homework4

zza

2025-07-01

1

```
ckm_nodes <- read.csv('ckm_nodes.csv')


noinfor <- which(is.na(ckm_nodes$adoption_date))


ckm_nodes <- ckm_nodes[-noinfor, ]


ckm_network <- read.table('ckm_network.dat')


ckm_network <- ckm_network[-noinfor, -noinfor]
```

2

```
num_doctors <- nrow(ckm_nodes)

num_months <- 17


result_df <- expand.grid(
  doctor_id = 1:num_doctors,
  month = 1:num_months
) %>%
```

```
  mutate(
    started_this_month = FALSE,
    already_started_before = FALSE,
    contacts_started_before = 0,
    contacts_started_by = 0
  )


for (i in 1:num_doctors) {
  adoption_date <- ckm_nodes$adoption_date[i]
  for (j in 1:num_months) {
    if (!is.na(adoption_date)) {

      result_df[(result_df$doctor_id == i) & (result_df$month == j),
                "started_this_month"] <- (adoption_date == j)

      result_df[(result_df$doctor_id == i) & (result_df$month == j),
                "already_started_before"] <- (adoption_date < j)


      contacts_before <- sum(ckm_network[i, ] & ckm_nodes$adoption_date <= j - 1)
      result_df[(result_df$doctor_id == i) & (result_df$month == j),
                "contacts_started_before"] <- contacts_before

      contacts_by <- sum(ckm_network[i, ] & ckm_nodes$adoption_date <= j)
      result_df[(result_df$doctor_id == i) & (result_df$month == j),
                "contacts_started_by"] <- contacts_by
    }
  }
}

# 查看结果
dim(result_df)
```

```
## [1] 2125    6
```

3

a

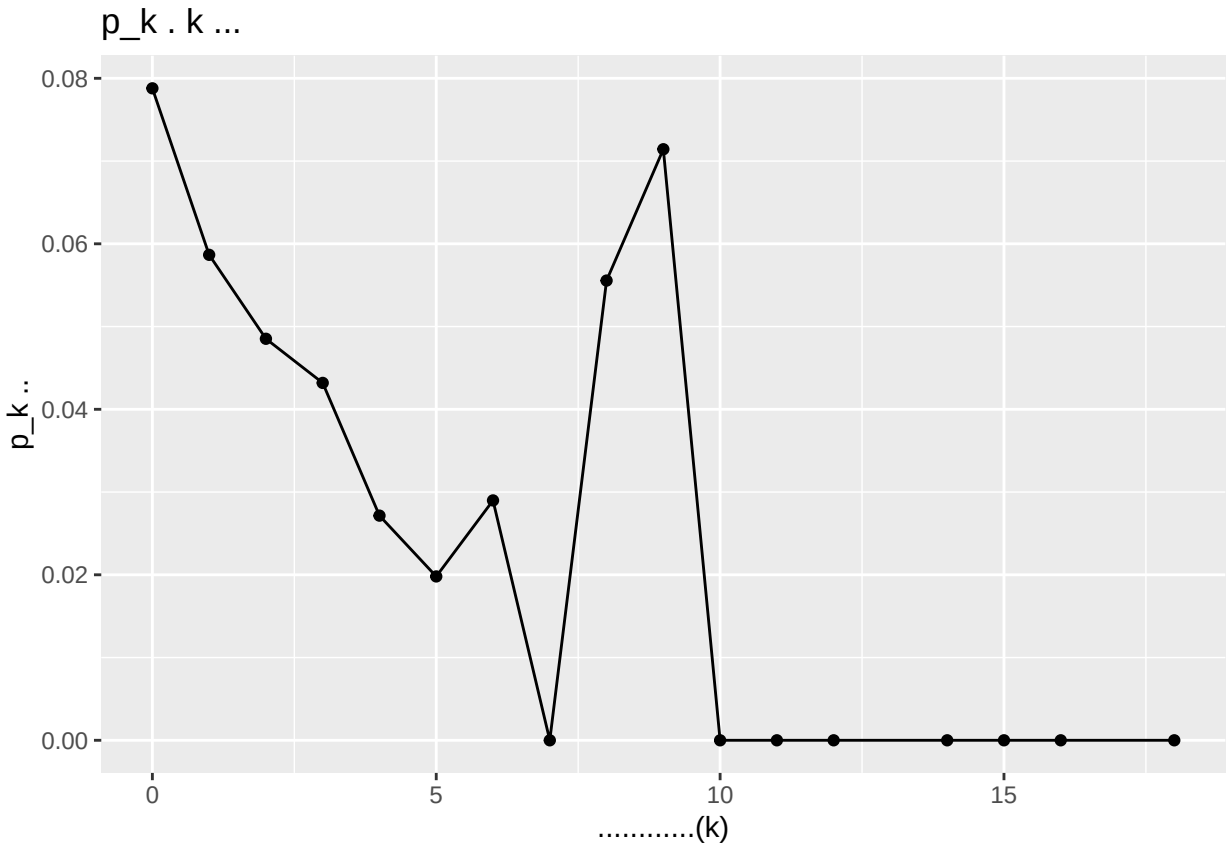
```
contacts_per_doctor <- rowSums(ckm_network)

max_contacts <- max(contacts_per_doctor)
cat(" 医生的最大社交连接数:", max_contacts, "\n")
```

```
## 医生的最大社交连接数: 20
```

