QUANTITATIVE METHODS: R HOMEWORK ASSIGNMENT

Pedro Iraburu Muñoz

29/11/2019

1.0 INTRODUCTION

This homework analyzes a database compound of variables that explains the customer's behavior in online shops (Kaggle, 2019). For this task, I will make a descriptive data analysis prior to a regression logit (and probit) model. I found the database in Kaggle, a webpage that provides a wide number of databases.

This database is provided by the master thesis of Mete Alpaslan Katircioglu, from Bahçeşehir University. It is made of 18 variables, eight of them are categorical and the rest are numerical. The information was collected during a year from 12,300 sessions (each one from a different individual).

"Administrative", "Informational" and "Product related", describe the number of pages of different type each user searched, with their following time spent on each.

"Bounce Rate", "Exit Rate" and "Page Value" represents distinct metric collected by "Google Analytics". The first one refers to the percentage of users that leave the page without visiting others from that website. On the contrary, "Exit Rate" measure the visitors who exit the website after visiting others pages from that same website.

Finally, the "Page Value" variable, is defined by Google as: "Page Value is the average value for a page that a user visited before landing on the goal page or completing an Ecommerce transaction (or both)" (Google, 2019).

The remaining variables are: "Special Day", that indicates the closeness of an important event like Valentine's Day; "Month", that measures the month the session took time in; "Operating System", "Browser" and "Weekend" are self-explanatory. "Traffic type" reports the type by which the visitor has arrived at the website (Katircioglu, 2018). Last, but not least, "Revenue" just tells us if a transaction has been conducted during the session.

Figure 1: Kaggle logo (source: https://www.kaggle.com/)

2.0 Descriptive data analysis

First of all, I will check if there are any NAs and eliminate the rows that contains any:

```r
# Checking if there are nas, and how many: missing count
sum(is.na(data))
```

```
## [1] 112
```

```r
# eliminating rows with nas
data <- data[complete.cases(data),]
```

Counting the number of current rows, and classifying each column by its type:

```r
# Checking if there are nas, and how many: missing count
sum(is.na(data))
```

```
## [1] 0
```

```r
# eliminating rows with nas
data <- data[complete.cases(data),]


# Class of each column
sapply(data,class)
```

```
##          Administrative Administrative_Duration           Informational
##               "integer"               "numeric"               "integer"
##  Informational_Duration          ProductRelated ProductRelated_Duration
##               "numeric"               "integer"               "numeric"
##             BounceRates               ExitRates              PageValues
##               "numeric"               "numeric"               "numeric"
##              SpecialDay                   Month        OperatingSystems
##               "numeric"                "factor"               "integer"
##                 Browser                  Region             TrafficType
##               "integer"               "integer"               "integer"
##             VisitorType                 Weekend                 Revenue
##                "factor"               "logical"               "logical"
```

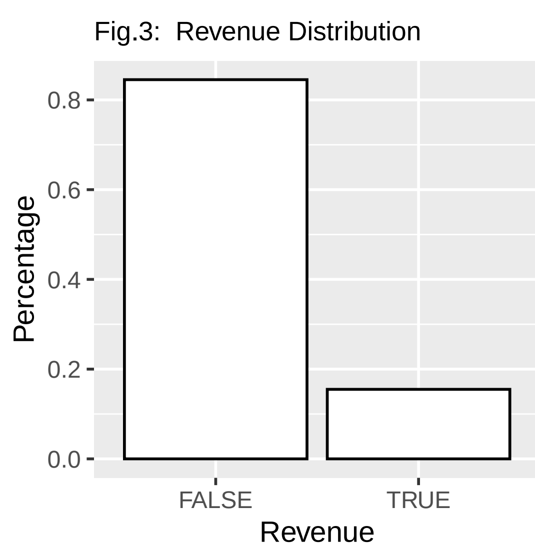
```r
# Classifying non-numeric variables
dummies <- c("Weekend", "Revenue")
categ <- c("VisitorType", "Month", "TrafficType", "Region", "Browser","OperatingSystems")

data.dummies <- data[,dummies]
data.categ <- data[,categ]
data.num <- data[,-c(10:18)] # selecting just the numeric variables
```

