Generalized Linear Models

library(readxl)  
library(car)

## Loading required package: carData

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

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

library(ggpubr)

## Warning: package 'ggpubr' was built under R version 4.0.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.4

library(stargazer)

## Warning: package 'stargazer' was built under R version 4.0.3

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

library(data.table)

## Warning: package 'data.table' was built under R version 4.0.3

library(MASS)  
library(AER)

## Warning: package 'AER' was built under R version 4.0.4

## Loading required package: sandwich

## Loading required package: survival

library(pscl)

## Warning: package 'pscl' was built under R version 4.0.4

## Classes and Methods for R developed in the  
## Political Science Computational Laboratory  
## Department of Political Science  
## Stanford University  
## Simon Jackman  
## hurdle and zeroinfl functions by Achim Zeileis

setwd("C:/Users/surya/Downloads")  
  
orc <- read\_excel("OnlineRetailCampaign.xlsx", sheet = 'Data')  
names(orc) <- tolower(colnames(orc))  
  
#NA values column wise  
sapply(orc, function(x) sum(is.na(x)))

## recency historysegment history mens womens   
## 0 0 0 0 0   
## zipcode newcustomer channel campaign visit   
## 0 0 0 0 0   
## conversion spend   
## 0 0

str(orc)

## tibble [64,000 x 12] (S3: tbl\_df/tbl/data.frame)  
## $ recency : num [1:64000] 10 6 7 9 2 6 9 9 9 10 ...  
## $ historysegment: chr [1:64000] "2) $100 - $200" "3) $200 - $350" "2) $100 - $200" "5) $500 - $750" ...  
## $ history : num [1:64000] 142.4 329.1 180.7 675.8 45.3 ...  
## $ mens : num [1:64000] 1 1 0 1 1 0 1 0 1 0 ...  
## $ womens : num [1:64000] 0 1 1 0 0 1 0 1 1 1 ...  
## $ zipcode : chr [1:64000] "Surburban" "Rural" "Surburban" "Rural" ...  
## $ newcustomer : num [1:64000] 0 1 1 1 0 0 1 0 1 1 ...  
## $ channel : chr [1:64000] "Phone" "Web" "Web" "Web" ...  
## $ campaign : chr [1:64000] "Womens E-Mail" "No E-Mail" "Womens E-Mail" "Mens E-Mail" ...  
## $ visit : num [1:64000] 0 0 0 0 0 1 0 0 0 0 ...  
## $ conversion : num [1:64000] 0 0 0 0 0 0 0 0 0 0 ...  
## $ spend : num [1:64000] 0 0 0 0 0 0 0 0 0 0 ...

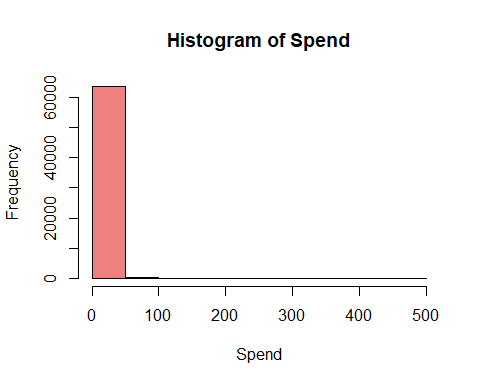
#Feature Engineering/Pre-processing  
orc$historysegment <- NULL  
orc$phone <- ifelse(orc$channel == 'Phone' | orc$channel == 'Multichannel', 1, 0)  
orc$web <- ifelse(orc$channel == 'Web' | orc$channel == 'Multichannel', 1, 0)  
orc$channel <- NULL  
orc$campaign <- as.factor(orc$campaign)  
orc$zipcode <- as.factor(orc$zipcode)  
orc$campaign <- relevel(orc$campaign, ref='No E-Mail')  
orc$spend <- round(orc$spend)  
orc\_rd <- orc  
orc <- orc[orc$spend > 0,]  
#orc\_woz <- orc  
  
