Due Monday, November 6, 5:00 PM

General instructions for labs: You are encouraged to work in pairs to complete the lab. Labs must be completed as an R Markdown file. Be sure to include your lab partner (if you have one) and your own name in the file. Give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used.

Agenda: Fitting models by optimization; transforming data from one representation to another; handling missing data

Many theories of the diffusion of innovations (new technologies, practices, beliefs, etc.) suggest that the fraction of members of a group who have adopted the innovation by time t, $p(t)$, should follow a logistic curve or logistic function,

$$p(t) = \frac{e^{b(t-t_0)}}{1 + e^{b(t-t_0)}}.$$

We will look at a classic data set on the diffusion of innovations, which is supposed to show such a curve. It concerns a survey of 246 doctors in four towns in Illinois in the early 1950s, and when they began prescribing (adopted) a then-new antibiotic, tetracycline, and how they became convinced that they should do so (from medical journals, from colleagues, etc.).

Load the file [http://faculty.ucr.edu/~jflegal/206/ckm_nodes.csv]. Each row is a doctor. The column adoption date shows how many months, after it became available, each doctor began prescribing tetracycline. Doctors who had not done so by the end of the survey, i.e., after month 17, have a value of `Inf` in this column. This information is not available (`NA`) for some doctors. There are twelve other variables which may also be `NA`.

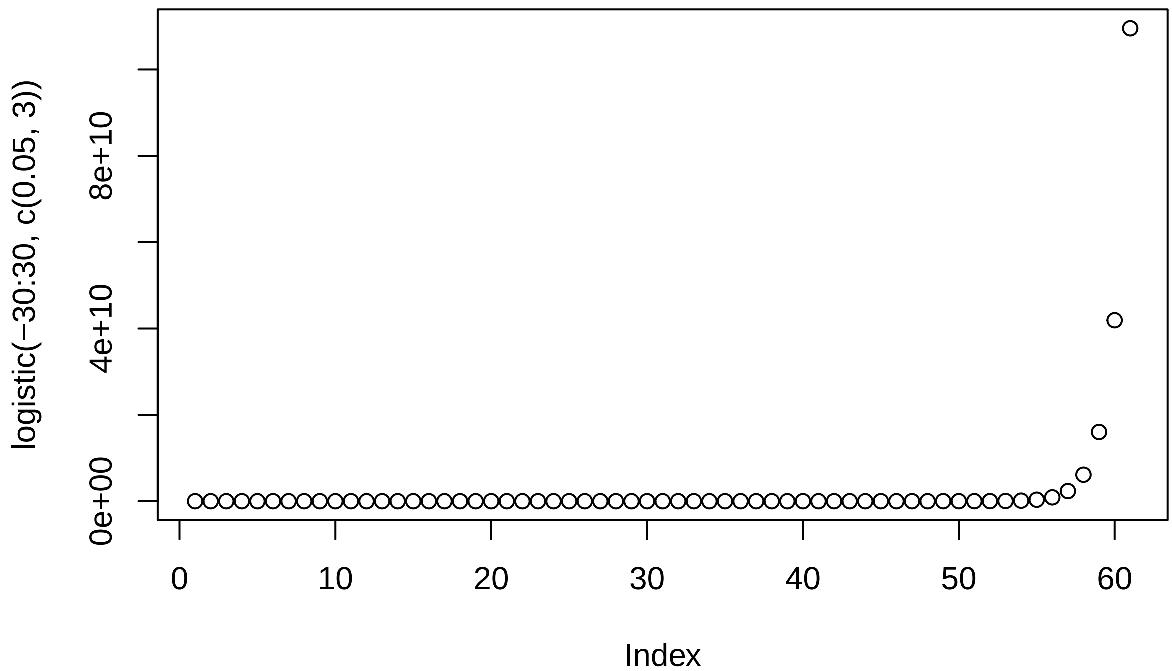
```
docs = read.csv("http://faculty.ucr.edu/~jflegal/206/ckm_nodes.csv", header = TRUE)
head(docs)
```

```
##     city adoption_date medical_school attend_meetings medical_journals
## 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
##   free_time_with discuss_medicine_socially club_with_drs
## 1    non-doctors                        no            no
## 2        doctors                       yes            no
## 3        doctors                        no            no
## 4    non-doctors                        no            no
## 5    non-doctors                       yes            no
## 6          split                       yes            no
##   drs_among_three_best_friends practicing_here office_visits_per_week
## 1                            0       20+ years                101--150
## 2                            3         1- year                 76--100
## 3                            2    10--20 years                 76--100
## 4                            0     5--10 years                  51--75
## 5                            1    10--20 years                  51--75
## 6                            0    10--20 years                101--150
##       proximity_to_other_drs    specialty
## 1     in_building_and_office pediatrician
## 2     in_building_and_office           GP
## 3 in_building_but_not_office    internist
## 4     in_building_and_office           GP
## 5 in_building_but_not_office           GP
## 6     in_building_and_office    internist
```

```
docs = subset(docs, docs$adoption_date <= 17 , select = TRUE)
#docs
```

