Assignment4\_stat359

Koki Itagaki

2023-03-20

#Q1. In a study examining smoking and lung cancer, a random sample of men #between the ages of 55 and 60 was obtained. The smoking and disease status of #each sampled subject was ascertained. For each subject, a ’1’ is assigned if #the subject had lung cancer (case) and a ’0’ if not. Similarly, a ’1’ #indicates that a subject is a smoker and a ’0’ indicates a nonsmoker. #The data are found in the Excel file ‘LungCancer’.

#• Read the data into R, and use table() function to produce a contingency #table summarizing these data.

LungCancer<-read.csv(file='~/Desktop/stat359/data/LungCancer.csv',header = TRUE)  
observed<-table(LungCancer)  
observed

## Smoker  
## Case 0 1  
## 0 60 650  
## 1 22 687

#• Assuming that there is no association between smoking and lung cancer,  
#compute a table of ‘expected’ counts.  
  
expected <- round(chisq.test(observed)$expected,2)  
expected

## Smoker  
## Case 0 1  
## 0 41.03 668.97  
## 1 40.97 668.03

#I also did this by hand.  
#• By hand, compute the observed value of the test statistic for testing  
#association between lung cancer and smoking.  
#The result is in a different file  
  
#• Assuming there is no association, what is the distribution of   
#the test statistic?  
  
#If there is no association,the distribution of   
#the test statistic follows a chi-squared distribution with degrees  
#of freedom equal to (r-1)(c-1) = (2-1)(2-1) = 1  
  
#• Using R, compute the p-value for a test of association, and give a   
#detailed conclusion based on the p-value and a comparison of the tables   
#observed and expected counts.  
1-pchisq(18.63,df = 1)

## [1] 1.587034e-05

#Since p-value = 1.587034e-05 << a = 0 .05, we reject Ho. There is a significant   
#evidence that there is an association between Smokers and Cases of lung cancer.  
#Also from 2 tables I created above, you can see that the number of cases of the  
#lung cancers whose patients are smokers is at least 10 times higher than the  
#number of cases of the lung cancers whose patients are non-smokers.  
#Therefore the number of cases of the lung cancers is associated with smoking.

#2. The following data are from a study examining the incidence of tuberculosis #in relation to blood groups in a sample of Eskimos. It is of interest to #determine if there is any association between the disease and blood group #within the ABO system. #Severity O A AB B #Moderate-advanced 7 7 7 13 #Minimal 27 34 12 18 #Not Present 55 52 11 24 #• Assuming that there is no association between disease and blood group, #compute a table of ‘expected’ counts.

data <- c(7,7,7,13,27,34,12,18,55,52,11,24)  
  
data <- c(7, 7, 7, 13, 27, 34, 12, 18, 55, 52, 11, 24)  
mat <- matrix(data, nrow = 3, ncol = 4, byrow = TRUE)  
  
# Calculate the expected counts  
row\_totals <- rowSums(mat)  
col\_totals <- colSums(mat)  
grand\_total <- sum(mat)  
e\_row <- matrix(row\_totals, nrow = nrow(mat), ncol = ncol(mat), byrow = TRUE)  
e\_col <- matrix(col\_totals, nrow = nrow(mat), ncol = ncol(mat), byrow = FALSE)  
mat\_expected <- e\_row \* e\_col / grand\_total  
  
print(mat\_expected)

## [,1] [,2] [,3] [,4]  
## [1,] 11.33333 18.74532 15.955056 11.84270  
## [2,] 31.69663 47.33333 7.003745 10.22472  
## [3,] 15.95506 11.84270 30.333333 29.25094

#• By hand, compute the observed value of the test statistic for testing association between  
#disease and blood group.  
  
#• Assuming there is no association, what is the distribution of the test statistic?  
#If there is no association between the disease and blood group,the distribution of   
#the test statistic follows a chi-squared distribution with degrees  
#of freedom equal to (r-1)(c-1) = (3-1)(4-1) = 6  
#• Using R, compute the p-value for a test of association, and give a detailed conclusion  
#based on the p-value and a comparison of the tables observed and expected counts.  
  
