HW1_group21

Group21

January 16, 2018

Loading libraries

```
library(knitr)
opts_chunk$set(tidy = TRUE)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(xtable)
options(digits=3)
theme_set(theme_bw())
```

Q1 BG model

Defining the log likelihood of BG

```
# log of P
log_P <- function(t, a, b) {
    lbeta(a + 1, b + t - 1) - lbeta(a, b)
}

# log of S
log_S <- function(t, a, b) {
    lbeta(a, b + t) - lbeta(a, b)
}

# log likelihood of Beta Geometric functions
```

```
LL <- function(pars, N, S) {
    a <- exp(pars[1])
    b <- exp(pars[2])
    Tobs <- length(N)
    t <- 1:Tobs
    LL_p <- sum(N * log_P(t, a, b))
    LL_s <- S * log_S(Tobs, a, b)
    LL_all <- LL_p + LL_s
    return(-LL_all)
}  #Scalar
```

Specifying the data

```
N0 <- 2132  #Starting Customers

Tobs <- 8  # No. of observations/renewal opportunities

full <- dplyr::data_frame(t = 1:8, S = c(1406, 1069, 894, 765, 656, 578, 525,
    482))

# S[j] = Survivors after period j-1 - survivors after period j
full <- mutate(full, N = c(N0 - S[1], -diff(S)))  #dataframe with interval, retained and churned custoe
calib <- filter(full, t <= 8)

pars.start <- c(1, 1)
res <- optim(pars.start, fn = LL, N = calib$N[1:Tobs], S = calib$S[Tobs])

res  # Optimized LL, a, b
```

```
## $par
## [1] -0.234  0.434
##
## $value
## [1] 3869
##
## $counts
## function gradient
##       67       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Optimizing a and b parameters of the Beta function

```
# Collecting the results
LL_mle <- -res$value
pars_mle <- res$par
a <- exp(pars_mle[1])
```

```r
b <- exp(pars_mle[2])
cat("LL = ", LL_mle, " a = ", a, "  b = ", b, "\n")
```

```
## LL =  -3869  a =  0.792   b =  1.54
```

```r
# Additional Info
log(a)
```

```
## [1] -0.234
```

```r
log(b)
```

```
## [1] 0.434
```

```r
c(a, b)
```

```
## [1] 0.792 1.543
```

```r
# Estimated mean churn
a/(a + b)
```

```
## [1] 0.339
```

Q2

(a)Probability that this customer will cancel service after only one month

```r
beta_n <- beta(a + 1, b + 1 - 1)
beta_d <- beta(a, b)
Churn1 <- beta_n/beta_d
Churn1
```

```
## [1] 0.339
```

```r
# Alternatively also given by exp(log_P(1,a,b)) #t=1.
# P(T=1|a,b)=(theta*(1-theta)^2)
```

(b)Probability that this customer will cancel service after 2 months

```r
beta_n2 <- beta(a + 1, b + 2 - 1)
beta_d2 <- beta(a, b)
Churn2 <- beta_n2/beta_d2
Churn2
```

```
## [1] 0.157
```

```r
# Alternatively also given by exp(log_P(2,a,b)) #t=2.
# P(T=2|a,b)=(theta*(1-theta)^2)
```

(c)Customer has renewed for February. What is the expected probability that he will renew for March?

```r
# Because we know that the customer has renwed for Feb, we need to calculate
# the value of posterior survival function after (n-1 = 1) renewal
# opportunities and will survive for another (t*=1)
Renew_March <- beta(a, b + 2 - 1 + 1)/beta(a, b + 2 - 1)
Renew_March
```

```
## [1] 0.763
```

(d)Expected renewal probability for a customer who remained active through August?

```r
# Customers who will have remained active through August will have 8 renewal
# opportunities. That they will still remain active is given by the
# posterior survival distribution of beta(a, b+n-1+t*)/beta(a, b+n-1), where
# (n-1=8) in this case
Renew_Sep <- beta(a, b + 8 - 1 + 1)/beta(a, b + 8 - 1)
Renew_Sep
```

```
## [1] 0.915
```

(e)How many members of the cohort do we expect to be active through the end of year?

