Homework 4

zhang zhuohan

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.

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.

```
library(tidyverse)
ckm_nodes <- read_csv("../data/ckm_nodes.csv")
noinfor <- which(is.na(ckm_nodes$adoption_date))
ckm_nodes <- ckm_nodes[-noinfor, ]
ckm_network <- read.table("../data/ckm_network.dat")
ckm_network <- ckm_network[-noinfor, -noinfor]
```

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.

```
library(dplyr)
new.ckm <- data.frame(doctor = rownames(ckm_nodes)) %>%
        slice(rep(1:n(), each=17)) %>%
        mutate(month = rep(1:17,length.out=n())) %>%
        mutate(whether_adoption = as.numeric(ckm_nodes[doctor,2]==month)) %>%
        mutate(adoption_before_month = as.numeric(ckm_nodes[doctor,2]<month))
new.ckm <- new.ckm %>%
        mutate(contacts_str_before = apply(new.ckm, 1, function(x){
            idx <- (ckm_network[as.numeric(x[1]),] == 1)
            idx <- unname(t(idx))[,1]
            return(sum(ckm_nodes[idx, 2] < as.numeric(x[2])))
        }))
new.ckm <- new.ckm %>%
        mutate(contacts_with_before = apply(new.ckm, 1, function(x){
            idx <- (ckm_network[as.numeric(x[1]),] == 1)
            idx <- unname(t(idx))[,1]
            return(sum(ckm_nodes[idx, 2] <= as.numeric(x[2])))
        }))
```

   - There are 6 variables, so there are 6 columns in the data frame.
   - There are 125 doctors and 17 months, so there are totally $125 \times 17 = 2125$ rows in the data frame.

3. With p_k and q_k defined as the original question, 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.

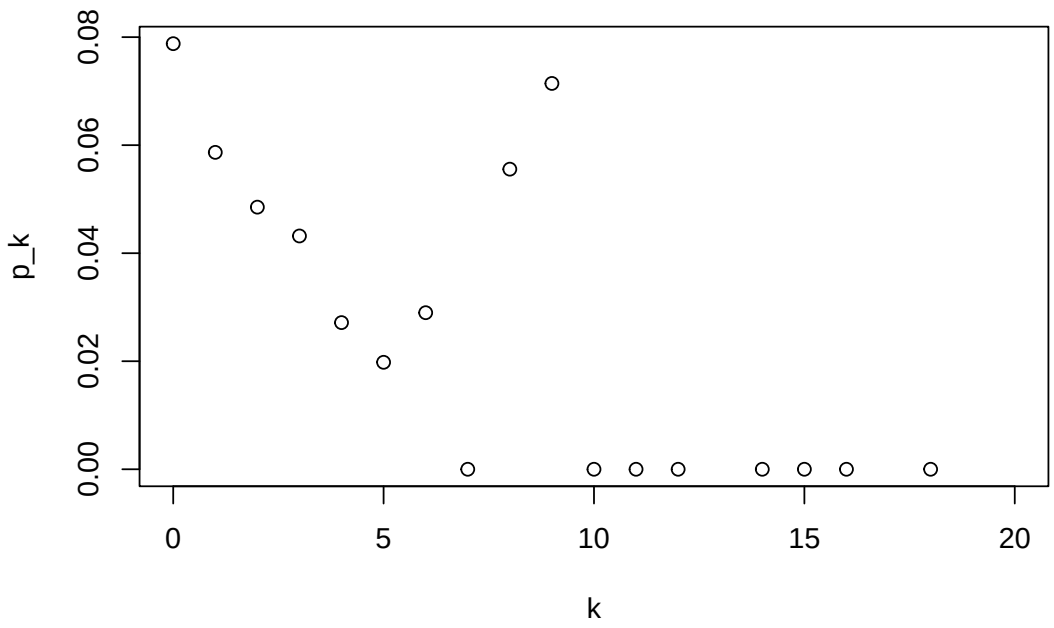
```
max(apply(ckm_network, 1, sum))
```

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

So a doctor has at most 20 contact, which means $0 \leqslant k \leqslant 20$.

