Homework 4: Diffusion of Tetracycline

2025 暑假短学期《数学实践》

任心然, 3220103492

2025/06/29

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]
```

```
head(ckm_nodes)
```

```
## # A tibble: 6 x 13
##   city   adoption_date medical_school attend_meetings medical_journals
##   <chr>          <dbl> <chr>          <chr>                      <dbl>
## 1 Peoria             1 1920--1929     specialty                      9
## 2 Peoria            12 1945+          none                           5
## 3 Peoria             8 1935--1939     general                        7
## 4 Peoria             9 1940--1944     general                        6
## 5 Peoria             9 1935--1939     general                        4
## 6 Peoria            10 1930--1934     none                           7
## # i 8 more variables: free_time_with <chr>, discuss_medicine_socially <chr>,
## #   club_with_drs <chr>, drs_among_three_best_friends <dbl>,
## #   practicing_here <chr>, office_visits_per_week <chr>,
## #   proximity_to_other_drs <chr>, specialty <chr>
```

```
head(ckm_network)
```

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.

```
df <- data.frame(doctor = rownames(ckm_nodes)) |>
  slice(rep(1:n(),each = 17)) |>
  mutate(month = rep(1 : 17 , length.out = n())) |>
  mutate(begin = as.numeric(ckm_nodes[doctor, 2] == month)) |>
  mutate(done = as.numeric(ckm_nodes[doctor, 2] < month))

countsum1 <- function(x){
    return(sum( ckm_nodes[ckm_network[as.numeric(x[1])] == 1, 2] < as.numeric(x[2])))
}
df <- df |>  mutate(contacts_prescribed_before = apply(df, 1, countsum1))

countsum2 <- function(x){
  return(sum(ckm_nodes[ckm_network[as.numeric(x[1])] == 1, 2] <= as.numeric(x[2])))
}
df <- df |>  mutate(contacts_prescribed_now_before = apply(df,1,countsum2))
```

```
# Verify the dimensions of the data frame
nrow(df)  # Should be 2125 (125 doctors * 17 months)
```

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

