homework4

3170102187　　孙晨旭

```
ckm_nodes <- read_csv("~/github/Rcourse_1/data/ckm_nodes.csv")
ckm_network <-read.table("~/github/Rcourse_1/data/ckm_network.dat")
```

1.

```
omit <- which(is.na(ckm_nodes$adoption_date))
ckm_nodes <- ckm_nodes[-omit,]
ckm_network <- ckm_network[-omit,-omit]
```

2.

```
#------------------------
temp <- function(x){return(x==1:17)}
temp2<-function(Fun,V,data = ckm_nodes$adoption_date){
  col1<-apply(array(data),1,Fun)
  col1<-data.frame(col1)
  col1<-col1 %>% gather()
  col1 <- cbind(1:17,col1)
  colnames(col1)<-c('date','doctor',V)
  return(col1)
}
#------------------------col1
ckm<-temp2(temp,'begin_prescribing')
#------------------------col2
temp <- function(x){return(x<=1:17)}
ckm<-full_join(ckm,temp2(temp,'adopted_before'),by = c('date','doctor'))# not strictly
#------------------------col3
```

```
temp3 <- function(x){
  if(is.na(x)){
    return(vector(length = 17))
  }else{
    return(x<1:17)
  }
}
temp4<-function(x){
  re<-apply(array(ckm_nodes$adoption_date[as.logical(ckm_network[x,])]),1,temp3)
  if(length(re)!=0){
    return(rowSums(re))
  }else{
    return(as.numeric(vector(length=17)))
  }
}
ckm<-full_join(ckm,temp2(temp4,'ncontact_strict_before',1:125),by = c('date','doctor'))
#------------------------col4
temp3 <- function(x){
  if(is.na(x)){
    return(vector(length = 17))
  }else{
    return(x<=1:17)
  }
}
ckm<-full_join(ckm,temp2(temp4,'ncontact_in_or_earlier',1:125),by = c('date','doctor'))
#------------------------
head(ckm)
```

```
##   date doctor begin_prescribing adopted_before ncontact_strict_before
## 1    1     X1              TRUE           TRUE                       0
## 2    2     X1             FALSE           TRUE                       1
## 3    3     X1             FALSE           TRUE                       1
## 4    4     X1             FALSE           TRUE                       2
## 5    5     X1             FALSE           TRUE                       3
```

```
## 6    6     X1              FALSE           TRUE                       3
##   ncontact_in_or_earlier
## 1                      1
## 2                      1
## 3                      2
## 4                      3
## 5                      3
## 6                      3
```

如向量 data.frame ckm 所示，四列数据和两列标签一共组成六列数据，而 125 为医生和 17 个月使得数据共有 125*17 = 2125 行。

3.a.

