Homework 4

Yuhao Mao

2020/7/12

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

1. Clean the data.

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

2. Reformat data.

```
data <- data.frame(month=rep(c(1:16,Inf),each=125),doctor=rep(as.numeric(rownames(ckm_nodes)),times=17)
data$this_month <- data$month==data$adoption_date
data$before_this_month <- data$month>data$adoption_date

cal.before <- function(vec){
  doc <- vec[2]
  month <- vec[1]
  contact <- which(ckm_network[doc,]==1)
  contact_adopt <- ckm_nodes$adoption_date[contact]
  return(sum(contact_adopt<month))
}
cal.before_or_eq <- function(vec){
  doc <- vec[2]
  month <- vec[1]
  contact <- which(ckm_network[doc,]==1)
  contact_adopt <- ckm_nodes$adoption_date[contact]
  return(sum(contact_adopt<=month))
}

data$num_contact_before <- apply(data, 1, cal.before)
data$num_contact_before_or_eq <- apply(data, 1, cal.before_or_eq)
data$adoption_date <- NULL
```

Because there are 125 doctors and 17 month values, total number of rows is $125 \times 17 = 2125$.

3. Probabilities.

+a.

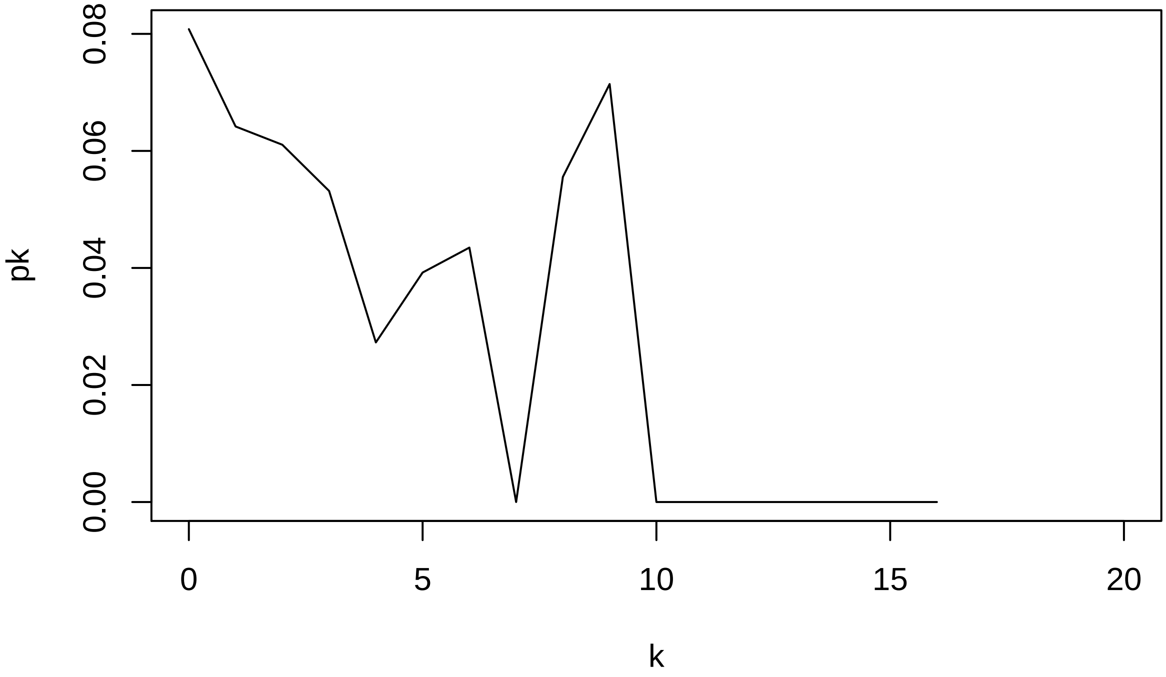
```
max(data[,c(5,6)])
```

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

Since the maximum value of k is 20, we can only estimate at most 21 values of k.

+b. Calculate p_k.

