logistic regression modeling

Justin Lo

2023-08-16

In this project, I explore logistic regression modeling using a dataset from the 2016 European Social Survey [(https://www.europeansocialsurvey.org)]

The data includes the following variables:

Table 1: ESS codebook

Variable	Description
`country_code`	The country of the respondent
`leave`	1 if the respondent would vote to leave the European Union in a referendum, 0 otherwise
`gender`	Whether the respondent is male or female
`age`	The age of the respondent (in years)
`years_education`	The number of years of education the respondent has completed
`unemployed`	1 if the respondent is unemployed, 0 otherwise
`income`	1 if the respondent earns above the median income in their country, 0 otherwise
`religion`	Categorical variable of the religion of the respondent
`trade_union`	1 if the respondent is a member of a trade union, 0 otherwise
`news_consumption`	Amount of time the respondent spends reading newspapers/online news each week (in minutes)
`trust_people`	The degree to which the respondent trusts other people (0 = low trust, 10 = high trust)
`trust_politicians`	The degree to which the respondent trusts politicians (0 = low trust, 10 = high trust)
`past_vote`	1 if the respondent voted in the last general election in their country, 0 otherwise
`immig_econ`	The respondent's view of the economic effects of immigration in their country (0 = Immigration is bad for the economy; 10 = Immigration is good for the economy)
`immig_culture`	The respondent's view of the cultural effects of immigration in their country (0 = Immigration undermines the country's culture; 10 = Immigration enriches the country's culture)
`country_attach`	The respondent's emotional attachment to their country (0 = Not at all emotionally attached; 10 = Very emotionally attached)
`climate_change`	How worried the respondent is about climate change (1 = Not at all worried; 5 = Very worried)
`imp_tradition`	How important the respondent feels it is to follow traditions and customs (1 = Very important; 6 = Not at all important)
`imp_equality`	How important the respondent feels it is people are treated equally and have equal opportunities (1 = Very important; 6 = Not at all important)
`eu_integration`	The respondent's views on European unification/integration (0 = "Unification has already gone too far"; 10 = "Unification should go much further")
`train`	A variable indicating whether the respondent should be used in the training set (`TRUE`) or the test set (`FALSE`).

Reading the dataset into r

```
ess <- read.csv("https://raw.githubusercontent.com/lse-me314/lse-me314.github.io/master/data/ess.csv")
```

Fitting a logistic regression model with 'leave' as response and with 'age', 'gender', 'years_education' and income as predictors.

```
logistic_model_1<- glm(leave ~ age + gender +years_education + income, family = 'binomial', data= ess)
summary(logistic_model_1)
```

```
##
## Call:
## glm(formula = leave ~ age + gender + years_education + income,
##     family = "binomial", data = ess)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.927116   0.125971  -7.360 1.84e-13 ***
## age              0.001681   0.001355   1.240   0.2148
## genderMale       0.100327   0.046347   2.165   0.0304 *
## years_education -0.048847   0.006433  -7.593 3.13e-14 ***
## income          -0.293550   0.050183  -5.850 4.93e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12188  on 13074  degrees of freedom
## Residual deviance: 12027  on 13070  degrees of freedom
## AIC: 12037
##
## Number of Fisher Scoring iterations: 4
```

Overall, the model indicates that gender, years_education, and income are statistically significant predictors of leave. However, age does not appear to have a statistically significant relationship with leave.

To calculate the predicted probability of voting to leave the EU for respondonts with the following characteristics 1. A 25-year old man, with above median income, and 10 years of education 2. A 25-year old woman, with above median income, and 15 years of education 3. A 65-year old woman, with below median income, and 8 years of education 4. A 65-year old man, with below median income, and 12 years of education

```
df1<- data.frame(age=25, income= 1, years_education=10, gender= 'Male')
df2<- data.frame(age=25, income= 1, years_education=15, gender= 'Female')
df3<- data.frame(age=65, income= 0, years_education=8, gender= 'Female')
df4<- data.frame(age=65, income= 0, years_education=12, gender= 'Male')
pd1<-predict(logistic_model_1, df1, type = 'response')
pd2<-predict(logistic_model_1, df2, type = 'response')
pd3<-predict(logistic_model_1, df3, type = 'response')
pd4<-predict(logistic_model_1, df4, type = 'response')
pd1
```

```
##         1
## 0.1726746
```

```
pd2
```

```
##         1
## 0.1288289
```

```
pd3
```

```
##         1
## 0.2299428
```

```
pd4
```

```
##         1
## 0.2135435
```

A 65-year old woman, with below median income, and 8 years of education has the highest probability of voting to leave the EU based on the model A 25-year old woman, with above median income, and 15 years of education has the lowest probability of voting to leave the EU based on the model

Now to calculate the predicted probability of every observations in the dataset.

```
prediction1<- predict(logistic_model_1, type = "response")
head(prediction1)
```

```
##         1         2         3         4         5         6
## 0.1903361 0.1782114 0.1484992 0.1216273 0.1624984 0.2190983
```

To update the model, including the country code

```
logistic_model_2<- glm(leave ~ age + gender +years_education + income + country_code, family = 'binomial
prediction2<- predict(logistic_model_2, type = "response")
head(prediction2)
```

```
##         1         2         3         4         5         6
## 0.2369347 0.2491320 0.1864658 0.1559886 0.2115764 0.2869778
```

```
count<- sum(prediction2>0.5)
count
```

```
## [1] 274
```

274 observations have probability greater than 0.5

```
high_probability_observations <- ess[prediction2 > 0.5, ]
high_probability_observations$country_code
```

```
##   [1] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
##  [16] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
##  [31] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
##  [46] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
##  [61] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
##  [76] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
##  [91] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [106] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [121] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [136] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [151] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [166] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [181] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [196] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [211] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [226] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [241] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [256] "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB" "GB"
## [271] "GB" "GB" "GB" "GB"
```

```
print(unique(high_probability_observations$country_code))
```

```
## [1] "GB"
```

GB is the only country with respondents with a probability of voting to leave of greater than 0.5