因此 k 不超过 21

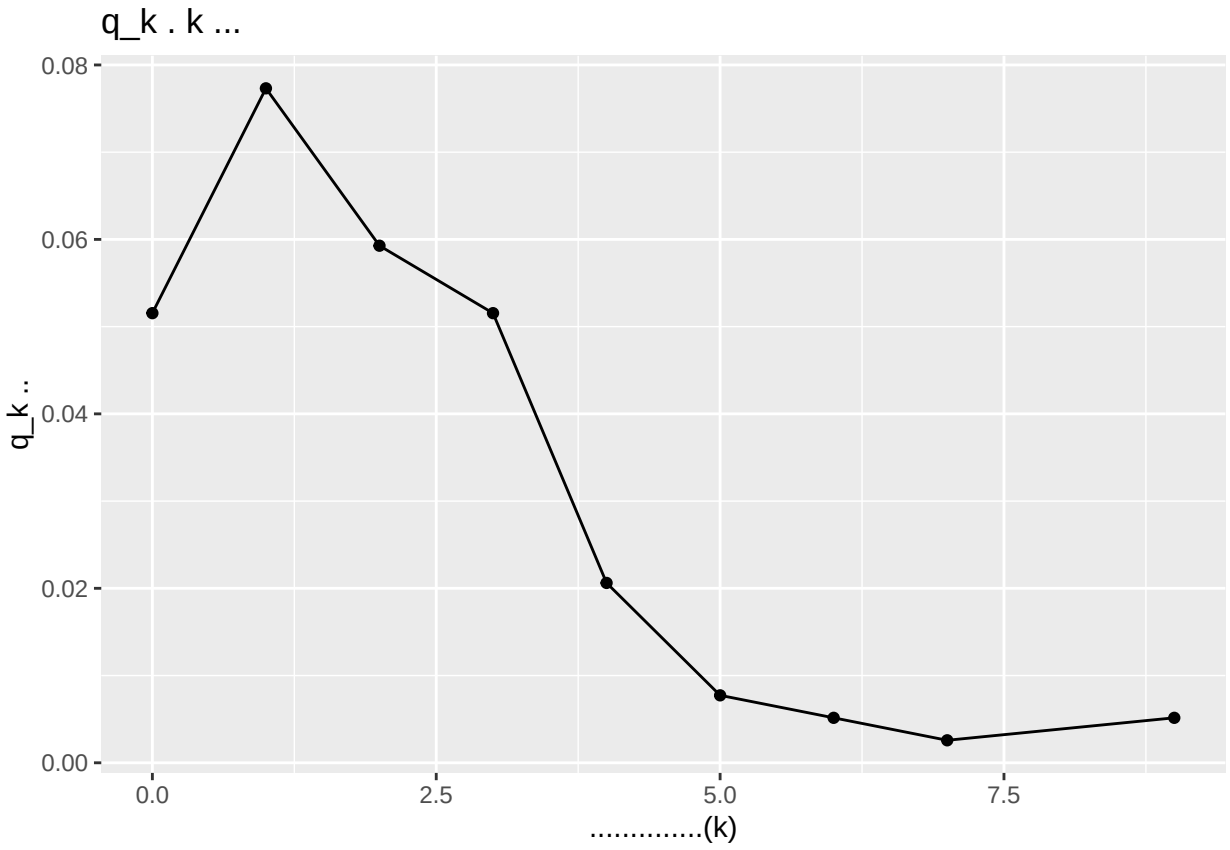
b

```
p_k_df <- result_df %>%
  group_by(contacts_started_before) %>%
  summarise(
    total = n(),
    success = sum(started_this_month),
    p_k = success / total
  ) %>%
  filter(!is.nan(p_k))

# 绘图: p_k 与 k 的关系
ggplot(p_k_df, aes(x = contacts_started_before, y = p_k)) +
  geom_point() + geom_line() +
  labs(x = " 当月前已采用的联系人数量（k）", y = "p_k 概率", title = "p_k 与 k 的关系")
```



