Homework4

Yiheng

7/17/2020

Diffusion of Tetracycline

1.

```
#load data
ckm_nodes <- read_csv('data/ckm_nodes.csv')
```

```
## Parsed with column specification:
## cols(
##   city = col_character(),
##   adoption_date = col_double(),
##   medical_school = col_character(),
##   attend_meetings = col_character(),
##   medical_journals = col_double(),
##   free_time_with = col_character(),
##   discuss_medicine_socially = col_character(),
##   club_with_drs = col_character(),
##   drs_among_three_best_friends = col_double(),
##   practicing_here = col_character(),
##   office_visits_per_week = col_character(),
##   proximity_to_other_drs = col_character(),
##   specialty = col_character()
## )
```

```
ckm_network <- read.table('data/ckm_network.dat')
#eliminate
anum <- which(!is.na(ckm_nodes$adoption_date))
ckm_nodes <- ckm_nodes[anum,]
ckm_network <- ckm_network[anum,anum]
```

2. There are 125 doctors in the study, and 17 months in the study period. To make a tidy data frame, there should be 17 rows for each doctor, 125*17 = 2125. 6 columns are “doctor”, “month” and 4 variables asked.

```
#create new dataframe and fill in doctors and months
records <- matrix(seq(1,125), nrow = 2125) %>% sort()
records <- data.frame("doctor.No"= records ,"month"=seq(1,17))
#columns 3&4
records <- data.frame(records,"begin.this.month" = (rep(ckm_nodes$adoption_date,times = 17)[records$doc
                      "began.before" =
(rep(ckm_nodes$adoption_date, times = 17)[records$doctor.No]< records$month))

#columns 5&6

a <- data.frame(which(ckm_network[ceiling(c(1:2125)/17),]==1 ,arr.ind = TRUE))
a <- a[order(a[,1]),]
# create a new column of the index we want to look up in records$began.before
# `ifelse()` function is capable of vector type condition
a<- mutate(a,no=ifelse(a$row%%17==0,a$col*17,(a$col-1)*17+a$row%%17))
x<- data.frame(rowname = as.character(c(1:2125)))
# `tapply()` function's most useful for avoiding for() loop
y <- data.frame(contact.before = tapply(records$began.before[a$no],INDEX = a$row,sum))
y<-rownames_to_column(y,var = "rowname")
contact.before <- left_join(x,y,by = "rowname")

z <- data.frame(contact.thismonth = tapply(records$begin.this.month[a$no],INDEX = a$row,sum))
z<- rownames_to_column(z,var = "rowname")
contact.thismonth<- left_join(x,z,by = "rowname")
records<-data.frame(records,contact.before = contact.before$contact.before,
                    contact.orbefore = contact.before$contact.before+contact.thismonth$contact.thismonth

# it's easy to use for() directly, but inefficient

# contact.strictlybefore.num =matrix()
# contact.before.num = matrix()
# for(i in 1:2125){
#   contact.strictlybefore.num[i] <- sum(ckm_nodes$adoption_date[which(ckm_network[ceiling(i/17),]==1)]
#   contact.before.num[i] <- sum(ckm_nodes$adoption_date[which(ckm_network[ceiling(i/17),]==1)] <= reco
#   }
# records <- data.frame(records,contact.strictlybefore.num,contact.before.num)
```

3.a.

