Homework 4: Diffusion of Tetracycline

Zhang YunMengGe_3170105497

2020/7/12

We continue examining the diffusion of tetracycline among doctors in Illinois in the early 1950s, building on our work in lab 6. You will need the data sets `ckm_nodes.csv` and `ckm_network.dat` from the labs.

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v dplyr   1.0.0
## v tibble  3.0.1     v stringr 1.4.0
## v tidyr   1.1.0     v forcats 0.5.0
## v purrr   0.3.4
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'purrr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(conflicted)
```

```
## Warning: package 'conflicted' was built under R version 4.0.2
```

```
conflict_prefer("filter", "dplyr")
```

```
## [conflicted] Will prefer dplyr::filter over any other package
```

```
conflict_prefer("select", "dplyr")
```

```
## [conflicted] Will prefer dplyr::select over any other package
```

```
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')
#ckm_network
```

1. Clean the data to eliminate doctors for whom we have no adoption-date information, as in the labs. Only use this cleaned data in the rest of the assignment.

```
nadate<-which(is.na(ckm_nodes["adoption_date"]))
ckm_nodes_clean<-ckm_nodes[-nadate,]
#ckm_nodes_clean
```

2. Create a new data frame which records, for every doctor, for every month, whether that doctor began prescribing tetracycline that month, whether they had adopted tetracycline before that month, the number of their contacts who began prescribing strictly *before* that month, and the number of their contacts who began prescribing in that month or earlier. Explain why the dataframe should have 6 columns, and 2125 rows. Try not to use any loops.

```
ckm_network_clean<-ckm_network[-nadate,-nadate]
dim(ckm_network_clean)
```

```
## [1] 125 125
```

```r
maxmonth<-unique(ckm_nodes_clean$adoption_date)
maxmonth<-sort(maxmonth)
maxmonth
```

```
##  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17 Inf
```

```r
sum(is.finite(maxmonth))
```

```
## [1] 17
```

We can see that there are 17 months except the INF.

```r
Doctor<-rep(1:125,each=17)
Month<-rep(1:17,times=125)
```

```r
pre_this_month<-rep(0,times=2125)
FUN1<-function(data,i=seq(1,2125,by=17)){
  m=(i-1)/17+1
  if(ckm_nodes_clean$adoption_date[m]!=Inf){
    data[i-1+ckm_nodes_clean$adoption_date[m]]=1
  }
  return(data)
}
pre_this_month<-FUN1(pre_this_month)
```

```r
pre_before_this_month<-array(pre_this_month,c(17,125) )
FUN2<-function(c){
  n<-length(c)
  index<-which(c==1)
  if(length(index)==1&&index!=17){
    c[index]<-0
    c[(index+1):n]<-1
  }
  else if(length(index)==1){
    c[index]<-0
  }
  return(c)
}
pre_before_this_month<-apply(pre_before_this_month, 2, FUN2)
pre_before_this_month<-as.vector(pre_before_this_month)
#pre_before_this_month
```

```r
count_contacts <- function(month_i, contacts){
    return(sum(month_i > ckm_nodes_clean[contacts, 2]))
}
FUN3 <- function(i){
    contacts <- which(ckm_network_clean[, i]==1)
    return(sapply(seq(1, 17), count_contacts, contacts=contacts))
}
num_before <- array(sapply(seq(1, 125), FUN3))
```

```
count_contacts_plus <- function(month_i, contacts){
    return(sum(month_i >= ckm_nodes_clean[contacts, 2]))
}
FUN4 <- function(i){
    contacts <- which(ckm_network_clean[, i]==1)
    return(sapply(seq(1, 17), count_contacts_plus, contacts=contacts))
}
num_before_this<- array(sapply(seq(1, 125), FUN4))
```

```
RECORDS<-data.frame(Doctor,Month,pre_this_month,pre_before_this_month,num_before,num_before_this)
head(RECORDS)
```

```
##   Doctor Month pre_this_month pre_before_this_month num_before num_before_this
## 1      1     1              1                      0          0               1
## 2      1     2              0                      1          1               1
## 3      1     3              0                      1          1               2
## 4      1     4              0                      1          2               3
## 5      1     5              0                      1          3               3
## 6      1     6              0                      1          3               3
```

From the question there are 6 columns. What is more, with 125 doctors and 17 months, the row number is 125*17 = 2125.

3. Let

   p_k = Pr(A doctor starts prescribing tetracycline this month | Number of doctor's contacts prescribing before this mon

   and

   q_k = Pr(A doctor starts prescribing tetracycline this month | Number of doctor's contacts prescribing this month = k)

We suppose that p_k and q_k are the same for all months. a. Explain why there should be no more than 21 values of k for which we can estimate p_k and q_k directly from the data.