```r
# For customers left at Sept, n-1=8, and we need to calculate if they will
# survive another (t*=3) periods
End_Year <- beta(a, b + 9 - 1 + 3)/beta(a, b + 9 - 1)
End_Year * 482
```

```
## [1] 387
```

```r
# Alternatively also given by: 2132*exp(log_S(11,a,b))
```

Q3

(a)Predicted counts from BG models

```r
##
sr1 <- beta(a, b + 1)/beta(a, b)  # Surviving customers after 0 renewals
sr2 <- beta(a, b + 2)/beta(a, b)  # Surviving customers after 1 renewals
sr3 <- beta(a, b + 3)/beta(a, b)  # Surviving customers after 2 renewals
sr4 <- beta(a, b + 4)/beta(a, b)  # Surviving customers after 3 renewals
sr5 <- beta(a, b + 5)/beta(a, b)  # Surviving customers after 4 renewals
sr6 <- beta(a, b + 6)/beta(a, b)  # Surviving customers after 5 renewals
sr7 <- beta(a, b + 7)/beta(a, b)  # Surviving customers after 6 renewals
sr8 <- beta(a, b + 8)/beta(a, b)  # Surviving customers after 7 renewals
sr9 <- beta(a, b + 9)/beta(a, b)  # Surviving customers after 8 renewals
sr10 <- beta(a, b + 10)/beta(a, b)  # Surviving customers after 9 renewals
sr11 <- beta(a, b + 11)/beta(a, b)  # Surviving customers after 10 renewals
sr12 <- beta(a, b + 12)/beta(a, b)  # Surviving customers after 11 renewals
```

```r
list_sr <- c(sr1, sr2, sr3, sr4, sr5, sr6, sr7, sr8, sr9, sr10, sr11)

survival <- 2132 * list_sr  # vector of surviving customers for t=1:11

Chart <- data.frame(Month = c("September", "October", "November", "December",
    "January", "February", "March", "April", "May", "June", "July", "August"),
    Sept = c(2132, survival[1:11]), Oct = c(0, 2132, survival[1:10]), Nov = c(0,
        0, 2132, survival[1:9]), Dec = c(0, 0, 0, 2132, survival[1:8]))
Chart <- mutate(Chart, total = Sept + Oct + Nov + Dec)
Chart
```