- 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.

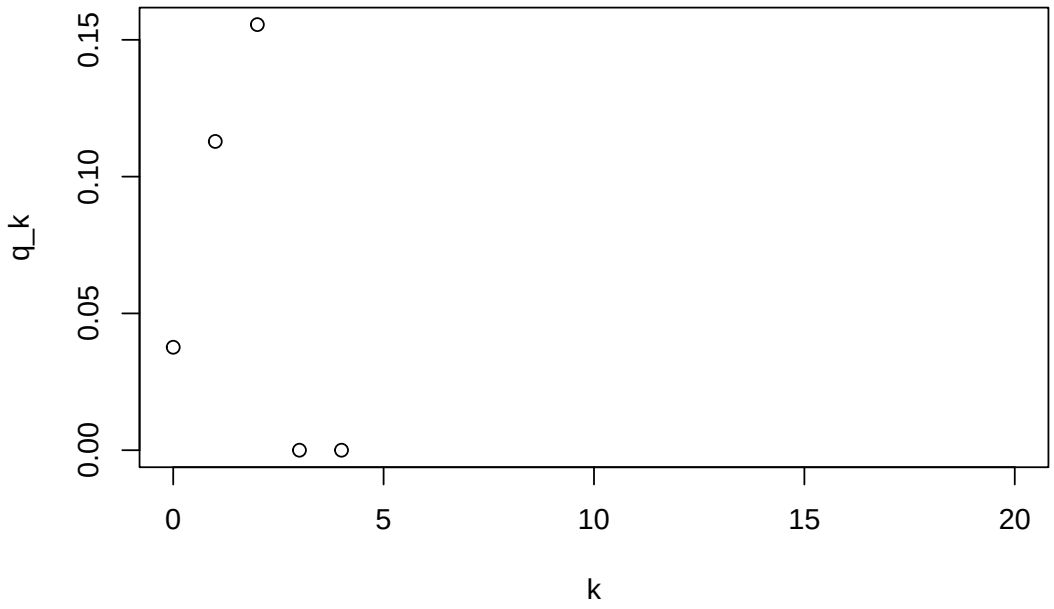
```
k.vec <- c(0:20)
pk.vec <- vector(mode = "numeric",length = 21)
for(k in 0:20){
    fil.ckm <- new.ckm %>% filter(contacts_str_before==k)
    all_count <- dim(fil.ckm)[1]
    fil.ckm <- fil.ckm %>% filter(whether_adoption==1)
    condi_count <- dim(fil.ckm)[1]
    pk.vec[k+1] <- condi_count/all_count
}
plot(k.vec, pk.vec, xlab="k", ylab="p_k")
```



- 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.

```
qk.vec <- vector(mode = "numeric",length = 21)
for(k in 0:20){
    fil.ckm <- new.ckm %>% filter(
            contacts_with_before-contacts_str_before==k)
```

```
    all_count <- dim(fil.ckm)[1]
    fil.ckm <- fil.ckm %>% filter(whether_adoption==1)
    condi_count <- dim(fil.ckm)[1]
    qk.vec[k+1] <- condi_count/all_count
}
plot(k.vec, qk.vec, xlab="k", ylab="q_k")
```



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.

```
df.p <- data.frame(k = k.vec, pk = pk.vec)
model.1 <- lm(pk~k, df.p)
summary(model.1)
```

```
##
## Call:
## lm(formula = pk ~ k, data = df.p)
##
## 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 ***
## k           -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
##   (4 observations deleted due to missingness)
## 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 $p_k = \frac{e^{a+bk}}{1+e^{a+bk}}$. 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.2 <- lm(pk.log ~ k, df.p %>%
            mutate(pk.log = ifelse(pk==0, log(0.0001/(1-0.0001)),
                log(pk/(1-pk)))))
summary(model.2)
```