c

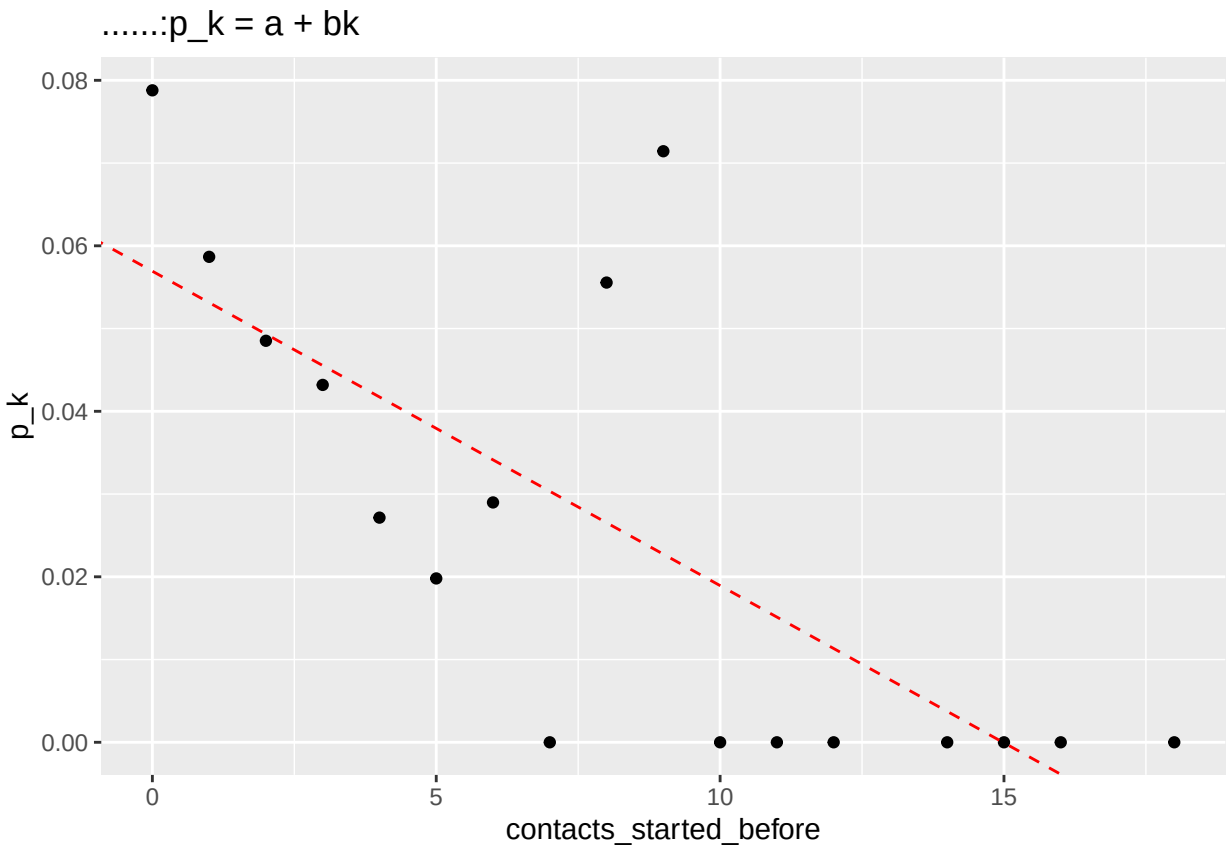
```
q_k_df <- result_df %>%
  filter(started_this_month == TRUE) %>%
  group_by(contacts_started_by) %>%
  summarise(
    total = nrow(filter(result_df, contacts_started_by == cur_group_id())),
    success = sum(started_this_month),
    q_k = success / total
  )


ggplot(q_k_df, aes(x = contacts_started_by, y = q_k)) +
  geom_point() + geom_line() +
  labs(x = " 当月及之前已采用的联系人数量（k）", y = "q_k 概率", title = "q_k 与 k 的关系")
```



