STAT 530 HW 5

Md Muhtasim Billah

4/21/2020

Question 1

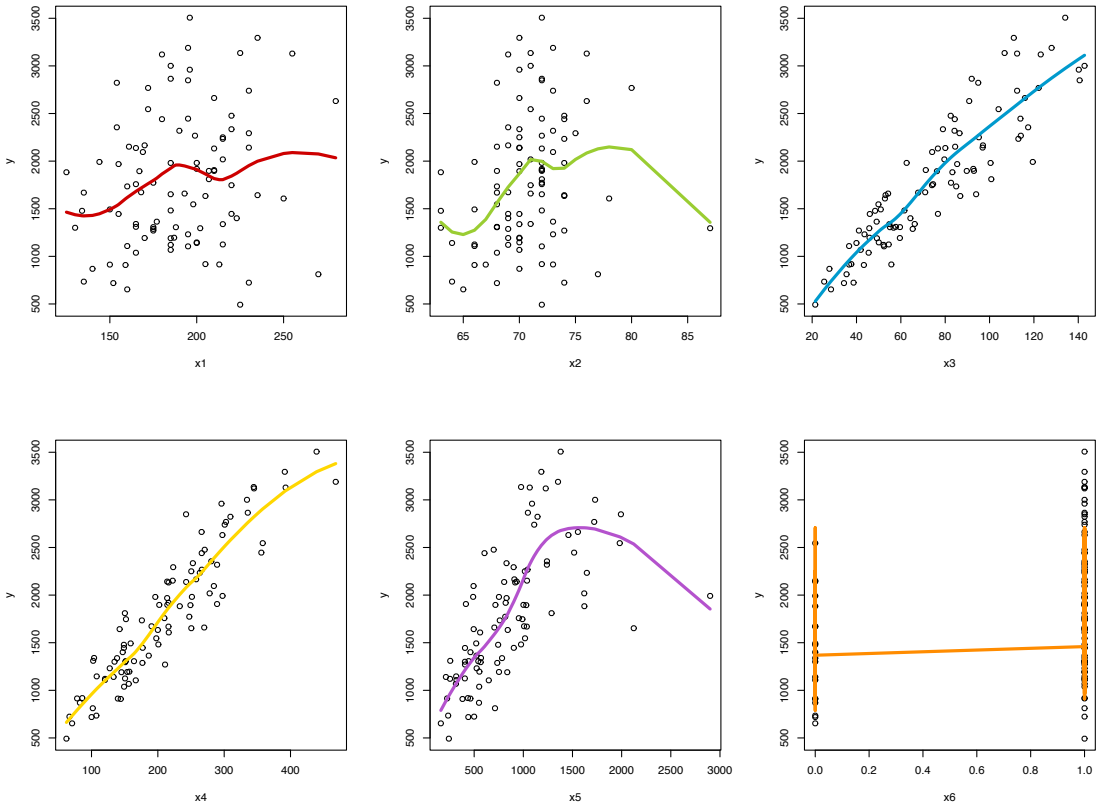
Use the data set you used in HW 3 and answer the following questions.

a. Do a LOESS fit for Calories using the variables: weight, height, protein, carbohydrates, calcium and gender and look at the relationship that LOWESS provides and comment.
b. Take the residuals versus predicted plot that you ran for HW 3 diagnostics and LOWESS it. Do you see any patterns? Should you?

Answer

a.

The LOWESS plots for the response variable, y (Calories) with 6 predictors x1, x2, x3, x4, x5 and x6 (weight, height, protein, carbohydrates, calcium and gender respectively) are provided below.



LOWESS is helpful for understanding the trend of relationship between two variables. Based on the LOWESS plots, following comments can be made.

Calories (y) vs Weight (x1) and Height (x2):

There is no clear linear or non-linear pattern noticed in the dataset when y is scatter-plotted against x1 or x2. Rather the data points look like randomly scattered on the map. Yet, linear regression might be applicable on x1 and non-linear on x2.

Calories (y) vs Protein (x3) and Carbohydrates (x4):

The relationship of y with x3 and x4 looks very linear as exhibited by the LOWESS plots. Multiple linear regression should prove to be the the most appropriate approach for analyzing the dataset based on these two variables.