Using “stargazer”, a library for R, we get the following table that includes basic descriptive statistics for the numerical variables of the database:

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Administrative	12,316	2.318	3.323	0	0	4	27
Administrative_Duration	12,316	80.906	176.860	−1	0	93.5	3,399
Informational	12,316	0.504	1.271	0	0	0	24
Informational_Duration	12,316	34.506	140.825	−1	0	0	2,549
ProductRelated	12,316	31.764	44.490	0	7	38	705
ProductRelated_Duration	12,316	1,196.037	1,914.373	−1	185	1,466.5	63,974
BounceRates	12,316	0.022	0.048	0.000	0.000	0.017	0.200
ExitRates	12,316	0.043	0.049	0.000	0.014	0.050	0.200
PageValues	12,316	5.896	18.578	0	0	0	362
SpecialDay	12,316	0.061	0.199	0	0	0	1

For the non numerical, I think a graphical approach is much better for visualizing the data, so I used histograms. As we can see the probability of not buying is much higher. I also could not find the names of the regions, so their categories are represented by numbers (as well as the traffic type)[1]:



[1] I searched in his Master Thesis pdf file, but could not find it.

Fig.4: Visitor Type Distribution

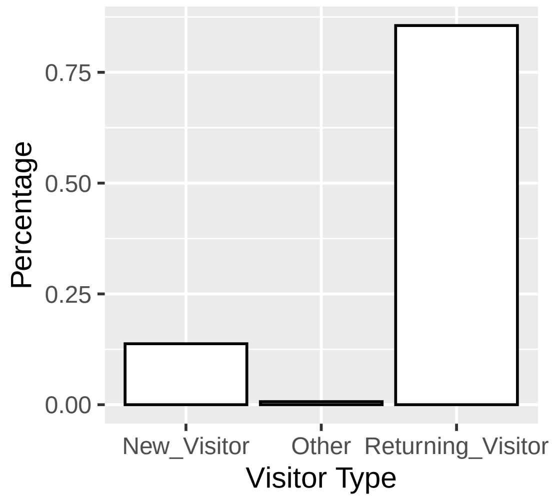


Fig.5: Months Distribution

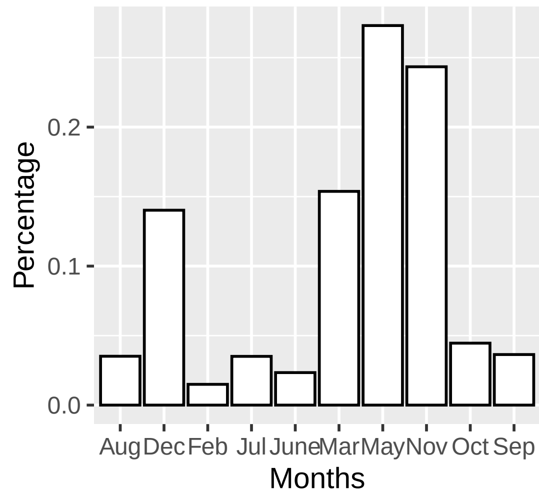


Fig.6: Traffic Type Distribution

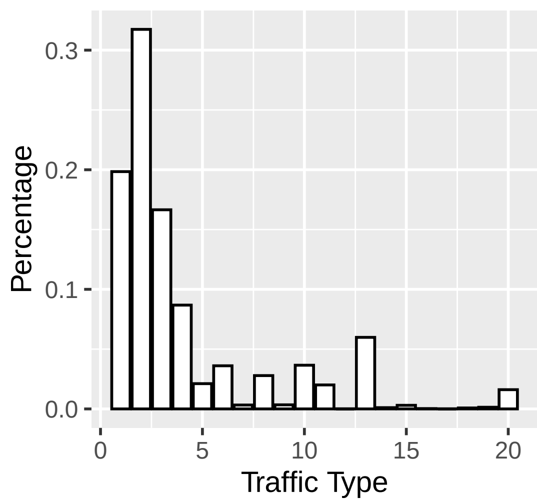


Fig.7: Region Distribution

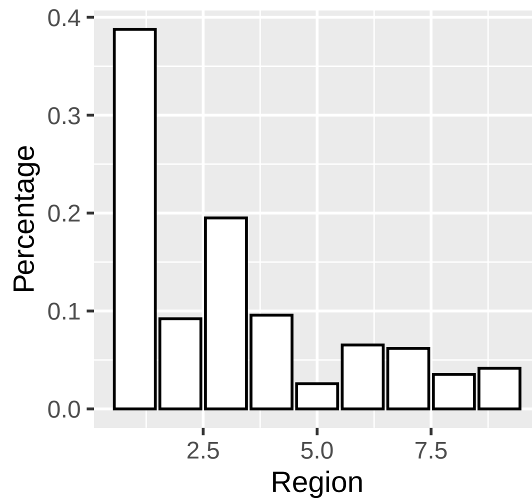


Fig.8: Browser Distribution

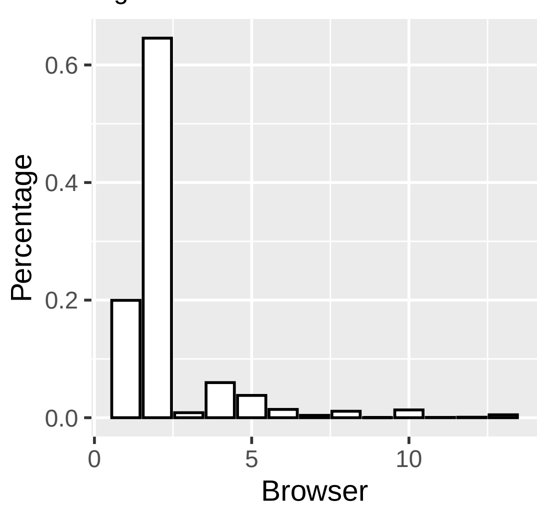
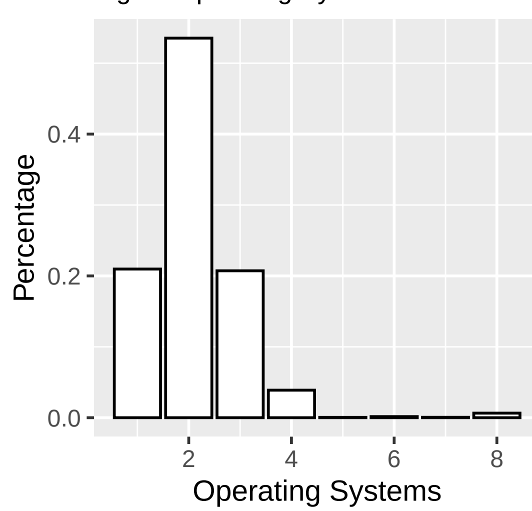


Fig.9: Operating Systems Distribution