1. The Model.
   a. Write a function, `logistic`, which calculates the logistic function. It should take two arguments, `t` and `theta`. The `theta` argument should be a vector of length two, the first component being the parameter b and the second component being t_0. Your function may not use any loops. Plot the curve of the logistic function with $b = 0.05$, $t_0 = 3$, from $t = -30$ to $t = 30$.

```
theta = c()
logistic = function(t, theta) {
  answer = exp(t - theta[2]) / (1 + exp(theta[1]*(t - theta[2])))
  return(answer)
}

plot(logistic(-30:30, c(.05, 3)))
```




```
b. Explain why $p(t_0)=0.5$, no matter what $b$ is. Use this to check your logistic function at multipl
```

```
#When we use p(t_0) our terms in the top and bottom turn into e^(0) / (1  + e^0)
#this always will turn out to be 1/2 which is 0.5
#essentially we are saying that t and t_0 are the same

logistic(.5, c(10, .5))
```

```
## [1] 0.5
```

```r
logistic(11, c(-10, 11))
```

```
## [1] 0.5
```

```r
logistic(-22, c(30, -22))
```

```
## [1] 0.5
```

```r
logistic(500, c(1000, 500))
```

```
## [1] 0.5
```

```
c. Explain why the slope of $p(t)$ at $t=t_0$ is $b/4$. (Hint: calculus.) Use this to check your `logis
```

```r
#When t=t_0 we have the same situation where the logistic function will return
#the derivative of the function will give the slop at a point. when t = t_0
#the derivative of the function is e^x / (e^x + 1)^2
#ultimately we have e^0 / (e^0 + 1)^2 -> 1/4. This means whenever p(t) = 1/2 then the slop = 4

logistic(.5, c(11, .5))
```

```
## [1] 0.5
```

```r
logistic(11, c(-22, 11))
```

```
## [1] 0.5
```

```r
logistic(-22, c(30, -22))
```

```
## [1] 0.5
```

```r
logistic(500, c(998877, 500))
```

```
## [1] 0.5
```

2. The Data.
   a. How many doctors in the survey had adopted tetracycline by month 5? Hint: Use `na.omit` carefully.

```r
#The first step here was to omit the na in the column for adoption date
#this puts the data from that column in a numeric vector
#then count how many items match 5 or under
#there are 51 matches for thiw

clean = na.omit(docs$adoption_date)
month_5_count = sum(clean <= 5)
month_5_count
```

```
## [1] 51
```

```
b. What proportion of doctors, for whom adoption dates are available, had adopted tetracycline by month
```

```r
#We know that there are 51 doctors who adapted by month 5, now we need to know how many doctors
#there are without na or inf in the adoption date column


infin = docs$adoption_date == "Inf"
infz = sum(infin, na.rm = TRUE)
naz = sum(is.na(docs$adoption_date))
totz = nrow(docs)
denom = totz - infz - naz
```

```
answer = 51 / (denom)
answer
```

```
## [1] 0.4678899
```

