13.1

Nikhil Gopal

1/26/2022

Problem 13.1

Fitting logistic regression to data: The folder NES contains the survey data of presidential preference and income for the 1992 election analyzed in Section 13.1, along with other variables including sex, ethnicity, education, party identification, and political ideology.

```
rm(list = ls())
library(rosdata)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
data <- data(nes)
```

13.1 (a)

Fit a logistic regression predicting support for Bush given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.

sex, ethnicity, education, party identification, and political ideology.

Fit logistic regression model:

```
logistic <- glm(rep_presvote ~ female + race + educ1 + educ2 + educ3, + partyid3 + ideo_feel,
             data = nes, family = binomial)

summary(logistic)
```

```
##
## Call:
## glm(formula = rep_presvote ~ female + race + educ1 + educ2 +
##     educ3, family = binomial, data = nes, weights = +partyid3 +
##     ideo_feel)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -15.054   -8.715    5.461    7.728   17.600
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.139763   0.010707  13.053  < 2e-16 ***
## female      -0.252220   0.005767 -43.733  < 2e-16 ***
## race        -0.363386   0.003807 -95.464  < 2e-16 ***
## educ1       -0.053504   0.010183  -5.254 1.48e-07 ***
## educ2        0.556423   0.011039  50.406  < 2e-16 ***
## educ3       -0.319287   0.010488 -30.443  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 718573  on 9730  degrees of freedom
## Residual deviance: 692846  on 9725  degrees of freedom
##   (25177 observations deleted due to missingness)
## AIC: 692858
##
## Number of Fisher Scoring iterations: 5
```

Normal regression:

```
ols <- lm(rep_presvote ~ female + race + educ1 + educ2 + educ3, + partyid3 + ideo_feel,
             data = nes)

summary(ols)
```

```
##
## Call:
## lm(formula = rep_presvote ~ female + race + educ1 + educ2 + educ3,
##     data = nes, subset = +partyid3 + ideo_feel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9158 -0.5178  0.2832  0.2937  0.6177
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.472876   0.036317   40.56   <2e-16 ***
## female      -0.135477   0.007320  -18.51   <2e-16 ***
## race        -0.705190   0.034693  -20.33   <2e-16 ***
## educ1        0.273049   0.013041   20.94   <2e-16 ***
## educ2       -0.198998   0.007311  -27.22   <2e-16 ***
## educ3              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4556 on 20124 degrees of freedom
##   (14779 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.07176,    Adjusted R-squared:  0.07158
## F-statistic:   389 on 4 and 20124 DF,  p-value: < 2.2e-16
```

Interactions:

```
interactions <- glm(rep_presvote ~ female + race + educ1 + educ2 + educ3, + partyid3 + ideo_feel + femal
             data = nes, family = binomial)

summary(interactions)
```

```
##
## Call:
## glm(formula = rep_presvote ~ female + race + educ1 + educ2 +
##     educ3, family = binomial, data = nes, weights = +partyid3 +
##     ideo_feel + female * race + educ1 * partyid3 + race * partyid3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.523   -9.249    5.922    8.195   18.696
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  0.134531   0.009979   13.482  < 2e-16 ***
## female      -0.263690   0.005403  -48.805  < 2e-16 ***
## race        -0.335920   0.003324 -101.056  < 2e-16 ***
## educ1       -0.046220   0.009562   -4.833 1.34e-06 ***
## educ2        0.567140   0.010198   55.615  < 2e-16 ***
## educ3       -0.330435   0.009644  -34.263  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 822779  on 9730  degrees of freedom
## Residual deviance: 793162  on 9725  degrees of freedom
##   (25177 observations deleted due to missingness)
## AIC: 793174
##
## Number of Fisher Scoring iterations: 7
```

13.1 (b)

Evaluate and compare the different models you have fit.

```
lrtest(logistic, interactions)
```

```
## Likelihood ratio test
##
## Model 1: rep_presvote ~ female + race + educ1 + educ2 + educ3
## Model 2: rep_presvote ~ female + race + educ1 + educ2 + educ3
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -346423
## 2   6 -396581  0 100316  < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Above I ran a likelihood ratio test to test model fit between the logistic regression model and the logistic regression model with interactions. I decided not to consider the normal regression model as it had an R^2 value of 0.07 and thus I could determine quickly that the model had poor fit.

Above, I tested the null hypothesis that the logistic regression model fit the data better than the interaction model. The test returned an extremely small near zero value below our traditional threshold of 0.05. Thus, we can reject the null hypothesis that the normal logistic regression model fits the data better and conclude that the interaction model fits the data best.

13.1 (c)

For your chosen model, discuss and compare the importance of each input variable in the prediction.

```
summary(interactions)
```

```
##
## Call:
## glm(formula = rep_presvote ~ female + race + educ1 + educ2 +
##     educ3, family = binomial, data = nes, weights = +partyid3 +
##     ideo_feel + female * race + educ1 * partyid3 + race * partyid3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.523   -9.249    5.922    8.195   18.696
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  0.134531   0.009979   13.482  < 2e-16 ***
## female      -0.263690   0.005403  -48.805  < 2e-16 ***
## race        -0.335920   0.003324 -101.056  < 2e-16 ***
## educ1       -0.046220   0.009562   -4.833 1.34e-06 ***
## educ2        0.567140   0.010198   55.615  < 2e-16 ***
## educ3       -0.330435   0.009644  -34.263  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 822779  on 9730  degrees of freedom
## Residual deviance: 793162  on 9725  degrees of freedom
##   (25177 observations deleted due to missingness)
## AIC: 793174
##
## Number of Fisher Scoring iterations: 7
```

```
confint(interactions)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %      97.5 %
## (Intercept)  0.11497504  0.15409052
```

```
## female      -0.27428049 -0.25310148
## race        -0.34244247 -0.32941218
## educ1       -0.06496455 -0.02748029
## educ2        0.54715319  0.58712730
## educ3       -0.34933726 -0.31153322
```

We see here that being female decreases the log odds of voting for bush by a a factor of 0.263690. For every 1 increase in cateogries of race, education1 and education 3, the log odds of voting for bush decrease by factor of -0.335920, -0.046220 and -0.330435 respectively. Finally, education2 turned out to be the most influential variable. For every 1 unit increase in education 2, the log odds of voting for Bush increased by a factor of 0.567140.