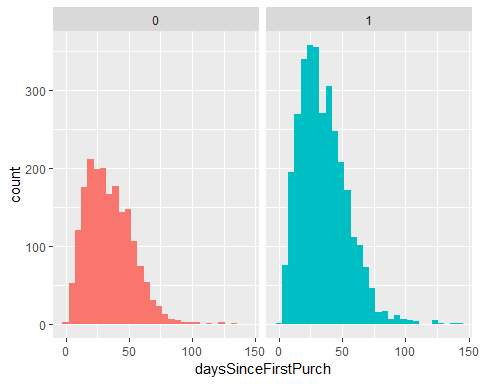
Survival\_analysis.R

#setwd("")  
library(ggplot2)  
##################################################################################  
#Applications of survival analysis  
##################################################################################  
#QUESTION : question that can be answered with survival analysis?  
#RESPONSE : After ordering for the first time in an online shop, when do customers place their second order?  
#COMMENT : In this case, you would model the time until the event of placing another order.  
##################################################################################  
#Data for survival analysis  
##################################################################################  
# work with data about customers of an online shop in order to practice survival analysis  
# Is about the time until the second order.  
dataNextOrder <- read.csv("survivalDataExercise.csv", stringsAsFactors = TRUE)  
# Look at the head of the data  
head(dataNextOrder)

## daysSinceFirstPurch shoppingCartValue gender voucher returned  
## 1 37 33.44 male 0 0  
## 2 63 31.71 male 1 0  
## 3 48 27.31 female 0 0  
## 4 17 41.07 male 0 0  
## 5 53 65.56 female 0 0  
## 6 11 38.44 female 0 0  
## boughtAgain  
## 1 0  
## 2 1  
## 3 0  
## 4 1  
## 5 0  
## 6 1

# Plot a histogram  
ggplot(dataNextOrder) +  
 geom\_histogram(aes(x = daysSinceFirstPurch,  
 fill = factor(boughtAgain))) +  
 facet\_grid( ~ boughtAgain) + # Separate plots for boughtAgain = 1 vs. 0  
 theme(legend.position = "none") # Don't show legend

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



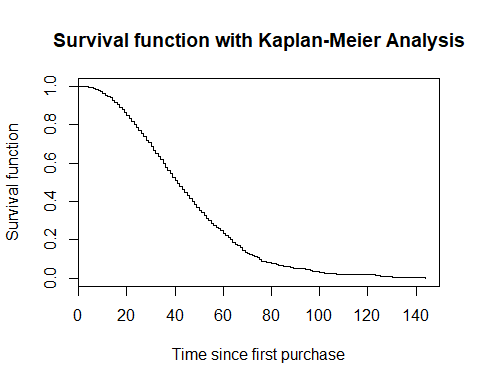
##################################################################################  
#Characteristics of survival analysis  
##################################################################################  
#QUESTION : Which of the following is a characteristic of survival analysis?  
#RESPONSE : Survival analysis is suited for situations where for some observations an event has not yet happened, but may happen at some point in time.  
#COMMENT : In survival analysis, each observation has one of two states: either an event occured, or it didn't occur  
##################################################################################  
#Survival function, hazard function and hazard rate  
##################################################################################  
#QUESTION : What is a survival Fucntion  
#RESPONSE : The survival function describes the proportion of observations who are still alive (or, for example, in a customer relationship, the proportion of customers who haven't churned yet), depending on their time under observation.  
##################################################################################  
#The survival object  
##################################################################################  
#Before you start any survival analysis, you need to transform your data into the right form, the survival object  
library(survival)  
# Create survival object  
survObj <- Surv(dataNextOrder$daysSinceFirstPurch, dataNextOrder$boughtAgain)  
# Look at structure  
str(survObj)

## 'Surv' num [1:5122, 1:2] 37+ 63 48+ 17 53+ 11 22 16 74+ 44 ...  
## - attr(\*, "dimnames")=List of 2  
## ..$ : NULL  
## ..$ : chr [1:2] "time" "status"  
## - attr(\*, "type")= chr "right"

