NBD, NBD with HCNB

2/9/2020

```
library('dplyr')
library('ggplot2')
library('knitr')
```

```
#inport data
data <- read.table("khakichinos.txt", header = TRUE)
options(digits = 3)
```

1.

Take the parameter estimates for the NBD, and NBD with HCNB models from the coffee creamer example. The penetration of a consumer good is the proportion of the target segment that purchases at least one unit during some period of time. It is the analog to "reach" from media metrics. (a) What do each of the models predict will be the penetration of coffee creamer purchases after 1 period? What about after 5 periods? 26 periods? (b) If you were the product manager for this particular brand of coffee creamer, and your manager asked you to estimate the maximum penetration that you could expect from this product, what one answer would you give? How would you justify the answer? The results from the NBD model are 0.407,0.553, and 0.668 respectively for periods 1, 5, and 26. The results from the NBD model with HCNB are 0.183, 0.248, and 0.300 respectively for periods 1,5, and 26.

```
#Question 1
##a)
###NBD model prediction
r_NBD <- 0.181
alpha_NBD <- 0.059
T <- c(1,5,26)
NBD_x0 <- (alpha_NBD/(alpha_NBD+T))^r_NBD
Reach_NBD <-1-NBD_x0
Reach_NBD
```

```
## [1] 0.407 0.553 0.668
```

```
###NBD with HCNB
r_HCNB <- 1.226
alpha_HCNB <- 0.215
pi <- 0.551
HCNB_x0 <- pi + (1-pi)*NBD_x0
Reach_HCNB <- 1-HCNB_x0
Reach_HCNB
```

```
## [1] 0.183 0.248 0.300
```

From the coffee creamer example, we knew that the prediction of the NBD model was not a good fit for the actual data. The first several counts were overestimated, and the later counts were underpredicted (based on the graph shown in Class slide P42). When there are such patterns in prediction results, it means that we may have a systematic error in our model. As a result, we introduced a new parameter pi to depict the hard-core-never-buy customers in the population. From the graph on class slide P48, we learned that the prediction results from the NBD model with HCNB aligned well with the actual data. Hence, we should use the estimates from the NBD with the HCNB model to estimate the maximum penetration, though the numbers from the NBD model were higher than the numbers from the NBD-HCNB model.

2.

The following table, taken from a paper by Bickart and Schmittlein (1999), shows the number of surveys filled out by a sample of 1,865 Americans in 1995. Fit an NBD model to these data, discuss and interpret the parameter estimates, and forecast how many respondents are expected to complete X = 0,1, 2,. . . ,10 surveys.

```
coffee_table <- tibble("Number of surveys"="Number of respondents","0"=1020, "1"=166,
                "2"=270,"3-5"=279,"6+"=130)
coffee_table <- kable(coffee_table, digits = 2, align = c("c","c","c"))
coffee_table
```

