Homework 4: Diffusion of Tetracycline

□ □ □ 3220103172

2024 年 7 月 4 日

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.

不用循环的方法：

```
library(dplyr)
c1 <- rep(1:17,times=125)
c2 <- rep(1:125,each=17)
c3<-c(1:17)
c<-data.frame(c1,c2)
b<-ckm_nodes$adoption_date
b<-rep(b,each=17)
c<-c %>%
mutate(whether_prescribe_that_month=(c1==b),whether_prescribe_before=(c1>b))
f<-function(i){
  return((as.numeric(apply(ckm_network[,-which(ckm_nodes$adoption_date>=rep(i,time=125))],1,sum))))
}
a<-rbind(f(1),f(2),f(3),f(4),f(5),f(6),f(7),f(8),f(9),f(10),f(11),f(12),f(13),f(14),f(15),f(16),f(17)
```

```
a<-data.frame(a)
a<-gather(a,value=number_of_prior_adopted_contacts)
f1<-function(i){
  return((as.numeric(apply(ckm_network[,-which(ckm_nodes$adoption_date>rep(i,time=125))],1,sum))))
}
a1<-rbind(f1(1),f1(2),f1(3),f1(4),f1(5),f1(6),f1(7),f1(8),f1(9),f1(10),f1(11),f1(12),f1(13),f1(14),f1
a1<-data.frame(a1)
a1<-gather(a1,value=number_of_prior_or_contemporary_adopted_contacts)
c<-cbind(c,a$number_of_prior_adopted_contacts,a1$number_of_prior_or_contemporary_adopted_contacts)
colnames(c) <- c('month','doctor','whether_prescribe_that_month','whether_prescribe_before','number_c
```

老方法:（使用循环）

```
library(dplyr)
c1 <- rep(1:17,times=125)
c2 <- rep(1:125,each=17)
c3<-c(1:17)
c<-data.frame(c1,c2)
b<-ckm_nodes$adoption_date
b<-rep(b,each=17)
c<-c %>%
mutate(whether_prescribe_that_month=(c1==b),whether_prescribe_before=(c1>b))
c5 <- c()
for(i in 1:nrow(ckm_nodes)){
 for (j in 1:17){
 c5 <- c(c5,sum(ckm_network[row.names(ckm_network) == i,] == 1 & ckm_nodes$adoption_date < j))
 }
}
c6 <- c()
for(i in 1:nrow(ckm_nodes)){
 for (j in 1:17){
 c6 <- c(c6,sum(ckm_network[row.names(ckm_network) == i,] == 1 & ckm_nodes$adoption_date <= j))
 }
}
c<-data.frame(c,c5,c6)
colnames(c) <- c('month','doctor','whether_prescribe_that_month','whether_prescribe_before','number_c
head(c)
```

```
##   month doctor whether_prescribe_that_month whether_prescribe_before
## 1     1      1                         TRUE                    FALSE
## 2     2      1                        FALSE                     TRUE
## 3     3      1                        FALSE                     TRUE
```

```
## 4     4      1                          FALSE                          TRUE
## 5     5      1                          FALSE                          TRUE
## 6     6      1                          FALSE                          TRUE
##   number_of_prior_adopted_contacts
## 1                                0
## 2                                1
## 3                                1
## 4                                2
## 5                                3
## 6                                3
##   number_of_prior_or_contemporary_adopted_contacts
## 1                                                1
## 2                                                1
## 3                                                2
## 4                                                3
## 5                                                3
## 6                                                3
```

由于一共有 125 名医生，月份有 17 个月，因此行数 =17*125=2125，列中包括了月份，医生，该医生当月是否开始开四环素处方，当月之前是否服用四环素，当月之前开始严格开处方的接触者人数，当月或更早开始开处方的接触者人数，因此列数 =6。