#Checking if DV is suitable for OLS  
mean(orc\_rd$spend)

## [1] 1.050891

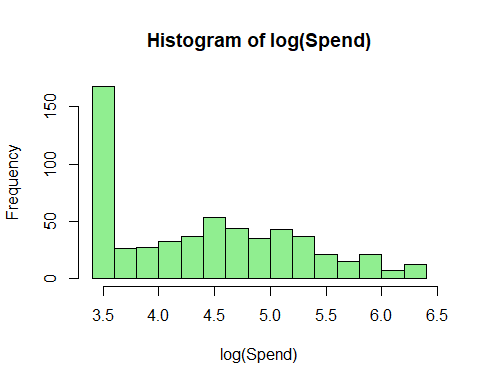
sd(orc\_rd$spend)

## [1] 15.03591

hist(orc\_rd$spend, col = 'lightcoral', main = "Histogram of Spend", xlab = 'Spend', ylab = 'Frequency')



hist(log(orc\_rd$spend), col = 'lightgreen', main = "Histogram of log(Spend)", xlab = 'log(Spend)', ylab = 'Frequency')



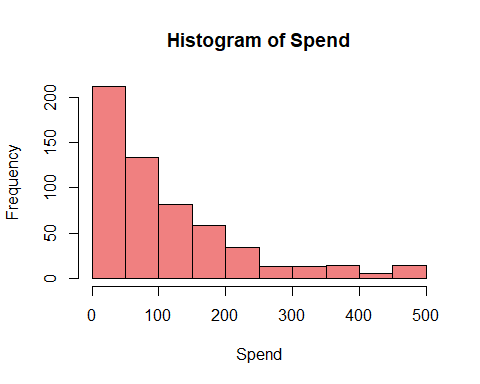
#Checking if filtering and rounding off changes anything  
mean(orc$spend)

## [1] 116.3616

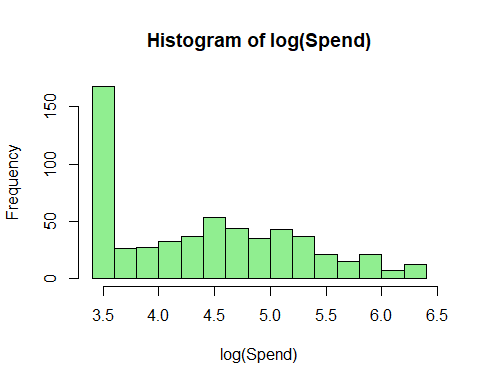
sd(orc$spend)

## [1] 107.8653

hist(orc$spend, col = 'lightcoral', main = "Histogram of Spend", xlab = 'Spend', ylab = 'Frequency')



hist(log(orc$spend), col = 'lightgreen', main = "Histogram of log(Spend)", xlab = 'log(Spend)', ylab = 'Frequency')



#Regression models  
m1 <- lm(log(spend) ~ recency + history + zipcode + newcustomer   
 + phone + web + campaign\*mens + campaign\*womens, data = orc)  
  
m2 <- glm(spend ~ recency + history + zipcode + newcustomer   
 + phone + web + campaign\*mens + campaign\*womens, data = orc, family=poisson (link=log))  
  
m3 <- glm(spend ~ recency + history + zipcode + newcustomer   
 + phone + web + campaign\*mens + campaign\*womens, data = orc, family=quasipoisson (link=log))  
  
m4 <- glm.nb(spend ~ recency + history + zipcode + newcustomer   
 + phone + web + campaign + mens + womens, data = orc)  
  
m4.A <- glm.nb(spend ~ recency + campaign\*history + campaign\*zipcode + campaign\*newcustomer   
 + campaign\*phone + campaign\*web + campaign\*mens + campaign\*womens, data = orc)  
  
m4.B <- glm.nb(spend ~ recency + history + zipcode + newcustomer   
 + phone + web + campaign + mens + womens, data = orc)  
  
m5 <- hurdle(spend ~ recency + campaign\*history + campaign\*zipcode + campaign\*newcustomer   
 + campaign\*phone + campaign\*web + campaign\*mens + campaign\*womens | visit + conversion, data=orc\_rd, link="logit", dist="negbin")  
  
m6 <- hurdle(spend ~ recency + history + zipcode + newcustomer   
 + phone + web + campaign + mens + womens | visit + conversion, data=orc\_rd, link="logit", dist="negbin")  
  
