《数学实践》作业四

许乐乐

```
library(magrittr)
library(tidyverse)
library(ggplot2)
```

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.

```
ckm_nodes<-read.csv("data/ckm_nodes.csv",header=TRUE)
ckm_network<-read.table("data/ckm_network.dat")
```

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.

```
index<-is.na(ckm_nodes$adoption_date)
ckm_nodes<-ckm_nodes[!index,]
ckm_network<-ckm_network[!index,!index]
```

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.

```
n<-dim(ckm_nodes)[1]
doctor_number<-rep(c(1:n),times=17)
month<-rep(c(1:17),each=n)
whether_began<-c(rep(ckm_nodes$adoption_date,times=17)==month)
whether_began_before<-c(rep(ckm_nodes$adoption_date,times=17)<month)
number_contact<-apply(
    matrix(whether_began,nrow=n)[,rep(1:17,each=n)]&ckm_network[,rep(1:n,times=17)],
    2,sum
)
number_contact_before<-apply(
    matrix(whether_began_before,nrow=n)[,rep(1:17,each=n)]&ckm_network[,rep(1:n,times=17)],
    2,sum
```

```
)
number_contact_orbefore<-number_contact+number_contact_before
record<-data.frame(doctor_number,month,whether_began,whether_began_before,number_contact_before,n
dim(record)
```

```
## [1] 2125    6
```

我们有 6 个变量，和 17 个月、125 个医生在清洗后的数据中，因此，record 有 17*125=2125 行和 6 列。

3. Let

$$p_k = \Pr(\text{A doctor starts prescribing tetracycline this month} \mid \text{Number of doctor's contacts prescribing before this month} = k)$$

and

$$q_k = \Pr(\text{A doctor starts prescribing tetracycline this month} \mid \text{Number of doctor's contacts prescribing this month} = k)$$

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(apply(ckm_network,2,sum))
```

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

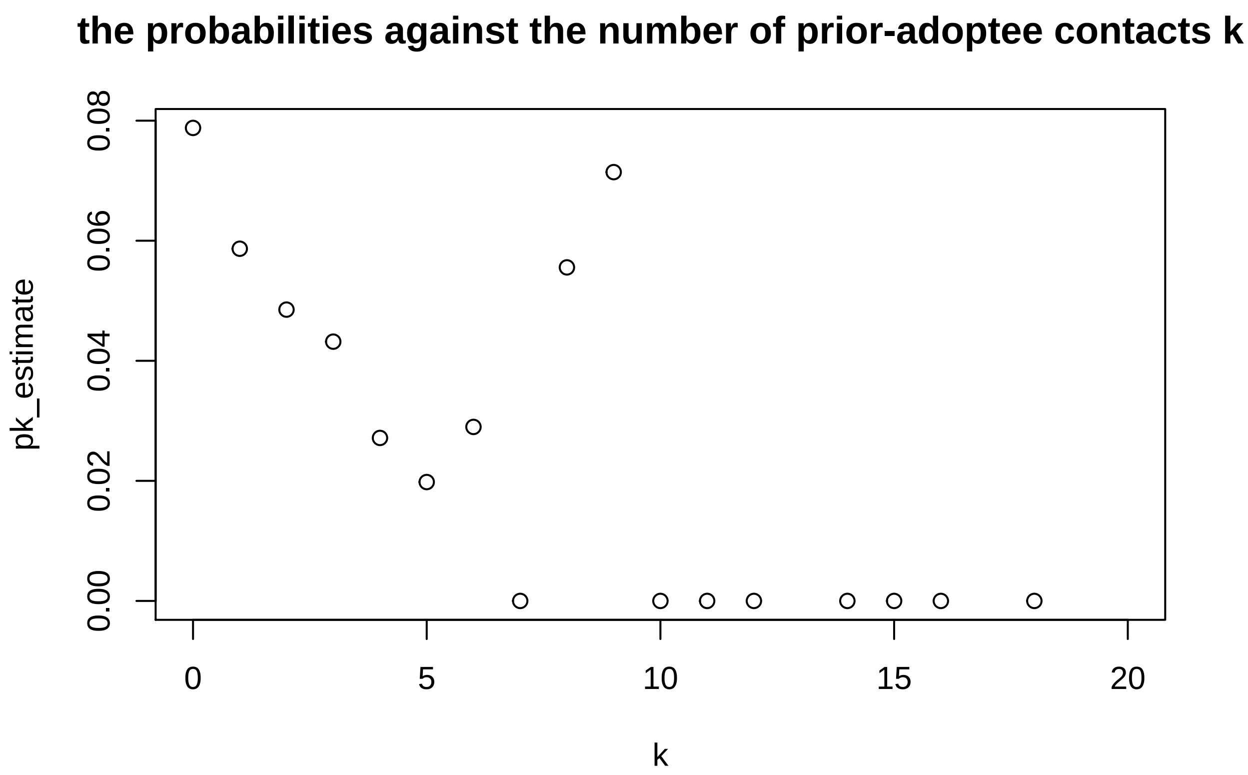
因为所有医生中最大联系数目为 20。