3. Let

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

and

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

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(colSums(ckm_network))
```

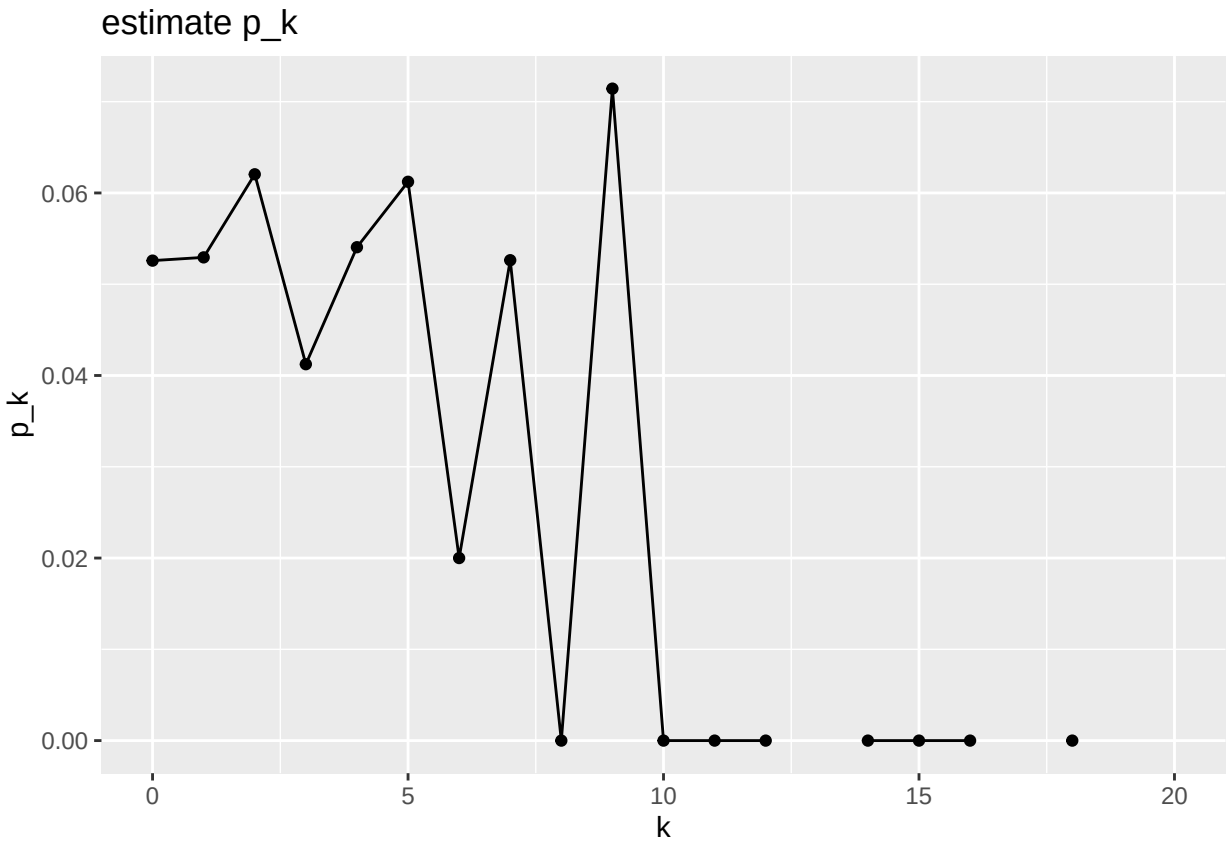
```
## [1] 20
```

所以 k 不会超过 21 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(tidyverse)
library(ggplot2)
total <- c()
for (k in 1:21){
 total[k]<-sum(c[c$whether_prescribe_that_month==TRUE,]$number_of_prior_adopted_contacts==k-1)/sum(c$
}
p_k<-data.frame("p_k"=total,"k"=c(0:20))
ggplot(data=p_k)+geom_point(aes(x = k,y = p_k))+labs(x="k",y="p_k",title="estimate p_k")+geom_line(ae
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```

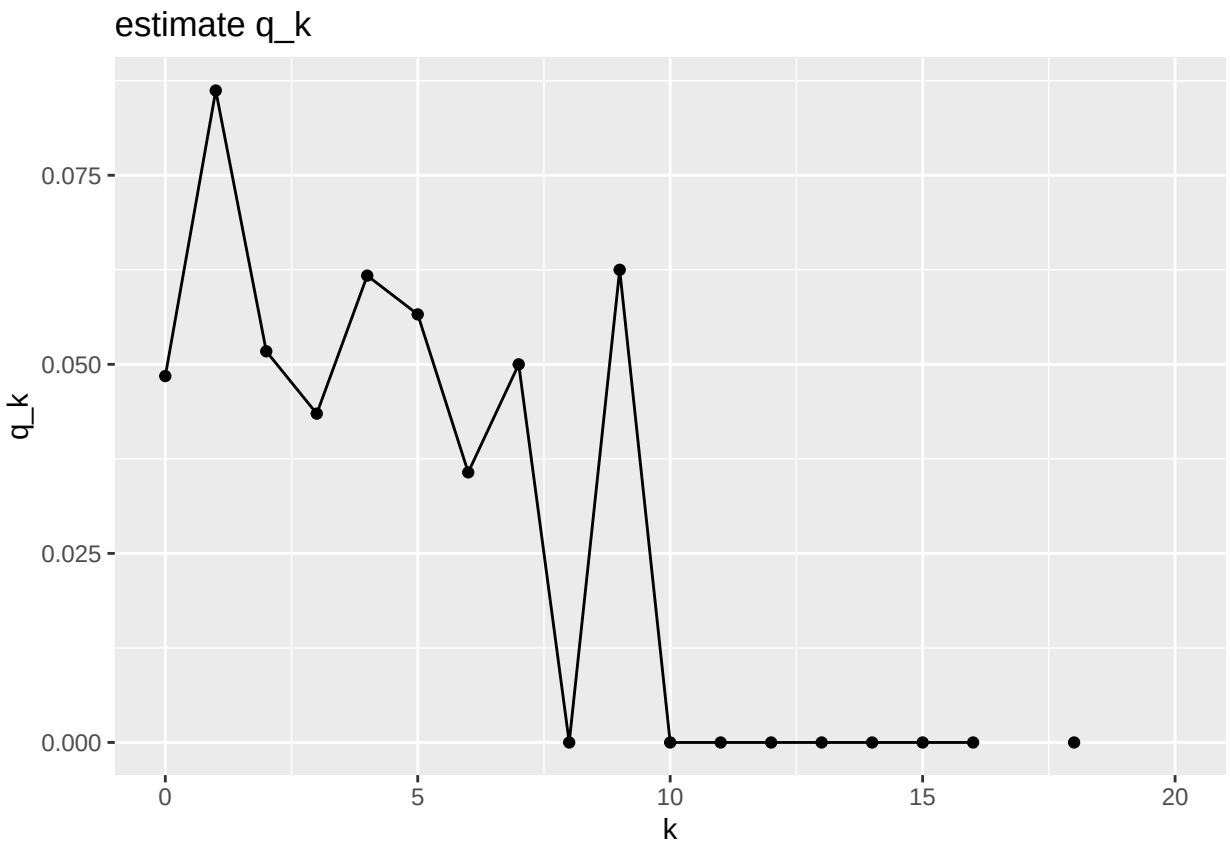


c. Create a vector of estimated q_k probabilities, using the data frame from (2). Plot the probabilities against the number of prior-or-contemporary-adoptee contacts k.