We can use R to check how many purchases each month had in the almost \$ 13,000\$ sessions, and observe that November and May gather more than half of the total:

```
summaryBy(Revenue ~ Month, FUN=sum, data=data)
```

```
##    Month Revenue.sum
## 1    Aug          76
## 2    Dec         216
## 3    Feb           3
## 4    Jul          66
## 5   June          29
## 6    Mar         192
## 7    May         365
## 8    Nov         760
## 9    Oct         115
## 10   Sep          86
```

3.0 Regression

Prior to making the model, I will "clean" the database to make it more suited to the logit/probit regression. For this purpose, I convert the "Visitor_type" variable into a dummy variable "visitor" which takes value 0 if the user is a returning one, and 0 otherwise. Also, after regressing the first model (you can check it in the .R file, I did not want to include many different models in this file), I just took the months that were relevant, and made two new variables: "relev_months_neg", which includes the months that had significant negative effects on the probability of the page getting a purchase; and "relev_months_pos", which incorporates the ones that had a positive effect.

```
# Assigning numerical values to the "Visitor" variable (from categorical to dummy)
data$visitor <- as.numeric(data$VisitorType=="Returning_Visitor")

# Not all months are relevant, we make a variable for those which are:
# Dec, feb, mar, may, nov

data$relev_months_neg <- as.numeric(data$Month=="Dec"|data$Month=="Feb"
                                    |data$Month=="Mar"|data$Month=="May")
data$relev_months_pos <- as.numeric(data$Month=="Nov")
```