```
max(unique(RECORDS$num_before))
```

```
## [1] 18
```

```
max(unique(RECORDS$num_before_this))
```

```
## [1] 18
```

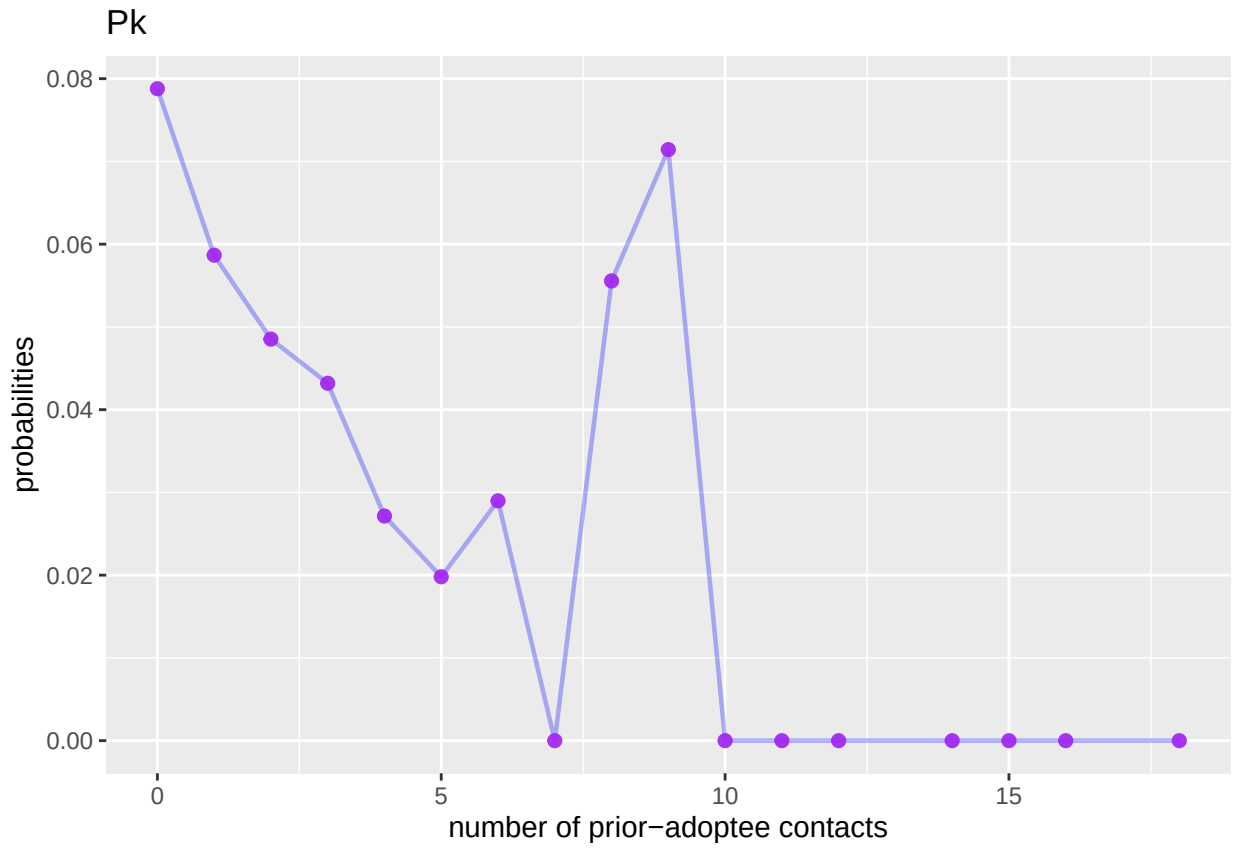
The number of contacts is no more than 21, So there should be no more than 21 values of k for which we can estimate p_k and q_k directly from the data.

b. Create a vector of estimated p_k probabilities, using the data frame from (2). Plot the probabilities against the number of prior-adoptee contacts k.

