Regressions USING MLE

VenkataYogeshGongati

November 20, 2016

# Installing the required Packages

requiredPackages <- c('bbmle')  
  
for (i in requiredPackages){  
 ifelse(i %in% rownames(installed.packages()),NA,install.packages(i) )  
}  
  
# to supress warnings globally  
options(warn=-1)  
  
# to supress package startup messages  
suppressPackageStartupMessages(library(data.table))  
  
# loading the packages  
library(bbmle)

## Loading required package: stats4

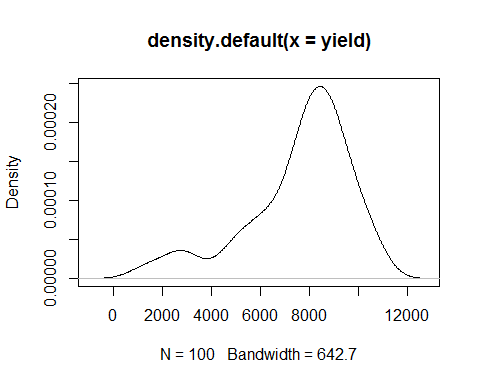
# removing the unused variables  
rm(i,requiredPackages)

# Linear Regression using MLE

# reading the corn data  
Corn <- read.csv("E:/Study/uconn/Fall2016/R/Assignments/Corn.csv")  
  
attach(Corn)  
reg1 <- lm(yield~nitrate)  
reg1

##   
## Call:  
## lm(formula = yield ~ nitrate)  
##   
## Coefficients:  
## (Intercept) nitrate   
## 6148.50 51.52

# saving the linear fitted values  
Corn$liner <- reg1$fitted.values  
  
  
#### von Bertalanffy growth model  
f1 <- function(a,beta,sigma){  
 y = a\*(1-exp(1-beta\*nitrate))  
 e = yield - y  
 LL = sum(dnorm(e,mean = 0,sd = sigma,log = T))  
 return(-1\*LL)   
}  
  
#looking at the distribution of yield  
plot(density(yield))



# using mle2 to find the paramtere values which maximizes the likelihood  
res = mle2(  
 minuslogl = f1,  
 start = list(a = mean(yield),  
 beta = 0,  
 sigma= 100),  
 lower = c(a=-Inf,beta=-Inf,sigma=0),   
 method = "L-BFGS-B"  
)  
  
# looking the result  
summary(res)

## Maximum likelihood estimation  
##   
## Call:  
## mle2(minuslogl = f1, start = list(a = mean(yield), beta = 0,   
## sigma = 100), method = "L-BFGS-B", lower = c(a = -Inf, beta = -Inf,   
## sigma = 0))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(z)   
## a 8.3274e+03 5.8972e-07 1.4121e+10 < 2.2e-16 \*\*\*  
## beta 2.2988e-01 1.4938e-02 1.5389e+01 < 2.2e-16 \*\*\*  
## sigma 1.8607e+03 7.9575e-11 2.3383e+13 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## -2 log L: 1789.531

Just like in any other MLE algorithm we are trying to optimize for the parameters of the function f1 which will give us the maximum likelihood for the sample data to come form the population using starting and lower values in mle2.

Assiging Mean value of the yield to the parameter 'a'(alpha) will be the most reasonable value we can assign to it. b(beta)'s best possible value that can be assigned is 0, which will bring our prediction values somewhere close to the parameter a(alpha)

The lower limit for sigma(i.e sd) should be no less than 0, as the sd can never be less than 0. Remaining parameters can have any value between -Inf to Inf.

