Homework 4: Diffusion of Tetracycline

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.

ckm\_nodes=read\_csv("data/ckm\_nodes.csv")

## Parsed with column specification:  
## cols(  
## city = col\_character(),  
## adoption\_date = col\_double(),  
## medical\_school = col\_character(),  
## attend\_meetings = col\_character(),  
## medical\_journals = col\_double(),  
## free\_time\_with = col\_character(),  
## discuss\_medicine\_socially = col\_character(),  
## club\_with\_drs = col\_character(),  
## drs\_among\_three\_best\_friends = col\_double(),  
## practicing\_here = col\_character(),  
## office\_visits\_per\_week = col\_character(),  
## proximity\_to\_other\_drs = col\_character(),  
## specialty = col\_character()  
## )

ckm\_network=read.table("data/ckm\_network.dat")  
ckm\_network=ckm\_network[which(!is.na(ckm\_nodes$adoption\_date)),which(!is.na(ckm\_nodes$adoption\_date))]

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.

ckm\_nodes=ckm\_nodes %>% filter(adoption\_date!='NA')

1. 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.

answer：因为有125个医生17个月，所以一共要125\*17=2125行；且医生和月份各占两列，然后题目中给出了4个数据类型，因此一共是6列

doctor=rep(1:125,each=17)  
month=rep(1:17,times=125)  
record\_month\_doctor=data.frame(doctor,month) %>%   
 mutate(wether\_began=(ckm\_nodes$adoption\_date==month)) %>%   
 mutate(wehter\_began\_before=(ckm\_nodes$adoption\_date<month)) %>%   
 group\_by(doctor,month) %>%   
 mutate(strict\_before=sum((ckm\_network[,doctor]==1)&(ckm\_nodes$adoption\_date<month))) %>%   
 mutate(before=sum((ckm\_network[,doctor]==1&(ckm\_nodes$adoption\_date<=month))))

1. Let

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

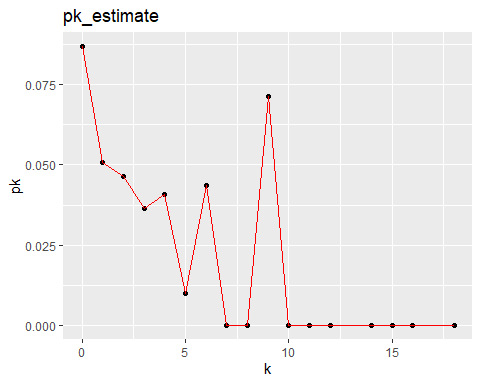
1. and

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

We suppose that and are the same for all months. a. Explain why there should be no more than values of for which we can estimate and directly from the data. answer:由2中的结果知，当月开始用药联系的医生最大值时18，不超过21

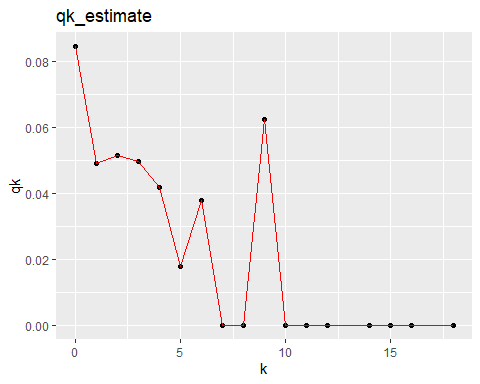
b. Create a vector of estimated $p\_k$ probabilities, using the data frame from (2). Plot the probabilities against the number of prior-adoptee contacts $k$.

k=unique(record\_month\_doctor$strict\_before)  
pk=vector(length = length(k))  
for(i in 1:length(k)){  
 cou=which(record\_month\_doctor$strict\_before==k[i])  
 pk[i]=sum(record\_month\_doctor$wether\_began[cou]==TRUE)/length(cou)  
}  
estimate\_pk=data.frame(k,pk)  
ggplot(data = estimate\_pk,aes(x=k,y=pk))+  
 geom\_point()+  
 geom\_line(col='red')+  
 labs(title = 'pk\_estimate')



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 $k$.

k=unique(record\_month\_doctor$strict\_before)  
qk=vector(length = length(k))  
for(i in 1:length(k)){  
 cou=which(record\_month\_doctor$before==k[i])  
 qk[i]=sum(record\_month\_doctor$wether\_began[cou]==TRUE)/length(cou)  
}  
estimate\_qk=data.frame(k,qk)  
ggplot(data = estimate\_pk,aes(x=k,y=qk))+  
 geom\_point()+  
 geom\_line(col='red')+  
 labs(title = 'qk\_estimate')



1. Because it only conditions on information from the previous month, is a little easier to interpret than . It is the probability per month that a doctor adopts tetracycline, if they have exactly contacts who had already adopted tetracycline.
   1. Suppose . 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.

pk\_model=lm(estimate\_pk$pk~estimate\_pk$k)  
summary(pk\_model)

##   
## Call:  
## lm(formula = estimate\_pk$pk ~ estimate\_pk$k)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.027632 -0.012410 -0.000246 0.006618 0.051407   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0542711 0.0094293 5.756 3.8e-05 \*\*\*  
## estimate\_pk$k -0.0038056 0.0009568 -3.977 0.00121 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.021 on 15 degrees of freedom  
## Multiple R-squared: 0.5133, Adjusted R-squared: 0.4808   
## F-statistic: 15.82 on 1 and 15 DF, p-value: 0.001214