```r
k_1<-sort(unique(RECORDS$num_before))

FUN5<-function(i){
  int<-which(RECORDS$num_before==i)
  N<-sum(RECORDS$pre_this_month[int]==1)
  M<-length(int)
  return(N/M)
}
pk<-sapply(k_1,FUN5)
Pdata<-data.frame(k_1,pk)
```

```r
Pdata %>% ggplot()+
  geom_line(aes(x =k_1, y =pk),color="blue",alpha=0.3,size=0.8) +
  geom_point(aes(x =k_1, y =pk),color="purple",alpha=0.9,size=2) +
labs(x = "number of prior-adoptee contacts",y = "probabilities",title = "Pk")
```



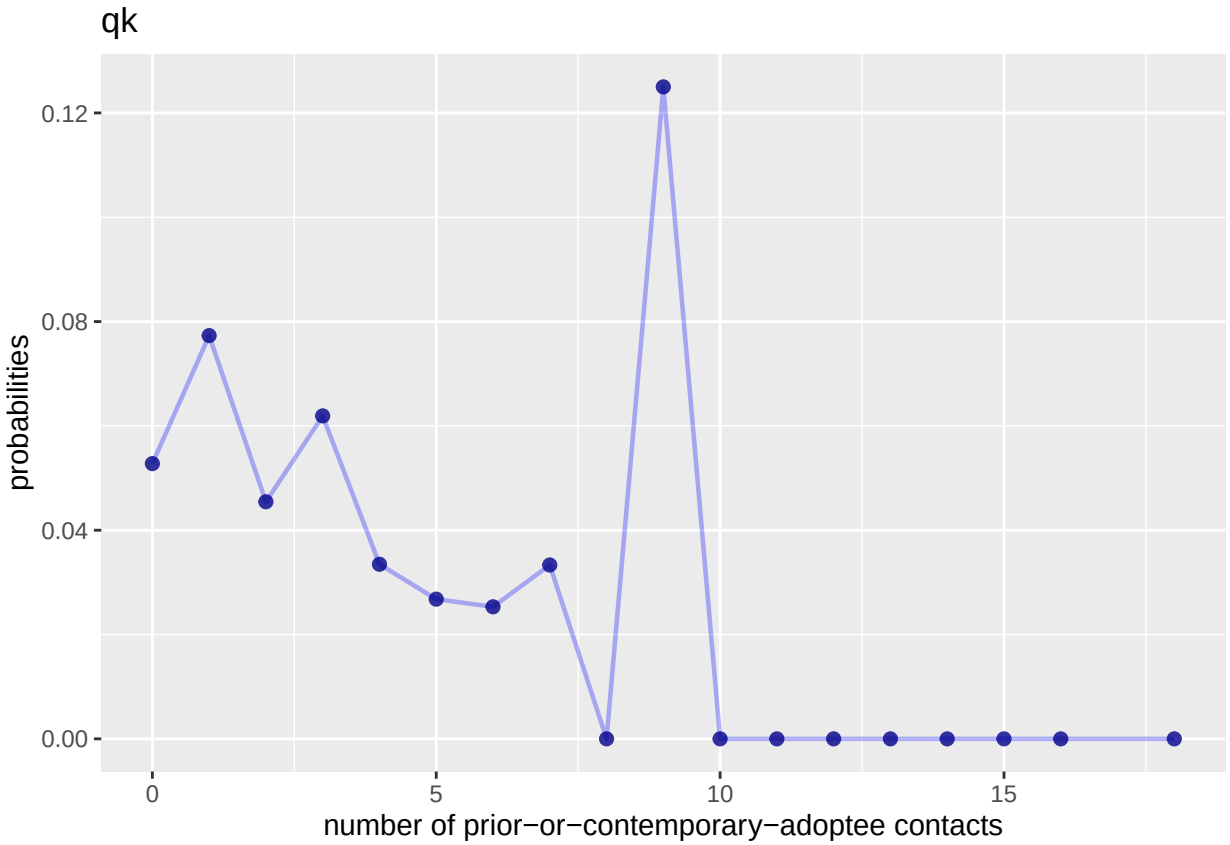
c. Create a vector of estimated q_k probabilities, using the data frame from (2). Plot the probabilities against the number of prior-or-contemporary-adoptee contacts k.