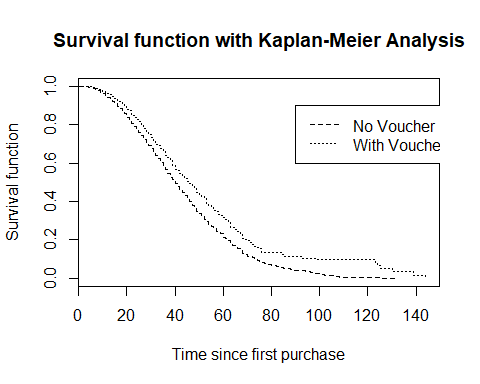
#For each observation there is the time under observation,   
#marked with a + if the second order has not been placed yet  
##################################################################################  
#Kaplan-Meier Analysis  
##################################################################################  
#In this exercise you are going to practice Kaplan-Meier Analysis - without and with a categorical covariate.  
# the data contains an additional covariate called voucher, which you will need in this exercise  
# categorical variable tells you if the customer used a voucher in her first order  
# Compute and print fit  
fitKMSimple <- survfit(survObj ~ 1)  
print(fitKMSimple)

## Call: survfit(formula = survObj ~ 1)  
##   
## n events median 0.95LCL 0.95UCL   
## 5122 3199 41 40 42

#N = 5122 number of customer  
#events = 3199 churn under the time of the observation  
#median = about 50% of the customers do not churn before they reach the duration of 70 days  
#median is when the line is cut by 0.5 on 1 of the survival function  
  
# Plot fit  
plot(fitKMSimple, conf.int = FALSE,  
 xlab = "Time since first purchase", ylab = "Survival function", main = "Survival function with Kaplan-Meier Analysis")



# Compute fit with covariate  
fitKMCov <- survfit(survObj ~ voucher, data = dataNextOrder)  
  
# Plot fit with covariate and add labels  
plot(fitKMCov, lty = 2:3,  
 xlab = "Time since first purchase", ylab = "Survival function", main = "Survival function with Kaplan-Meier Analysis")  
legend(90, .9, c("No Voucher", "With Voucher"), lty = 2:3)



#Customers using a voucher seem to take longer to place their second order.  
  
##################################################################################  
#Proportional hazard assumption  
##################################################################################  
#QUESTION : What does the proportional hazard assumption mean?  
#RESPONSE : The influence of the predictors does not change over time.  
#COMMENT : For example, it would not be allowed that the gender "male" has as positive effect on the survival time after a short time under observation, but a large negative effect after a longer time under observation.  
  
##################################################################################  
#Cox Proportional Hazard Model  
##################################################################################  
#Your data stored in dataNextOrder now contains four additional variables:  
#shoppingCartValue, voucher, returned, gender  
library(rms)

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## Loading required package: SparseM

##   
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':  
##   
## backsolve

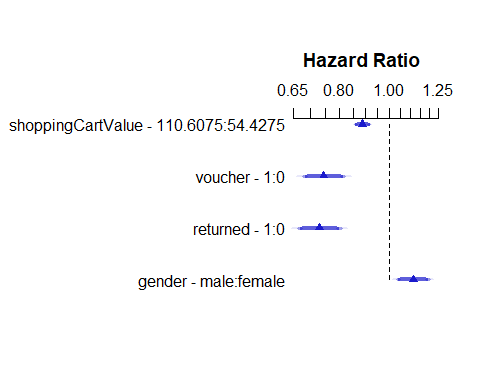
# Determine distributions of predictor variables  
dd <- datadist(dataNextOrder)  
options(datadist = "dd")  
# Compute Cox PH Model and print results  
fitCPH <- cph(Surv(daysSinceFirstPurch, boughtAgain) ~ shoppingCartValue + voucher + returned + gender,  
 data = dataNextOrder,  
 x = TRUE, y = TRUE, surv = TRUE)  
print(fitCPH)