3.1 Model

After a few more different models (again, in the .R file) I end up with this model:

$$P(Y=1|x_1,...,x_5) = \frac{exp(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5)}{1+exp(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5)} \tag{1}$$

Where $P(Y=1/x_1,...,x_5)$ is the probability of a page getting revenue, given the βs and the vector of x. The variables in this final model are: "Informational Duration", "Exit Rates", "Page Values" and "Relevant Months", both positive and negative. The "visitors" variable appeared to not be relevant at a $\alpha = 0.05$ significance level. The βs are estimated using ML, and in R this is done as it follows:

```
# Logit model
logit <- glm(Revenue ~ Informational_Duration +  ExitRates +
PageValues + relev_months_pos + relev_months_neg,
data = data, family = "binomial"(link = "logit"))
summary(logit)
```

```
##
## Call:
## glm(formula = Revenue ~ Informational_Duration + ExitRates +
##     PageValues + relev_months_pos + relev_months_neg, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0688  -0.4717  -0.3452  -0.1572   3.4371
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.688e+00  8.261e-02 -20.435  < 2e-16 ***
## Informational_Duration  6.185e-04  1.612e-04   3.837 0.000125 ***
```

```
## ExitRates              -2.108e+01  1.637e+00 -12.881  < 2e-16 ***
## PageValues              8.156e-02  2.376e-03  34.325  < 2e-16 ***
## relev_months_pos        6.141e-01  8.538e-02   7.193 6.36e-13 ***
## relev_months_neg       -5.918e-01  8.394e-02  -7.050 1.79e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10620  on 12315  degrees of freedom
## Residual deviance:  7235  on 12310  degrees of freedom
## AIC: 7247
##
## Number of Fisher Scoring iterations: 7
```

```
# Probit model
probit <- glm(Revenue ~ Informational_Duration +  ExitRates
+ PageValues + relev_months_pos + relev_months_neg,
data = data, family = binomial(link = "probit"))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(probit)
```

```
##
## Call:
## glm(formula = Revenue ~ Informational_Duration + ExitRates +
##     PageValues + relev_months_pos + relev_months_neg, family = binomial(link = "probit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.4860  -0.3474  -0.1340   3.7160
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -9.861e-01  4.313e-02 -22.862  < 2e-16 ***
## Informational_Duration  3.592e-04  9.201e-05   3.904 9.47e-05 ***
## ExitRates              -1.052e+01  7.914e-01 -13.287  < 2e-16 ***
## PageValues              3.931e-02  1.155e-03  34.027  < 2e-16 ***
## relev_months_pos        3.185e-01  4.580e-02   6.955 3.52e-12 ***
## relev_months_neg       -3.306e-01  4.355e-02  -7.591 3.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10620.1  on 12315  degrees of freedom
## Residual deviance:  7385.7  on 12310  degrees of freedom
## AIC: 7397.7
##
## Number of Fisher Scoring iterations: 14
```

I also included a probit model, even though I will only analyze the logit one. As we can see, although there are differences between the coefficients of both models (probit assumes a normal distribution, and logit a logarithmic one), the final probability given by the models should be very similar. About the coefficients, we can interpret its sign (we cannot give an interpretation of its values because of the non-linear relationship between the value and the outcome probability).

