homework4 report

王梓

2024-07-04

1.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
ckm_nodes <- read_csv('data/ckm_nodes.csv')
```

```
## Rows: 246 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────────────
## Delimiter: ","
## chr (10): city, medical_school, attend_meetings, free_time_with, discuss_med...
## dbl  (3): adoption_date, medical_journals, drs_among_three_best_friends
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
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.

`ckm_nodes` 中记录了125个医生在17个月中的数据，因此行数应为 17 * 125 = 2150。需要记录的信息有4项，再加上医生序号及月份，故共需要6列。

```
n_doc <- nrow(ckm_nodes)
n_mon <- max(ckm_nodes$adoption_date[is.finite(ckm_nodes$adoption_date)])

doc_mon <- data.frame("doctor" = rep(1:n_doc, each = n_mon),
                           "month" = rep(1:n_mon, times = n_doc))
```

```
doc_mon$began_prescribing <-
  ckm_nodes$adoption_date[doc_mon$doctor] == doc_mon$month
doc_mon$began_prescribing_before <-
  ckm_nodes$adoption_date[doc_mon$doctor] < doc_mon$month

tb <- t(ckm_nodes$adoption_date)[rep(1, nrow(doc_mon)), ]
doc_mon$n_began_prescribing_before <-
  rowSums(ckm_network[doc_mon$doctor, ] & (tb < doc_mon$month))
doc_mon$n_began_prescribing_or_before <-
  rowSums(ckm_network[doc_mon$doctor, ] & (tb <= doc_mon$month))

head(doc_mon)
```

```
##   doctor month began_prescribing began_prescribing_before
## 1      1     1              TRUE                    FALSE
## 2      1     2             FALSE                     TRUE
## 3      1     3             FALSE                     TRUE
## 4      1     4             FALSE                     TRUE
## 5      1     5             FALSE                     TRUE
## 6      1     6             FALSE                     TRUE
##   n_began_prescribing_before n_began_prescribing_or_before
## 1                          0                             1
## 2                          1                             1
## 3                          1                             2
## 4                          2                             3
## 5                          3                             3
## 6                          3                             3
```

3.

a.