1-pchisq(q = 16.1427,df=6)

## [1] 0.01300819

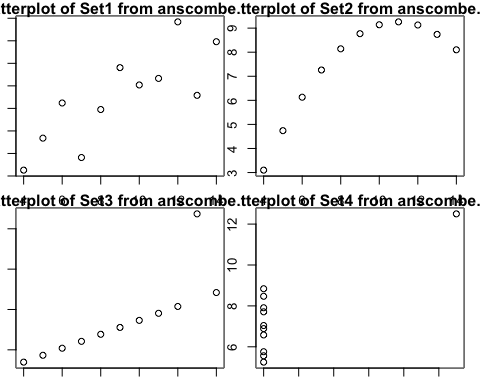
#Since p-value = 0.01300819 < a = 0 .05, we reject Ho. There is a significant   
#evidence that there is an association between disease and blood group.  
#Also from 2 tables,When I see the row Minimal, The number of cases, which severty  
# is Minimal, of people whose blood types arebO and A have almost two time as much as   
#people whose blood type is AB or B.

#3. The file ‘Anscombe’ contains 4 diferent datasets, each of which are based on a response Y, #and a covariate X.

ancombe<-read.csv('~/Desktop/stat359/data/anscombe.csv',header=TRUE)  
  
  
  
#(a) Produce 4 scatter plots (one for each dataset), on the same page,   
#illustrating the relationship between Y and X. Describe each of these briefly,   
#and state if you think a linear  
#model of the form yi = a + bxi + ≤i would be appropriate.  
#4  
ancombe[45:55,2]

## [1] "6.58" "5.76" "7.71" "8.84" "8.47" "7.04" "5.25" "12.5" "5.56" "7.91"  
## [11] "6.89"

par(mar = c(1,1,1,1))  
par(mfrow = c(2,2))  
plot(ancombe[2:12,1],ancombe[2:12,2], xlab = "The value of x",  
 ylab = "The value of y", main = "Scatterplot of Set1 from anscombe.csv",  
 sub = "Written by Koki Itagaki")  
plot(ancombe[16:26,1],ancombe[16:26,2], xlab = "The value of x",  
 ylab = "The value of y", main = "Scatterplot of Set2 from anscombe.csv",  
 sub = "Written by Koki Itagaki")  
plot(ancombe[30:40,1],ancombe[30:40,2], xlab = "The value of x",  
 ylab = "The value of y", main = "Scatterplot of Set3 from anscombe.csv",  
 sub = "Written by Koki Itagaki")  
plot(ancombe[45:55,1],ancombe[45:55,2], xlab = "The value of x",  
 ylab = "The value of y", main = "Scatterplot of Set4 from anscombe.csv",  
 sub = "Written by Koki Itagaki")



#According to the graphs, we can see the different trend of the data.  
#Graph 1 shows that there is a positive linear relationships between x and y.  
#However, graph 2 is a quadric equation.  
#The data from graph 3 also have a positive relationships between x and y with  
#a outlier at arount x = 13 and y = 14.  
#There is not a linear relationships in graph 4.The x value of all data is 8   
#except one outlier. it means it does not show any correlation between x and y.  
#Therefore, in my opinion,Graph1 and Graph 3 can be shown as yi = a + bxi + error  
  
  
#(b) Perform 4 separate simple linear regressions (one for each dataset)and   
#produce a table (in your text editor (ie. word)) that shows the R2 value.   
#Discuss what is happening here (hint: for simple linear regression, R2 is just  
#the square of the sample correlation coe±cient).  
  
set1\_x<-ancombe[2:12,1]  
set1\_x<-as.numeric(set1\_x)  
set1\_y<-ancombe[2:12,2]  
set1\_y<-as.numeric(set1\_y)  
re\_set1<-lm(set1\_x~set1\_y)  
re\_set1