## Cox Proportional Hazards Model  
##   
## cph(formula = Surv(daysSinceFirstPurch, boughtAgain) ~ shoppingCartValue +   
## voucher + returned + gender, data = dataNextOrder, x = TRUE,   
## y = TRUE, surv = TRUE)  
##   
## Model Tests Discrimination   
## Indexes   
## Obs 5122 LR chi2 155.68 R2 0.030   
## Events 3199 d.f. 4 Dxy 0.116   
## Center -0.2808 Pr(> chi2) 0.0000 g 0.238   
## Score chi2 140.57 gr 1.269   
## Pr(> chi2) 0.0000   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## shoppingCartValue -0.0021 0.0003 -7.56 <0.0001   
## voucher -0.2945 0.0480 -6.14 <0.0001   
## returned -0.3145 0.0495 -6.36 <0.0001   
## gender=male 0.1080 0.0363 2.97 0.0029   
##

# Interpret coefficients  
exp(fitCPH$coefficients)

## shoppingCartValue voucher returned gender=male   
## 0.9978601 0.7449362 0.7301667 1.1140891

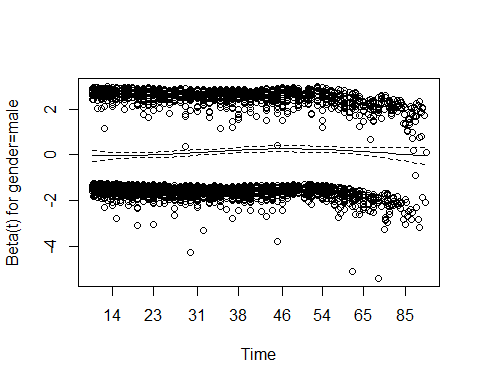
#shopping cart value increase of 1 dollar decreases   
#the hazard to buy again by a factor of only slightly below 1 -  
#but the coefficient is significant, as are all coefficients  
  
#For customers who used a voucher, the hazard of buying again is 0.74 times lower,   
#and for customers who returned any of the items, the hazard of buying again is 0.73 times lower.  
#Being a man compared to a women increases the hazard of buying again by the factor 1.11.  
  
# Plot results  
plot(summary(fitCPH), log = TRUE)



#survplot(fitCPH, boughtAgain, label.curves = list(keys=1:4))  
##################################################################################  
#Interpretation of coefficients  
##################################################################################  
#QUESTION : You computed a Cox PH model and got a coefficient of 0.8 for your continuous predictor X.   
#What is the correct interpretation?  
#RESPONSE : A one-unit increase in X increases the hazard by a factor of about 2.23.  
#COMMENT : The effect is multiplicative and you took the exponential.  
  
##################################################################################  
#Violation of the PH assumption  
##################################################################################  
#QUESTION : What can you do if the proportional hazard assumption is violated for a predictor?  
#RESPONSE : Stratify the sample according to this predictor and analyse the strata separately.  
#COMMENT : You can divide the data into strata and estimate the model separately within each stratum.  
  
##################################################################################  
#Model assumptions  
##################################################################################  
#look at the Cox PH model  
#Check the proportional hazard assumption of the model using cox.zph()  
# Check proportional hazard assumption and print result  
testCPH <- cox.zph(fitCPH)  
print(testCPH)

## rho chisq p  
## shoppingCartValue -0.0168 0.907 0.3409  
## voucher -0.0155 0.770 0.3803  
## returned 0.0261 2.182 0.1397  
## gender=male 0.0390 4.922 0.0265  
## GLOBAL NA 8.528 0.0740

#If p < 0.05 we reject the hypothesis that this given variable meet the proportion of the hazard assumption  
#if the PH assumption is violeted "stratefied analysis" is relevant.  
  