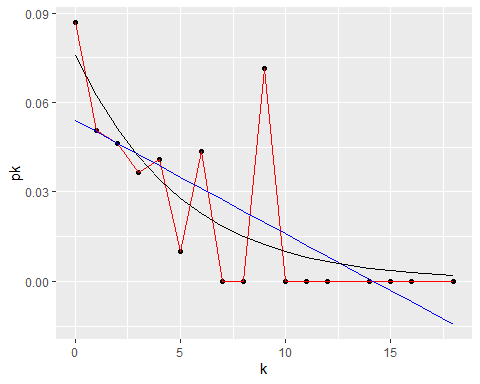
b. Suppose $p\_k = e^{a+bk}/(1+e^{a+bk})$. Explain, in words, what this model would imply about the impact of adding one more adoptee friend on a given doctor's probability of adoption. (You can suppose that $b > 0$, if that makes it easier.) Estimate the model by least squares, using the values you constructed in (3b).  
answer:当k增加一个时，pk的估计值会减小，但随着K的增大，减小的量越来越小

pk\_nl\_model=nls(estimate\_pk$pk~exp(a+b\*estimate\_pk$k)/(1+exp(a+b\*estimate\_pk$k)),start = c(a=0.05,b=0.05))  
summary(pk\_nl\_model)

##   
## Formula: estimate\_pk$pk ~ exp(a + b \* estimate\_pk$k)/(1 + exp(a + b \*   
## estimate\_pk$k))  
##   
## Parameters:  
## Estimate Std. Error t value Pr(>|t|)   
## a -2.50187 0.20346 -12.297 3.09e-09 \*\*\*  
## b -0.21749 0.06648 -3.272 0.00515 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01901 on 15 degrees of freedom  
##   
## Number of iterations to convergence: 8   
## Achieved convergence tolerance: 3.389e-06

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?  
answer:非线性的模型更加好，当K趋向于很大时，能够更好的拟合

a1=0.054  
b1=-0.0038  
a2=-2.5  
b2=-0.21  
ggplot(data = estimate\_pk,aes(x=k,y=pk))+  
 geom\_point()+  
 geom\_line(col='red')+  
 geom\_line(aes(x=k,y=a1+b1\*k),col='blue')+  
 geom\_line(aes(x=k,y=exp(a2+b2\*k)/(1+exp(a2+b2\*k))))



labs(title = 'pk\_estimate')

## $title  
## [1] "pk\_estimate"  
##   
## attr(,"class")  
## [1] "labels"

*For quibblers, pedants, and idle hands itching for work to do*: The values from problem 3 aren’t all equally precise, because they come from different numbers of observations. Also, if each doctor with adoptee contacts is independently deciding whether or not to adopt with probability , then the variance in the number of adoptees will depend on . Say that the actual proportion who decide to adopt is . A little probability (exercise!) shows that in this situation, , but that , where 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 . Find the , and then re-do the estimation in (4a) and (4b) where the squared error for is divided by . How much do the parameter estimates change? How much do the plotted curves in (4c) change? answer:结果和之前的（4）中的一模一样

k=unique(record\_month\_doctor$strict\_before)  
var\_pk=vector(length = length(k))  
for(i in 1:length(k)){  
 cou=which(record\_month\_doctor$strict\_before==k[i])  
 var\_pk[i]=var(record\_month\_doctor$wether\_began[cou])  
}  
var\_pk[14]=0  
pk\_new=(1-sqrt(1-4\*var\_pk))/2  
estimate\_pk\_new=data.frame(k,pk\_new)  
pk\_new\_model=lm(estimate\_pk\_new$pk~estimate\_pk\_new$k)  
summary(pk\_new\_model)

##   
## Call:  
## lm(formula = estimate\_pk\_new$pk ~ estimate\_pk\_new$k)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.028081 -0.012839 -0.000618 0.006214 0.056963   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.054756 0.009897 5.533 5.74e-05 \*\*\*  
## estimate\_pk\_new$k -0.003811 0.001004 -3.794 0.00176 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02204 on 15 degrees of freedom  
## Multiple R-squared: 0.4897, Adjusted R-squared: 0.4557   
## F-statistic: 14.4 on 1 and 15 DF, p-value: 0.001763

pk\_new\_nl\_model=nls(estimate\_pk\_new$pk~exp(a+b\*estimate\_pk\_new$k)/(1+exp(a+b\*estimate\_pk\_new$k)),start = c(a=0.05,b=0.05))  
summary(pk\_new\_nl\_model)

##   
## Formula: estimate\_pk\_new$pk ~ exp(a + b \* estimate\_pk\_new$k)/(1 + exp(a +   
## b \* estimate\_pk\_new$k))  
##   
## Parameters:  
## Estimate Std. Error t value Pr(>|t|)   
## a -2.50727 0.21661 -11.575 7.06e-09 \*\*\*  
## b -0.21074 0.06862 -3.071 0.00776 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02033 on 15 degrees of freedom  
##   
## Number of iterations to convergence: 8   
## Achieved convergence tolerance: 2.912e-06

a1\_new=0.054  
b1\_new=-0.0038  
a2\_new=-2.50  
b2\_new=-0.21  
ggplot(data = estimate\_pk\_new,aes(x=k,y=pk\_new))+  
 geom\_point()+  
 geom\_line(col='red')+  
 geom\_line(aes(x=k,y=a1+b1\*k),col='blue')+  
 geom\_line(aes(x=k,y=exp(a2+b2\*k)/(1+exp(a2+b2\*k))))+  
 geom\_line(aes(x=k,y=a1\_new+b1\_new\*k),col='green')+  
 geom\_line(aes(x=k,y=exp(a2\_new+b2\_new\*k)/(1+exp(a2\_new+b2\_new\*k))),col='yellow')+  
 labs(title = 'pk\_estimate')