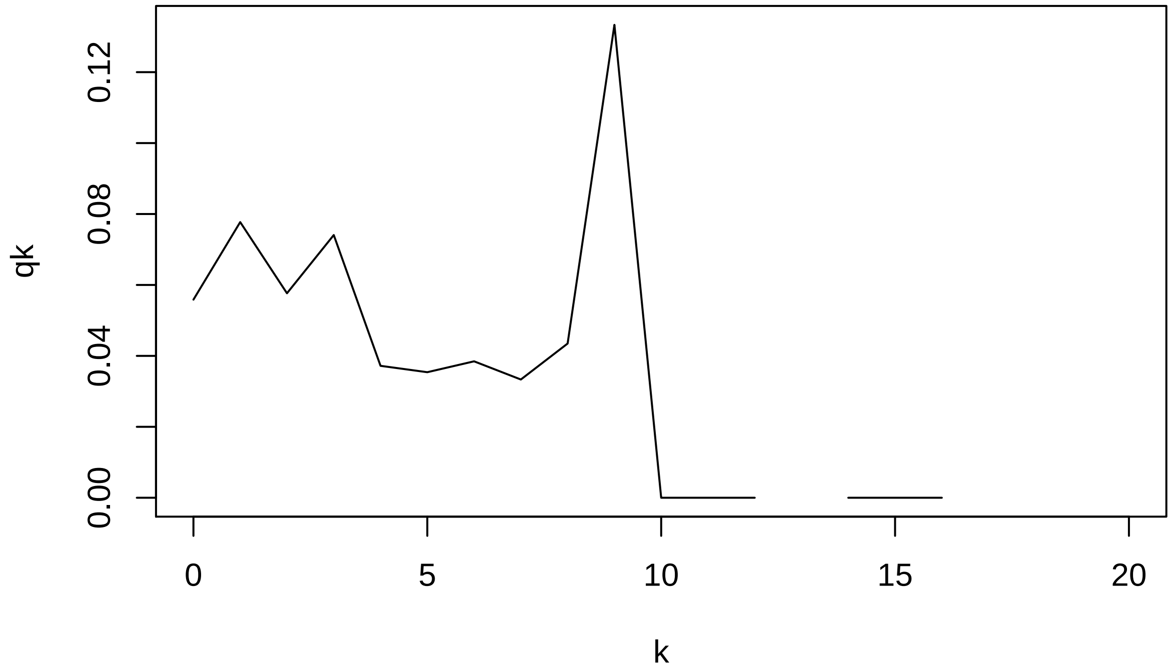
```
pk <- c()
for(k in 0:20){
  joint <- sum(data$this_month&data$num_contact_before==k)
  cond <- sum(data$num_contact_before==k)
  pk <- c(pk, joint/cond)
}
plot(0:20, pk, type='l', xlab='k')
```



+c. Calculate q_k.

```
qk <- c()
for(k in 0:20){
  joint <- sum(data$this_month&data$num_contact_before_or_eq==k)
  cond <- sum(data$num_contact_before_or_eq==k)
  qk <- c(qk, joint/cond)
}

plot(0:20, qk, type='l', xlab='k')
```



4. Interpretation of p_k.

+a. Linear model.

```
pk <- data.frame(pk=pk,k=0:20)
pk <- pk %>% filter(!is.na(pk))
linear.result <- lm(pk~k, data=pk)
linear.result$coefficients
```

```
##  (Intercept)            k
##  0.065802017 -0.004469464
```

+b. Logistic model.

```
logit.result <- glm(pk~k,data=pk,family = "binomial")
```

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

```
summary(logit.result)
```