```
 b. Create a vector of estimated $p_k$ probabilities, using the data frame from (2).  Plot the pro
```

```
pk_estimate<-c()
for(i in 0:20){
    pk_estimate[i+1]<-sum(record$number_contact_before==i&record$whether_began==TRUE)/sum(record$i
}
pk_estimate
```

```
##  [1] 0.07878788 0.05866667 0.04852321 0.04318937 0.02714932 0.01980198
##  [7] 0.02898551 0.00000000 0.05555556 0.07142857 0.00000000 0.00000000
## [13] 0.00000000        NaN 0.00000000 0.00000000 0.00000000        NaN
## [19] 0.00000000        NaN        NaN
```

```
plot(pk_estimate~c(0:20),main="the probabilities against the number of prior-adoptee contacts k",
```

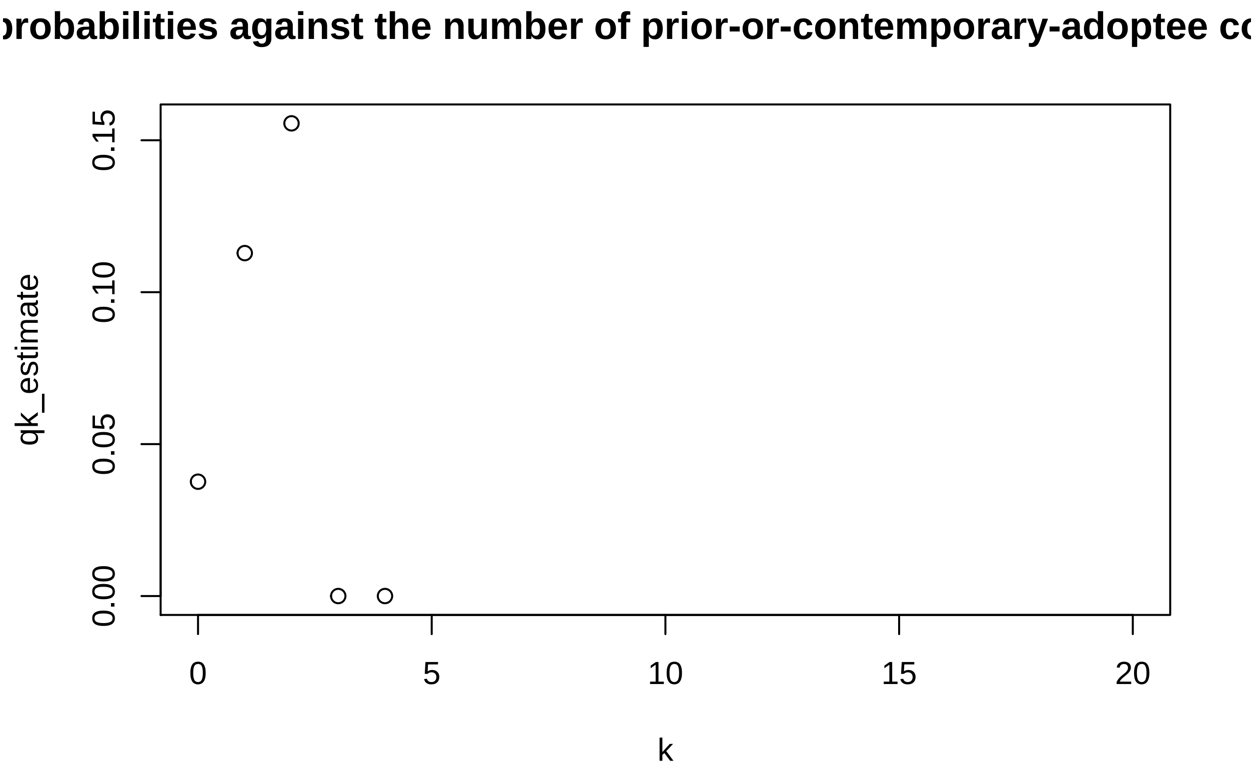


c. Create a vector of estimated q_k probabilities, using the data frame from (2). Plot the pro

```
qk_estimate<-c()
for(i in 0:20){
    qk_estimate[i+1]<-sum(number_contact==i&record$whether_began==TRUE)/sum(number_contact==i)
}
qk_estimate
```

```
##  [1] 0.03762828 0.11285266 0.15555556 0.00000000 0.00000000        NaN
##  [7]        NaN        NaN        NaN        NaN        NaN        NaN
## [13]        NaN        NaN        NaN        NaN        NaN        NaN
## [19]        NaN        NaN        NaN
```

```
plot(qk_estimate~c(0:20),main="the probabilities against the number of prior-or-contemporary-adop
```



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.