```
#c<-c %>% mutate(c,number_of_contemporary_adopted_contacts=number_of_prior_or_contemporary_adopted_co
total1 <- c()
for (k in 1:21){
 total1[k]<-sum(c[c$whether_prescribe_that_month==TRUE,]$number_of_prior_or_contemporary_adopted_cont
}
q_k<-data.frame("q_k"=total1,"k"=c(0:20))
ggplot(data=q_k)+geom_point(aes(x = k,y = q_k))+labs(x="k",y="q_k",title="estimate q_k")+geom_line(ae
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



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.

```
fit1 <- lm(p_k~k,data=p_k)
summary(fit1)
```

```
##
## Call:
## lm(formula = p_k ~ k, data = p_k)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028716 -0.012696 -0.003810  0.009318  0.046718
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0607557  0.0085563   7.101 3.62e-06 ***
## k           -0.0040050  0.0008682  -4.613 0.000338 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01905 on 15 degrees of freedom
##   (因为不存在，4个观察量被删除了)
## Multiple R-squared:  0.5865, Adjusted R-squared:  0.559
## F-statistic: 21.28 on 1 and 15 DF,  p-value: 0.0003383
```

由此得到的线性回归模型为 $p_k = -0.0040050k + 0.0607557$ 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).

```
m<-nls(p_k ~ exp(a+b*k)/(1+exp(a+b*k)),start=list(a=0.01,b=0.01),data=p_k)
summary(m)
```

```
##
## Formula: p_k ~ exp(a + b * k)/(1 + exp(a + b * k))
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a -2.62654    0.21115 -12.439 2.64e-09 ***
## b -0.13332    0.04507  -2.958  0.00977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02036 on 15 degrees of freedom
##
## Number of iterations to convergence: 12
## Achieved convergence tolerance: 4.089e-06
##   (因为不存在，4个观察量被删除了)
```