Number of surveys	0	1	2	3-5	6+
Number of respondents	1020	166	270	279	130

We used the data from Bickart and Schmittlein's paper and generated an NBD model. Part of the data given was censored, but to keep the accuracy of the prediction, we did not censor other parts of the data and keep the data as it was. The resulting parameters for the NBD models are r=0.391, alpha=0.248. Since r and alpha are both less than 1, the prediction curve is L-shaped, which is aligned with the data given. The mean number of services taken is r/alpha=1.581, and the heterogeneity of the distribution is captured by a variance of 1.581 services. The forecasted numbers of respondents by the number of services is shown below.

```
#Question 2
##get parameterss for the NBD model
###pars = c(log r,log theta)
LL_NBD_special <- function(pars){
  r <- exp(pars[1])
  alpha <- exp(pars[2])
  p0 <- (gamma(r+0)/(gamma(r)*factorial(0)))*((alpha/(alpha+1))^r)*((1/(alpha+1))^0)
  p1 <- (gamma(r+1)/(gamma(r)*factorial(1)))*((alpha/(alpha+1))^r)*((1/(alpha+1))^1)
  p2 <- (gamma(r+2)/(gamma(r)*factorial(2)))*((alpha/(alpha+1))^r)*((1/(alpha+1))^2)
  p3 <- (gamma(r+3)/(gamma(r)*factorial(3)))*((alpha/(alpha+1))^r)*((1/(alpha+1))^3)
  p4 <- (gamma(r+4)/(gamma(r)*factorial(4)))*((alpha/(alpha+1))^r)*((1/(alpha+1))^4)
  p5 <- (gamma(r+5)/(gamma(r)*factorial(5)))*((alpha/(alpha+1))^r)*((1/(alpha+1))^5)
  p6plus <- 1-(p0+p1+p2+p3+p4+p5)
  LL_x0 <- 1020*log(p0)
  LL_x1 <- 166*log(p1)
  LL_x2 <- 270*log(p2)
  LL_x3to5 <- 279*log(p3+p4+p5)
  LL_6plus <- 130*log(p6plus)
  LL_sum <- LL_x0 + LL_x1 + LL_x2 + LL_x3to5 + LL_6plus
  return(-LL_sum)
}

start <- c(log(1),log(1))
opt_NBD_special <- optim(start,fn=LL_NBD_special)
r2 <- exp(opt_NBD_special[["par"]][1])
alpha2 <- exp(opt_NBD_special[["par"]][2])
mean_Lambda = r2/alpha2
c(r2,alpha2,mean_Lambda)
```

```
## [1] 0.391 0.248 1.581
```

We also fit our NBD model to these data, discussed and interpreted the parameter estimates, and forecasted how many respondents are expected to complete X = 0,1,2, . . . ,10 surveys. Please see the table below.

```r
##Prediction
p_NBD <- function(pars,x,T){
  r <- pars[1]
  alpha <- pars[2]
  p_x <- (gamma(r+x)/(gamma(r)*factorial(x)))*((alpha/(alpha+T))^r)*((T/(alpha+T))^x)
  return(p_x)
}

X <- 0:10
T <- 1
pars2 <- c(r2,alpha2)
N <- 1865
Respondents <- p_NBD(pars = pars2,x = X,T)*N
table <- tibble(X=0:10, pre_Respondents = Respondents)
table <- kable(table, digits = 0, align = c("c","c"))
table
```

X	pre_Respondents
0	990
1	311
2	173
3	111
4	75
5	53
6	38
7	28
8	21
9	15
10	12

3.

khakichinos.txt, that contains Internet visit data for a sample of 2,728 comScore/Media Metrix panelists who visited one particular site (with the disguised name of khakichinos.com) in July, 2014. Ignore the covariate data (the demographic information) for now.

(a) Fit an NBD model to the data. What do the parameter estimates tell us about the different kinds of users in the Khakichinos website user base?
(b) Plot the expected reach of the Khakichinos website as a function of time, from 0 to 36 months. What is the expected reach after 12 months?

This time, we used data from Canvas, and built an NBD model. The estimated parameters for the gamma distribution are r=0.134 and alpha=0.141. Both parameters are less than one, which gives us a L-shaped prediction distribution. The mean exposure rate taken is r/alpha=0.949, and the heterogeneity of the distribution is captured by a variance of 0.949.

```r
#Question 3
data3 <- select(data, ID, Visits)

##a)
###get the parameters
LL_NBD <- function(pars,T,x){
  r <- exp(pars[1])
  alpha <- exp(pars[2])
```