Table 2: Results

	Dependent variable:	
	Revenue	
	logistic	*probit*
	(1)	(2)
Informational_Duration	0.001^{***}	0.0004^{***}
	(0.0002)	(0.0001)
ExitRates	-21.080^{***}	-10.516^{***}
	(1.637)	(0.791)
PageValues	0.082^{***}	0.039^{***}
	(0.002)	(0.001)
relev_months_pos	0.614^{***}	0.319^{***}
	(0.085)	(0.046)
relev_months_neg	-0.592^{***}	-0.331^{***}
	(0.084)	(0.044)
Constant	-1.688^{***}	-0.986^{***}
	(0.083)	(0.043)
Observations	12,316	12,316
Log Likelihood	-3,617.511	-3,692.848
Akaike Inf. Crit.	7,247.021	7,397.696
Note:	*p<0.1; **p<0.05; ***p<0.01	

3.2 Interpretation of the model

For the interpretation of the model I will use a table that showcases the changes in probability when a variable shifts to another value, and the partial changes in probability: also called the marginal effects[2].

Marginal effects

Using R, I compute the marginal effects at the mean of the independent variables. These values tell us how much does the probability of purchase changes if the independent variable increases one unit. For example, if "PageValues" increases one unit, the probability will grow in 0.6996%.

Table 3:

Variable	**Marginal Effect**
Informational_Duration	0.000053
ExitRates	-1.808000
PageValues	0.006996
relev_months_pos	0.052680
relev_months_neg	-0.050760

Discrete change in probability

These are different from the marginal effects. They can be interpreted as: "[...] for a change in the variable x_k from x_k to $x_k + \delta$, the predicted probability of an event changes by $\triangle Pr(y = 1|x)/ \triangle x_k$ holding all other variables constant." (Mariel, 2019). The benchmark I used takes the median values of the numerical variables, and 0 for the dummies. Computing in R we get the following changes in probabilities:

Table 4:

Variable	**Discrete change in probability**
Informational_Duration	0.007985
ExitRates	-0.060507
PageValues	0.233081
relev_months_pos	0.069304
relev_months_neg	0.041357

[2]You can find the code used to get these results in the $.R$ file.

3.3 Tests

I also conducted a few tests, like the Wald test for the Operating systems variable:

```
# Wald test for Operating systems
logit.w <- glm(Revenue ~ Informational_Duration +  ExitRates + PageValues
             +OperatingSystems+relev_months_pos+relev_months_neg,
             data = data, family = "binomial"(link = "logit"))
wald.test(b=coef(logit.w), Sigma=vcov(logit.w), Terms = 5)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 3.8, df = 1, P(> X2) = 0.052
```

