ASSN 3 – Sketch Solutions

Q1:

> lung<-read.csv("LungCancer.csv", header=T)

> table(lung)

Smoker

Case 0 1

0 60 650

1 22 687

The table of expected counts is obtained as the product of marginal row and columns totals divided by the overall total:

Smoker

Case 0 1

0 41 669

1 41 668

The observed value of the test stat is obtained as chisq = 18.63

Under the null the sampling distribution of the test statistic is chi-squared with DOF = (r-1)(c-1) = 1

Q2:

**Use R to compute the table of expected counts:**

> data <- c(7,7,7,13,27,34,12,18,55,52,11,24)

> A<-matrix(data,nrow=3,ncol=4,byrow=TRUE)

> A.expected <- A

> for (i in 1:nrow(A))

+ {

+ for (j in 1:ncol(A))

+ {

+ A.expected[i,j] <- rowSums(A)[i]\*colSums(A)[j]/sum(A)

+ }

+ }

>

> A.expected

[,1] [,2] [,3] [,4]

[1,] 11.33333 11.84270 3.820225 7.003745

[2,] 30.33333 31.69663 10.224719 18.745318

[3,] 47.33333 49.46067 15.955056 29.250936

**The Observed value of the test statistic is:**

> sum((A-A.expected)^2/A.expected)

[1] 16.14270

Under the null the sampling distributions of the test statistics is chi-squared with DOF=6

> 1-pchisq(16.1427,df=6)

[1] 0.01300819

Computing the p-value with respect to the above distribution we obtain a p-value = 0.01301, suggesting some evidence against the null hypothesis of no association between blood type and disease severity. Testing at a level alpha=0.05 we reject H0.

Q3:

Anscombe’s Data

anscombe<-read.csv(“anscombe.csv”, header=T)

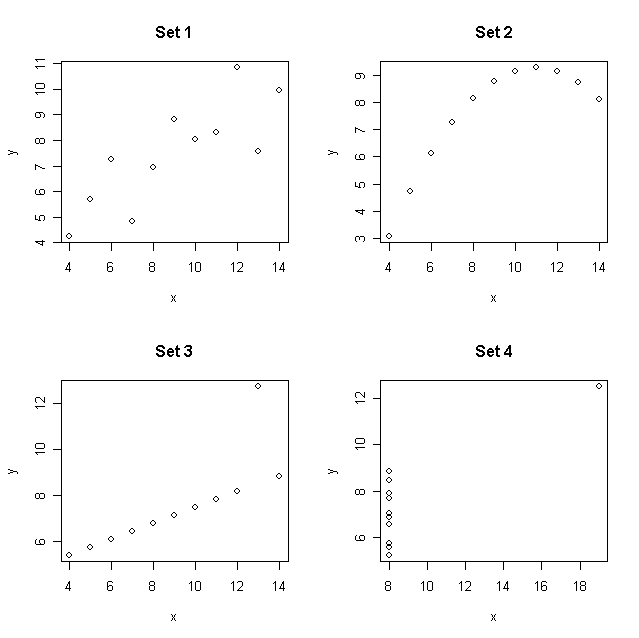
par(mfrow=c(2,2))

plot(anscombe[,1],anscombe[,2],main=”Set 1”,xlab=”x”, ylab=”y”)

plot(anscombe[,3],anscombe[,4] ,main=”Set 2” ,xlab=”x”, ylab=”y”)

plot(anscombe[,5],anscombe[,6] ,main=”Set 3” ,xlab=”x”, ylab=”y”)

plot(anscombe[,7],anscombe[,8] ,main=”Set 4” ,xlab=”x”, ylab=”y”)



a) Set 1 looks like a typical linear relationship between x and y. Set 2 looks like it should be modeled with a quadratic equation, with little or no variability, but is clearly not a linear relationship. Set 3 looks linear, but with an outlier. Set 4 does not look like a linear relationship but has an outlier that will probably be influential in driving a linear relationship between y and x.

b) Linear models

**fit1<-lm(anscombe[,2]~anscombe[,1])**

**fit2<-lm(anscombe[,4]~anscombe[,3])**

**fit3<-lm(anscombe[,6]~anscombe[,5])**

**fit4<-lm(anscombe[,8]~anscombe[,7])**

|  |  |  |
| --- | --- | --- |
| **Data** | **R2** |  |
| 1 | 0.67 |  |
| 2 | 0.67 |  |
| 3 | 0.67 |  |
| 4 | 0.67 |  |

R2 for each of these models is identical. However, R2 is a measure of the *linear association* between two variables. It also does not distinguish between a strong linear relationship with an outlier (Set 3) and a moderate linear relationship (Set1). Clearly, one must use caution when using R2 as a measure of the fit of the model.

