Homework 4

Shen Dingtao 3170104764

1.

```
ckm_nodes<-read_csv("data/ckm_nodes.csv")
vec<-which(ckm_nodes$adoption_date!="NA")
ckm_nodes<-ckm_nodes[vec,]
```

```
ckm_network<-read.table("data/ckm_network.dat")
ckm_network<-ckm_network[vec,vec]
```

2.

```
doctor<-(rep(1:125,each=17))
month<-(rep(1:17,125))
newckm<-data.frame(doctor,month)
newckm<-newckm %>% mutate(begin_the_month=(ckm_nodes$adoption_date[doctor]==month))
newckm<-newckm %>% mutate(before_the_month=(ckm_nodes$adoption_date[doctor]<month))
newckm<-newckm %>% group_by(doctor,month) %>%
                    mutate(con_str_before=sum((ckm_network[,doctor]==1)&
                        (ckm_nodes$adoption_date<month))) %>%
                   ungroup()
newckm<-newckm %>% group_by(doctor,month) %>%
                    mutate(con_before=sum((ckm_network[,doctor]==1)&
                        (ckm_nodes$adoption_date<=month))) %>%
                   ungroup()
```

```
newckm
```

```
## # A tibble: 2,125 x 6
##    doctor month begin_the_month before_the_month con_str_before con_before
##     <int> <int> <lgl>           <lgl>                      <int>      <int>
##  1      1     1 TRUE            FALSE                          0          1
##  2      1     2 FALSE           TRUE                           1          1
##  3      1     3 FALSE           TRUE                           1          2
##  4      1     4 FALSE           TRUE                           2          3
##  5      1     5 FALSE           TRUE                           3          3
##  6      1     6 FALSE           TRUE                           3          3
##  7      1     7 FALSE           TRUE                           3          3
##  8      1     8 FALSE           TRUE                           3          3
##  9      1     9 FALSE           TRUE                           3          3
## 10      1    10 FALSE           TRUE                           3          3
## # ... with 2,115 more rows
```

```
dim(newckm)
```

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

The data frame has 6 columns as required, and the records has `125*17=2125` rows for 17 month and 125 doctors.

3.