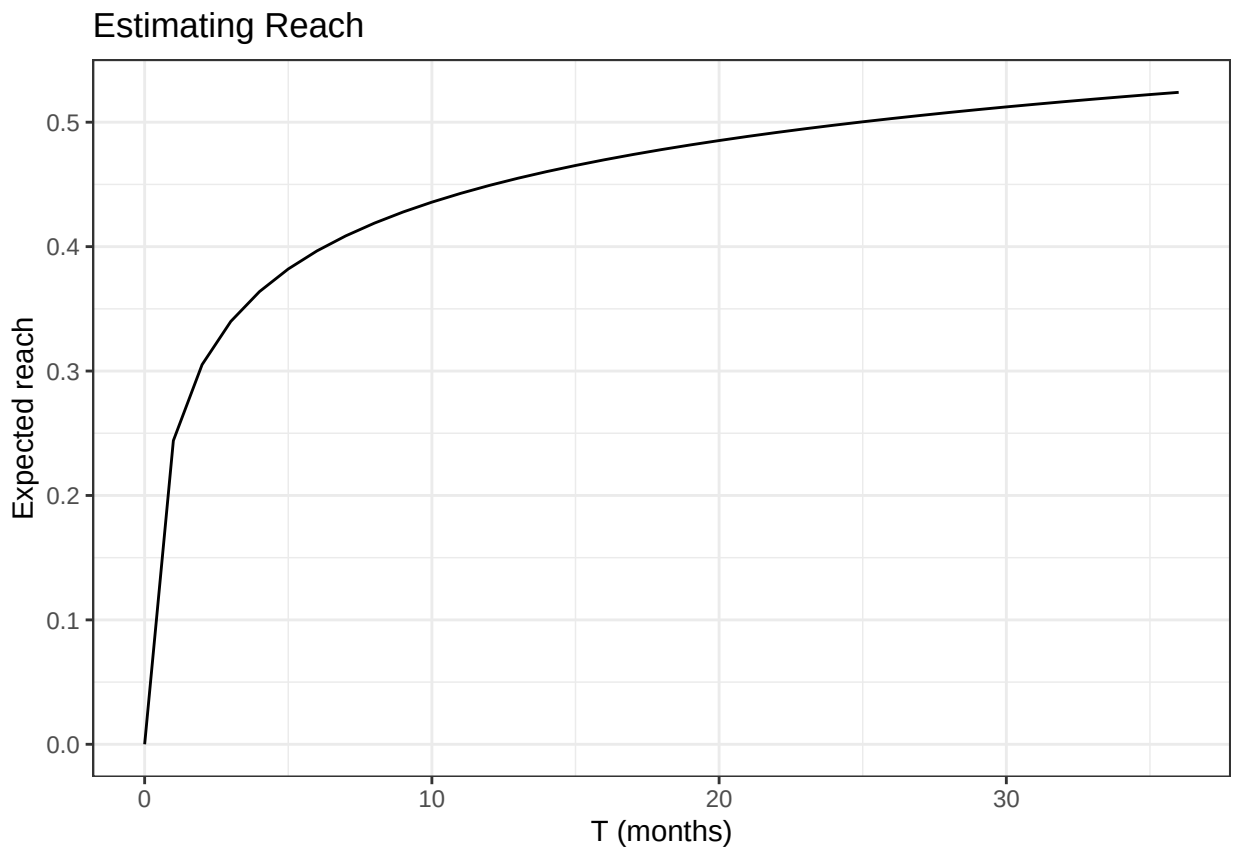
```
  p_x <- (gamma(r+x)/(gamma(r)*factorial(x)))*((alpha/(alpha+T))^r)*((T/(alpha+T))^x)
  LL_individual <- log(p_x)
  LL_sum <- sum(LL_individual)
  return(-LL_sum)
}

T <- rep(1,length(data3$ID))
start <- c(log(1),log(1))
opt_NBD <- optim(start,fn=LL_NBD,T=T,x=data3$Visits)
##T also need to be a array, as long as data3$Visits
r3 <- exp(opt_NBD[["par"]][1])
alpha3 <- exp(opt_NBD[["par"]][2])
Var3 <- r3/alpha3
c(r3,alpha3,Var3)
```

```
## [1] 0.134 0.141 0.949
```

We plot the expected reach of the Khakichinos website as a function of time, from 0 to 36 months which shows below.

```
##b)
T <- 0:36
Reach <- 1-(alpha3/(alpha3+T))^r3
table <- tibble(T=T,Reach=Reach)
plot <- table %>% ggplot(aes(x=T,y=Reach)) %>%
  + geom_line() %>%
  + theme_bw() %>%
  + ggtitle("Estimating Reach")%>%
  + scale_x_continuous("T (months)") %>%
  + scale_y_continuous("Expected reach")
plot
```



We also calculated the expected reach after 12 months is 0.0748

```
Reach_36month <- 1-(alpha3/(alpha3+36))^r3
Reach_12month <- 1-(alpha3/(alpha3+12))^r3
Reach_after12months <- Reach_36month - Reach_12month
Reach_after12months
```

```
## [1] 0.0748
```

4.