```
ncol(df)  # Should be 6
```

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

- There are six variables so it has 6 columns;
- There are 125 doctors and 17 months, so it has 125 * 17 = 2125 rows.

3. Let

$$
\begin{aligned}
p_k = \Pr(&\text{A doctor starts prescribing tetracycline this month} \mid \\
&\text{Number of doctor's contacts prescribing before this month} = k)
\end{aligned}
\tag{1}
$$

and

$$
\begin{aligned}
q_k = \Pr(&\text{A doctor starts prescribing tetracycline this month} \mid \\
&\text{Number of doctor's contacts prescribing this month} = k)
\end{aligned}
\tag{2}
$$

We suppose that p_k and q_k are the same for all months.

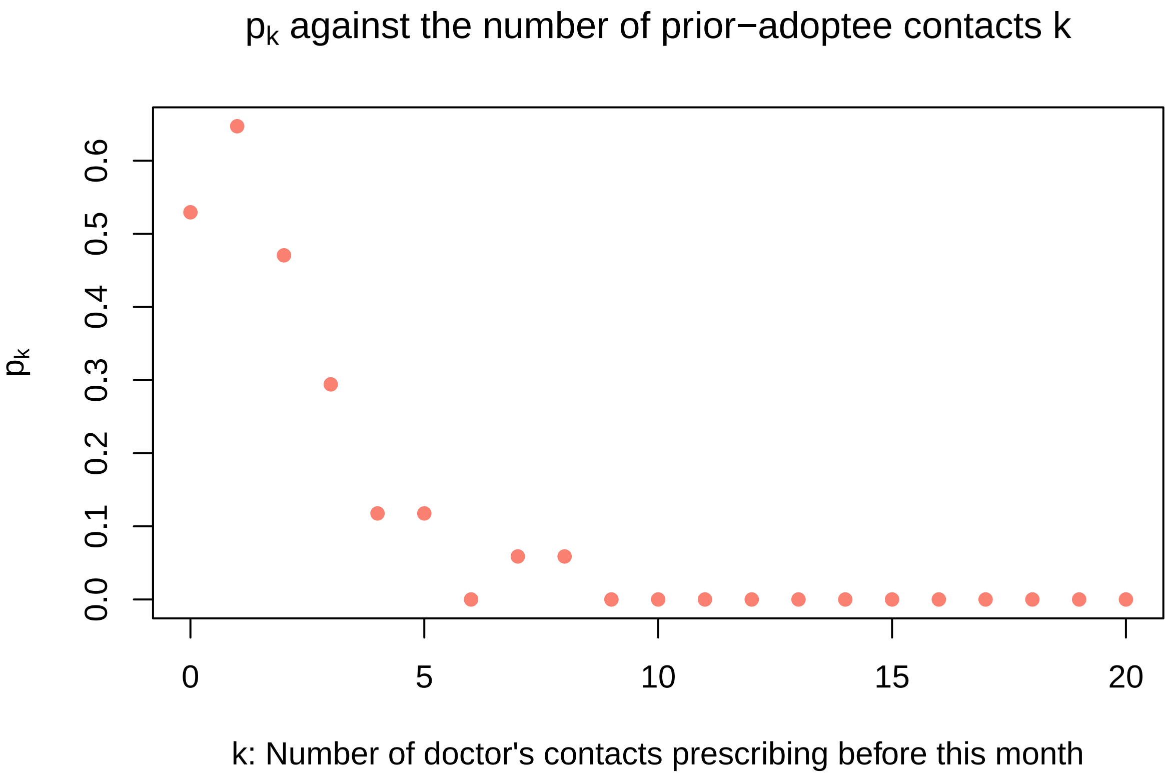
a. Explain why there should be no more than 21 values of k for which we can estimate $p_k$ and q_k directly from the data.

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

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

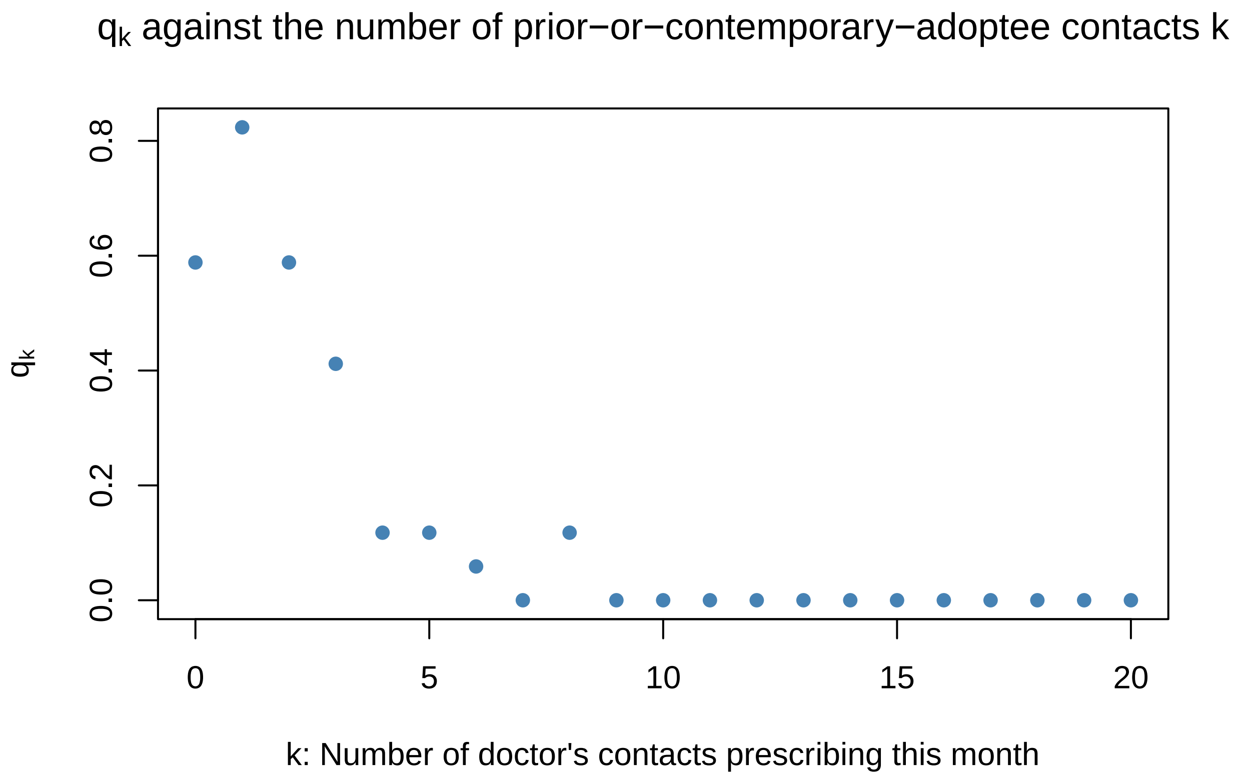
b. Create a vector of estimated $p_k$ probabilities, using the data frame from (2). Plot the probabilities against the number of prior-adoptee contacts k.