#vif test fail  
m7 <- zeroinfl(spend ~ recency + campaign\*history + campaign\*zipcode + campaign\*newcustomer   
 + campaign\*phone + campaign\*web + campaign\*mens + campaign\*womens | visit + conversion, data=orc\_rd, link="logit", dist="negbin")

## Warning: glm.fit: algorithm did not converge

## Warning in value[[3L]](cond): system is computationally singular: reciprocal  
## condition number = 3.18666e-18FALSE

m8 <- zeroinfl(spend ~ recency + history + zipcode + newcustomer   
 + phone + web + campaign + mens + womens | visit + conversion, data=orc\_rd, link="logit", dist="negbin")

## Warning: glm.fit: algorithm did not converge

## Warning in value[[3L]](cond): system is computationally singular: reciprocal  
## condition number = 5.8928e-18FALSE

#Dispersion test  
#dispersiontest(m4)  
  
#Stargazer  
stargazer(m4, m6, m8, type='text', single.row = TRUE)

##   
## =====================================================================  
## Dependent variable:   
## -----------------------------------------------  
## spend   
## negative hurdle zero-inflated  
## binomial count data   
## (1) (2) (3)   
## ---------------------------------------------------------------------  
## recency 0.008 (0.010) 0.008 (0.010) 0.008   
## history 0.00004 (0.0001) 0.00004 (0.0001) 0.00004   
## zipcodeSurburban 0.139 (0.094) 0.139 (0.095) 0.139   
## zipcodeUrban 0.090 (0.094) 0.090 (0.095) 0.090   
## newcustomer -0.005 (0.074) -0.005 (0.074) -0.005   
## phone -0.090 (0.102) -0.090 (0.104) -0.090   
## web -0.072 (0.102) -0.072 (0.105) -0.072   
## campaignMens E-Mail 0.003 (0.088) 0.003 (0.089) 0.003   
## campaignWomens E-Mail 0.104 (0.095) 0.104 (0.096) 0.104   
## mens 0.136 (0.103) 0.137 (0.102) 0.136   
## womens -0.128 (0.104) -0.128 (0.101) -0.128   
## Constant 4.666\*\*\* (0.176) 4.664\*\*\* (0.174) 4.666   
## ---------------------------------------------------------------------  
## Observations 578 64,000 64,000   
## Log Likelihood -3,294.330 -3,292.655 -3,293.330   
## theta 1.581\*\*\* (0.086)   
## Akaike Inf. Crit. 6,612.660   
## =====================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

stargazer(m4, m4.A, m4.B, m5, type='text', single.row = TRUE)

