# Candies – Logistic Regression and probability

Made by majafoi

Updated 4/18/2021, made in R

```
library (readr) # to load the package readr from the library
halloween<-read.csv ("candy-data.csv") # to read the CSV file
cs = complete.cases (halloween) #since there is a possibility that this original dataset has some mis sing data or NAs (not available data), I will use the function complete.cases in order to remove that part of the dataset.
Halloween = halloween[cs,]
attach (halloween) #I will only work with a complete dataset, or dataset without missing data and NAs.
```

Hello, everyone!

This time I'm performing **logistic regression**, which is fairly simple to conduct. Also, I won't be doing summaries (aka describing them) and graphic notions of all variables, as I have done those a lot of times in my past research, here on GitHub and overall.

As said on Kaggle, data was collected by creating a website where the participants were shown two candies and they were asked to click on one they would prefer to receive for Halloween.

**Acknowledgements:**

Link to Kaggle: https://www.kaggle.com/fivethirtyeight/the-ultimate-halloween-candy-power-ranking. This dataset is Copyright (c) 2014 ESPN Internet Ventures and distributed under an MIT license. Check out the analysis and write-up here: The Ultimate Halloween Candy Power Ranking. Thanks to Walt Hickey for making the data available.

The main goal here is to try to predict if one of those candies was a chocolate one, based on other variables in the dataset.

You can see the variable's description in the README file in the repository.

All of the variables, beside pricepercent, sugarpercent and winpercent, have a value of a 0 or a 1.

```
summary (pricepercent)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.2600  0.4700  0.4691  0.6500  0.9800

summary (sugarpercent)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.2200  0.4700  0.4782  0.7300  0.9900

summary (winpercent)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.45   39.14   47.83   50.32   59.86   84.18
```

Next, we have the formula for logistic regression - we are trying to predict a probability of a binary event occurring (male vs. female, 0 vs. 1, etc).

First, let's make a model and include all other variables in it as x values, whereas the chocolate variable will be our y variable.

```
m=glm (chocolate~fruity,family=binomial) # this is the formula for logistic regression between ch
ocolate and fruity variables (x~y).
summary(m)

##
## Call:
## glm(formula = chocolate ~ fruity, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.7043  -0.2309  -0.2309   0.7302   2.6972  #here we can see the distribution of residuals betwe
en those variables (residuals are difference between normal model and predictive one – difference)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.1856     0.3445   3.441 0.000579 ***
## fruity       -4.7965     1.0704  -4.481 7.42e-06 *** #this p-value shows us that the fruity variable
is significant in this model (it should be lower than 0.05 value which I have selected for alpha value)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 116.407  on 84  degrees of freedom
## Residual deviance:  60.395  on 83  degrees of freedom
## AIC: 64.395
##
## Number of Fisher Scoring iterations: 6

Newdata = data.frame(fruity=1) #if the fruity variable is 1
predict(m,type="response",newdata=newdata)

##         1
## 0.02631579 #probability that the candy is a chocolate one (2%)

newdata=data.frame(fruity=0) #if the fruity variable is 0
predict(m,type="response",newdata=newdata)

##         1
## 0.7659574 #probability that the candy is a chocolate one (76.59%)
```

The fruity variable is significant in this regression model, because the p value is set to be 0.05. If the value of p is lower than that in the model itself, than the variable in the model is significant (has meaning) and we can proceed.

Next, we want to predict if the candy is chocolate one, based on the fruity variable. If the candy is fruity, then the probability of a chocolate candy is 2%. If not, the probability is 76.59%. That makes sense, because… When was the last time you ate a chocolate candy and it being fruity as well? :)

Next we have the caramel in the candy-variable.

```
m=glm (chocolate~caramel, family=binomial) #formula for the x~y aka chocolate and caramel.
summary(m)

##
## Call:
## glm(formula = chocolate ~ caramel, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.5829  -0.9782  -0.9782   1.3906   1.3906
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4884     0.2445  -1.998   0.0458 *
## caramel       1.4046     0.6401   2.194   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.41  on 84  degrees of freedom
## Residual deviance: 111.07  on 83  degrees of freedom
```

```
## AIC: 115.07
##
## Number of Fisher Scoring iterations: 4

newdata=data.frame(caramel=1)
predict(m,type="response",newdata=newdata)

##       1
## 0.7142857

newdata=data.frame(caramel=0)
predict(m,type="response",newdata=newdata)

##       1
## 0.3802817
```

Again, caramel variable is significant and we can proceed with prediction. The probability of a candy being chocolate one, if there is caramel in the candy, is 71.42%. If there is no caramel in the candy, the probability is 38%.