Calories (y) vs Calcium (x5):

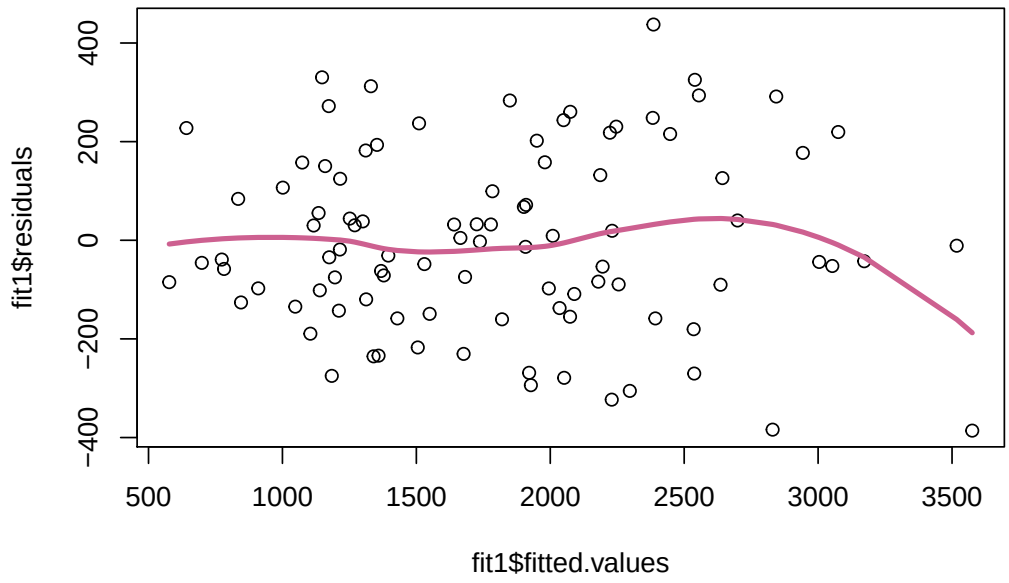
If the huge outlier is included in the model, from the LOWESS plot, it is apparent that a non-linear relationship between y and x5 might exist. A non-linear regression would be more appropriate if this variable is considered.

Calories (y) vs Gender (x6):

This is a categorical variable and LOWESS smoothing doesn't bear any physical meaning.

b.

LOWESS was performed on the residuals versus predicted plot from HW 3 diagnostics which is given below.



From the above plot, no pattern among the residuals is visible, rather all we see is a random scatter. It is further ensured by the pretty flat-looking LOWESS line which indicates that linear regression would be appropriate on this dataset. If a more curvilinear pattern was visible, non-linear regression would be applicable.

Question 2

You are given the data on time of exposure (x) and photolytic damage (y) for an exposed surface on the web as HW5_nlindata.xls. The scientists believe that the model relating damage and time is given by:

$$\text{damage} = \text{Theta1} - \text{Theta2} * \exp(-\text{Theta3} * \text{time})$$

a. Estimate the parameters theta1, theta2 and theta3 and test for the significances of the parameters. Use your estimates to predict y given x. What can you say about the model?
b. Use bootstrapping to verify your results.
c. Do a LOWESS plot to compare your linear and non-linear fit.

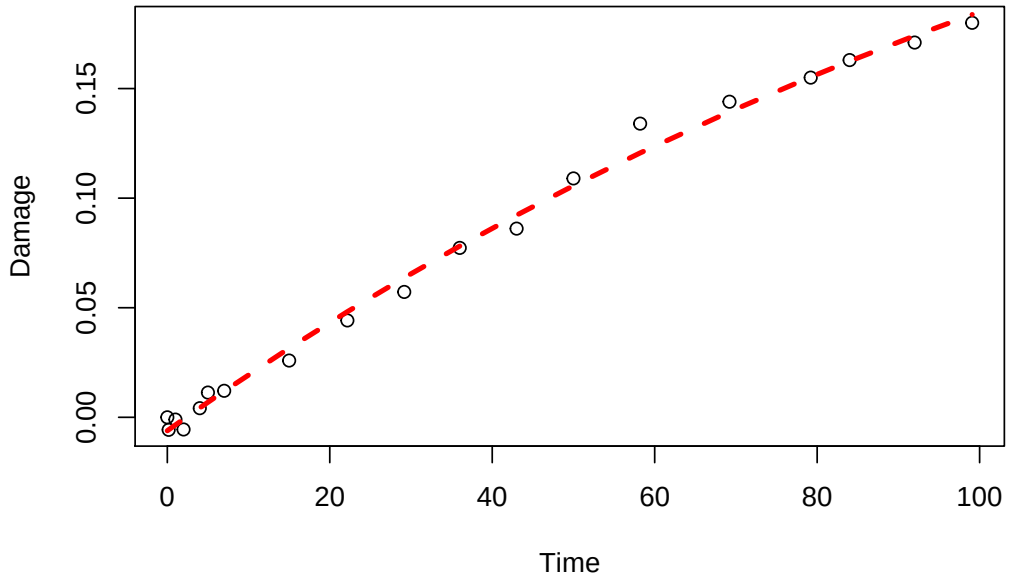
Answer

a.