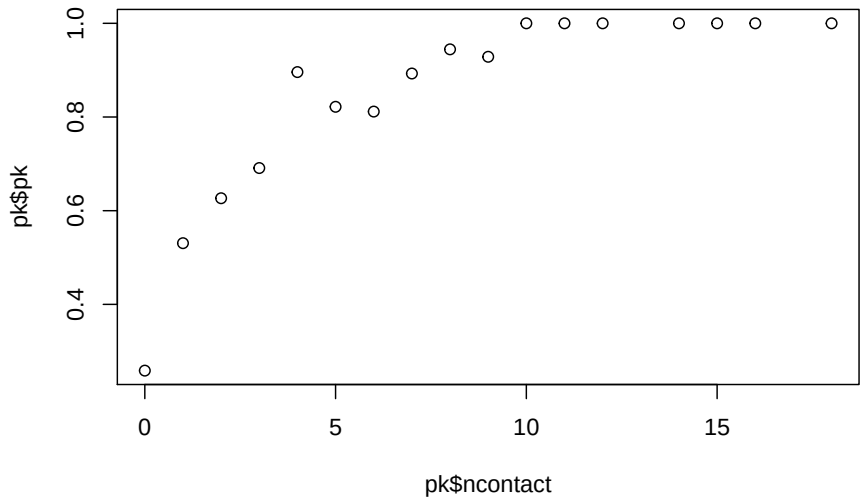
```
max(rowSums(ckm_network))
```

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

因为每个医生的社交圈内至多有 20 人，所以 k=0~20，至多有 21 种取法。

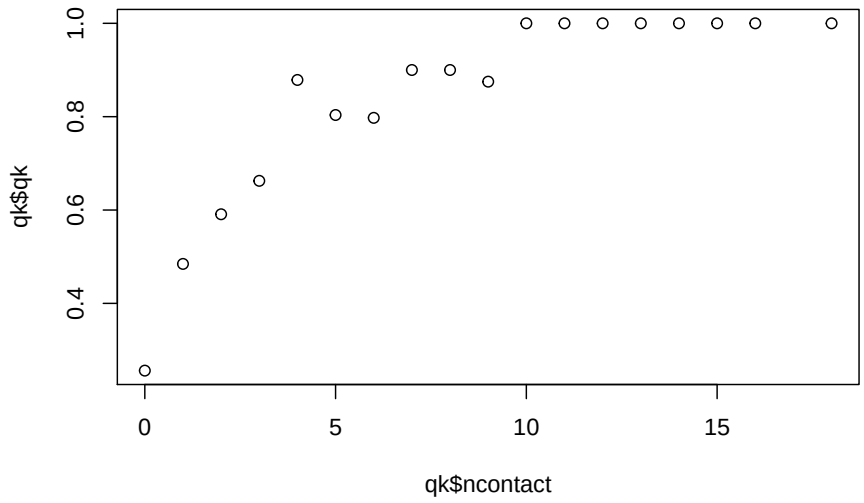
b.

```
T<- ckm %>% group_by(ncontact_strict_before) %>% mutate(pk = sum(adopted_before))
T<-table(T$pk,T$ncontact_strict_before)
pk <- data.frame(vector(length = ncol(T)))
pk[,1]<- as.numeric(colnames(T))
pk[1,2]<-1
colnames(pk) <- c('ncontact','pk')
for(i in 1:ncol(T)){
  pk[i,2] <- as.numeric(names(T[,i][(T[,i] != 0)]))/T[,i][(T[,i] != 0)]
}
plot(pk$pk~pk$ncontact)
```



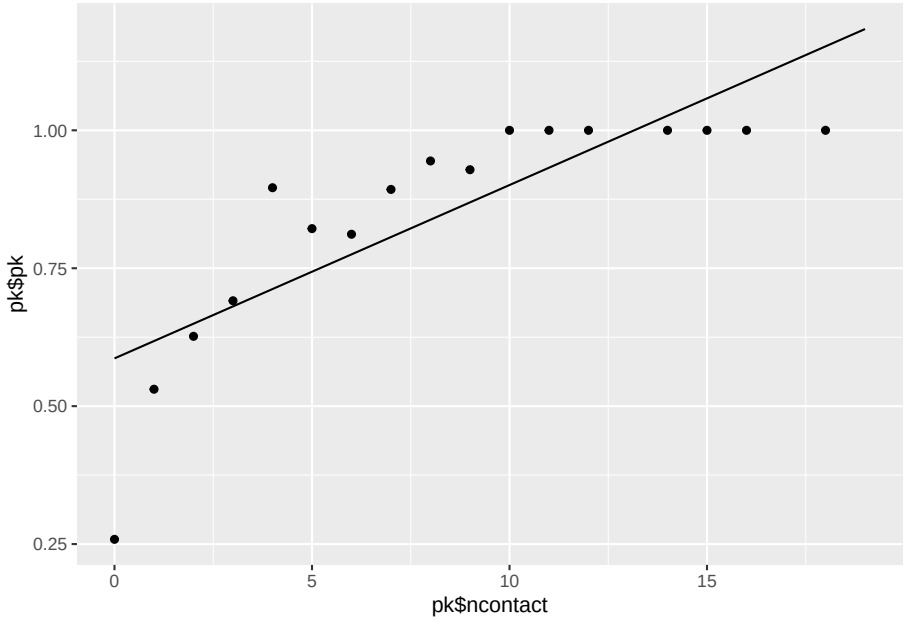
c.

```
T<- ckm %>% group_by(ncontact_in_or_earlier) %>% mutate(qk = sum(adopted_before))
T<-table(T$qk,T$ncontact_in_or_earlier)
qk <- data.frame(vector(length = ncol(T)))
qk[,1]<- as.numeric(colnames(T))
qk[1,2]<-1
colnames(qk) <- c('ncontact','qk')
for(i in 1:ncol(T)){
  qk[i,2] <- as.numeric(names(T[,i][(T[,i] != 0)]))/T[,i][(T[,i] != 0)]
}
plot(qk$qk~qk$ncontact)
```