# predict the yield using the VBGrowth model built using mle  
Corn$VBGrowth <- 8327.4\*(1-exp(1-(0.22988\*nitrate)))  
  
# function to evaluate the metric for model comparison   
metricsFun <- function(a,m){  
 metrics <- c(MAD = 0, MSE = 0, MAPE = 0, MPSE = 0, R2 = 0, TMAD = 0, P90 = 0)  
 metrics["MAD"] <- mean(abs(a-m))  
 metrics["MSE"] <- mean((a-m)^2)  
 metrics["MAPE"] <- mean(abs((a-m)/a))  
 metrics["MPSE"] <- mean(((a-m)/a)^2)  
 #for R2  
 SSE <- sum((a-m)^2)  
 SST <- sum((a-mean(a))^2)  
 metrics["R2"] <- 1-(SSE/SST)  
 metrics["TMAD"] <- mean(abs(a-m),trim=0.05)  
 metrics["P90"] <- quantile(abs(a-m),probs = 0.9)  
 metrics   
}  
  
# metric evaluation for linear model  
metricsFun(yield,Corn$liner)

## MAD MSE MAPE MPSE R2   
## 1.516551e+03 3.766379e+06 3.415469e-01 6.135421e-01 1.930015e-01   
## TMAD P90   
## 1.413319e+03 3.318568e+03

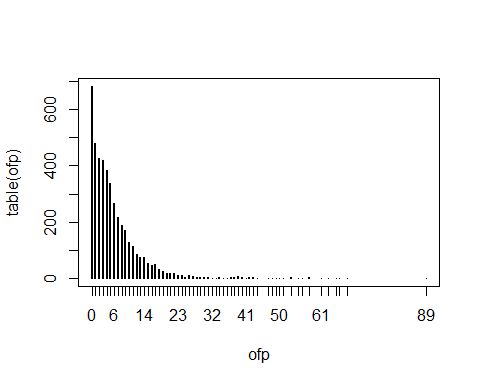
# metric evaluation for von Bertalanffy model  
metricsFun(yield,Corn$VBGrowth)

## MAD MSE MAPE MPSE R2   
## 1.443848e+03 3.462270e+06 3.032808e-01 3.556969e-01 2.581611e-01   
## TMAD P90   
## 1.351837e+03 3.229212e+03

From the above results we can se that the Rsquare for VBGrowht model is 0.25 and more than that of linear model, so we can say that VBGrowth model is better predictor model than linear model.

# Zero Inflated Poisson Regression using MLE

# reading the health data  
health <- read.csv("E:/Study/uconn/Fall2016/R/Assignments/Health.csv")  
  
attach(health)  
  
plot(table(ofp))



# we can see the dataset contains lot of observations with ofp = 0's   
# so to predict ofp(count values) we can use ZIP regression   
  
# function to calculate the likelihood  
f2 <- function(a0,a1,a2,b0,b1,b2,b3){  
 y1 <- a0+a1\*numchron+a2\*male  
 y2 <- b0+b1\*numchron+b2\*employed+b3\*married  
 p <- exp(y1)/(1+exp(y1))  
 lamb <- exp(y2)  
 LL <- sum(log(ifelse(ofp == 0 , p+((1-p)\*dpois(x = ofp,lambda = lamb)),(1-p)\*dpois(x = ofp,lambda = lamb))))  
 return(-1\*LL)  
}  
  
# using mle2 to find the paramtere values which maximizes the likelihood  
res1 = mle2(  
 minuslogl = f2,  
 start = list(a0 = 0,  
 a1 = 0,  
 a2 = 0,  
 b0 = mean(ofp),  
 b1 = 0,  
 b2 = 0,  
 b3 = 0))  
  
summary(res1)

## Maximum likelihood estimation  
##   
## Call:  
## mle2(minuslogl = f2, start = list(a0 = 0, a1 = 0, a2 = 0, b0 = mean(ofp),   
## b1 = 0, b2 = 0, b3 = 0))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(z)   
## a0 -1.1971637 0.0730858 -16.3803 < 2.2e-16 \*\*\*  
## a1 -0.5429373 0.0427499 -12.7003 < 2.2e-16 \*\*\*  
## a2 0.3734568 0.0866203 4.3114 1.622e-05 \*\*\*  
## b0 1.6770432 0.0129344 129.6573 < 2.2e-16 \*\*\*  
## b1 0.1495035 0.0042856 34.8851 < 2.2e-16 \*\*\*  
## b2 0.0490210 0.0213314 2.2981 0.02156 \*   
## b3 -0.0579783 0.0126693 -4.5763 4.733e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## -2 log L: 33519.83

