####

####The Framingham Heart Study https://www.framinghamheartstudy.org/ is the longest on-going

####study of risk factors from heart disease and many other types of chronic diseases. It is a national

####treasure started in 1948.

####Framingham data contains observations for medical test on systolic blood pressure

####VARIABLE NAMES and DEFINITIONS:

####COL 01 -- OBS = observation number (1-1615)

####COL 02 -- AGE = age at exam 2

####COL 03 -- SBP21 = first systolic blood pressure at exam 2

####COL 04 -- SBP22 = second systolic blood pressure at exam 2

####COL 05 -- SBP31 = first systolic blood pressure at exam 3

####COL 06 -- SBP32 = second systolic blood pressure at exam 3

####COL 07 -- SMOKE = present smoking at exam 1

####COL 08 -- CHOLEST2 = serum cholesterol at exam 2

####COL 09 -- CHOLEST3 = serum cholesterol at exam 3

####COL 10 -- FIRSTCHD = indicator of first evidence of CHD occurring at exam 3

####through 6, i.e., within an eight-year follow-up period

####to exam 2. CHD = coronary heart disease

####GENERAL INFORMATION:

####1. The data are for MALES only.

####2. The data contain complete records only.

####This exercise is to study whether systolic blood pressure and serum cholesterol are related, and

####whether smoking status (and later on age) are factors in systolic blood pressure (it is).

####Vivek Kumar Gupta , Stat 689 Assignment 6

####

####Set the working director for the assignment

setwd("F:/OneDrive/Learning/DataScience/Statistics Texas A&M University/689/Assignment/Assignment6")

#Clear the working spaace and load in the libraries

rm(list = ls())

set.seed(1234)

library("HRW")

library("mgcv")

library("nlme")

## Read the data into a frame called framingham

framingham = read.csv("Framingham(1).csv")

####There are four systolic blood pressure measurements. Take their average and create the variable

####Average the two cholesterol measurements and take their logarithm

framingham$LSBP = log(apply(framingham[,c("SBP22" , "SBP21" ,"SBP31" ,"SBP32")], 1,mean) - 50)

framingham$Lcholest = log(apply(framingham[,c("Cholest2" , "Cholest3" )], 1,mean))

#####Reset the dataframe

framingham = framingham[, c("Age" , "Smoker" , "LSBP" , "Lcholest" , "CHD")]

####Convert Smoker column to factors and check the reference

framingham$Smoker = as.factor(framingham$Smoker)

unique(framingham$Smoker)

####Attach the frame so that it is easier to reference

attach(framingham)

**################Answer to Question 1 ####################**

####Fit a multiple linear regression of LSBP on Lcholest and smoker using lm. Since the

####smoking variable is binary, this is an ordinary ANCOVA without an interaction. You will

####notice that the Rsquared is quite low. Produce a table of estimates, standard errors and p-

####values.

plot(Lcholest , LSBP , col = as.numeric(Smoker) + 1 , main = "Plot of Framingham dataset" )

legend("bottomright",

legend = c("NonSmoker" ,"Smoker"),

col = 2:3,

cex = 1.1,

text.col = 2:3

)

#ANCOVA model

framingham.mlr.fit = lm (LSBP ~ Lcholest + Smoker)

summary(framingham.mlr.fit)

####Coefficients:

####Estimate Std. Error t value Pr(>|t|)