##   
## Call:  
## lm(formula = set1\_x ~ set1\_y)  
##   
## Coefficients:  
## (Intercept) set1\_y   
## -0.9975 1.3328

set2\_x<-ancombe[16:26,1]  
set2\_x<-as.numeric(set2\_x)  
set2\_y<-ancombe[16:26,2]  
set2\_y<-as.numeric(set2\_y)  
re\_set2<-lm(set2\_x~set2\_y)  
re\_set2

##   
## Call:  
## lm(formula = set2\_x ~ set2\_y)  
##   
## Coefficients:  
## (Intercept) set2\_y   
## -0.9948 1.3325

set3\_x<-ancombe[30:40,1]  
set3\_x<-as.numeric(set3\_x)  
set3\_y<-ancombe[30:40,2]  
set3\_y<-as.numeric(set3\_y)  
re\_set3<-lm(set3\_x~set3\_y)  
re\_set3

##   
## Call:  
## lm(formula = set3\_x ~ set3\_y)  
##   
## Coefficients:  
## (Intercept) set3\_y   
## -1.000 1.333

set4\_x<-ancombe[45:55,1]  
set4\_x<-as.numeric(set4\_x)  
set4\_y<-ancombe[45:55,2]  
set4\_y<-as.numeric(set4\_y)  
re\_set4<-lm(set4\_x~set4\_y)  
re\_set4

##   
## Call:  
## lm(formula = set4\_x ~ set4\_y)  
##   
## Coefficients:  
## (Intercept) set4\_y   
## -1.004 1.334

summary(re\_set1)

##   
## Call:  
## lm(formula = set1\_x ~ set1\_y)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6522 -1.5117 -0.2657 1.2341 3.8946   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.9975 2.4344 -0.410 0.69156   
## set1\_y 1.3328 0.3142 4.241 0.00217 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.019 on 9 degrees of freedom  
## Multiple R-squared: 0.6665, Adjusted R-squared: 0.6295   
## F-statistic: 17.99 on 1 and 9 DF, p-value: 0.00217

summary(re\_set2)

##   
## Call:  
## lm(formula = set2\_x ~ set2\_y)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.8516 -1.4315 -0.3440 0.8467 4.2017   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.9948 2.4354 -0.408 0.69246   
## set2\_y 1.3325 0.3144 4.239 0.00218 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.02 on 9 degrees of freedom  
## Multiple R-squared: 0.6662, Adjusted R-squared: 0.6292   
## F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002179

summary(re\_set3)

##   
## Call:  
## lm(formula = set3\_x ~ set3\_y)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9869 -1.3733 -0.0266 1.3200 3.2133   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.0003 2.4362 -0.411 0.69097   
## set3\_y 1.3334 0.3145 4.239 0.00218 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.019 on 9 degrees of freedom  
## Multiple R-squared: 0.6663, Adjusted R-squared: 0.6292   
## F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002176

summary(re\_set4)

##   
## Call:  
## lm(formula = set4\_x ~ set4\_y)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7859 -1.4122 -0.1853 1.4551 3.3329   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.0036 2.4349 -0.412 0.68985   
## set4\_y 1.3337 0.3143 4.243 0.00216 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.018 on 9 degrees of freedom  
## Multiple R-squared: 0.6667, Adjusted R-squared: 0.6297   
## F-statistic: 18 on 1 and 9 DF, p-value: 0.002165

#Dataset R^2  
# 1. 0.67  
# 2. 0.67  
# 3 0.67  
# 4. 0.67  
  
#The Rs of all datasets is the same even though the shapes of 4 graphs are  
#totally different. For examle, the first graph shows that there is a linear   
#moderate relationship and the third graph shows that there is a strong linear  
#relationships with only one outlier. This means if the data set has at least   
#one outlier, the correlation rate between x and y changes considerably.  
#Moreover, the data set 4 shows that quadratic curve and the correlation rate is  
#also the same as the dataset 1 and 3.

