Logistic Regression

Truong Thi An Hai

12/11/2020

Create training data

```
#Import survey.csv
data = read.csv("C:/Users/hp/Desktop/survey.csv")
#create the following additional columns in the survey table:
#price20 - will have the value 1 if the price is $20, 0 otherwise
data$price20 = as.numeric(data$Price==20)
#price30 - will have the value 1 if the price is $30, 0 otherwise
data$price30 = as.numeric(data$Price==30)
#Remove column Price
data = subset(data,select = -c(Price))
#View 5 first rows
head(data,5)
```

```
##   MYDEPV Income Age price20 price30
## 1      1     33  37       0       0
## 2      0     21  55       1       0
## 3      1     59  55       0       1
## 4      1     76  44       1       0
## 5      0     24  37       0       1
```

a.Create a logistic regression model and display only the coefficients of the independent variables.

```
library(ISLR)
#Create logistic regression model
logisModel <- glm(MYDEPV ~ Income + Age + price20 + price30, data = data, family = binomial)
#View summary() of the model
summary(logisModel)
```

```
##
## Call:
## glm(formula = MYDEPV ~ Income + Age + price20 + price30, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0388  -0.5581  -0.2434   0.4178   3.2377
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.02116    0.53244 -11.309  < 2e-16 ***
## Income       0.12876    0.00923  13.950  < 2e-16 ***
## Age          0.03506    0.01179   2.974  0.00294 **
## price20     -0.74418    0.26439  -2.815  0.00488 **
## price30     -2.21028    0.31108  -7.105  1.2e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.81  on 749  degrees of freedom
## Residual deviance:  534.17  on 745  degrees of freedom
## AIC: 544.17
##
## Number of Fisher Scoring iterations: 6
```

Coefficients of the independent variables:

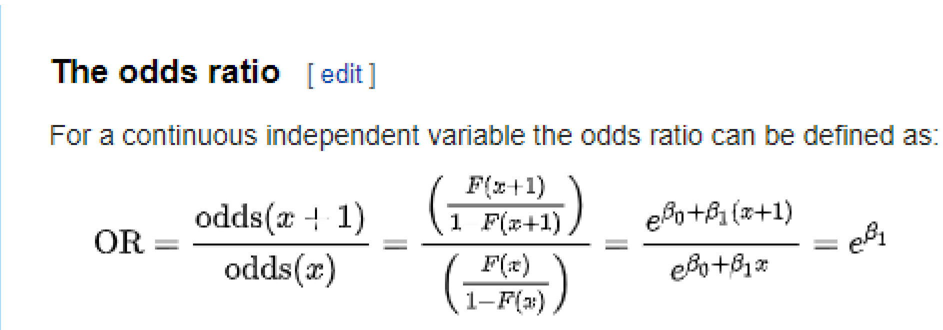
- Intercept: -6.0211606
- Income: 0.1287594
- Age: 0.0350638
- price20: -0.7441775
- price30: -2.2102805

$$p = \frac{1}{1 - \ e^{-(-6.021+0.129*Income+0.035*Age-0.744*price20-2.21*price30)}}$$

Formula of Logistic Regression Model

b. For every unit increase in income while all other independent variables remain constant, by what percentage does the odds-ratio increase/decrease?

For a one unit increase in income, the log odds of accepting an offer increases by 0.129.



The odds ratio [edit]

For a continuous independent variable the odds ratio can be defined as:

$$\mathrm{OR} = \frac{\mathrm{odds}(x+1)}{\mathrm{odds}(x)} = \frac{\left(\frac{F(x+1)}{1\ F(x+1)}\right)}{\left(\frac{F(x)}{1-F(x)}\right)} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

Formula of OR
=> OR = exp(0.129) = 1.1376901

=> For a one unit increase in income, the odds-ratio increases by 13.7690124%

c. If the price of a row of data were to increase from $10 to $30 while all other independent variables remained constant, by what percentage would the odds-ratio increase/decrease?

The price of a row of data were to increase from $10 to $30 while all other independent variables remained constant => changes the log odds of accepting offer by -2.2102805.

=> OR = exp(coefficient_price30) = exp(-2.2102805) = 0.1096699

The odds-ratio decrease by 89.0330115%

Use our logistic regression model to make predictions on the probability of success for each of the rows of data in the survey table.

```
#Add a column named Prob to the survey table that calculates the probability of success of each row of data.
data$Prob <- predict(logisModel, type = "response")
#Add a column named Pred
data$Pred <- ifelse(data$Prob > 0.5,1,0)
#View first 10 rows
head(data,10)
```

```
##    MYDEPV Income Age price20 price30       Prob Pred
## 1       1     33  37       0       0 0.38349464    0
## 2       0     21  55       1       0 0.10594158    0
## 3       1     59  55       0       1 0.78480150    1
## 4       1     76  44       1       0 0.98967880    1
## 5       0     24  37       0       1 0.02096206    0
## 6       0     22  32       1       0 0.05675454    0
## 7       1     28  32       0       0 0.21520067    0
## 8       1     49  38       0       0 0.83486027    1
## 9       0     76  43       0       1 0.95529838    1
## 10      1     59  55       1       0 0.94047412    1
```

d. Test the rule that the probability mass equals the counts. Use the survey table and take the sums of the mydepv column and the prediction column. Are the values equal? Excluding rounding errors, does probability mass equal count?

Sums of the mydepv column

```
Sum_Mydepv = sum(data$MYDEPV)
Sum_Mydepv
```

```
## [1] 324
```

Sum of probability column

```
Sum_Prob = sum(data$Prob)
Sum_Prob
```

```
## [1] 324
```

The sums of mydepv and probability are equal. It proves that the probability mass equals the count, at least in this case.

e. Using the logistic model we have created, what is the likelihood of a person who is 25 years old with an income of $58,000 accepting an offer with a price of $20?

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
x <- data.frame(Income = 58, Age = 25, price20 = 1, price30 =0)
p <- predict(logisModel,x, type = "response")
p
```

```
##         1
## 0.8291054
```

We see that there is 82.9105381% chance that this person will accept an offer with a price of $20