####(Intercept) 3.55569 0.17029 20.880 < 2e-16 \*\*\*

####Lcholest 0.15540 0.03140 4.949 8.22e-07 \*\*\*

####Smoker1 -0.03796 0.01251 -3.034 0.00246 \*\*

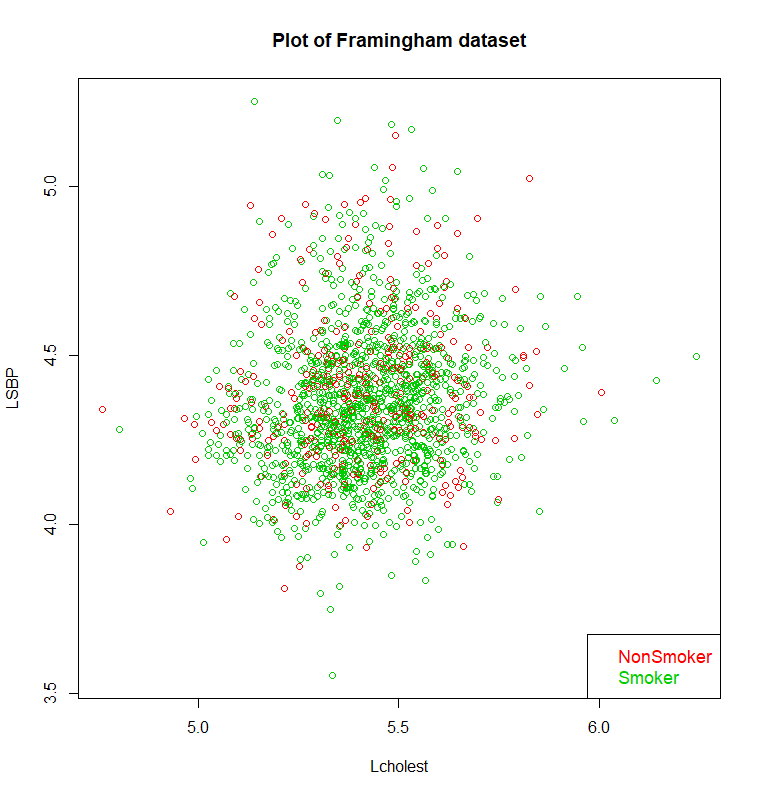
####---

####Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

####Residual standard error: 0.2107 on 1612 degrees of freedom

####Multiple R-squared: 0.02036, Adjusted R-squared: 0.01915

####F-statistic: 16.75 on 2 and 1612 DF, p-value: 6.299e-08



**################Answer to Question 2 ####################**

####Do a little bit of a web search about whether smokers have higher or lower blood pressure

####than nonsmokers. Does the analysis in (1) agree?

####Conclusion from the above model ####

####http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/MakeChangesThatMatter/Smoking-High-Blood-Pressure-and-Your-Health\_UCM\_301886\_Article.jsp#.Wn-KTujwaUk

####From the above results ( negative slope of Smoker, significant pvalue of 0.00232) , it seems that the model believes

####that as Smoking level goes from 0 to 1 controlling from Level of cholestrol, in logscale , systolic blood pressure (logscale)

#### decreases. replicated experiments have shown otherwise.

####The model seems incorrect.

**################Answer to Question 3 ####################**

####Answered in Question 2.

**################Answer to Question 4 ####################**

#### Multiple Linear Regression model with Interaction term

framingham.mlr.fit.Int = lm (LSBP ~ Lcholest + Smoker + Smoker : Lcholest)

summary(framingham.mlr.fit.Int)

####Coefficients:

#### Estimate Std. Error t value Pr(>|t|)

####(Intercept) 3.55075 0.33525 10.591 <2e-16 \*\*\*

####Lcholest 0.15632 0.06191 2.525 0.0117 \*

####Smoker1 -0.03130 0.38907 -0.080 0.9359

####Lcholest:Smoker1 -0.00123 0.07184 -0.017 0.9863

####---

####Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

####Residual standard error: 0.2108 on 1611 degrees of freedom

####Multiple R-squared: 0.02036, Adjusted R-squared: 0.01854

####F-statistic: 11.16 on 3 and 1611 DF, p-value: 2.993e-07

#### While the Adjusted R2 value is very less , multiple linear factor intercation model does not find

#### any interaction between smoking status and Log of cholestrol.

####Comparing models

anova(framingham.mlr.fit , framingham.mlr.fit.Int)

####Analysis of Variance Table

####Model 1: LSBP ~ Lcholest + Smoker

####Model 2: LSBP ~ Lcholest + Smoker + Smoker:Lcholest

####Res.Df RSS Df Sum of Sq F Pr(>F)

####1 1612 71.579

####2 1611 71.579 1 1.3019e-05 3e-04 0.9863

####With a high pvalue we dont have evidence that an interaction exists.