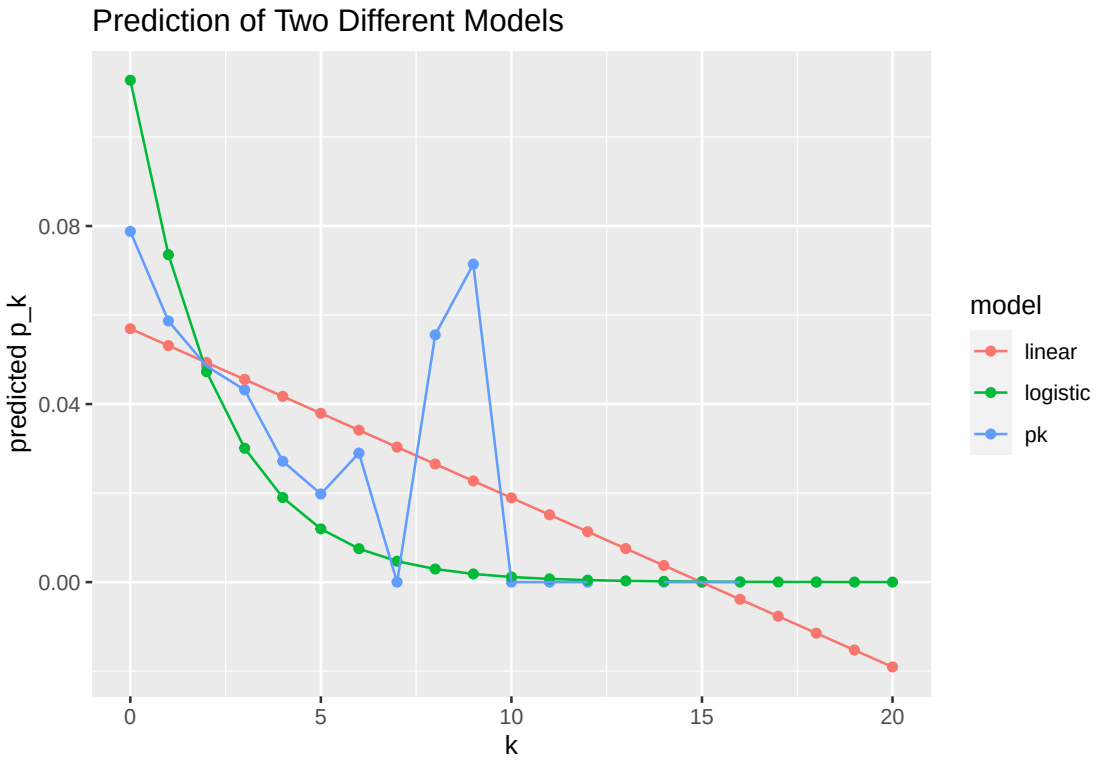
```
##
## Call:
## lm(formula = pk.log ~ k, data = df.p %>% mutate(pk.log = ifelse(pk ==
##     0, log(1e-04/(1 - 1e-04)), log(pk/(1 - pk)))))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8553 -0.5633  0.0276  0.5124  3.7306
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.06297    0.85768  -2.405   0.0295 *
## k           -0.47029    0.08703  -5.404 7.32e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.91 on 15 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6606, Adjusted R-squared:  0.638
## F-statistic:  29.2 on 1 and 15 DF,  p-value: 7.319e-05
```

The logistic model suggests that growth is approximately exponential in the initial stages as k grows; then, as saturation begins, growth slows to linear, and at last, growth stops.

- 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?

```
library(ggplot2)
library(tidyr)
pred.1 <- predict(model.1, data.frame(k=c(0:20)))
temp <- predict(model.2, data.frame(k=c(0:20)))
pred.2 <- exp(temp)/(1+exp(temp))
df.p <- df.p %>% mutate(linear=pred.1, logistic=pred.2)
```

```
df.p %>% gather(model, pred, -k) %>%
        ggplot() + geom_point(aes(k, pred, color=model)) +
        geom_line(aes(k, pred, color=model)) +
        labs(x = "k", y = "predicted p_k",
             title = "Prediction of Two Different Models")
```



Obviously, the logistic model is better.