##   
## ============================================================================================================  
## Dependent variable:   
## ---------------------------------------------------------------------  
## spend   
## negative hurdle   
## binomial   
## (1) (2) (3) (4)   
## ------------------------------------------------------------------------------------------------------------  
## recency 0.008 (0.010) 0.003 (0.010) 0.008 (0.010) 0.003 (0.010)   
## history 0.00004 (0.0001) -0.0001 (0.0003) 0.00004 (0.0001) -0.0001 (0.0002)   
## zipcodeSurburban 0.139 (0.094) 0.428\*\* (0.196) 0.139 (0.094) 0.429\*\* (0.199)   
## zipcodeUrban 0.090 (0.094) 0.201 (0.198) 0.090 (0.094) 0.202 (0.203)   
## newcustomer -0.005 (0.074) -0.279 (0.183) -0.005 (0.074) -0.279 (0.184)   
## phone -0.090 (0.102) -0.308 (0.229) -0.090 (0.102) -0.309 (0.235)   
## web -0.072 (0.102) -0.284 (0.227) -0.072 (0.102) -0.285 (0.231)   
## campaignMens E-Mail 0.003 (0.088) 0.183 (0.450) 0.003 (0.088) 0.183 (0.438)   
## campaignWomens E-Mail 0.104 (0.095) 0.696 (0.475) 0.104 (0.095) 0.698 (0.471)   
## mens 0.136 (0.103) 0.530\*\* (0.255) 0.136 (0.103) 0.531\*\* (0.246)   
## womens -0.128 (0.104) 0.253 (0.248) -0.128 (0.104) 0.254 (0.239)   
## campaignMens E-Mail:history 0.0001 (0.0003) 0.0001 (0.0003)   
## campaignWomens E-Mail:history 0.0003 (0.0004) 0.0003 (0.0003)   
## campaignMens E-Mail:zipcodeSurburban -0.386 (0.242) -0.387 (0.246)   
## campaignWomens E-Mail:zipcodeSurburban -0.317 (0.255) -0.317 (0.259)   
## campaignMens E-Mail:zipcodeUrban -0.268 (0.245) -0.269 (0.250)   
## campaignWomens E-Mail:zipcodeUrban 0.097 (0.254) 0.097 (0.260)   
## campaignMens E-Mail:newcustomer 0.345 (0.211) 0.346 (0.214)   
## campaignWomens E-Mail:newcustomer 0.306 (0.222) 0.307 (0.224)   
## campaignMens E-Mail:phone 0.183 (0.274) 0.183 (0.279)   
## campaignWomens E-Mail:phone 0.335 (0.284) 0.336 (0.299)   
## campaignMens E-Mail:web 0.214 (0.273) 0.215 (0.276)   
## campaignWomens E-Mail:web 0.272 (0.285) 0.273 (0.297)   
## campaignMens E-Mail:mens -0.300 (0.293) -0.301 (0.285)   
## campaignWomens E-Mail:mens -0.865\*\*\* (0.316) -0.867\*\*\* (0.317)  
## campaignMens E-Mail:womens -0.199 (0.288) -0.199 (0.278)   
## campaignWomens E-Mail:womens -0.928\*\*\* (0.312) -0.930\*\*\* (0.309)  
## Constant 4.666\*\*\* (0.176) 4.438\*\*\* (0.394) 4.666\*\*\* (0.176) 4.436\*\*\* (0.383)   
## ------------------------------------------------------------------------------------------------------------  
## Observations 578 578 578 64,000   
## Log Likelihood -3,294.330 -3,283.781 -3,294.330 -3,282.181   
## theta 1.581\*\*\* (0.086) 1.631\*\*\* (0.089) 1.581\*\*\* (0.086)   
## Akaike Inf. Crit. 6,612.660 6,623.562 6,612.660   
## ============================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

stargazer(m4.A, m4.B, m5, m6, type='text', single.row = TRUE)