Q4:

In each case I choose the starting values from the data by solving 3 equations for the three unknowns a, b, and c. Using properties of the growth curves as well as the data, the equations are based on

(1): a = max(height)

(2): f(0) = min(height)

(3): f(mean(time)) = mean(height)

y <- read.table(file="./growth.txt",header=TRUE)$height

t <- read.table(file="./growth.txt",header=TRUE)$t

## fit nonlinear growth curves

##logistic

a.start <- max(y)

b.start <- a.start/(min(y))

c.start <- -log((a.start-mean(y))/(b.start\*mean(y)))/mean(t)

logistic <- nls(y~a/(1+b\*exp(-c\*t)),start=list(a=a.start,b=b.start,c=c.start),trace=TRUE)

## gompertz

a.start <- max(y)

b.start <- -log(min(t)/a.start)

c.start <- -log(-log(mean(y)/a.start)/b.start)/mean(t)

gompertz <- nls(y~a\*exp(-b\*exp(-c\*t)),start=list(a=a.start,b=b.start,c=c.start),trace=TRUE)

##VB

a.start <- max(y)

b.start <- -log((mean(y)-a.start)/(min(y)-a.start))/mean(t)

c.start <- -log((a.start-min(y))/a.start)/b.start

vb <- nls(y~a\*(1-exp(-b\*(t+c))),start=list(a=a.start,b=b.start,c=c.start),trace=TRUE )

plot(t,y,xlab="time in years",ylab = "height in meters")

H.vb <- predict(vb,list(t=seq(min(t),max(t),0.01)))

H.gompertz <- predict(gompertz,list(t=seq(min(t),max(t),0.01)))

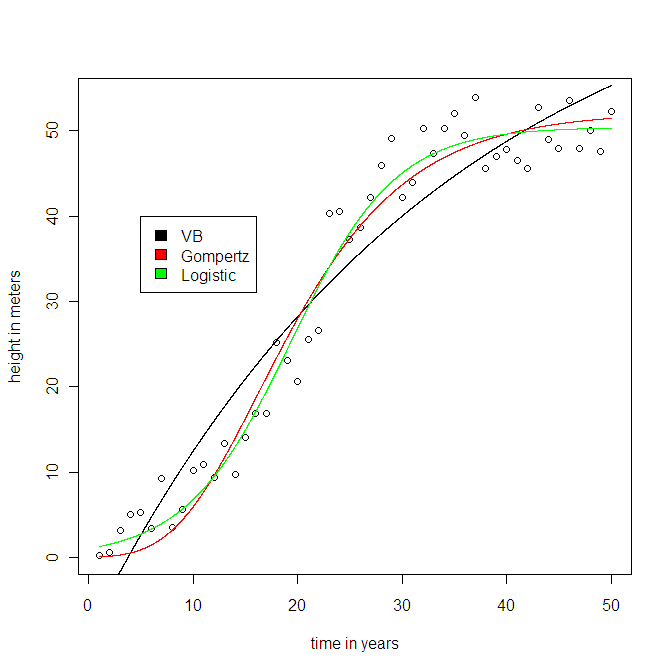
H.logistic <- predict(logistic,list(t=seq(min(t),max(t),0.01)))

lines( seq(min(t),max(t),0.01), H.vb,col="black")

lines( seq(min(t),max(t),0.01), H.gompertz,col="red")

lines( seq(min(t),max(t),0.01), H.logistic,col="green")

legend(x=5,y=40,legend = c("VB","Gompertz","Logistic"), fill = c("black","red","green"))



In each case the limit is just the parameter “a”, which represents the asymptotic or maximum height. Confidence intervals based on the normal distribution are:

> rbind(

+ c(74.411938 -1.96\*9.950934,74.411938 +1.96\*9.950934),

+ c(50.4208-1.96\*0.8473,50.4208 +1.96\*0.8473),

+ c(52.21789 -1.96\*1.33361,52.21789 +1.96\*1.33361 ))

[,1] [,2]

[1,] 54.90811 93.91577

[2,] 48.76009 52.08151

[3,] 49.60401 54.83177

All the models have three parameters so we choose the one with smallest RSS. This is the logistic model in this case. Starting with the estimated logistic growth curve, one takes the derivative with respect to time using the chain rule. The estimated growth rate curve is depicted below:

>t.plot <- seq(min(t),max(t),.01)

>a <- 50.4208

>b <- 47.1377

>c <- 0.1993

>Y.deriv <- a\*b\*c\*exp(-c\*t.plot)/((1+b\*exp(-c\*t.plot))^2)