(a) With the following code

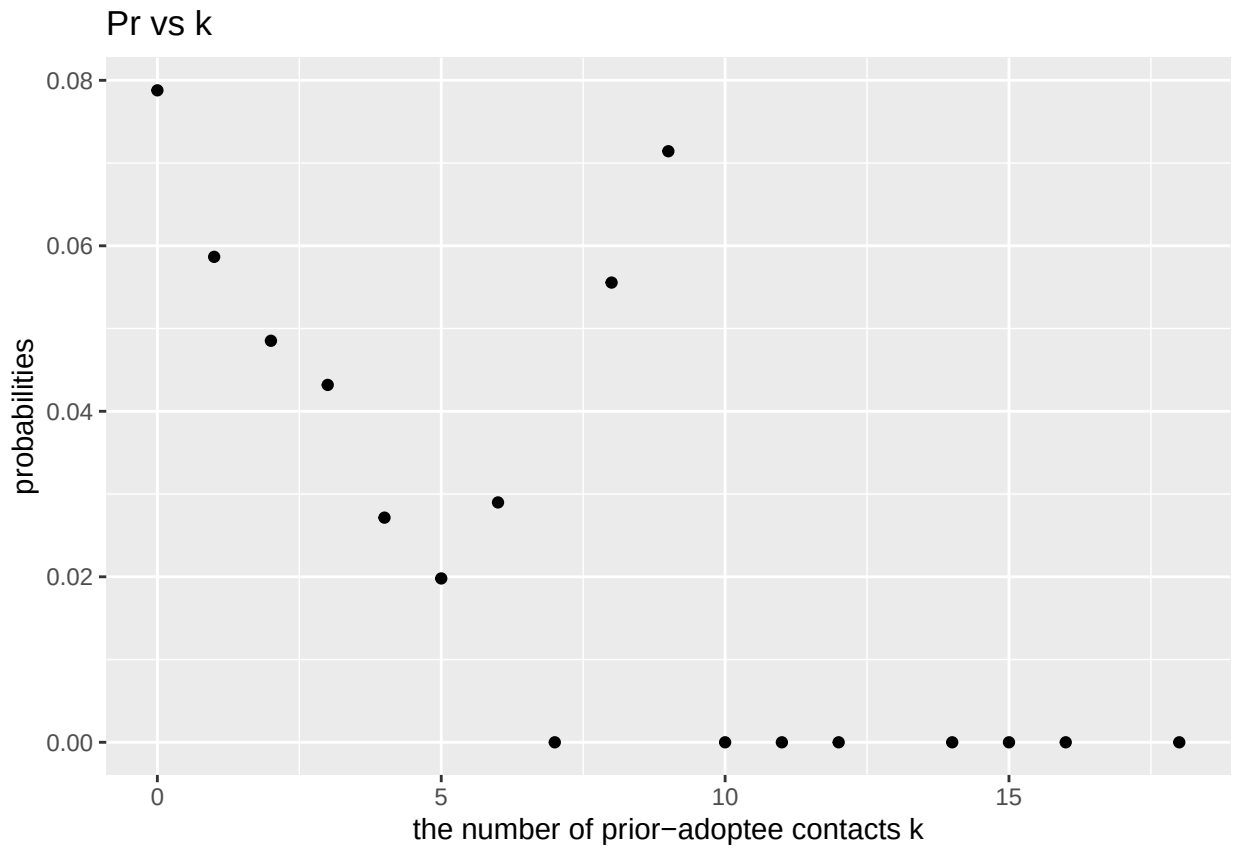
```
max(colSums(ckm_network))
```

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

We can know that any doctor contacts at most 20 doctors, so `k` can be `0,1,2,...,20`, which means no more than 21 values for `k`.

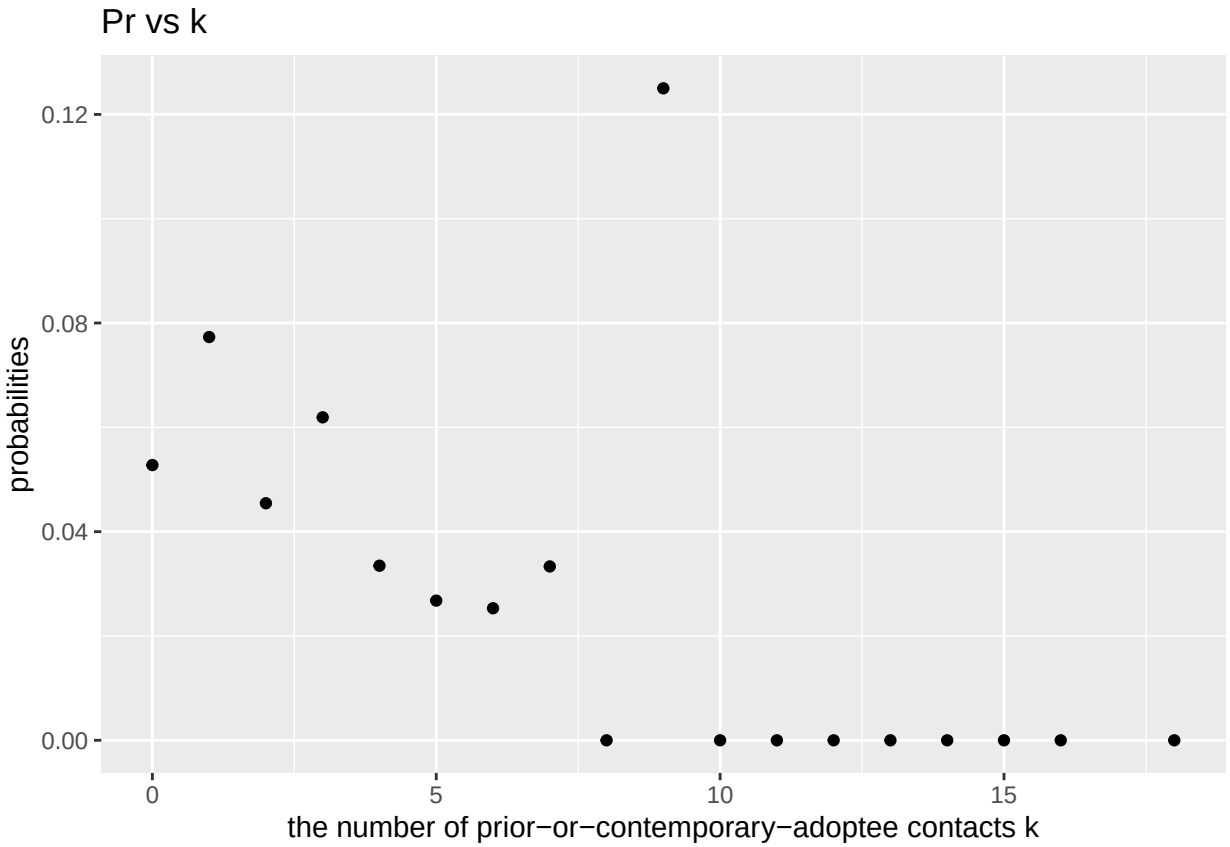
(b)

```
p_k<-c()
p_a_c<-c()
for(k in 0:20){
    if(nrow(newckm %>% dplyr::filter(con_str_before==k))!=0){
        tmp<-sum((newckm %>%
        dplyr::filter(con_str_before==k))$begin_the_month)/
        nrow(newckm %>% dplyr::filter(con_str_before==k))
        p_k<-c(p_k,tmp)
        p_a_c<-c(p_a_c,k)
    }
}
ggplot()+
    geom_point(aes(x=p_a_c,y=p_k))+
    labs(x="the number of prior-adoptee contacts k",
         y="probabilities",title="Pr vs k")
```