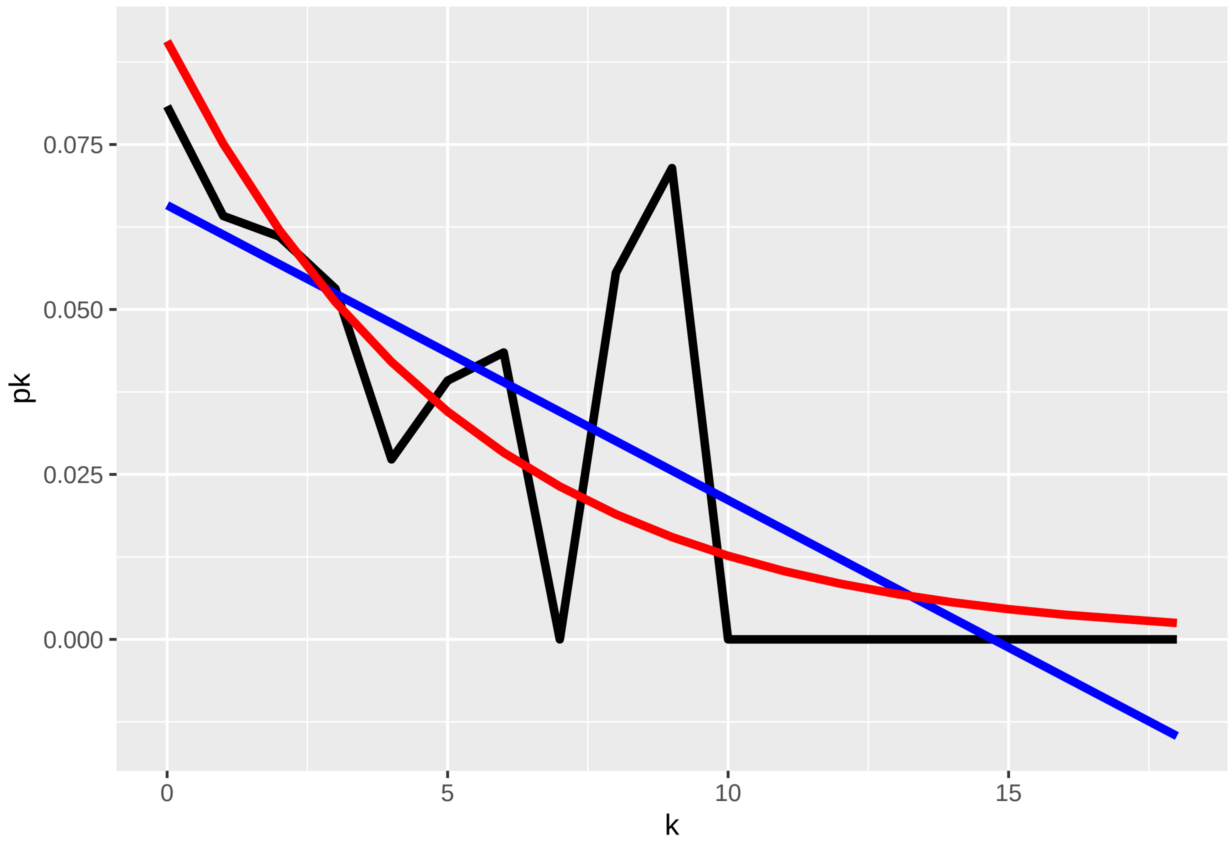
```
##
## Call:
## glm(formula = pk ~ k, family = "binomial", data = pk)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.21655  -0.11467  -0.07448   0.00584   0.33116
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3057     2.1056  -1.095    0.274
## k            -0.2051     0.3673  -0.558    0.577
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.74624  on 17  degrees of freedom
## Residual deviance: 0.33125  on 16  degrees of freedom
## AIC: 5.0192
##
## Number of Fisher Scoring iterations: 7
```

Therefore, if k increases by 1, then $a + bk$ decreases because its coefficient is negative and e^{a+bk} is smaller. By the monotonicity of logit function, we find p_k is smaller.

+c. Plot fitted values.

```
pk %>% ggplot() + geom_line(aes(k, pk), col="black", size=1.5) +
  geom_line(aes(k, linear.result$fitted.values), col="blue", size=1.5) +
  geom_line(aes(k, logit.result$fitted.values), col="red", size=1.5)
```



I prefer the logistic model because it seems to capture the trend when k increases while the linear model exihibts some kinds of deviation.