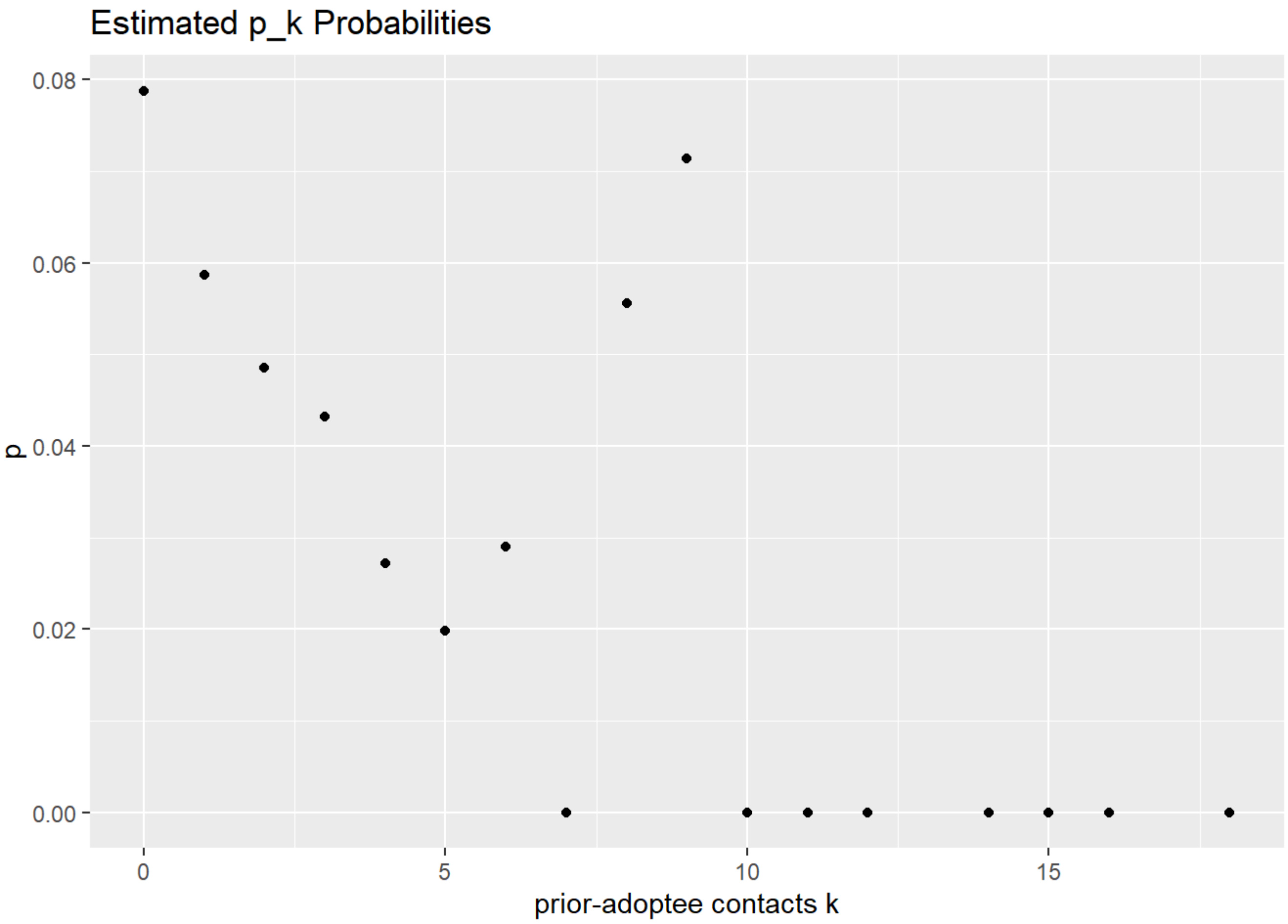
```
max(rowSums(ckm_network))
```

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

数据中一个医生最多有 20 个联系人，因此 k 只能取 0 ~ 20 共21个量。

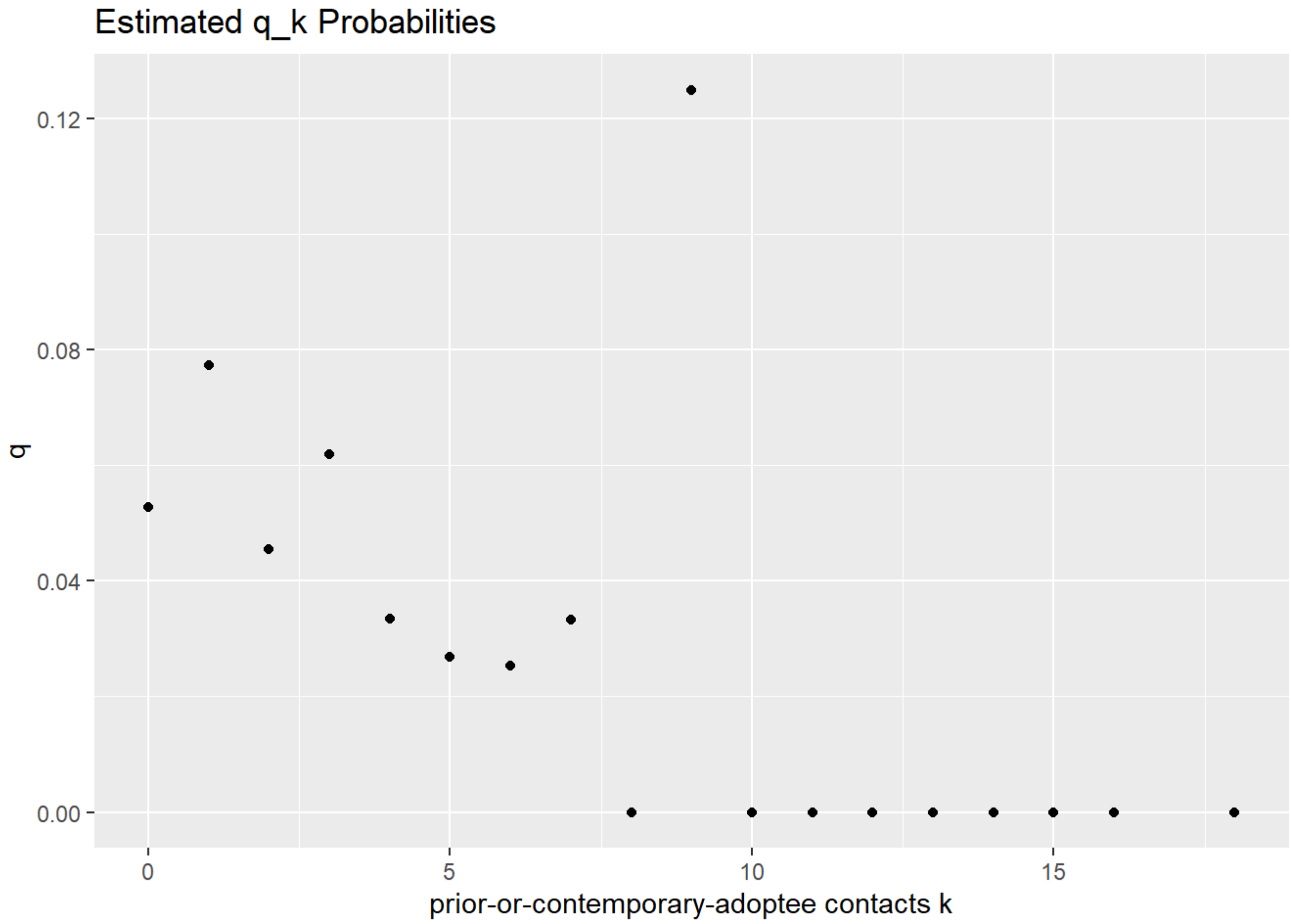
b.

```
p_data <- doc_mon |>
  group_by(n_began_prescribing_before) |>
  summarise(num = sum(began_prescribing), total = n()) |>
  mutate(p_k = num / total) |>
  rename(k = n_began_prescribing_before)

ggplot(data = p_data) +
  geom_point(aes(x = k, y = p_k)) +
  labs(title = ″Estimated p_k Probabilities″,
       x = ″prior-adoptee contacts k″,
       y = ″p″)
```



c.

```
q_data <- doc_mon |>
  group_by(n_began_prescribing_or_before) |>
  summarise(num = sum(began_prescribing), total = n()) |>
  mutate(q_k = num / total) |>
  rename(k = n_began_prescribing_or_before)

ggplot(data = q_data) +
  geom_point(aes(x = k, y = q_k)) +
  labs(title = ″Estimated q_k Probabilities″,
       x = ″prior-or-contemporary-adoptee contacts k″,
       y = ″q″)
```



4.

a.

```
a.model <- lm(p_k ~ k, data = p_data)
a.model$coefficients
```

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

b.

若 b>0，则起初时随着 k 的增大，p的增长速率会逐渐上升，当k增大到一定程度时，p的增长速率放缓，最终停止增长。

```
b.func <- function(para, X) {
  return(1 - 1 / (1 + exp(para[1] + para[2] * X)))
}

b.model <- nls(p_k ~ b.func(para, k),
              data = p_data,
              start = list(para = c(0, 0)))
summary(b.model)$coefficients[c(1, 2)]
```