(c)

```
q_k<-c()
q_pca_c<-c()
for(k in 0:20){
    if(nrow(newckm %>% dplyr::filter(con_before==k))!=0){
        tmp<-sum((newckm %>%
        dplyr::filter(con_before==k))$begin_the_month)/
        nrow(newckm %>% dplyr::filter(con_before==k))
        q_k<-c(q_k,tmp)
        q_pca_c<-c(q_pca_c,k)
    }
}
ggplot()+
    geom_point(aes(x=q_pca_c,y=q_k))+
    labs(x="the number of prior-or-contemporary-adoptee contacts k",
         y="probabilities",title="Pr vs k")
```



4.

(a) Using the vector `p_k` and `p_a_c` in 3(b)

```
p_k_coe_lm<-coefficients(lm(p_k~p_a_c))
p_k_coe_lm
```

```
##  (Intercept)        p_a_c
##  0.056932428 -0.003799739
```

(b) We select starting value: a=-2.5, b=-0.1

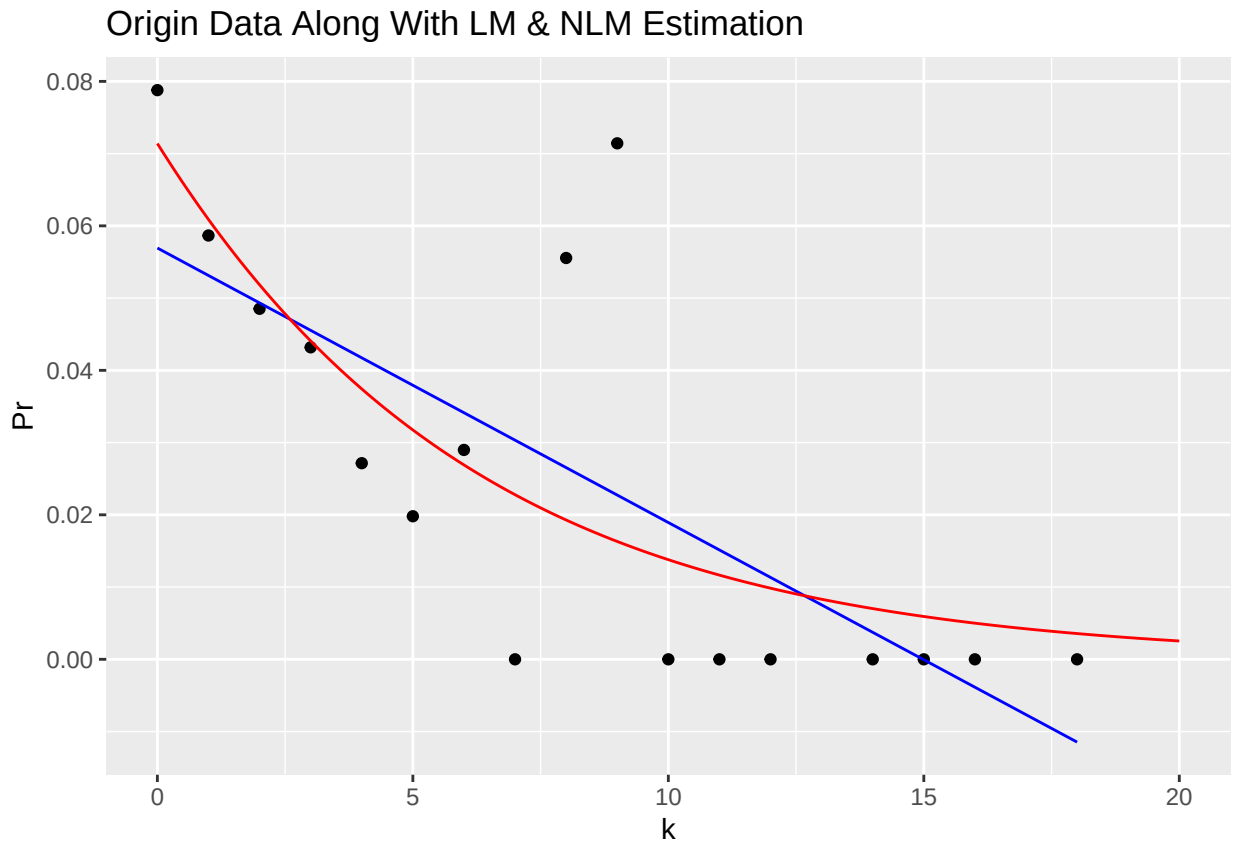
```
m_pk<-function(coe){
    p_k_e<-exp(coe[1]+coe[2]*p_a_c)/(1+exp(coe[1]+coe[2]*p_a_c))
    return(sum((p_k_e-p_k)^2)/length(p_a_c))
}

p_k_fit<-nlm(m_pk,c(-2.5,-0.1))
p_k_coe_nlm<-p_k_fit$estimate
sprintf("Estimation: a=%8.7f,b=%8.7f",p_k_coe_nlm[1],p_k_coe_nlm[2])
```

```
## [1] "Estimation: a=-2.5653733,b=-0.1704526"
```

(c)

```r
x_t<-seq(0,20,0.1)
ggplot()+
    geom_point(aes(x=p_a_c,y=p_k),color="black")+
    geom_line(aes(x=p_a_c,y=p_k_coe_lm[1]+p_k_coe_lm[2]*p_a_c),color="blue")+
    geom_line(aes(x=x_t,y=exp(p_k_coe_nlm[1]+
                                  p_k_coe_nlm[2]*x_t)/
                (1+exp(p_k_coe_nlm[1]+p_k_coe_nlm[2]*x_t)))
              ,color="red")+
    labs(x="k",y="Pr",title="Origin Data Along With LM & NLM Estimation")
```



I prefer the non-linear model $p_k = e^{a+bk}/(1+e^{a+bk})$ in 4(b), because its curve fits the origin data better and the changing trends in the plot.