As we can see, the variable is non relevant at $\alpha = 0.05$ significance level I will now check the same null hypothesis using now the LR test:

```
# We can test the same H_0 by LR test #

# Full model #
NR <- glm(Revenue ~ Informational_Duration+  ExitRates+PageValues+OperatingSystems+relev_months_pos+rele
data = data, family = "binomial"(link = "logit"))

# Restricted model #
R <-  glm(Revenue ~ Informational_Duration+  ExitRates+PageValues+relev_months_pos+relev_months_neg,
data = data, family = "binomial"(link = "logit"))

# LR test #
library(lmtest)
lrtest.default(NR,R)
```

```
## Likelihood ratio test
##
## Model 1: Revenue ~ Informational_Duration + ExitRates + PageValues + OperatingSystems +
##     relev_months_pos + relev_months_neg
## Model 2: Revenue ~ Informational_Duration + ExitRates + PageValues + relev_months_pos +
##     relev_months_neg
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -3615.6
## 2   6 -3617.5 -1 3.8393    0.05006 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can observe, the same conclusion holds. Lastly, I will check if there is multicollinearity within the independent variables, using a vif function I made myself. As the output shows, there are no hints of multicollinearity.

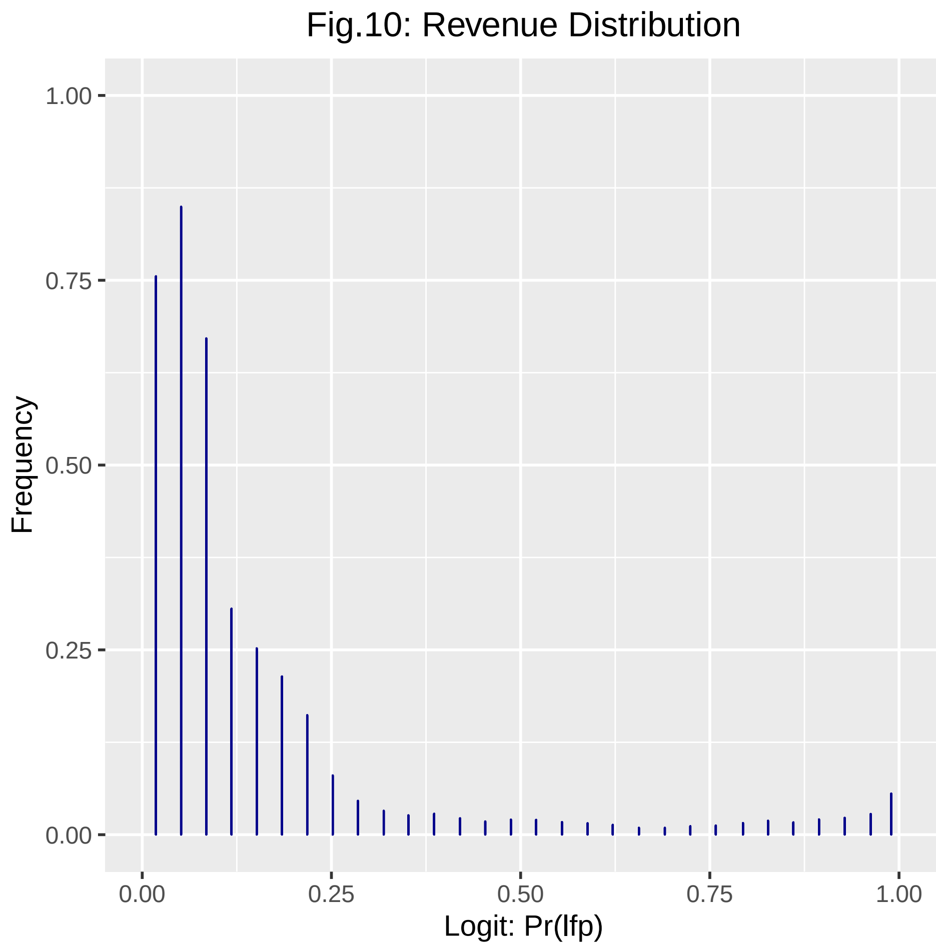
```
vif <-function(regression){
n1=length(regression$coefficients)-1
n2=length(regression$residuals)
```

```r
n3=length(regression$coefficients)-2
output<-matrix(,n2,n1)
r<-matrix(,n1)
i=2
for(var in 1:n1){
  a=unlist(model.frame(regression)[i], use.names=FALSE)
  output[,var]=a
  i=i+1
}
for(j in 1:n1){
  output_2=output[,-j]
  if(n3==1){
    reg=paste("lm(formula = output[,",j,"]~output_2)")
  }
  if(n3>1){
    reg=paste("lm(formula = output[,",j,"]~output_2[,1]")
    for(k in 2:n3){
      reg=paste(reg,"+output_2[,",k,"]")
      if(k==n3){
        reg=paste(reg,")")
      }
    }
  }
  r[j]=summary(eval(parse(text=reg)))$r.squared
}
vif=matrix(,n1)
for(k in 1:n1){
  vif[k]=1/(1-r[k])
}
return(vif)
}
vif(logit)
```