#4. The file ‘growth’ gives data on the height of a white spruce #tree measured annually for 50 years. Letting Yt denote the height of the #tree at year t > 0, we consider describing the growth of the tree over time #with a non-linear model Yt = f(t) + ≤t, ≤t iidª N(0;, æ2). Three growth #curves are considered for f(t) #(a) Logistic: f(t) = a/(1 + b § exp{°ct}) #(b) Gompertz: f(t) = a exp{°b exp{°ct}} #(c) Von BertalanÆy: f(t) = a ° a exp{°b(t + c)} #• Fit all three models using the non-linear least squares function nls() in R. #Explain how you are choosing the starting values for nls() in each case. #Produce a figure depicting the estimated curves all on the same plot, along #with the observed data. Be sure to include a legend to distinguish the #diÆerent curves.

data <- read.table(file='~/Desktop/stat359/data/growth.txt',header=TRUE)  
y <- data$height  
t <- data$t  
a.start <- max(y)  
  
# Logistic  
b.start <- a.start/(min(y))  
c.start <- -log((a.start-mean(y))/(b.start\*mean(y)))/mean(t)  
  
# Fit logistic model using nls()  
logistic <- nls(y ~ a/(1+b\*exp(-c\*t)),   
 start=list(a=a.start, b=b.start, c=c.start),   
 trace=TRUE)

## 2180.554 (1.84e+00): par = (53.95014 269.7507 0.2328016)  
## 846.6886 (9.29e-01): par = (53.10289 109.7646 0.2177264)  
## 495.3645 (3.38e-01): par = (51.69629 64.09383 0.2088176)  
## 445.9083 (8.97e-02): par = (50.37352 47.11803 0.2013761)  
## 442.5841 (1.02e-02): par = (50.45075 46.50495 0.1984343)  
## 442.5424 (1.38e-03): par = (50.41649 47.14147 0.1993367)  
## 442.5417 (1.68e-04): par = (50.42135 47.12783 0.1992812)  
## 442.5417 (2.05e-05): par = (50.42068 47.13808 0.1992959)  
## 442.5417 (2.50e-06): par = (50.42076 47.13769 0.1992949)

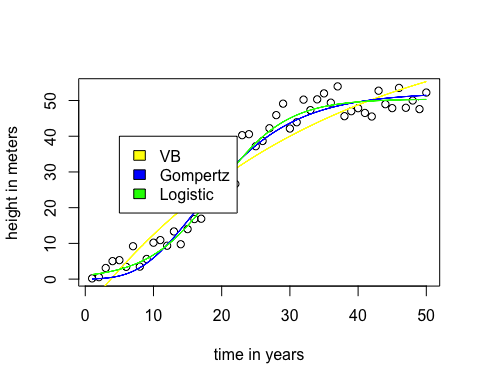
# Gompertz  
b.start <- -log(min(t)/a.start)  
c.start <- -log(-log(mean(y)/a.start)/b.start)/mean(t)  
  
# Fit gompertz model using nls()  
gompertz <- nls(y ~ a\*exp(-b\*exp(-c\*t)),   
 start=list(a=a.start, b=b.start, c=c.start),   
 trace=TRUE)

## 1517.220 (1.21e+00): par = (53.95014 3.98806 0.07855924)  
## 971.6970 (7.44e-01): par = (48.08192 5.698813 0.1185395)  
## 626.9119 (1.35e-01): par = (52.89492 6.410667 0.1143171)  
## 616.5275 (4.82e-02): par = (52.14363 7.238899 0.1233726)  
## 614.9412 (1.27e-02): par = (52.29073 7.417474 0.1237323)  
## 614.8192 (3.98e-03): par = (52.22642 7.534484 0.1246399)  
## 614.8067 (1.22e-03): par = (52.22317 7.563496 0.1248037)  
## 614.8055 (3.79e-04): par = (52.21911 7.574479 0.1248795)  
## 614.8054 (1.18e-04): par = (52.21833 7.577643 0.1248995)  
## 614.8054 (3.67e-05): par = (52.21801 7.578676 0.1249064)  
## 614.8054 (1.14e-05): par = (52.21792 7.57899 0.1249084)  
## 614.8054 (3.54e-06): par = (52.21789 7.579089 0.1249091)