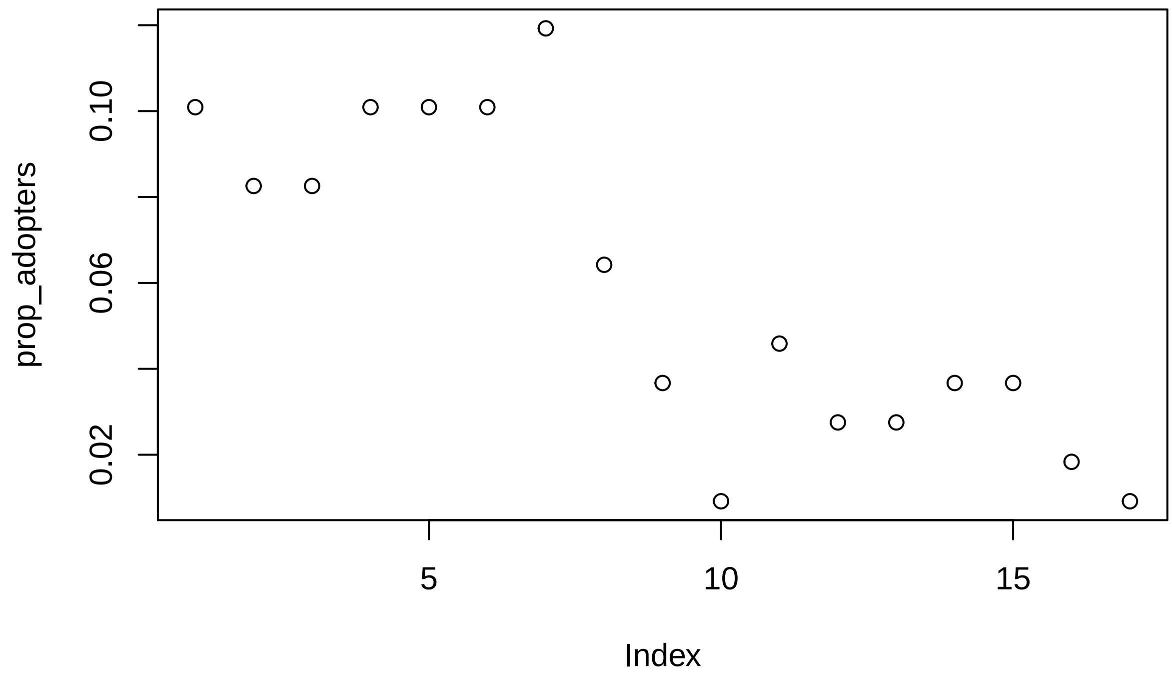
c. Create a vector, `prop_adopters`, storing the proportion of doctors who have adopted by each month.

```
prop_adopters = c()

prop_month = function(month) {

  numer = sum(clean == month)
  answer = numer / denom
  #prop_adopters = c(prop_adopters, answer)
  #return(prop_adopters)

}

for(x in 1:17){
  prop_adopters = c(prop_adopters, prop_month(x))
prop_adopters <<- prop_adopters
}

prop_adopters
```

```
##  [1] 0.100917431 0.082568807 0.082568807 0.100917431 0.100917431
##  [6] 0.100917431 0.119266055 0.064220183 0.036697248 0.009174312
## [11] 0.045871560 0.027522936 0.027522936 0.036697248 0.036697248
## [16] 0.018348624 0.009174312
```

d. Make a scatter-plot of the proportion of adopters over time.

```
#from the plot it looks like many doctors adopted early and as time went the last few joined as well

plot(prop_adopters)
```




```
e. Make rough guesses about $t_0$ and $b$ from the plot, and from your answers in problem 1.
```

```
#after substituting a lot of numbers in for b and t_0 in the first plot in the lab assignment
#I think a good rough guesses are: b = 1, t_0 = 8
```

3. The Fit.
   a. Write a function, `logistic_mse`, which calculates the mean squared error of the logistic model on this data set. It should take a single vector, `theta`, and return a single number. This function cannot contain any loops, and must use your `logistic` function.

```
#For MSE we must get the squared deviation from out actual data to the logistic curve depending
#on which theta we use / number of rows

#Note: when i use .05, 3 for theta the mse is huge, when i use my guess of 1,8 it is much smaller

logistic_mse = function(x) {
  return(mean((prop_adopters - logistic(1:17, c(x[1], x[2]))^2)))
}

logistic_mse(c(1, 8))
```

```
## [1] -0.4411849
```

```
b. Use `optim` to minimize `logistic_mse`, starting from your rough guess in problem 2e. Report the loca
```

```
#I am naming the object fit to hold the optimization results
#in this case the location is -359.8, -337.9
#and the
```

```r
fit = optim(par = c(1,8), fn = logistic_mse)
fit
```

```
## $par
## [1] -359.7578 -337.8914
##
## $value
## [1] -1.222979e+307
##
## $counts
## function gradient
##      307       NA
##
## $convergence
## [1] 10
##
## $message
## NULL
```

c. Add a curve of the fitted logistic function to your scatterplot from Problem 2d. Does it seem like a

```r
#the function parameters I have found do not match the data, I believe the error is in my optim functio
#but i'm having a hard time figuring out what i did wrong

plot(prop_adopters)
lines(logistic(-30:30, c(fit$par[1],fit$par[2])))
```