```
library(latex2exp)
pk.estimated <- numeric(21)
for (k in 0:20){
    df_select_k <- df |>  filter(contacts_prescribed_before == k)
    if (nrow(df_select_k) == 0){
      pk.estimated[k] <- 0
    }
    else {
      pk.month = 0
      for (i in 1:17){
        df_select_k_month <- df_select_k |>  filter(month == i)
        pk.month <- pk.month + (sum(df_select_k_month[,3]) > 0)
      }
      pk.estimated[k] <- pk.month/17
    }
}
plot(0 : 20, pk.estimated,
     xlab = "k: Number of doctor's contacts prescribing before this month",
     ylab = TeX("$p_k$"),
     main = TeX("$p_k$ against the number of prior-adoptee contacts k"),
     pch = 16, col = "salmon")
```



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.

```r
qk.estimated <- numeric(21)
for (k in 0 : 20){
  df_select_qk <- df |>  filter(contacts_prescribed_now_before == k)
  if (nrow(df_select_qk) == 0){
    qk.estimated[k] <- 0
  }
  else{
    qk.month <- 0
    for (i in 1 : 17){
      df_select_qk_month <- df_select_qk |>  filter(month == i)
      qk.month <- qk.month + (sum(df_select_qk_month[, 3]) > 0)
    }
    qk.estimated[k] <- qk.month / 17
  }
}
plot(0 : 20, qk.estimated,
     xlab = "k: Number of doctor's contacts prescribing this month",
     ylab = TeX("$q_k$"),
     main = TeX("$q_k$ against the number of prior-or-contemporary-adoptee contacts k"),
     pch = 16, col = "steelblue")
```



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 <- 0:20
df.pk <- data.frame(pk = pk.estimated, K = 0:20)
lm.p <- lm(pk.estimated ~ K, data = df.pk)
summary(lm.p)
```

```
##
## Call:
## lm(formula = pk.estimated ~ K, data = df.pk)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20550 -0.10924 -0.01299  0.08327  0.32124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.349885   0.056771   6.163 6.36e-06 ***
## K           -0.024064   0.004856  -4.955 8.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1348 on 19 degrees of freedom
## Multiple R-squared:  0.5638, Adjusted R-squared:  0.5408
## F-statistic: 24.56 on 1 and 19 DF,  p-value: 8.784e-05
```

- Parameter estimates: $\hat{a}$ = 0.349885, $\hat{b}$ = -0.024064.

b. Suppose $p_k = e^{a+bk}/(1+e^{a+bk})$. 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).

- Explanation: Suppose $b > 0$. As k grows, the initial phase exhibits near-exponential (geometric) growth. Upon approaching saturation, the growth transitions to linear (arithmetic) scaling before finally stabilizing at maturity.

```r
nls.p = nls(pk.estimated ~ exp(a+b*K)/(1+exp(a+b*K)),
            start = list(a = 0, b = 0.1), data = df.pk)
summary(nls.p)
```

```
##
## Formula: pk.estimated ~ exp(a + b * K)/(1 + exp(a + b * K))
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a  0.69207    0.15681   4.413 0.000298 ***
## b -0.53926    0.06157  -8.759 4.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04973 on 19 degrees of freedom
##
## Number of iterations to convergence: 7
## Achieved convergence tolerance: 1.459e-06
```