4

a

```
# 基于 3.b 得到的 p_k_df, 拟合线性模型
linear_fit <- lm(p_k ~ contacts_started_before, data = p_k_df)


a_linear <- coef(linear_fit)[1]
b_linear <- coef(linear_fit)[2]


ggplot(p_k_df, aes(x = contacts_started_before, y = p_k)) +
  geom_point() +
  geom_abline(intercept = a_linear, slope = b_linear, color = "red", linetype = "dashed") +
  labs(title = " 线性模型拟合: p_k = a + bk")
```



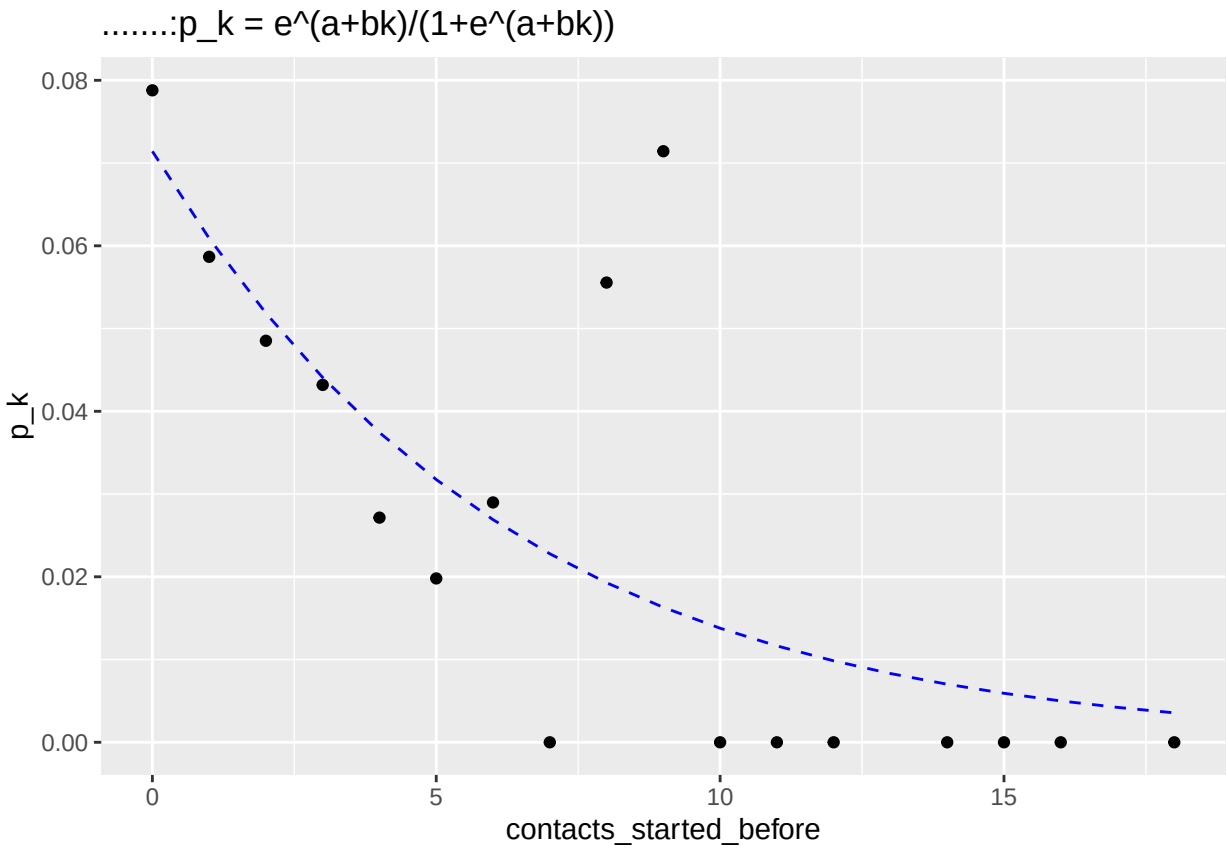
b

```
logistic_model <- function(k, a, b) {
  exp(a + b * k) / (1 + exp(a + b * k))
}


nonlinear_fit <- nls(
  p_k ~ logistic_model(contacts_started_before, a, b),
  data = p_k_df,
  start = list(a = a_linear, b = b_linear)  # 用线性模型结果做初始值
)


summary(nonlinear_fit)
```

```
##
## Formula: p_k ~ logistic_model(contacts_started_before, a, b)
```

```
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a -2.56508    0.20610 -12.446 2.62e-09 ***
## b -0.17051    0.05371  -3.174  0.00628 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01957 on 15 degrees of freedom
##
## Number of iterations to convergence: 7
## Achieved convergence tolerance: 1.592e-07
```

```r
a_logistic <- coef(nonlinear_fit)["a"]