```
k=c(0:20)
p1<-pk_estimate[!is.na(pk_estimate)]
k1<-k[!is.na(pk_estimate)]
fit1<-lm(p1~k1)
summary(fit1)
```

```
##
## Call:
## lm(formula = p1 ~ k1)
##
## 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 ***
## k1          -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
```

```
b. Suppose $p_k = e^{a+bk}/(1+e^{a+bk})$.  Explain, in words, what this model would imply about t
```

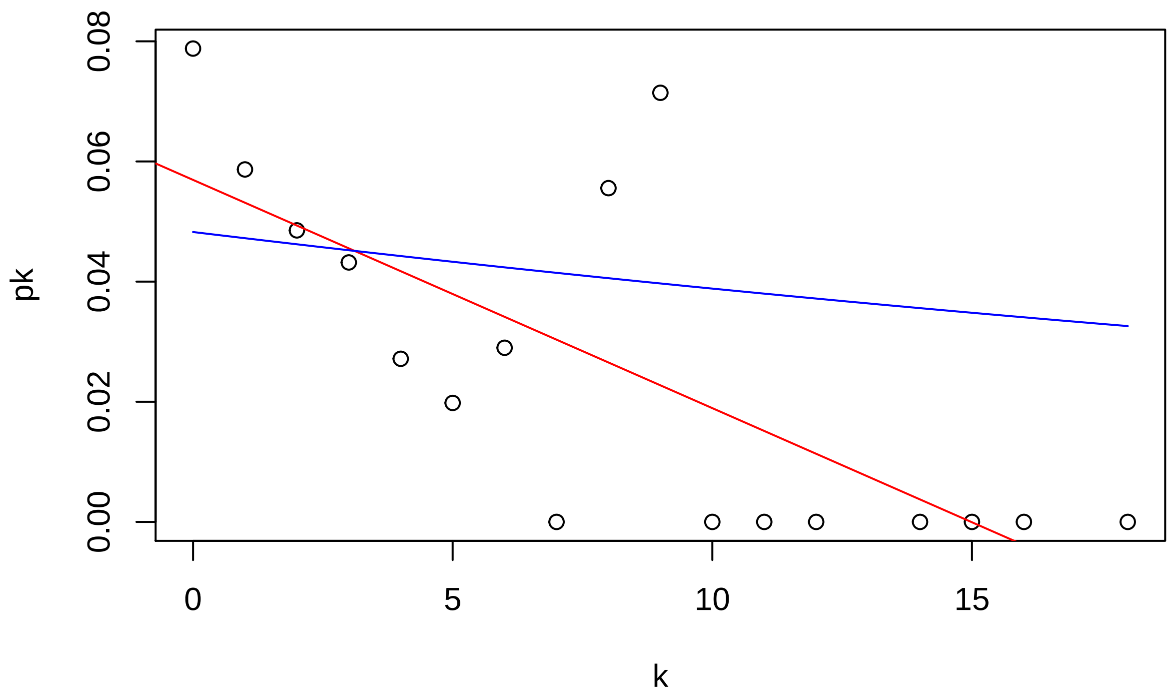
```r
index=(p1!=0)
p2<-p1[index]
k2<-k1[index]
y=log(p2/(1-p2))
fit2<-lm(y~k2)
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ k2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80666 -0.39353  0.05123  0.33021  0.62118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.98180    0.30592  -9.747 2.53e-05 ***
## k2          -0.02270    0.05974  -0.380    0.715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5193 on 7 degrees of freedom
## Multiple R-squared:  0.02022,    Adjusted R-squared:  -0.1198
## F-statistic: 0.1444 on 1 and 7 DF,  p-value: 0.7152
```

```
c. Plot the values from (3b) along with the estimated curves from (4a) and (4b).  (You should hav
```

```r
plot(k1,p1,xlab="k",ylab="pk",xlim=c(0,18))
abline(fit1,col="red")
```

```
a=fit2$coefficients[1]
b=fit2$coefficients[2]
curve(exp(a+b*x)/(1+exp(a+b*x)),from=0,to=18,add=TRUE,col="blue")
```



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?

```
nk<-rep(NA,19)
for(k in 0:18){
    nk[k+1]<-length(which(record$number_contact_before==k))
}
vk=p1*(1-p1)/nk
vk
```

```
##  [1] 1.466270e-04 1.472664e-04 9.740233e-05 1.372892e-04 1.195124e-04
##  [6] 1.921768e-04 4.079036e-04 0.000000e+00 2.914952e-03 4.737609e-03
## [11] 0.000000e+00 0.000000e+00 0.000000e+00          NaN 0.000000e+00
## [16] 0.000000e+00 0.000000e+00          Inf 2.761244e-02
```