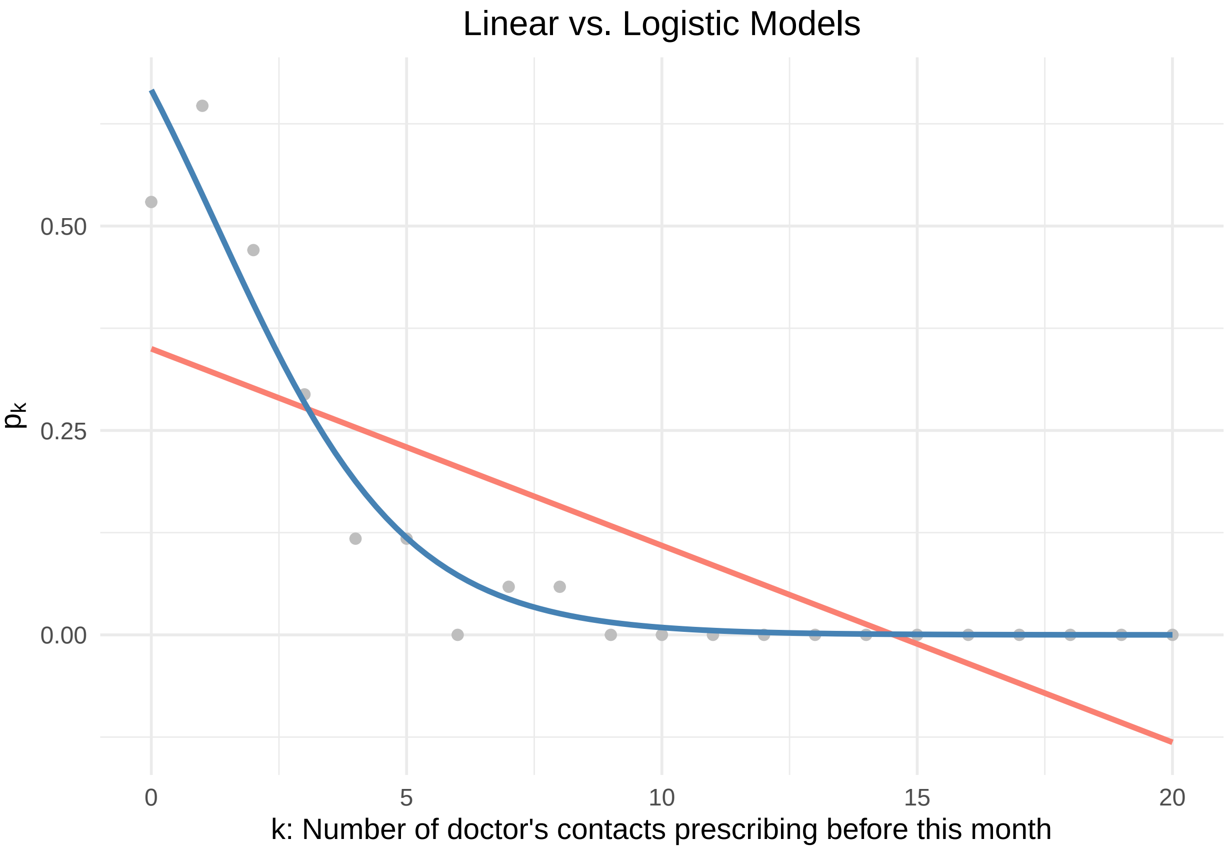
- Estimates: $\hat{a}$ = 0.69207, $\hat{b}$ = -0.53926.

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?

```
a.linear <- 0.349885
b.linear <- -0.024064

a.logical <- 0.69207
b.logical <- -0.53926

lm <- function(pk){a.linear + b.linear * pk}
logic <- function(pk){exp(a.logical + b.logical * pk) / (1 + exp(a.logical + b.logical * pk))}
df.pk |> ggplot(aes(x = K, y = pk)) +
  geom_point(size = 1.5, color = "gray") +
  stat_function(fun = lm, col = "salmon", linewidth = 1) +
  stat_function(fun = logic, col = "steelblue", linewidth = 1) +
  theme_minimal()+
  labs(x = "k: Number of doctor's contacts prescribing before this month",
       y = TeX("$p_k$"), title = "Linear vs. Logistic Models") +
  theme(
  plot.title = element_text(hjust = 0.5)
  )
```



- I favor the logistic model due to its significantly better fit with the empirical data.

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?

- While the variance-weighted approach would indeed provide more precise estimates, I must defer this analysis due to time constraints. The current unweighted models serve as reasonable first approximations given these limitations.