4.

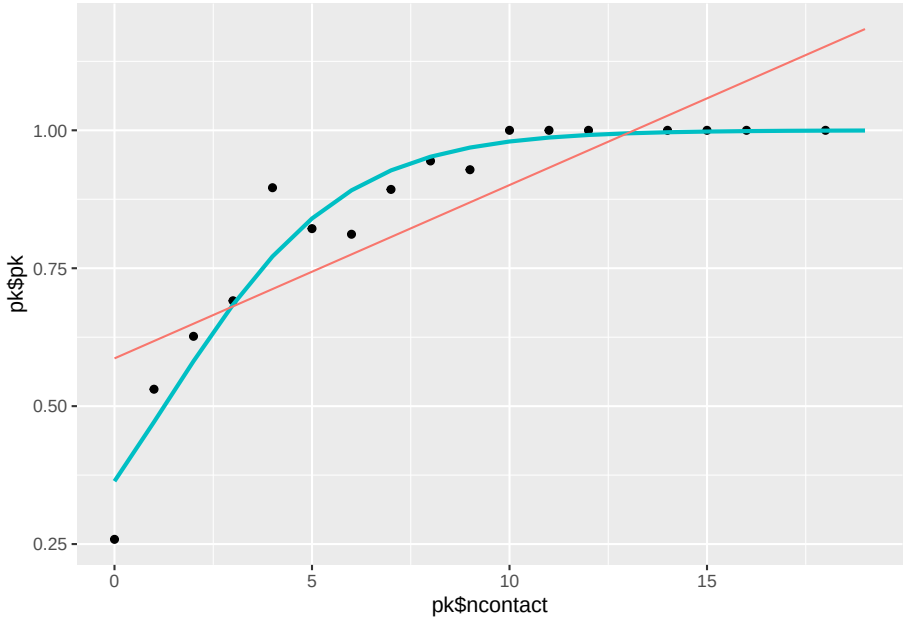
```
k = 0:19
L <- lm(pk$pk~pk$ncontact)
y1 <- coef(L)[1] + coef(L)[2]*k
ggplot()+
  geom_point(aes(x = pk$ncontact,y = pk$pk))+
  geom_line(aes(x = k,y = y1))
```




```
Lo <- glm(pk$pk~pk$ncontact,family = binomial())
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
y2 <- exp(coef(Lo)[1] + coef(Lo)[2]*k)/(1+exp(coef(Lo)[1] + coef(Lo)[2]*k))
ggplot()+
  geom_point(aes(x = pk$ncontact,y = pk$pk))+
  geom_line(aes(x = k,y = y2,col = "green"),size = 1)+
  geom_line(aes(x = k,y = y1,col = "blue"))+
  theme(legend.position = "none")
```



从图像上看，第二种模型明显更合适。而且，当使用线性模型预测 k>18 的情况时，得到结果明显超过 1，但该结果表示的实际意义是某个概率，大于 1 就是不合理的。