The provided data were fit into a nonlinear regression model using the provided non-linear equation. Based on the initial values, it took 12 iteration for the model to converge with a tolerance of $1.493e^{-06}$ The estimated parameters are privided below along with their confidence intervals and the test of significance.

Parameter	Estimate	Conf. Interval	Std. Error	t-value	Pr(>|t|)	Significant?
Theta1	0.3789	(0.2450, 0.5128)	0.0632	5.999	1.85e-05	YES
Tehta2	0.3850	(0.2533, 0.5168)	0.0621	6.196	1.28e-05	YES
Theta3	0.0069	(0.0037, 0.0101)	0.0015	4.539	0.000335	YES

Based on the fitted model, y can be estimated given x which is shown by the red regression line in the figure below.



Though there is a linear trend in the dataset, performing a non-linear regression on this dataset gives better fit of the regression line. From the fitted line, it can be said that the model is pretty accurate for explaining the variance in the response.

b.

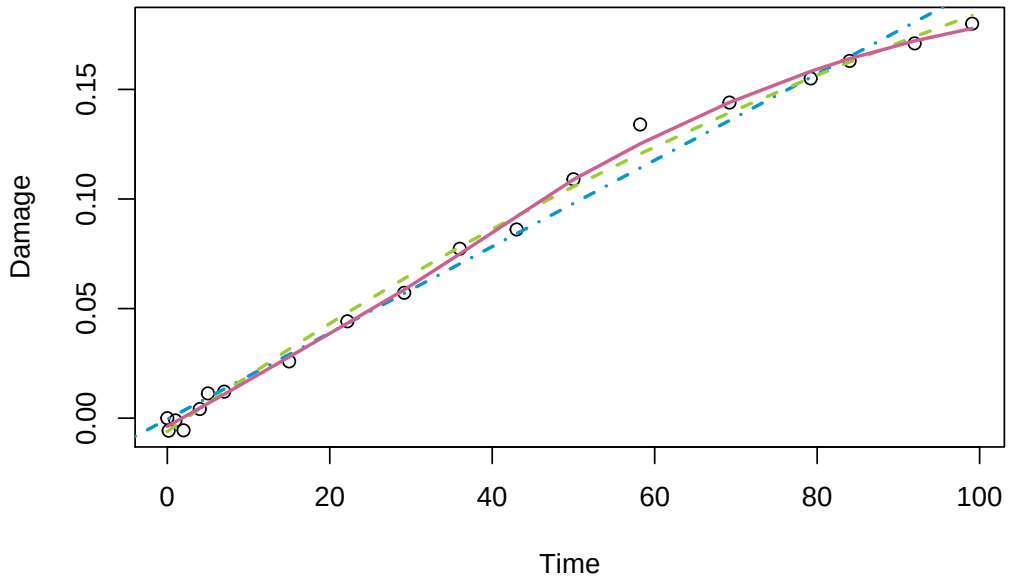
The parameter estimations can be verified by bootstrapping. It was found that the same estimations are provided by the bootstrap. The total number of iterations, estimations for the parameters, their standard deviations and biases are provided below.

```
models
```

```
$n
[1] 994

$a
[1] 0.3789092

$b
[1] 0.3850283

$c
[1] 0.006858974

$SD_a
[1] 0.06418601

$SD_b
[1] 0.06330094

$SD_c
[1] 0.001310201

$Bias_a
[1] -0.01269746

$Bias_b
[1] -0.01277335

$Bias_c
[1] 2.177338e-05
```

c.

The LOWESS plot (solid pink line) is provided to compare the linear (dash-dotted skyblue line) and non-linear (dashed olive line) fit. Comparing the fitted lines with the LOWESS, it seems that a non-linear regression is more suitable for this dataset.