**################Answer to Question 5 ####################**

####Semi paramteric regression model without any interaction

framingham.gam.fit = gam (LSBP ~ Smoker + s(Lcholest , bs = "cr" , k = 23) , method = "REML")

summary(framingham.gam.fit)

####Parametric coefficients:

####Estimate Std. Error t value Pr(>|t|)

####(Intercept) 4.39713 0.01100 399.735 < 2e-16 \*\*\*

####Smoker1 -0.03799 0.01251 -3.036 0.00244 \*\*

####Approximate significance of smooth terms:

#### edf Ref.df F p-value

####s(Lcholest) 1.064 1.126 22.27 1.73e-06 \*\*\*

####R-sq.(adj) = 0.0192 Deviance explained = 2.04%

####-REML = -214.85 Scale est. = 0.044402 n = 1615

**################Answer to Question 6 ####################**

####Display a plot of the two lines (fitted values in 2 groups), but without the data.

####Setup a grid of values of predictor , Lcholest

sq = 1000

Lcholest\_pr = seq(min(Lcholest) , max(Lcholest) , length = sq)

####Get the predictions from the fit 1. For non smoker and 2. for smoker along with the SE

LSBP\_pr\_nsmoker = predict(framingham.gam.fit , newdata = data.frame("Lcholest" = Lcholest\_pr , "Smoker" = as.factor(rep(0,sq ))) , se = T)

LSBP\_pr\_smoker = predict(framingham.gam.fit , newdata = data.frame("Lcholest" = Lcholest\_pr , "Smoker" = as.factor(rep(1,sq ))) , se = T)

####Set the base plot

plot(Lcholest,LSBP,type="n",

xlab="Cholestrol(Log)",

ylab="Log Syst. Blood Pressure , fit",

main = "LSBP vs LCholestrol",

xlim = c(min(Lcholest\_pr) , max(Lcholest\_pr)),

ylim = c(4.1 , 4.7))

####Overlay the lines on the plot

lines(Lcholest\_pr , LSBP\_pr\_nsmoker$fit , type = "l" , col = 2 , lwd = 2)

lines(Lcholest\_pr , LSBP\_pr\_smoker$fit , type = "l" , col = 3 , lwd = 2)

legend("bottomright",

legend = c("NonSmoker" ,"Smoker"),

col = 2:3,

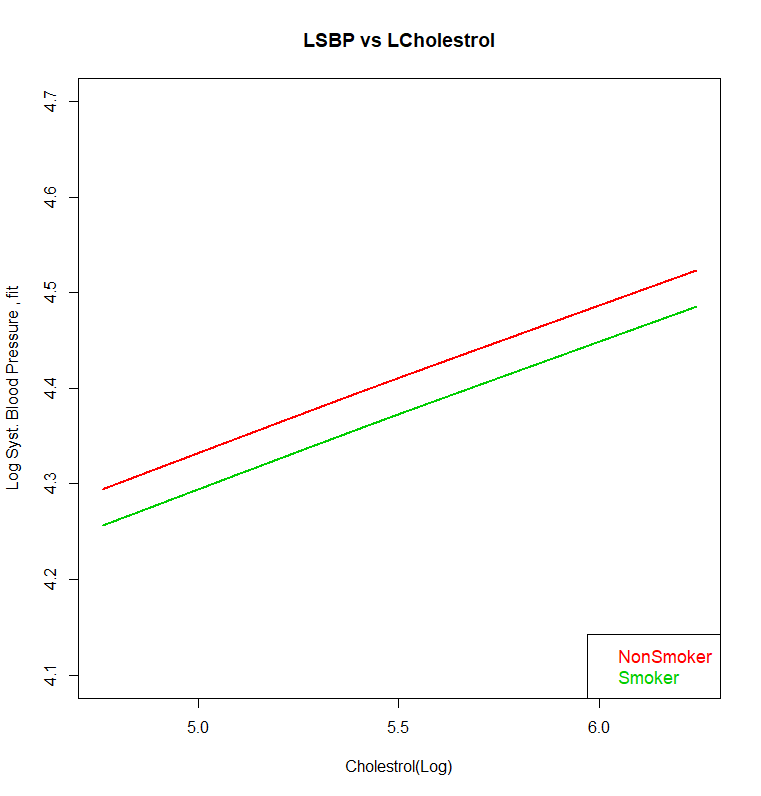
cex = 1.1,

text.col = 2:3

)