Next we have peanuts, peanut butter or almonds in the candy-option.

```
m=glm (chocolate~peanutyalmondy, family=binomial)
summary(m)

##
## Call:
## glm(formula = chocolate ~ peanutyalmondy, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.9728  -0.9317  -0.9317   1.4449   1.4449
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.6098     0.2485  -2.454  0.01413 *
## peanutyalmondy   2.4015     0.8032   2.990  0.00279 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 116.41  on 84  degrees of freedom
## Residual deviance: 103.60  on 83  degrees of freedom
## AIC: 107.6
##
## Number of Fisher Scoring iterations: 4

newdata=data.frame(peanutyalmondy=1)
predict(m,type="response",newdata=newdata)
```

```
##         1
## 0.8571429
```

```
newdata=data.frame(peanutyalmondy=0)

predict(m,type="response",newdata=newdata)
```

```
##         1
## 0.3521127
```

The variable is significant in the model. If a candy has peanuts, peanut butter or almonds in it, the probability of the candy being a chocolate one is 85.71%, if not, only 35.21%.

Next is the nougat option.

```
m=glm (chocolate~nougat, family=binomial)
summary(m)

##
## Call:
## glm(formula = chocolate ~ nougat, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.973  -1.007  -1.007   1.358   1.358
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4162     0.2314  -1.799   0.0721 .
## nougat        2.2079     1.1046   1.999   0.0456 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.41  on 84  degrees of freedom
## Residual deviance: 110.57  on 83  degrees of freedom
## AIC: 114.57
##
## Number of Fisher Scoring iterations: 4

newdata=data.frame(nougat=1) #if the candy has nougat in it
predict(m,type="response",newdata=newdata)

##         1
## 0.8571429

newdata=data.frame(nougat=0) #if the candy has no nougat in it
predict(m,type="response",newdata=newdata)

##         1
## 0.3974359
```

The variable is significant, although the p-value being very close to 0.05 border. If the candy contains nougat, the probability of a candy being a chocolate one is 85.71%, if not, only 39.74%.

What about the hardiness of the candy?

```
m=glm (chocolate~hard, family=binomial)
summary(m)

##
## Call:
## glm(formula = chocolate ~ hard, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.2018  -1.2018  -0.3715   1.1532   2.3272
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05716    0.23914   0.239   0.8111
## hard        -2.69622    1.06236  -2.538   0.0112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 116.41  on 84  degrees of freedom
## Residual deviance: 104.33  on 83  degrees of freedom
## AIC: 108.33
##
## Number of Fisher Scoring iterations: 5

newdata=data.frame(hard=1) #if the candy is hard
predict(m,type="response",newdata=newdata)

##         1
## 0.06666667

newdata=data.frame(hard=0) #if the candy isn't hard
predict(m,type="response",newdata=newdata)

##         1
## 0.5142857
```

Hardiness of the candy is also a significant variable in the model. If the candy is hard, the probability of it being chocolate is 6%, if not, the probability goes up to 51.42%. Most chocolate candies in Croatia (like chocolates) are hard ones, so this is an unusual finding. But again, it depends on the country from which the contestants come from, since it was an online challenge.

At last, what happens with the probability if the candy is in a shape of a bar?

```
m=glm (chocolate~bar, family=binomial)
summary(m)

##
## Call:
## glm(formula = chocolate ~ bar, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.4676  -0.7858  -0.7858   0.3124   1.6283
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.017      0.283  -3.593 0.000327 ***
## bar            4.013      1.063   3.775 0.000160 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 116.407  on 84  degrees of freedom
## Residual deviance:  82.135  on 83  degrees of freedom
## AIC: 86.135
##
## Number of Fisher Scoring iterations: 5

newdata=data.frame(bar=1) #if the candy has a shape of a bar
predict(m,type="response",newdata=newdata)

##         1
## 0.952381

newdata=data.frame(bar=0) #if the candy doesn't have a shape of a bar
predict(m,type="response",newdata=newdata)

##         1
## 0.265625
```

If a candy is a bar, then the probability of it being chocolate one also is 95%. If not, that probability is still high enough - 26%.