# Plot time-dependent beta  
plot(testCPH, var = "gender=male")



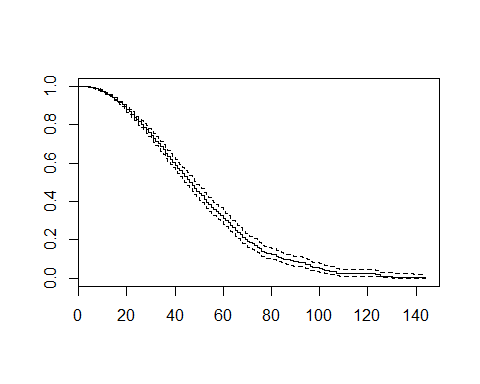
#Beta and T show that overtime the Beta is moving  
  
# Load rms package  
library(rms)  
  
# Validate model  
validate(fitCPH, method = "crossvalidation",  
 B = 10, dxy = TRUE, pr = FALSE)

## index.orig training test optimism index.corrected n  
## Dxy 0.1159 0.1159 0.1152 0.0007 0.1152 10  
## R2 0.0299 0.0300 0.0290 0.0011 0.0289 10  
## Slope 1.0000 1.0000 0.9837 0.0163 0.9837 10  
## D 0.0032 0.0033 0.0042 -0.0009 0.0041 10  
## U 0.0000 0.0000 0.0002 -0.0002 0.0002 10  
## Q 0.0032 0.0033 0.0040 -0.0007 0.0040 10  
## g 0.2380 0.2383 0.2330 0.0053 0.2328 10

#Method Crossvalidation with 10 folds  
  
#Unfortunately, the explanatory power of your model is rather low.   
#You could try to collect more explanatory variables.  
#R2 0.29  
  
##################################################################################  
#Predictions  
##################################################################################  
#Now you are going to predict the survival curve for a new customer from the Cox Proportional Hazard model  
# Create data with new customer  
newCustomer <- data.frame(daysSinceFirstPurch = 21, shoppingCartValue = 99.90, gender = "female", voucher = 1, returned = 0, stringsAsFactors = FALSE)  
  
# Make predictions  
pred <- survfit(fitCPH, newdata = newCustomer)  
print(pred)

## Call: survfit(formula = fitCPH, newdata = newCustomer)  
##   
## n events median 0.95LCL 0.95UCL   
## 5122 3199 47 44 49

plot(pred)



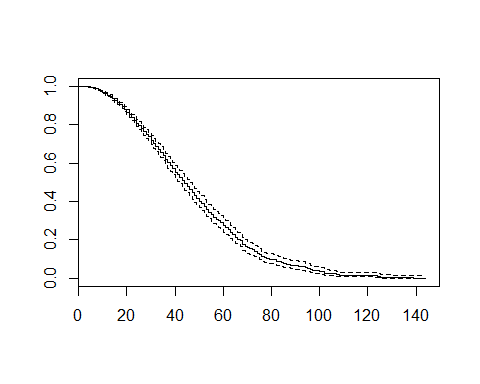
#predicted median time until the second order is 47  
  
#Observe the hazard probability of survivce of "newCustomer" on a given timestamps "X"  
str(survest(fitCPH, newdata = newCustomer, times = 50))

## List of 5  
## $ time : num 50  
## $ surv : num 0.441  
## $ std.err: num 0.0398  
## $ lower : num 0.408  
## $ upper : num 0.477

#On time 50 the estimate survival probability that this customer not chrun on time 50 would is 44%  
  
# Correct the customer's gender  
newCustomer2 <- newCustomer  
newCustomer2$gender <- "male"  
  
# Redo prediction  
pred2 <- survfit(fitCPH, newdata = newCustomer2)  
print(pred2)

## Call: survfit(formula = fitCPH, newdata = newCustomer2)  
##   
## n events median 0.95LCL 0.95UCL   
## 5122 3199 44 42 47

plot(pred2)



#the correction of the gender decreased the predicted median time   
#until the second order from 47 to 44 days.