```r
k_2<-sort(unique(RECORDS$num_before_this))

FUN6<-function(i){
  int<-which(RECORDS$num_before_this==i)
  N<-sum(RECORDS$pre_this_month[int]==1)
```

```
  M<-length(int)
  return(N/M)
}
qk<-sapply(k_2,FUN6)
Qdata<-data.frame(k_2,qk)
```

```
Qdata %>% ggplot()+
  geom_line(aes(x =k_2, y =qk),color="blue",alpha=0.3,size=0.8) +
  geom_point(aes(x =k_2, y =qk),color="darkblue",alpha=0.8,size=2) +
labs(x = "number of prior-or-contemporary-adoptee contacts",y = "probabilities",title = "qk")
```



4. Because it only conditions on information from the previous month, p_k is a little easier to interpret than q_k. It is the probability per month that a doctor adopts tetracycline, if they have exactly k contacts who had already adopted tetracycline.

a. Suppose $p_k = a + bk$. This would mean that each friend who adopts the new drug increases the probability of adoption by an equal amount. Estimate this model by least squares, using the values you constructed in (3b). Report the parameter estimates.

```
model_pk<-lm(Pdata$pk~Pdata$k_1,data=Pdata)
summary(model_pk)
```

```
##
```

```
## Call:
## lm(formula = Pdata$pk ~ Pdata$k_1, data = Pdata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030334 -0.014584 -0.002344  0.005534  0.048694
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0569324  0.0090507   6.290 1.45e-05 ***
## Pdata$k_1   -0.0037997  0.0009184  -4.137 0.000877 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02015 on 15 degrees of freedom
## Multiple R-squared:  0.533,  Adjusted R-squared:  0.5018
## F-statistic: 17.12 on 1 and 15 DF,  p-value: 0.0008773
```

In our model, estimated $\bar{a} = 0.0569324$, estimated $\bar{b} = -0.0037997$.

The model for $p_k = a + bk$:

$$p_k = 0.0569324 - 0.0037997k$$

b. Suppose $p_k = e^{a+bk}/(1 + e^{a+bk})$. Explain, in words, what this model would imply about the impact of adding one more adoptee friend on a given doctor's probability of adoption. (You can suppose that $b > 0$, if that makes it easier.) Estimate the model by least squares, using the values you constructed in (3b).

```
model_pk_2<-nls(pk ~ exp(a+b*k_1)/(1+exp(a+b*k_1)),Pdata,start = list(a = 0.01,b = 0.01))
summary(model_pk_2)
```