For a typical consumer good, maintain the standard assumptions of the NBD model, where the purchase rate λ is expressed in units per month, and the mixing distribution of λ is gamma(r; α). (a) How many units do you predict a randomly-chosen household will make in a typical month? (b) What is the posterior distribution of λ|x, T for a customer who purchased x units in the last T months? (For this question, I am essentially asking you to derive the Bayes update for the NBD model. We did not do this in class, but the same principles that hold for other models will hold here as well). (c) The following table includes visit counts to the Khakichinos website (from Question 3) for five randomly selected users, during the past three months. Complete the table by estimating, for each user, the expected number of visits for the next month, the month after, and the two months after that. That is, if we are at the end of Month 3, what are the expected visit counts for Month 4 alone, Month 5 alone, and Months 6 and 7 combined? Use your parameter estimates from above to answer this question. To be clear, the only updating of beliefs occurs at the end of Month 3.

Months	→	Observed counts 1	2	3	Expected counts 4	5	6 and 7
	A	0	0	0			
	B	2	0	0			
User	C	0	5	4			
	D	0	0	1			
	E	6	5	4			

The units we can predict a randomly-chosen household will make in a typical month is:

$$E(X(T=1)) = r/\alpha$$

Also,we know

$$Conditional = \frac{(\lambda T)^x e^{-\lambda T}}{x!}$$

$$Prior = \frac{(\alpha)^r}{\Gamma(r)} \lambda^{r-1} e^{-\alpha\lambda}$$

$$Marginal = \frac{\Gamma(r+x)}{\Gamma(r)x!} (\frac{\alpha}{\alpha+T})^r (\frac{T}{\alpha+T})^x$$

$$Posterior = Prior * \frac{Conditinal}{Marginal}$$

So We can calculate mathematical formula for the posterior distribution of λ|x; T for a customer who purchased x units in the last T months:

$$gamma(\lambda|X(T)=x,r,\alpha) = \frac{(\alpha)^r}{\Gamma(r)} \lambda^{r-1} e^{-\alpha\lambda} * \frac{\frac{(\lambda T)^x e^{-\lambda T}}{x!}}{\frac{\Gamma(r+x)}{\Gamma(r)x!} (\frac{\alpha}{\alpha+T})^r (\frac{T}{\alpha+T})^x} = gamma(r+x, \alpha+T)$$

Assuming we are at the end of Month 3, the expected visit counts for Month 4 alone, Month 5 alone, and Months 6 and 7 combined shows below:

```
##c)

data4 <- tibble(User=c("A","B","C","D","E"),"1"=c(0,2,0,0,6), "2"=c(0,0,5,0,5),
                "3"=c(0,0,4,1,4))
data5 <- mutate(data4, sum_m123 = rowSums(data4[,c("1","2","3")]))
Expected_4 <- round(1*((r3+data5$sum_m123)/(alpha3+3)),2)
Expected_5 <- round(1*((r3+Expected_4)/(alpha3+1)),2)
Expected_6and7 <- round(2*((r3+Expected_5)/(alpha3+1)),2)

table <- mutate(data4,"4"=Expected_4, "5"=Expected_5,
                "6 and 7" = Expected_6and7)
table <- kable(table, digits = 2, align = c("c","c","c"))
table
```

User	1	2	3	4	5	6 and 7
A	0	0	0	0.04	0.15	0.50
B	2	0	0	0.68	0.71	1.48
C	0	5	4	2.91	2.67	4.91
D	0	0	1	0.36	0.43	0.99
E	6	5	4	4.82	4.34	7.84