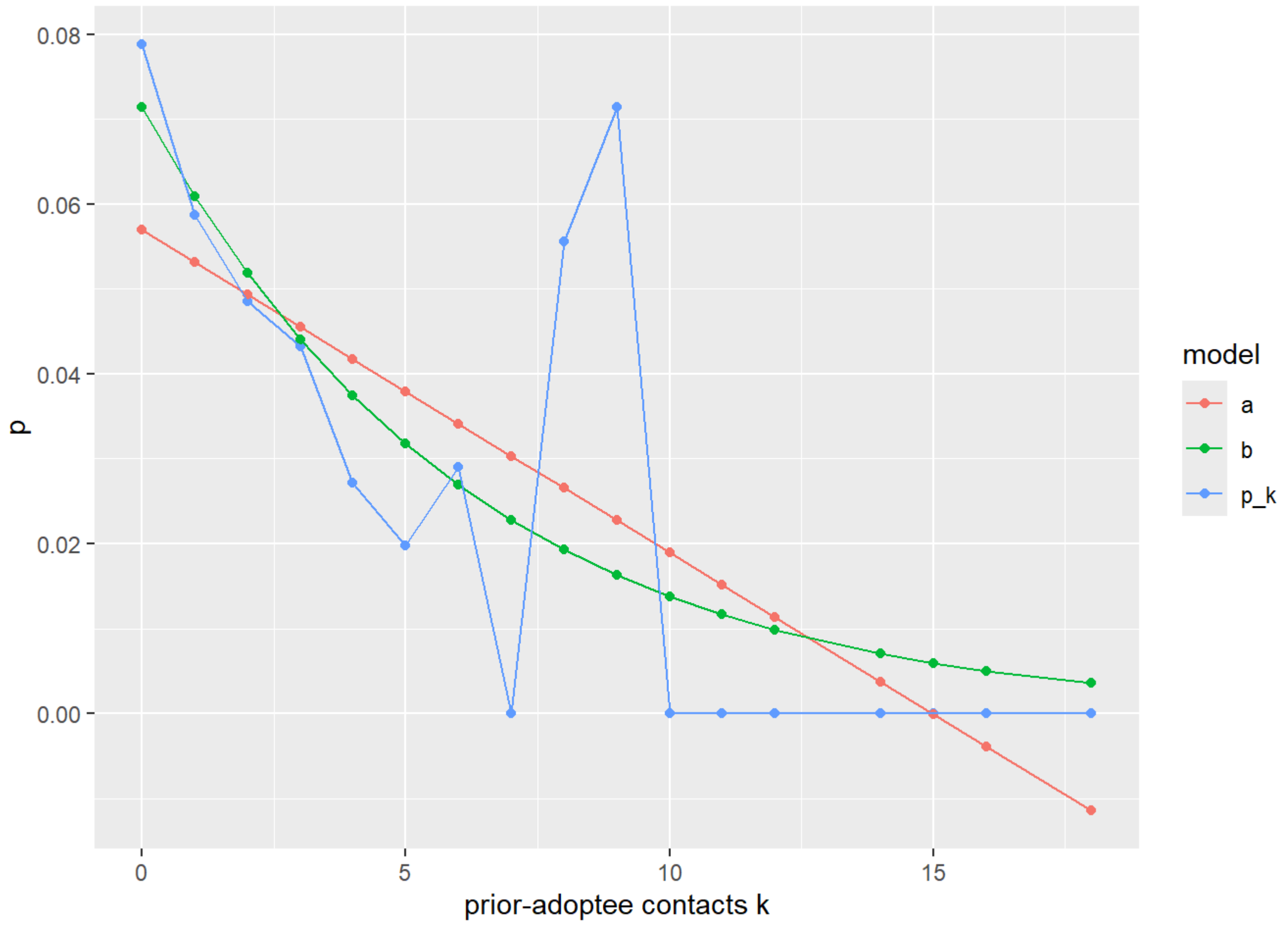
```
## [1] -2.5650784 -0.1705091
```

c.

```
library(dplyr)

p_data <- p_data |>
  mutate(a = predict(a.model, p_data), b = predict(b.model, p_data))

p_tidy <- p_data |>
  dplyr::select(-num, -total) |>
  gather(key = model, value = p, -k)
```

```
## Warning: attributes are not identical across measure variables; they will be
## dropped
```

```
p_tidy |>
  ggplot(aes(x = k, y = p, col = model)) +
  geom_point() +
  geom_line() +
  labs(x = "prior-adoptee contacts k")
```



4(b) 中的模型似乎更加符合计算出的 p_k。 模型 a 中 p 的取值为 [-inf, inf]，而模型 b 中 p 的取值为[0,1]，更加贴合实际。