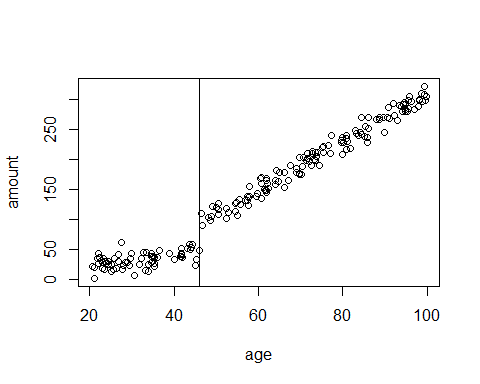
detach(health)

As the a0,a1,a2 are the parameters used for forming the equation to predict the number visits of people who are not sick, i.e usualy 0, we can use 0 for all of the three parameters including the a0(mean of hospital visits people who are not sick is 0). Similarly using 0 for b1,b2,b3 is also reasonable starting point as these are coefficients used in the linear equation. Where as for b0 we can use the mean value of (ofp) as these is a constant value(mean is the best starting constant value we can use for a linear model).

The p value in the results suggest that all the varibales that we have used in ZIP regression are signifficant.

# Condition Threshold regression using MLE

# reading the spending data  
spending <- read.csv("E:/Study/uconn/Fall2016/R/Assignments/spending.csv")  
attach(spending)  
  
# plotting the data  
plot(spending)  
  
abline(v = 46)



f3 <- function(l,a,b,sig1,sig2){  
 y1 <- a  
 y2 <- a+b\*(age-l)  
 LL<-sum(ifelse(age<=l,dnorm(amount-y1,mean = 0,sd = sig1,log=T),dnorm(amount-y2,mean = 0,sd = sig2 ,log=T)))  
 return(-1\*LL)  
}  
  
# using mle2 to find the paramtere values which maximizes the likelihood  
res2 <- mle2(minuslogl = f3,  
 start = list(  
 l=min(age),  
 a=mean(amount),  
 b=0,  
 sig1=sd(amount),  
 sig2=sd(amount)  
 ),  
 lower = c(l = 0,a=0,b=0,sig1=10,sig2=10),   
 method = "L-BFGS-B"  
)  
  
summary(res2)

## Maximum likelihood estimation  
##   
## Call:  
## mle2(minuslogl = f3, start = list(l = min(age), a = mean(amount),   
## b = 0, sig1 = sd(amount), sig2 = sd(amount)), method = "L-BFGS-B",   
## lower = c(l = 0, a = 0, b = 0, sig1 = 10, sig2 = 10))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(z)   
## l 33.900350 0.454930 74.5177 < 2.2e-16 \*\*\*  
## a 27.736016 1.575662 17.6028 < 2.2e-16 \*\*\*  
## b 4.338348 0.049323 87.9586 < 2.2e-16 \*\*\*  
## sig1 11.346102 1.267307 8.9529 < 2.2e-16 \*\*\*  
## sig2 13.591473 0.759709 17.8904 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## -2 log L: 1597.133

detach(spending)

For assigning statrting value of l(threshold on age) equal to minimum value of age, as threshold cannot be less than that but can be greater than minimum. For a(alpha) we are assigning the mean of amount as the value as thats the most reasonable value can think of assigning the constant parameter to in linear regression, and for b(beta) we assign 0 (zero) as it is a reasonable value to the slope(coefficient of varibale) For sig1,sig2 we can assign the standard deviation of the data itself.

Coming to the lower possible values, we can use 0 as the lower value for l(threshold),a(alpha),b(beta) as these things can never be negative. As the sd of any random varibale cannot be less than zero we give the sig1 and sig2 equals to 10 based one the original standard deviation

Looking at the results we can say that there are two different linear regressions formed parted by age varibale at age = 33.90. Also all the parameters seem to significant (lloking at the parameter).