b_logistic <- coef(nonlinear_fit)["b"]



k_range <- seq(min(p_k_df$contacts_started_before), max(p_k_df$contacts_started_before), by = 1)
predicted <- logistic_model(k_range, a_logistic, b_logistic)

ggplot(p_k_df, aes(x = contacts_started_before, y = p_k)) +
  geom_point() +
  geom_line(data = data.frame(k = k_range, p = predicted),
            aes(x = k, y = p), color = "blue", linetype = "dashed") +
  labs(title = " 非线性模型拟合: p_k = e^(a+bk)/(1+e^(a+bk))")
```



c

```
pred_linear <- a_linear + b_linear * k_range
pred_logistic <- logistic_model(k_range, a_logistic, b_logistic)


ggplot(p_k_df, aes(x = contacts_started_before, y = p_k)) +
  geom_point(color = "black") +
  geom_line(data = data.frame(k = k_range, p = pred_linear),
            aes(x = k, y = p), color = "red", linetype = "dashed", size = 1) +
  geom_line(data = data.frame(k = k_range, p = pred_logistic),
            aes(x = k, y = p), color = "blue", linetype = "dashed", size = 1) +
  labs(
    x = " 联系人数量（k）",
    y = " 采用概率（p_k）",
    title = " 线性 vs 非线性模型拟合对比",
    caption = " 红色：线性模型；蓝色：逻辑斯蒂模型"
  )
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```