##   
## ============================================================================================================  
## Dependent variable:   
## ---------------------------------------------------------------------  
## spend   
## negative hurdle   
## binomial   
## (1) (2) (3) (4)   
## ------------------------------------------------------------------------------------------------------------  
## recency 0.003 (0.010) 0.008 (0.010) 0.003 (0.010) 0.008 (0.010)   
## campaignMens E-Mail 0.183 (0.450) 0.003 (0.088) 0.183 (0.438) 0.003 (0.089)   
## campaignWomens E-Mail 0.696 (0.475) 0.104 (0.095) 0.698 (0.471) 0.104 (0.096)   
## history -0.0001 (0.0003) 0.00004 (0.0001) -0.0001 (0.0002) 0.00004 (0.0001)  
## zipcodeSurburban 0.428\*\* (0.196) 0.139 (0.094) 0.429\*\* (0.199) 0.139 (0.095)   
## zipcodeUrban 0.201 (0.198) 0.090 (0.094) 0.202 (0.203) 0.090 (0.095)   
## newcustomer -0.279 (0.183) -0.005 (0.074) -0.279 (0.184) -0.005 (0.074)   
## phone -0.308 (0.229) -0.090 (0.102) -0.309 (0.235) -0.090 (0.104)   
## web -0.284 (0.227) -0.072 (0.102) -0.285 (0.231) -0.072 (0.105)   
## mens 0.530\*\* (0.255) 0.136 (0.103) 0.531\*\* (0.246) 0.137 (0.102)   
## womens 0.253 (0.248) -0.128 (0.104) 0.254 (0.239) -0.128 (0.101)   
## campaignMens E-Mail:history 0.0001 (0.0003) 0.0001 (0.0003)   
## campaignWomens E-Mail:history 0.0003 (0.0004) 0.0003 (0.0003)   
## campaignMens E-Mail:zipcodeSurburban -0.386 (0.242) -0.387 (0.246)   
## campaignWomens E-Mail:zipcodeSurburban -0.317 (0.255) -0.317 (0.259)   
## campaignMens E-Mail:zipcodeUrban -0.268 (0.245) -0.269 (0.250)   
## campaignWomens E-Mail:zipcodeUrban 0.097 (0.254) 0.097 (0.260)   
## campaignMens E-Mail:newcustomer 0.345 (0.211) 0.346 (0.214)   
## campaignWomens E-Mail:newcustomer 0.306 (0.222) 0.307 (0.224)   
## campaignMens E-Mail:phone 0.183 (0.274) 0.183 (0.279)   
## campaignWomens E-Mail:phone 0.335 (0.284) 0.336 (0.299)   
## campaignMens E-Mail:web 0.214 (0.273) 0.215 (0.276)   
## campaignWomens E-Mail:web 0.272 (0.285) 0.273 (0.297)   
## campaignMens E-Mail:mens -0.300 (0.293) -0.301 (0.285)   
## campaignWomens E-Mail:mens -0.865\*\*\* (0.316) -0.867\*\*\* (0.317)   
## campaignMens E-Mail:womens -0.199 (0.288) -0.199 (0.278)   
## campaignWomens E-Mail:womens -0.928\*\*\* (0.312) -0.930\*\*\* (0.309)   
## Constant 4.438\*\*\* (0.394) 4.666\*\*\* (0.176) 4.436\*\*\* (0.383) 4.664\*\*\* (0.174)  
## ------------------------------------------------------------------------------------------------------------  
## Observations 578 578 64,000 64,000   
## Log Likelihood -3,283.781 -3,294.330 -3,282.181 -3,292.655   
## theta 1.631\*\*\* (0.089) 1.581\*\*\* (0.086)   
## Akaike Inf. Crit. 6,623.562 6,612.660   
## ============================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

stargazer(m5, m7, m8, type='text', single.row = TRUE)