```
##        Month Sept  Oct  Nov  Dec total
## 1  September 2132    0    0    0  2132
## 2    October 1409 2132    0    0  3541
## 3   November 1075 1409 2132    0  4616
## 4   December  878 1075 1409 2132  5494
## 5    January  748  878 1075 1409  4110
## 6   February  655  748  878 1075  3356
## 7      March  584  655  748  878  2865
## 8      April  528  584  655  748  2515
## 9        May  484  528  584  655  2251
## 10      June  447  484  528  584  2043
## 11      July  415  447  484  528  1874
## 12    August  389  415  447  484  1734
```

(b) Retention rates

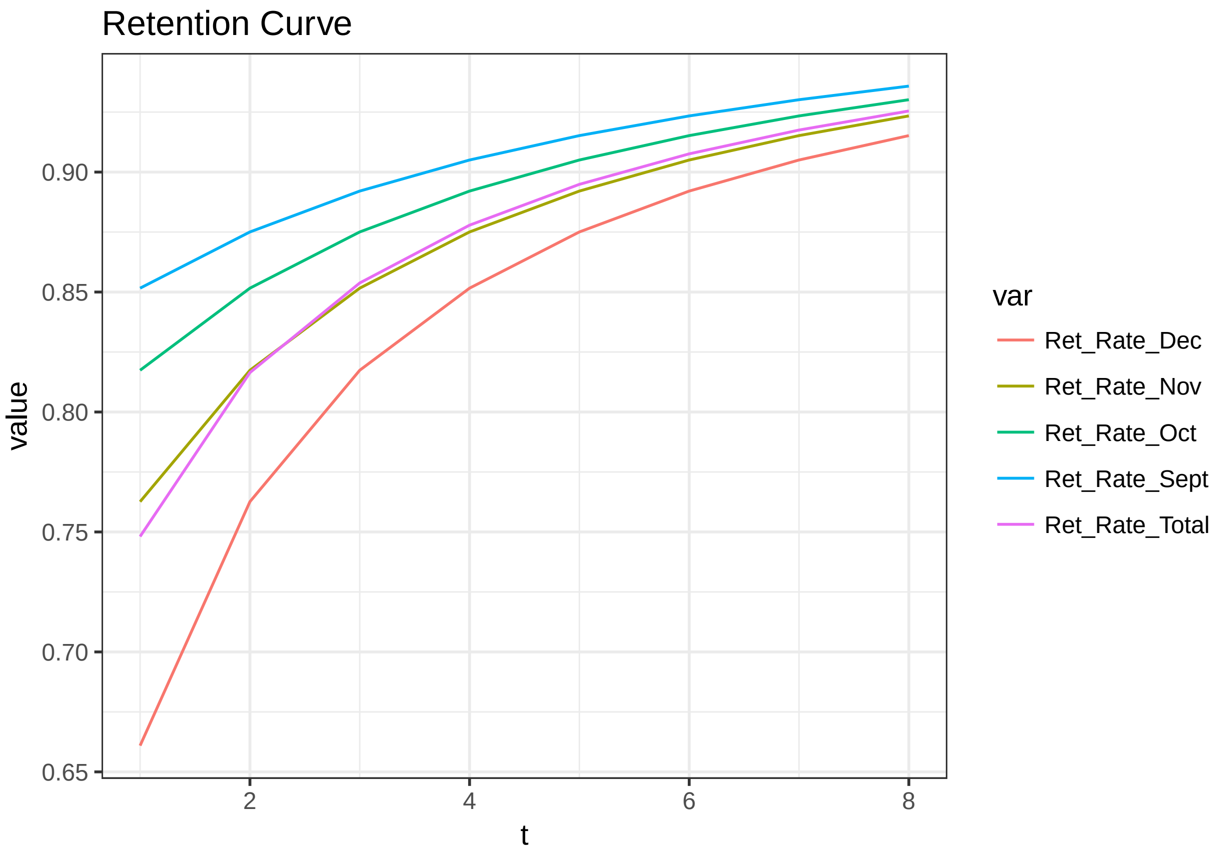
```r
options(digits = 3)

retention <- data.frame(t = c(1:8), Ret_Rate_Sept = Chart$Sept[5:12]/Chart$Sept[4:11],
    Ret_Rate_Oct = Chart$Oct[5:12]/Chart$Oct[4:11], Ret_Rate_Nov = Chart$Nov[5:12]/Chart$Nov[4:11],
    Ret_Rate_Dec = Chart$Dec[5:12]/Chart$Dec[4:11], Ret_Rate_Total = Chart$total[5:12]/Chart$total[4:11
retention  # retention rates from Jan to August
```

```
##   t Ret_Rate_Sept Ret_Rate_Oct Ret_Rate_Nov Ret_Rate_Dec Ret_Rate_Total
## 1 1         0.852        0.817        0.763        0.661          0.748
## 2 2         0.875        0.852        0.817        0.763          0.816
## 3 3         0.892        0.875        0.852        0.817          0.854
## 4 4         0.905        0.892        0.875        0.852          0.878
## 5 5         0.915        0.905        0.892        0.875          0.895
## 6 6         0.923        0.915        0.905        0.892          0.908
## 7 7         0.930        0.923        0.915        0.905          0.918
## 8 8         0.936        0.930        0.923        0.915          0.925
```

```r
plot1 <- gather(retention, var, value, -t)


Rplot1 <- ggplot(plot1, aes(x = t, y = value, group = var, col = var)) %>% +geom_line() %>%
    +ggtitle("Retention Curve")
Rplot1
```



When comparing retention rates for each cohort, you can see in the chart below that the September cohort has a much higher retention rate than the December cohort. September also has a higher rate than October and November, but the gap between September and each month closes as they get closer to September as more loyal customers are retained. Because we are examining retention rates of each cohort from January to August, this makes sense. The December cohort was just acquired the previous month. We do not have any prior information on this cohort, and since they have recently been acquired, they will have a much less retention rate in January and the following months compared to the September cohort that has already been active for a few months. As customers survive more renewal opportunities, the expected churn probability decreases. This is why September will have the highest retention rate, and each month following will have a lower retention rate compared to the month before.