###We see visually that the 2 groups means are different given value of the covariate, LCholest and that

####there isn’t visual evidence of an interaction



**################Answer to Question 7 ####################**

##Show the fitted lines along with Se bands in 2 separate plots

dev.off()

par(mfrow = c(2,1))

plot(Lcholest\_pr,LSBP\_pr\_nsmoker$fit,type="l",

xlab="Cholestrol",

ylab="Log Syst. Blood Pressure",

main = "Mean function of LSBP , NonSmoker",

xlim = c(min(Lcholest\_pr) , max(Lcholest\_pr)),

ylim = c(4.1 , 4.7) , col = 2)

####Get the bands from the fit

upper = LSBP\_pr\_nsmoker$fit + (1.96 \* LSBP\_pr\_nsmoker$se.fit)

lower = LSBP\_pr\_nsmoker$fit - (1.96 \* LSBP\_pr\_nsmoker$se.fit)

####Draw a polygon

polygon(x=c(Lcholest\_pr, rev(Lcholest\_pr)), y=c(upper, rev(lower)),

col="gray", border=NA)

####Add the mean function on top of the polygon

lines(Lcholest\_pr , LSBP\_pr\_nsmoker$fit , type = "l" , col = 2 , lwd = 2)

####Add Legends for dislplay

legend("bottomright",

legend = c("NonSmoker" ),

col = 2,

cex = 1.1,

text.col = 2

)

plot(Lcholest\_pr,LSBP\_pr\_smoker$fit,type="l",

xlab="Cholestrol",

ylab="Log Syst. Blood Pressure",

main = "Mean function of LSBP , Smoker",

xlim = c(min(Lcholest\_pr) , max(Lcholest\_pr)),

ylim = c(4.1 , 4.7) , col = 3)

upper = LSBP\_pr\_smoker$fit + (1.96 \* LSBP\_pr\_smoker$se.fit)

lower = LSBP\_pr\_smoker$fit - (1.96 \* LSBP\_pr\_smoker$se.fit)

polygon(x=c(Lcholest\_pr, rev(Lcholest\_pr)), y=c(upper, rev(lower)),

col="gray", border=NA)

legend("bottomright",

legend = c("Smoker" ),

col = 3,

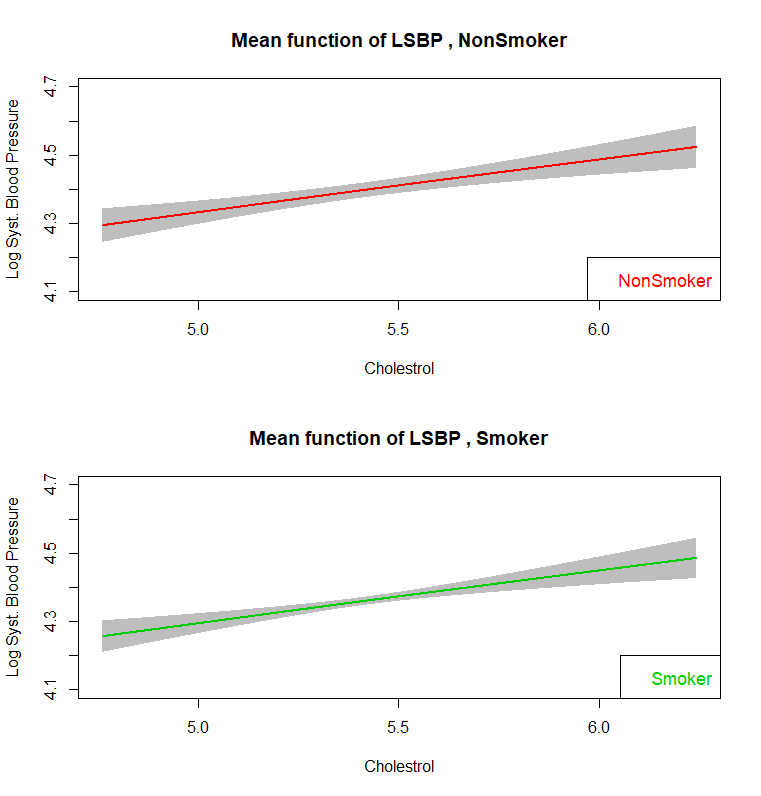
cex = 1.1,

text.col = 3

)

lines(Lcholest\_pr , LSBP\_pr\_smoker$fit , type = "l" , col = 3 , lwd = 2)

