysis of Deviance Table
##
## Model 1: death ~ sex + education + alcoholfreq + exercise + race
## Model 2: death ~ sex + education + alcoholfreq
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1128     1024.0
## 2      1131     1029.6 -3  -5.5617    0.135
```

The first reduced model used the variables that were significant from chi-squared test produced from the ANOVA test. This reduced model was then compared to the full model using ANOVA to see if it was a better fit. This produces a p-value of $0.135 > 0.05$. This means that the reduced model is not a better fit than the full model.

Second Reduced Model

```
## Analysis of Deviance Table
##
## Model 1: death ~ sex + education + alcoholfreq + exercise + race
## Model 2: death ~ sex + education
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1128     1024.0
## 2      1135     1048.3 -7  -24.315 0.001003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The second reduced model used the variables had all levels significant in the full model. Again, this reduced model was compated to the full model using ANOVA to see if it was a better fit. This produced a p-valued of $0.0010028 < 0.05$. This means that the reduced model is a better fit than the full model, and will be used for prediction comparison.

```
##
## Call:
## glm(formula = death ~ sex + education, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0610  -0.6291  -0.5353  -0.4142   2.2457
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.2801     0.1503  -1.864 0.062329 .
## sex1         -0.5424     0.1595  -3.401 0.000671 ***
## education2   -0.6969     0.2057  -3.388 0.000704 ***
## education3   -1.3669     0.2005  -6.819 9.18e-12 ***
## education4   -1.6155     0.3802  -4.248 2.15e-05 ***
## education5   -1.5902     0.3179  -5.003 5.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1127.4  on 1140  degrees of freedom
## Residual deviance: 1048.3  on 1135  degrees of freedom
## AIC: 1060.3
##
## Number of Fisher Scoring iterations: 4
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: death
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       1140     1127.3
## sex        1   15.895      1139     1111.5 6.695e-05 ***
## education  4   63.144      1135     1048.3 6.329e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary of the reduced model and the anova test confirm that all levels of sex and education are significant in predicting the likelihood of death.

Prediction

The full model prediction accuracy is 0.8094262.
The reduced model prediction accuracy is 0.8053279.