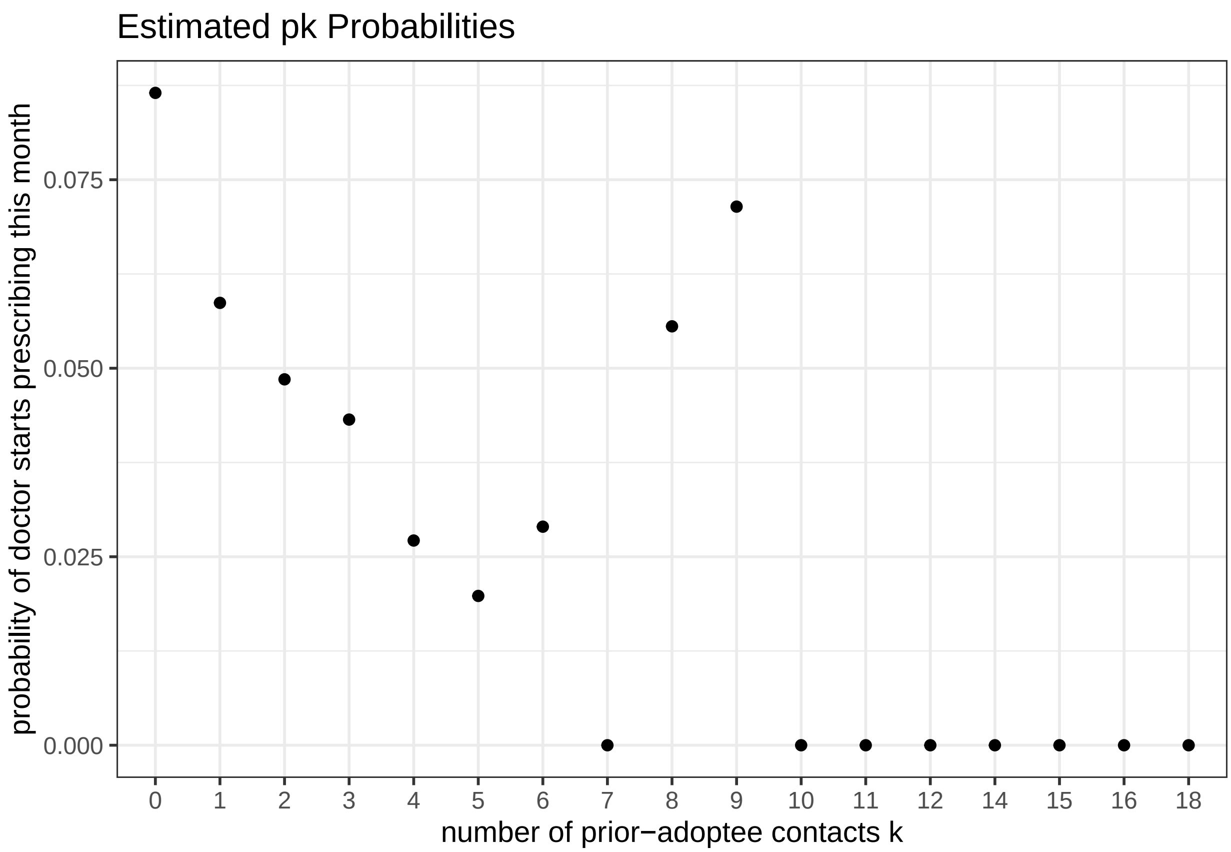
```
max(apply(ckm_network,1,sum))
```

```
## [1] 20
```

⇒ There should be no more than 21 values of k as the doctors has atmost 20 contacts.