####Plot on the next page



**################Answer to Question 8 ####################**

####Add an interaction term in the semi parametric model

framingham.gam.fit.Int = gam (LSBP ~ Smoker\*Lcholest + s(Lcholest , bs = "cr" , k = 23) , method = "REML")

summary(framingham.gam.fit.Int)

####Parametric coefficients:

####Estimate Std. Error t value Pr(>|t|)

####(Intercept) 0.0000000 0.0000000 NA NA

####Smoker1 -0.0327238 0.3891274 -0.084 0.933

####Lcholest 0.8121214 0.0020324 399.592 <2e-16 \*\*\*

####Smoker1:Lcholest -0.0009734 0.0718495 -0.014 0.989

####---

####Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

####Approximate significance of smooth terms:

####edf Ref.df F p-value

####s(Lcholest) 1.067 1.131 98.79 <2e-16 \*\*\*

####---

####Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

####Rank: 25/26

####R-sq.(adj) = 0.0186 Deviance explained = 2.04%

####-REML = -210.81 Scale est. = 0.04443 n = 1615

####Pvalues of the coefficients and value of coefficients are cited above with a non significant interaction (p value 0.989).

**################Answer to Question 9 ####################**

####Display fits of the interaction model

dev.off()

LSBP\_pr\_int\_nsmoker = predict(framingham.gam.fit.Int , newdata = data.frame("Lcholest" = Lcholest\_pr , "Smoker" = as.factor(rep(0,sq ))) , se = T)

LSBP\_pr\_int\_smoker = predict(framingham.gam.fit.Int , newdata = data.frame("Lcholest" = Lcholest\_pr , "Smoker" = as.factor(rep(1,sq ))) , se = T)

plot(Lcholest,LSBP,type="n",

xlab="Cholestrol(Log)",

ylab="Log Syst. Blood Pressure , fit",

main = "Mean function of LSBP with an Interaction Model",

xlim = c(min(Lcholest\_pr) , max(Lcholest\_pr)),

ylim = c(4.1 , 4.7))

lines(Lcholest\_pr , LSBP\_pr\_int\_nsmoker$fit , type = "l" , col = 2 , lwd = 2)

lines(Lcholest\_pr , LSBP\_pr\_int\_smoker$fit , type = "l" , col = 3 , lwd = 2)