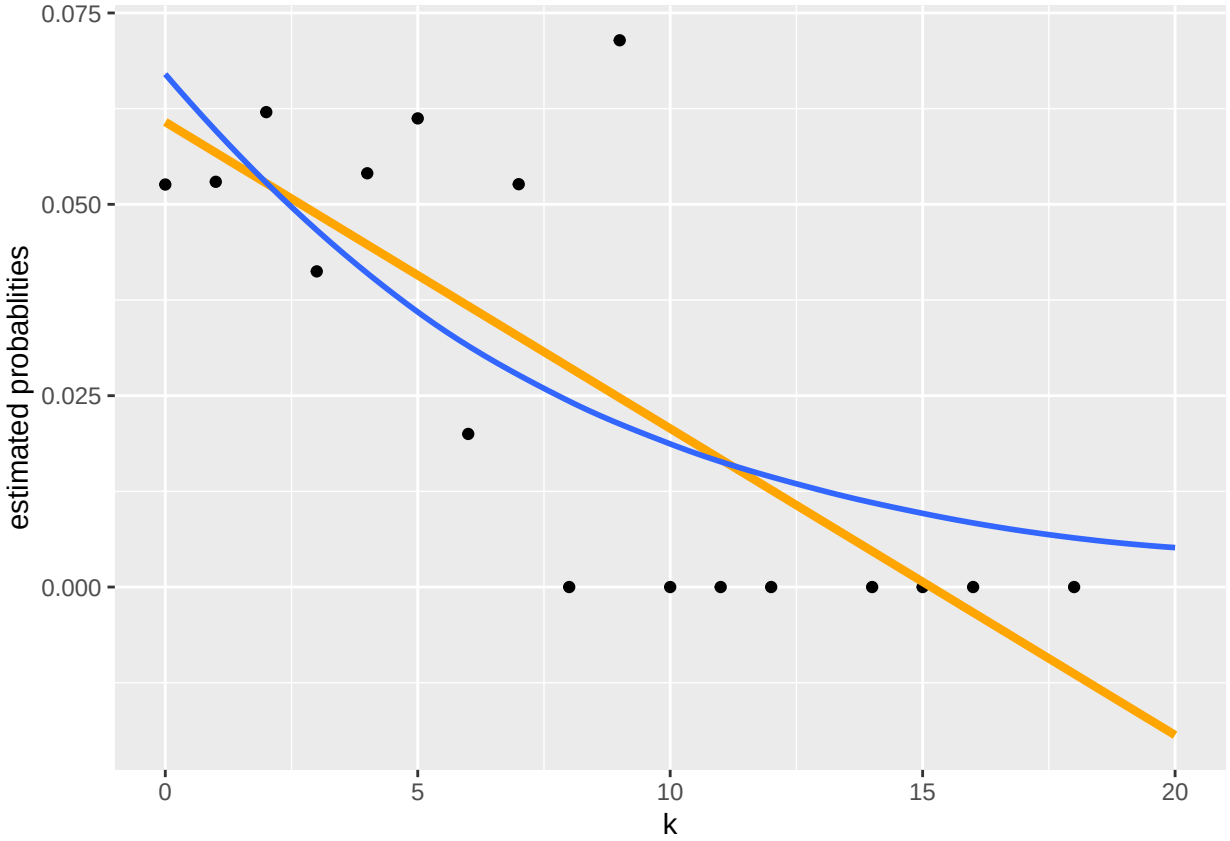
所以非线性回归的结果为 $p_k = e^{-2.62654-0.13332k}/\left(1+e^{-2.62654-0.13332k}\right)$ 该结果显示的是每增加一个用药，医生使用该药的可能性会下降。c. Plot the values from (3b) along with the estimated curves from (4a) and (4b). (You should have one plot, with k on the horizontal axis, and probabilities on the vertical axis .) Which model do you prefer, and why?

```
lm0<- function(x){
 y<- 0.0607557-0.0040050*x
 return(y)
```

```
}
nls0 <- function(x){
 y <- exp(-2.62654-0.13332*x)/(1+exp(-2.62654-0.13332*x))
 return(y)
}
ggplot(data=p_k)+geom_point(aes(x=k,y=p_k))+geom_line(aes(x=k,y=lm0(k)),color="orange",size=1.5)+geom
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



由图可知，非线性拟合模型的效果更好，更能反映数据的趋势。*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?