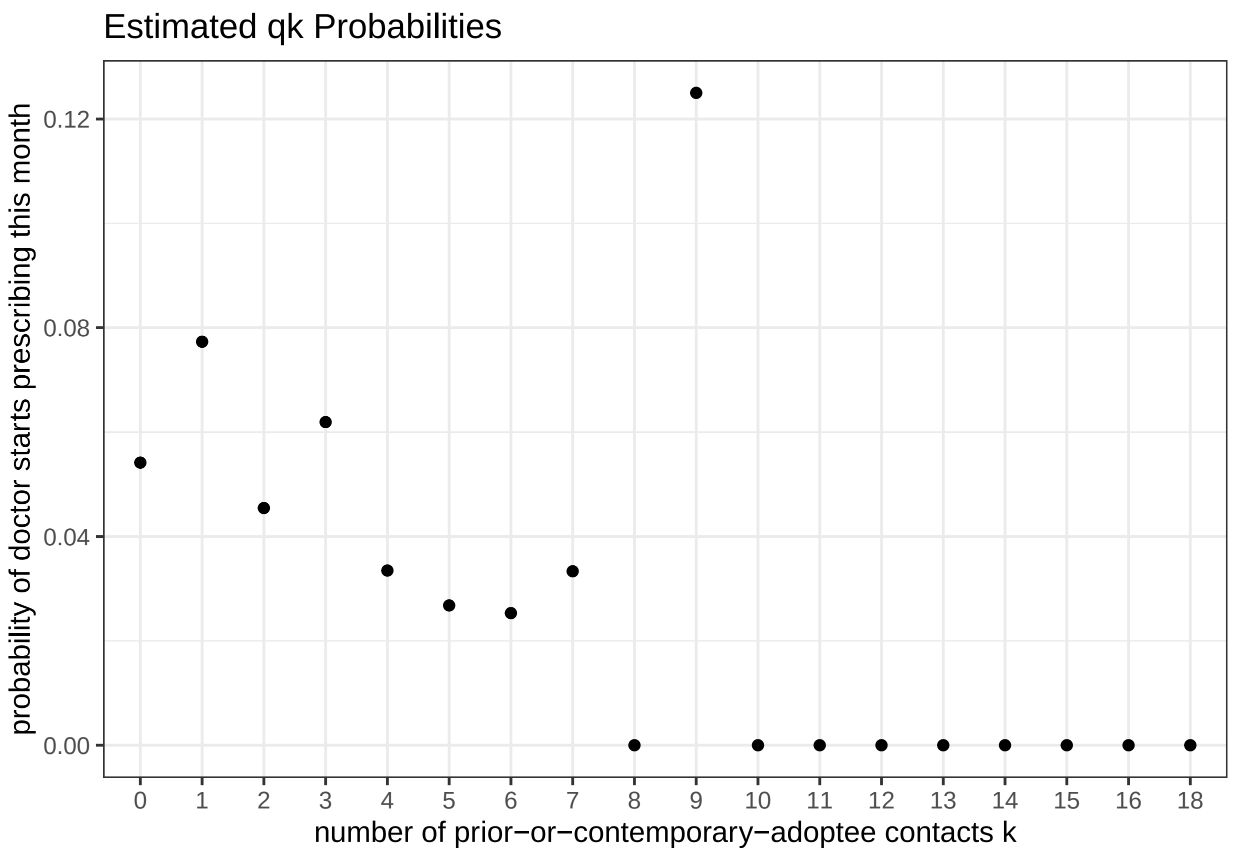
b.

```
pk_estimate <- data.frame(tapply(records$begin.this.month,
                                 INDEX = records$contact.before,sum)
                          /table(records$contact.before))
names(pk_estimate) <- c("k","pk")
ggplot(data = pk_estimate)+
  geom_point(aes(x = k,y = pk))+
  labs(x = "number of prior-adoptee contacts k",
       y = "probability of doctor starts prescribing this month",
       title = "Estimated pk Probabilities")+
  theme_bw()
```



c.

```
qk_estimate <- data.frame(tapply(records$begin.this.month,
                                 INDEX = records$contact.orbefore,sum)
                          /table(records$contact.orbefore))
names(qk_estimate) <- c("k","qk")
ggplot(data = qk_estimate)+
  geom_point(aes(x = k,y = qk))+
  labs(x = "number of prior-or-contemporary-adoptee contacts k",
       y = "probability of doctor starts prescribing this month",
       title = "Estimated qk Probabilities")+
  theme_bw()
```



4.a.

```
pk <- c(pk_estimate$pk)
k <- c(pk_estimate$k)
mod1 <- lm(pk~1+k)
summary(mod1)
```

```
##
## Call:
## lm(formula = pk ~ 1 + k)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030196 -0.012898 -0.004249  0.004400  0.049882
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.064792   0.010393   6.234  1.6e-05 ***
## k           -0.004325   0.001014  -4.264 0.000679 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02049 on 15 degrees of freedom
## Multiple R-squared:  0.5479, Adjusted R-squared:  0.5178
## F-statistic: 18.18 on 1 and 15 DF,  p-value: 0.0006794
```