```
##          [,1]
## [1,] 1.012408
## [2,] 1.047968
## [3,] 1.032422
## [4,] 1.812465
## [5,] 1.818885
```

3.4 Graphical representation: a dotplot and the CDF

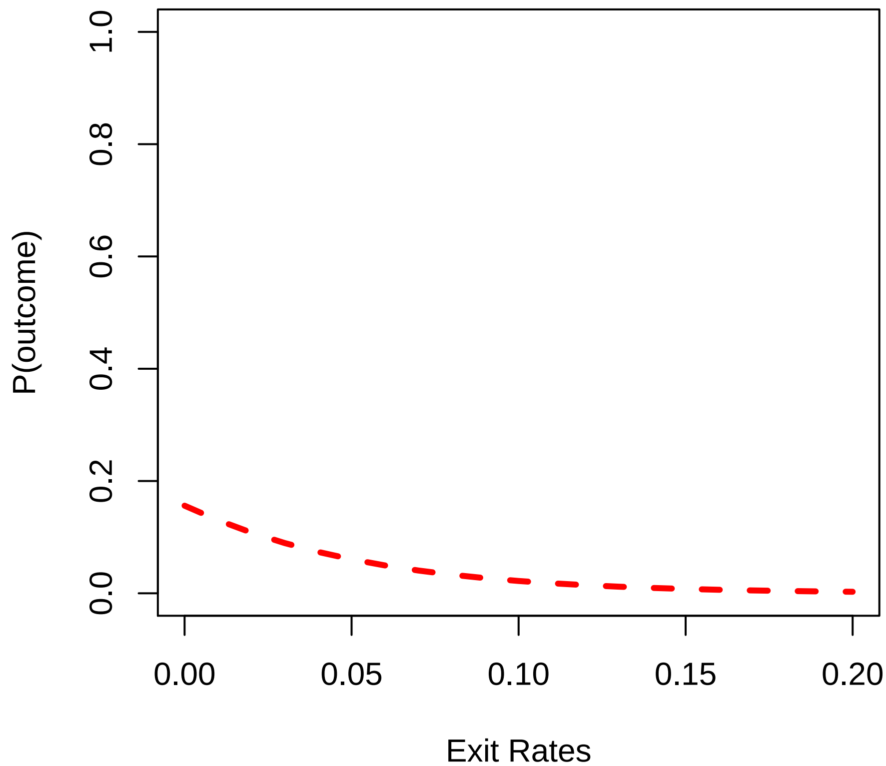
For the last part of the homework I will draw a dotplot that showcases the predicted probabilities of our sample, and the CDF of the "ExitRates" variable.



Note that representing the actual count is very difficult (for me at least) in R using "ggplot2".

The plot clearly shows that a great part of the observations have predicted probabilities concentrated near the $0 - 0.25$ range. Regarding the CDF, the plot clearly shows that a great part of the observations have predicted probabilities concentrated near the $0 - 0.05$ range. As we can see, the probability of revenue diminishes as the exit rates increases, as it has a negative impact on the dependent variable. Regarding the CDF, the benchmark sets the median for numeric variables and 0 for the dummies. As the overall probability of buying, within the sample, is rare, the CDF has this functional form.

Fig.11 Probability of Revenue



4.0 Conclusion

From the database analysis we can get several conclusions, with the more inmediate one being that the majority of the purchases were made outside weekends. It is also really clear that a minority of the sessions ended in a purchase, which makes sense, as you usually do not buy something impulsively. This statement is reforced with the percentage of returning visitors over the total, which computes about 80%. As I said in the introduction, the majority of the shopping took place in November, prior to the holidays, I suppose.

From the results of the regression we can conclude that the probability of a session ending in a purchase is rare, with few of the initials variables being relevant. There is no multicollinearity in the variables, and that some of the coefficients are very small because of some values of the variables. A rescale could be a good idea for a better understanding. All the coefficients made sense according to the relation with the independent variable (with "Exit Rates" having the most significant one [3]).

[3] Notice that the units of this variable are very small.

Bibliography

Mariel, P. (2019): *Qualitative Dependent Variables*

Google (2019): *How Page Value is calculated* Retrieved from: https://support.google.com/analytics/answer/2695658?hl=en

Kaggle (2019): *Online Shopper's Intention* Retrieved from: https://www.kaggle.com/roshansharma/online-shoppers-intention

Katircioglu, M. (2018): *Predicting Commercial Intent of Online Consumers using Machine Learning Techniques* Retrieved from: http://acikerisim.bahcesehir.edu.tr:8080/xmlui/bitstream/handle/123456789/1221/139125.pdf?sequence=1&isAllowed=y