```
##
## Formula: pk ~ exp(a + b * k_1)/(1 + exp(a + b * k_1))
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a -2.56508    0.20610 -12.446 2.62e-09 ***
## b -0.17051    0.05371  -3.174  0.00628 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01957 on 15 degrees of freedom
##
## Number of iterations to convergence: 7
## Achieved convergence tolerance: 1.743e-07
```

To make it easier, suppose that $b > 0$, when adding one more adoptee friend, the given doctor's probability of adoption will rise. It reflects that a friend's treatment will have a positive impact on the doctor's medication possibilities.

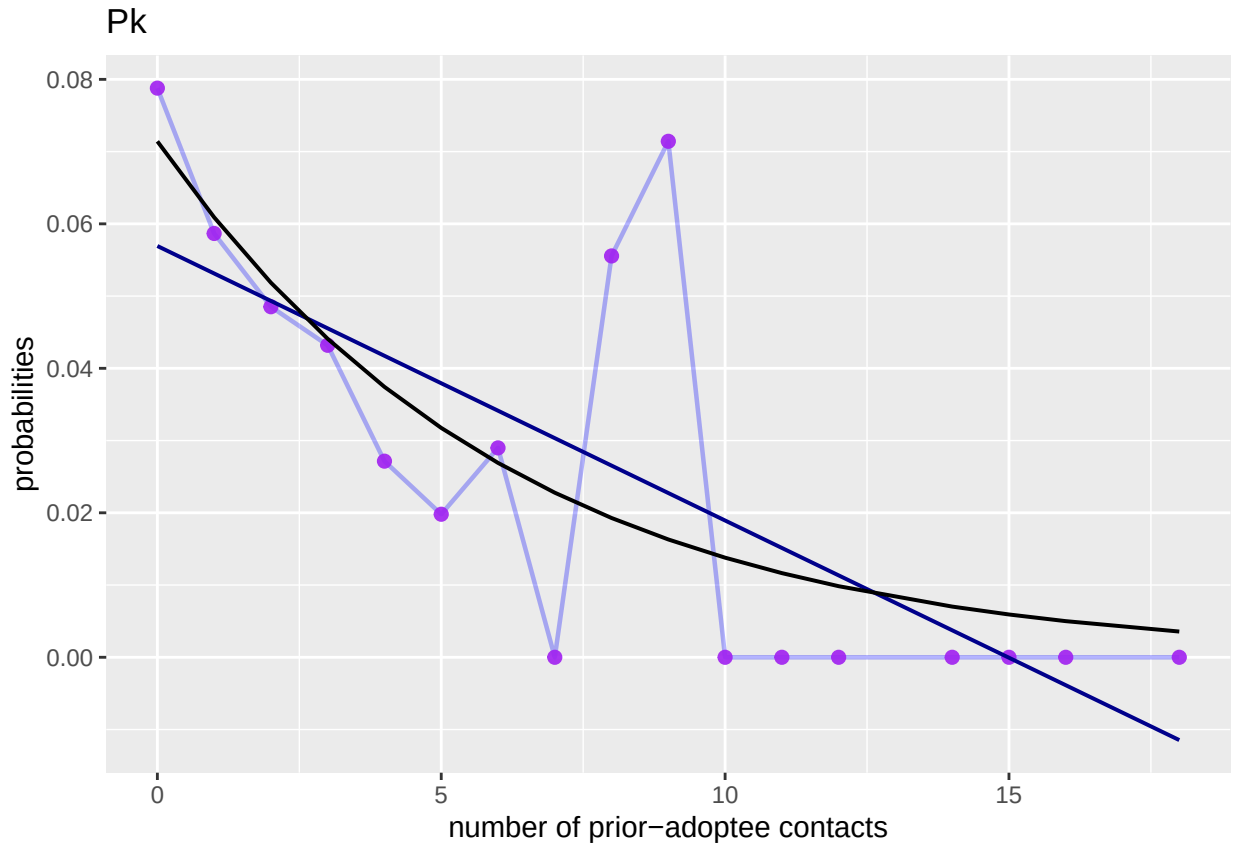
In our model, estimated $\bar{a} = -2.56508$, estimated $\bar{b} = -0.17051$.

The model for $p_k = e^{a+bk}/(1 + e^{a+bk})$:

$$p_k = e^{-2.56508-0.17051k}/(1+e^{-2.56508-0.17051k})$$

c. Plot the values from (3b) along with the estimated curves from (4a) and (4b). (You should have one plot, with k on the horizontal axis, and probabilities on the vertical axis .) Which model do you prefer, and why?

```
Pdata %>% ggplot()+
  geom_line(aes(x =k_1, y =pk),color="blue",alpha=0.3,size=0.8) +
  geom_point(aes(x =k_1, y =pk),color="purple",alpha=0.9,size=2) +
  labs(x = "number of prior-adoptee contacts",y = "probabilities",title = "Pk")+
  geom_line(aes(x = k_1, y =0.0569324 -0.00379978*k_1 ), col = "darkblue", size = 0.7)+
geom_line(aes(x = k_1, y =  exp(-2.56508-0.17051*k_1)/(1+exp(-2.56508-0.17051*k_1))), col = "black", si
```



I prefer $p_k = e^{-2.56508-0.17051k}/(1+e^{-2.56508-0.17051k})$.

For quibblers, pedants, and idle hands itching for work to do: The p_k values from problem 3 aren't all equally precise, because they come from different numbers of observations. Also, if each doctor with k adoptee contacts is independently deciding whether or not to adopt with probability p_k, then the variance in the number of adoptees will depend on p_k. Say that the actual proportion who decide to adopt is $\hat{p}_k$. A little probability (exercise!) shows that in this situation, $\mathbb{E}[\hat{p}_k] = p_k$, but that $\mathrm{Var}[\hat{p}_k] = p_k(1-p_k)/n_k$, where n_k is the number of doctors in that situation. (We estimate probabilities more precisely when they're really extreme [close to 0 or 1], and/or we have lots of observations.) We can estimate that variance as $\hat{V}_k = \hat{p}_k(1-\hat{p}_k)/n_k$. Find the $\hat{V}_k$, and then re-do the estimation in (4a) and (4b) where the squared error for p_k is divided by $\hat{V}_k$. How much do the parameter estimates change? How much do the plotted curves in (4c) change?