```
library(ggplot2)
n<-c(1:9)
p_k1<-p_k[-which((is.na(p_k))|(p_k$p_k==0)),]
wt<-p_k1[,1]*(1-p_k1[,1])/n
m.wt<-lm(p_k~k,data=p_k1,weight=1/wt)
summary(m.wt)
```

```
##
## Call:
## lm(formula = p_k ~ k, data = p_k1, weights = 1/wt)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46895  0.05269  0.09060  0.14280  0.28679
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0408263  0.0193426   2.111   0.0727 .
## k           0.0006647  0.0032846   0.202   0.8454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2331 on 7 degrees of freedom
## Multiple R-squared:  0.005816,   Adjusted R-squared:  -0.1362
## F-statistic: 0.04095 on 1 and 7 DF,  p-value: 0.8454
```

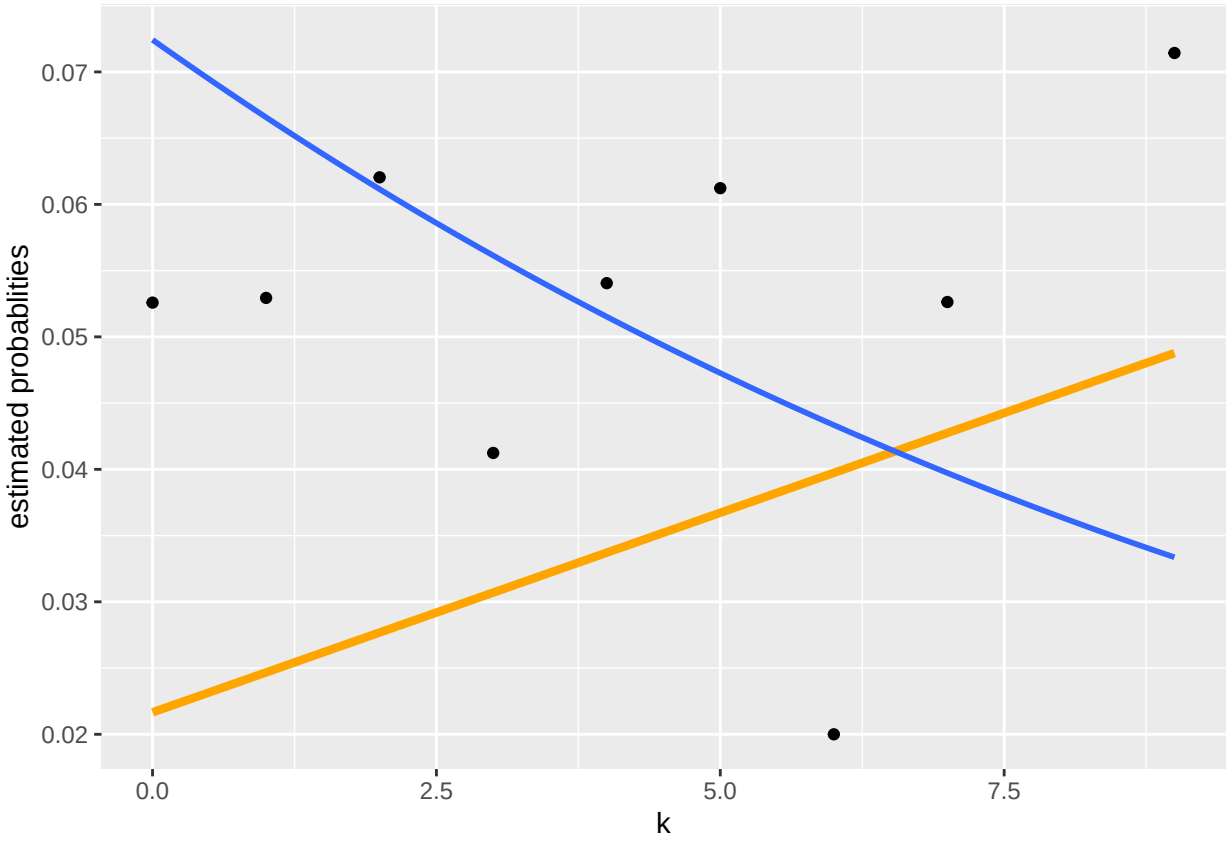
```
n.w2<-nls(p_k ~ exp(a+b*k)/(1+exp(a+b*k)),start=list(a=0.01,b=0.01),data=p_k1,weight=wt)
summary(n.w2)
```

```
##
## Formula: p_k ~ exp(a + b * k)/(1 + exp(a + b * k))
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a -2.89449    0.07652 -37.828 2.35e-09 ***
## b  0.01557    0.02137   0.729     0.49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001095 on 7 degrees of freedom
##
## Number of iterations to convergence: 6
```

```
## Achieved convergence tolerance: 4.19e-06
```

```
lm1 <- function(x){
 y<- 0.021658+0.003013*x
 return(y)
}
nls1 <- function(x){
 y <- exp(-2.55002-0.09070*x)/(1+exp(-2.55002-0.09070*x))
 return(y)
}
ggplot(data=p_k1)+geom_point(aes(x=k,y=p_k))+geom_line(aes(x=k,y=lm1(k)),color="orange",size=1.5)+geo
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



“‘