# VB  
b.start <- -log((mean(y)-a.start)/(min(y)-a.start))/mean(t)  
c.start <- -log((a.start-min(y))/a.start)/b.start  
  
# Fit VB model using nls()  
vb <- nls(y ~ a\*(1-exp(-b\*(t+c))),   
 start=list(a=a.start, b=b.start, c=c.start),   
 trace=TRUE)

## 3380.398 (1.23e+00): par = (53.95014 0.03424298 0.1084606)  
## 2144.571 (7.68e-01): par = (75.10794 0.02491852 -4.467943)  
## 1384.782 (1.55e-01): par = (70.49888 0.03103404 -3.733227)  
## 1354.408 (3.20e-02): par = (74.91996 0.02883085 -3.728334)  
## 1353.121 (5.14e-03): par = (74.15239 0.0295587 -3.772171)  
## 1353.094 (1.40e-03): par = (74.48107 0.02933935 -3.757477)  
## 1353.093 (4.14e-04): par = (74.39064 0.02940495 -3.761695)  
## 1353.092 (1.23e-04): par = (74.41844 0.02938538 -3.760418)  
## 1353.092 (3.68e-05): par = (74.41021 0.02939122 -3.760798)  
## 1353.092 (1.10e-05): par = (74.41267 0.02938948 -3.760684)  
## 1353.092 (3.32e-06): par = (74.41194 0.02939 -3.760718)

#I basically find the unknown parameters:a,b,and c for each cases and pass the   
#values to the nls function to calculate   
  
  
# Define a function to generate predictions from a model  
predict\_model <- function(model, t\_values) {  
 predicted\_values <- predict(model, list(t = t\_values))  
 return(predicted\_values)  
}  
  
plot(t, y, xlab = "time in years", ylab = "height in meters")  
# predict values for different models  
t\_seq <- seq(min(t), max(t), 0.01)  
H.vb <- predict(vb, list(t = t\_seq))  
H.gompertz <- predict(gompertz, list(t = t\_seq))  
H.logistic <- predict(logistic, list(t = t\_seq))  
  
# add lines to the plot for each model  
lines(t\_seq, H.vb, col = "yellow")  
lines(t\_seq, H.gompertz, col = "blue")  
lines(t\_seq, H.logistic, col = "green")  
  
# add a legend to the plot  
legend(x = 5, y = 40, legend = c("VB", "Gompertz", "Logistic"), fill = c("yellow", "blue", "green"))



#• For each of the three models, give a 95% confidence interval for limt!1f(t).  
#What doest his represent?  
z<-1.96  
paste("95% CI for the first model is: ",74.411938 -z\*9.950934 , 74.411938 +z\*9.950934,  
 "95% CI for the second model is: ",50.4208-z\*0.8473 , 50.4208 +z\*0.8473,  
 "95% CI for the third model is: ",52.21789 -z\*1.33361 , 52.21789 +z\*1.33361 )

## [1] "95% CI for the first model is: 54.90810736 93.91576864 95% CI for the second model is: 48.760092 52.081508 95% CI for the third model is: 49.6040144 54.8317656"

t.plot <- seq(min(t), max(t), 0.01)  
  
  
  
#• Select the best of the three models, and plot an estimate of the  
#derivative df(t) dt , which represents the rate of growth over time.  
  
# define variables a, b, and c  
a <- 50.4208; b <- 47.1377; c <- 0.1993  
  
# calculate the derivative of Y with respect to time  
deriva <- a\*b\*c\*exp(-c \* t.plot) / ((1 + b \* exp(-c \* t.plot))^2)  
  
plot(deriva,main = "estimate of the derivative")