##   
## ==============================================================================  
## Dependent variable:   
## ---------------------------------------  
## spend   
## hurdle zero-inflated   
## count data   
## (1) (2) (3)   
## ------------------------------------------------------------------------------  
## recency 0.003 (0.010) 0.003 0.008   
## campaignMens E-Mail 0.183 (0.438) 0.183 0.003   
## campaignWomens E-Mail 0.698 (0.471) 0.697 0.104   
## history -0.0001 (0.0002) -0.0001 0.00004   
## zipcodeSurburban 0.429\*\* (0.199) 0.428 0.139   
## zipcodeUrban 0.202 (0.203) 0.201 0.090   
## newcustomer -0.279 (0.184) -0.279 -0.005   
## phone -0.309 (0.235) -0.308 -0.090   
## web -0.285 (0.231) -0.284 -0.072   
## mens 0.531\*\* (0.246) 0.530 0.136   
## womens 0.254 (0.239) 0.254 -0.128   
## campaignMens E-Mail:history 0.0001 (0.0003) 0.0001   
## campaignWomens E-Mail:history 0.0003 (0.0003) 0.0003   
## campaignMens E-Mail:zipcodeSurburban -0.387 (0.246) -0.386   
## campaignWomens E-Mail:zipcodeSurburban -0.317 (0.259) -0.317   
## campaignMens E-Mail:zipcodeUrban -0.269 (0.250) -0.269   
## campaignWomens E-Mail:zipcodeUrban 0.097 (0.260) 0.097   
## campaignMens E-Mail:newcustomer 0.346 (0.214) 0.345   
## campaignWomens E-Mail:newcustomer 0.307 (0.224) 0.306   
## campaignMens E-Mail:phone 0.183 (0.279) 0.183   
## campaignWomens E-Mail:phone 0.336 (0.299) 0.335   
## campaignMens E-Mail:web 0.215 (0.276) 0.214   
## campaignWomens E-Mail:web 0.273 (0.297) 0.272   
## campaignMens E-Mail:mens -0.301 (0.285) -0.300   
## campaignWomens E-Mail:mens -0.867\*\*\* (0.317) -0.865   
## campaignMens E-Mail:womens -0.199 (0.278) -0.199   
## campaignWomens E-Mail:womens -0.930\*\*\* (0.309) -0.928   
## Constant 4.436\*\*\* (0.383) 4.438 4.666   
## ------------------------------------------------------------------------------  
## Observations 64,000 64,000 64,000   
## Log Likelihood -3,282.181 -3,282.781 -3,293.330  
## ==============================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

#vif(m4)  
vif(m4.A)

## GVIF Df GVIF^(1/(2\*Df))  
## recency 1.170848 1 1.082057  
## campaign 623.684330 2 4.997367  
## history 7.847132 1 2.801273  
## zipcode 22.618768 2 2.180807  
## newcustomer 7.640325 1 2.764114  
## phone 12.074134 1 3.474785  
## web 11.209804 1 3.348105  
## mens 14.900916 1 3.860170  
## womens 13.540534 1 3.679746  
## campaign:history 28.997876 2 2.320553  
## campaign:zipcode 263.354920 4 2.007094  
## campaign:newcustomer 20.231510 2 2.120836  
## campaign:phone 72.719776 2 2.920204  
## campaign:web 89.385311 2 3.074798  
## campaign:mens 91.961872 2 3.096720  
## campaign:womens 113.294180 2 3.262510

vif(m4.B)

## GVIF Df GVIF^(1/(2\*Df))  
## recency 1.104338 1 1.050875  
## history 1.600778 1 1.265218  
## zipcode 1.019798 2 1.004913  
## newcustomer 1.202321 1 1.096504  
## phone 2.305083 1 1.518250  
## web 2.212866 1 1.487570  
## campaign 1.054499 2 1.013355  
## mens 2.362954 1 1.537190  
## womens 2.296710 1 1.515490

vif(m5)

## Warning in vif.default(m5): No intercept: vifs may not be sensible.

## GVIF Df GVIF^(1/(2\*Df))  
## recency 3.558708 1 1.886454  
## campaign 2828.549789 2 7.292744  
## history 14.096475 1 3.754527  
## zipcode 130.093484 2 3.377255  
## newcustomer 13.844696 1 3.720846  
## phone 27.676793 1 5.260874  
## web 30.603797 1 5.532070  
## mens 31.469302 1 5.609751  
## womens 32.387879 1 5.691035  
## campaign:history 43.420619 2 2.566989  
## campaign:zipcode 876.995869 4 2.332785  
## campaign:newcustomer 33.293455 2 2.402092  
## campaign:phone 147.570908 2 3.485380  
## campaign:web 200.282004 2 3.761928  
## campaign:mens 163.997799 2 3.578570  
## campaign:womens 211.939442 2 3.815513

#vif(m6)  
  
dwtest(m5)

##   
## Durbin-Watson test  
##   
## data: m5  
## DW = 2.0062, p-value = 0.7824  
## alternative hypothesis: true autocorrelation is greater than 0