Appendix

The R code used for this homework is provided below.

```
setwd("/Users/muhtasim/Desktop/STAT530/HWs/HW5")
data=read.csv("HW3_2020_clean.csv",header=T)
##problem1a
y=data$Calories
x1=data$Weight.lbs
x2=data$Height.Inches
x3=data$Protein
x4=data$Carbohydrates
x5=data$Calcium
x6=data$gender

par(mfrow=c(3,2))

lw1=loess(y ~ x1,data, span=.6)
plot(y ~ x1, data=data)
j1 <- order(x1)
lines(x1[j1],lw1$fitted[j1],col="red3",lwd=3)

lw2=loess(y ~ x2,data, span=.65)
plot(y ~ x2, data=data)
j2 <- order(x2)
lines(x2[j2],lw2$fitted[j2],col="olivedrab3",lwd=3)

lw3=loess(y ~ x3,data, span=.65)
plot(y ~ x3, data=data)
j3 <- order(x3)
lines(x3[j3],lw3$fitted[j3],col="deepskyblue3",lwd=3)

lw4=loess(y ~ x4,data, span=.65)
```

```
plot(y ~ x4, data=data)
j4 <- order(x4)
lines(x4[j4],lw4$fitted[j4],col="gold",lwd=3)

lw5=loess(y ~ x5,data, span=.65)
plot(y ~ x5, data=data)
j5 <- order(x5)
lines(x5[j5],lw5$fitted[j5],col="mediumorchid3",lwd=3)

lw6=loess(y ~ x5,data, span=.65)
plot(y ~ x6, data=data)
j6 <- order(x6)
lines(x6[j6],lw6$fitted[j6],col="darkorange",lwd=3)

par(mfrow=c(1,1))

##problem1b
#MLR
fit1 = lm(Calories ~ Weight.lbs + Height.Inches + Protein + Carbohydrates
          + Calcium + gender, data = data)
#residual vs fitted plot
lwfit=loess(fit1$residuals ~ fit1$fitted.values, data, span=.65)
plot(fit1$fitted.values, fit1$residuals)
jfit = order(fit1$fitted.values)
lines(fit1$fitted.values[jfit],lwfit$fitted[jfit],col="hotpink33",lwd=3)


##problem2a
library("readxl")
nlindata=read_excel("Hw5_Nlindata.xls")
y=nlindata$damage
x=nlindata$time
##starting values a=intercept, b=slope/intercept
model=nls(y~a-b*exp(-c*x),start=list(a=15,b=10,c=0.5))
plot(x,y)
lines(x,predict(model),col="red",lty=2,lwd=3)
summary(model)
library(nlstools)
cf=confint2(model,level=0.95)
cf


##problem2b
#bootstrap by Vasili
dat<-read.csv("Hw5_Nlindata.csv", header = TRUE, sep = ",",
              na.strings = " ", colClasses = c('numeric', 'numeric'))
library(data.table)
library(boot)
model.fun<- function(dt){
  fit<-nls(damage~a-(b*exp(-c*time)), start = list(a=0,b=0.01, c=0.01), data = dt)
  df<-dt
  df<-data.frame(df,fitted(fit),residuals(fit))
  colnames(df)<-c("time", "damage", "fitted", "resid")
```

```
  fun<-function(df, inds){

    library(data.table)
    bootDamage=df$fitted +df$resid[inds]
    df<-data.frame(df,bootDamage)
    colnames(df)<-c("time", "damage", "fitted", "resid", "bootDamage")
    tryCatch(coef(nls(bootDamage~a-(b*exp(-c*time)),
                      start=list(a=0,b=0.01,c=0.01), data=df)),
             error=function(e)c("a"= NA, "b"= NA, "c"= NA))
  }


  kk<-boot(df, fun, R=1000)
  print(boot.ci(kk, type="bca", index = 1) )
  print(boot.ci(kk, type="bca", index = 2) )
  print(boot.ci(kk, type="bca", index = 3) )
  res0<-kk$t0
  res1<- apply(kk$t, 2, sd, na.rm= TRUE)
  res2<- res0 - colMeans(kk$t, na.rm = TRUE)
  return(as.list(setNames(c(sum(!is.na(kk$t[,1])), res0, res1, res2),
                          c("n","a","b","c","SD_a","SD_b","SD_c","Bias_a", "Bias_b", "Bias_c"))))

}
models<-model.fun(dat)
models


##problem2c
#linear fit
lin=lm(y~x,data=nlindata)
#lowess
lw=loess(y ~ x,nlindata, span=.75)
j=order(x)
#plotting
plot(x,y, xlab="Time", ylab="Damage")
abline(lin,col="deepskyblue3",lty=4,lwd=2)   #linear
lines(x,predict(model),col="olivedrab3",lty=2,lwd=2)  #nonlinear
lines(x[j],lw$fitted[j],col="hotpink3",lty=1,lwd=2)   #lowess
```