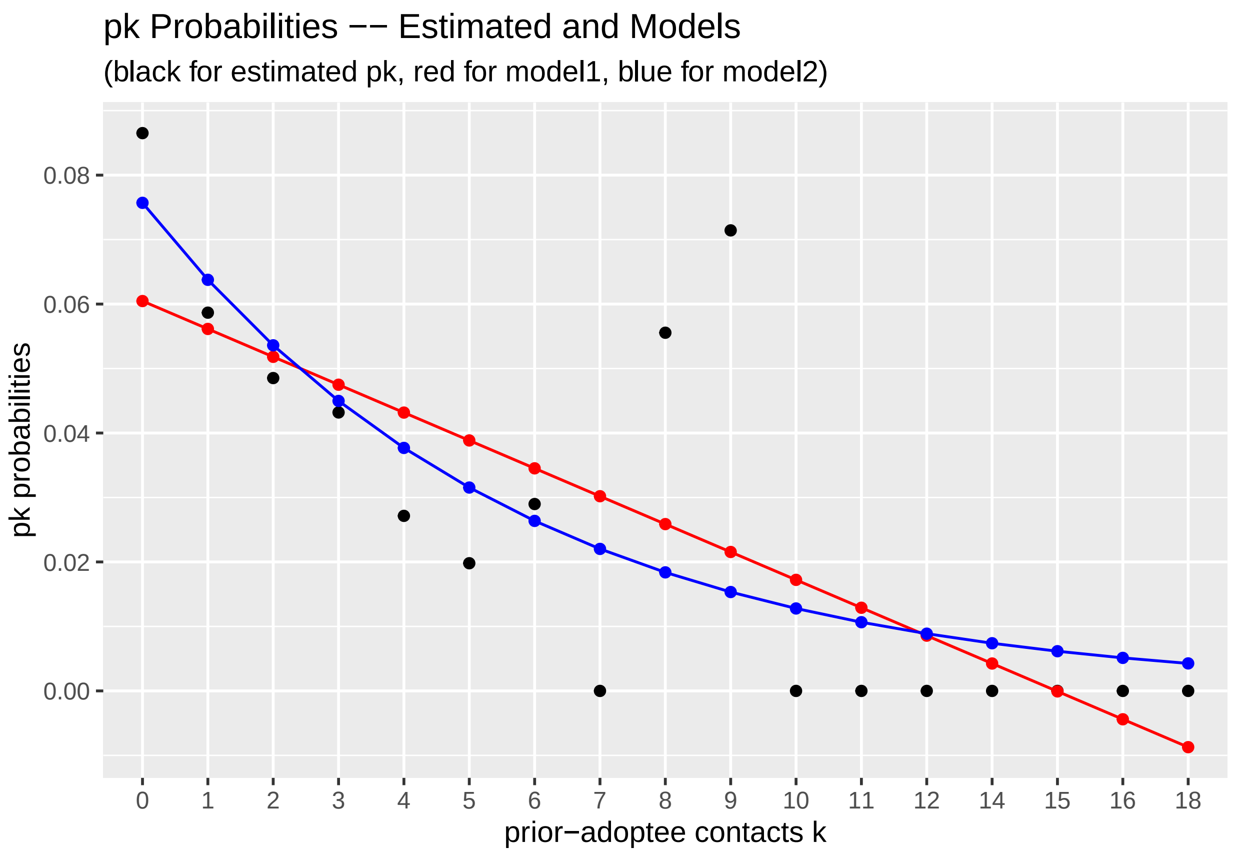
b.

```r
mod2 <- nls(pk~exp(a+b*k)/(1+exp(a+b*k)),start = list(a = -2,b = -0.2))
summary(mod2)
```

```
##
## Formula: pk ~ exp(a + b * k)/(1 + exp(a + b * k))
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a -2.31783    0.24285  -9.544  9.2e-08 ***
## b -0.18443    0.05569  -3.312  0.00474 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.01989 on 15 degrees of freedom
##
## Number of iterations to convergence: 4
## Achieved convergence tolerance: 5.281e-07
```

c.

```r
ggplot(pk_estimate,aes(k,pk,group = 1)) +
         geom_point() +
         geom_point(aes(k,fitted(mod1)),color = "red") +
  geom_line(aes(k,fitted(mod1)),color = "red") +
         geom_point(aes(k,fitted(mod2)),color = "blue") +
  geom_line(aes(k,fitted(mod2)), color = "blue") +
  labs(x = "prior-adoptee contacts k",
       y = "pk probabilities",
       title = "pk Probabilities -- Estimated and Models",
       subtitle = "(black for estimated pk, red for model1, blue for model2) ")
```


pk Probabilities −− Estimated and Models
(black for estimated pk, red for model1, blue for model2)
pk probabilities
0.08
0.06
0.04
0.02
0.00
0
1
2
3
4
5
6
7
8
9
10
11
12
14
15
16
18
prior−adoptee contacts k