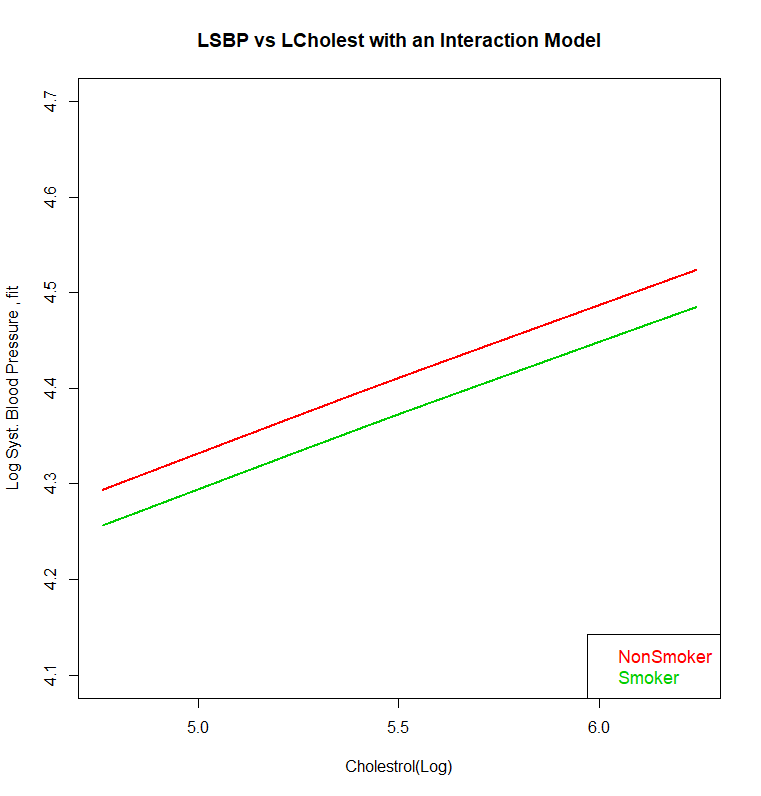
legend("bottomright",

legend = c("NonSmoker" ,"Smoker"),

col = 2:3,

cex = 1.1,

text.col = 2:3)



**################Answer to Question 10 ####################**

####What does having an interaction mean in the case when the factors are binary?

####Interaction in general

####When the mean of the response at two different levels of a factor at a given value of a covariate (or level of another factor)

####is not the same at a different value of covariate(or another factor) , the facor is said to interact with covariate (or another factor)

####When both the factors are binary, the difference in mean response of a factor at two levels is not the same at two ####different levels of the second factor.

####Algebraically, mu\_ab - mu\_a`b != mu\_ab` - mu\_a`b` where Factor 1 is at levels a and a`` while Factor 2 is at level b and b`

**################Answer to Question 11 ####################**

IndSmoker = as.numeric(Smoker == 1)

IndNSmoker = as.numeric(Smoker == 0)

framingham.gam.fit.Smoker = gam (LSBP ~ s(Lcholest, IndSmoker ) , method = "REML" )

framingham.gam.fit.NSmoker = gam (LSBP ~ s(Lcholest, IndNSmoker ) , method = "REML")

anova(framingham.gam.fit.Smoker , framingham.gam.fit.NSmoker)

####Analysis of Deviance Table

####Model 1: LSBP ~ s(Lcholest, IndSmoker)

####Model 2: LSBP ~ s(Lcholest, IndNSmoker)

####Resid. Df Resid. Dev Df Deviance

####1 1612 71.579

####2 1612 71.579 -4.5475e-13 0

#### We see that the 2 models and hence the two fits for smoker and nonsmokers are statistically significant.

**################Answer to Question 12 ####################**

y = framingham$LSBP

x = framingham$Age

numIntKnots = 23

intKnots = quantile(unique(x) , seq( 0 , 1 , length = (numIntKnots + 2))[ -c( 1 , (numIntKnots + 2))])

a= 1.01 \* min(x) - 0.01 \* max(x)

b= 1.01 \* max(x) - 0.01 \* min(x)

Z = ZOSull(x , range.x = c(a , b) , intKnots = intKnots)

dummyID = factor( rep(1 , length(x)))

fit = lme(y ~ x , random = list( dummyID = pdIdent( ~ -1 + Z)))

betaHat = fit$coef$fixed

uHat = unlist(fit$coef$random)

sigsqepsHat = fit$sigma ^ 2

sigsquHat = as.numeric(VarCorr(fit)[1, 1])

ng = 1001

xg = seq(a , b , length = ng)

Xg = cbind(rep(1 , ng) , xg)

Zg = ZOSull(xg , range.x = c(a , b) , intKnots = intKnots)

fHatg = as.vector((Xg %\*%betaHat + Zg%\*%uHat))

**################Answer to Question 13 ####################**

plot( x , y , bty = "l" , xlab = "Age" , ylab = "LSBP" ,

col = "dodgerblue" , cex.lab = 1.5 , cex.axis = 1.5)

plot(xg , fHatg , col = 3 , lwd = 2 , ylim = c(4.1, 4.6),

xlab = "Age" , ylab = "LSBP" ,

main = "Fit via NLME/LME ")

Cg = cbind(rep(1 , ng) , xg, Zg)

C = cbind( rep(1 , length(y)) , x , Z)

D = diag(c(0 , 0 , rep(1 , ncol(Z))))

sdg = sqrt(sigsqepsHat) \* sqrt(diag(Cg %\*% solve(crossprod(C) + (sigsqepsHat/sigsquHat)\*D , t(Cg) )))

CIlowg = fHatg - 2 \* sdg

CIuppg = fHatg + 2 \* sdg

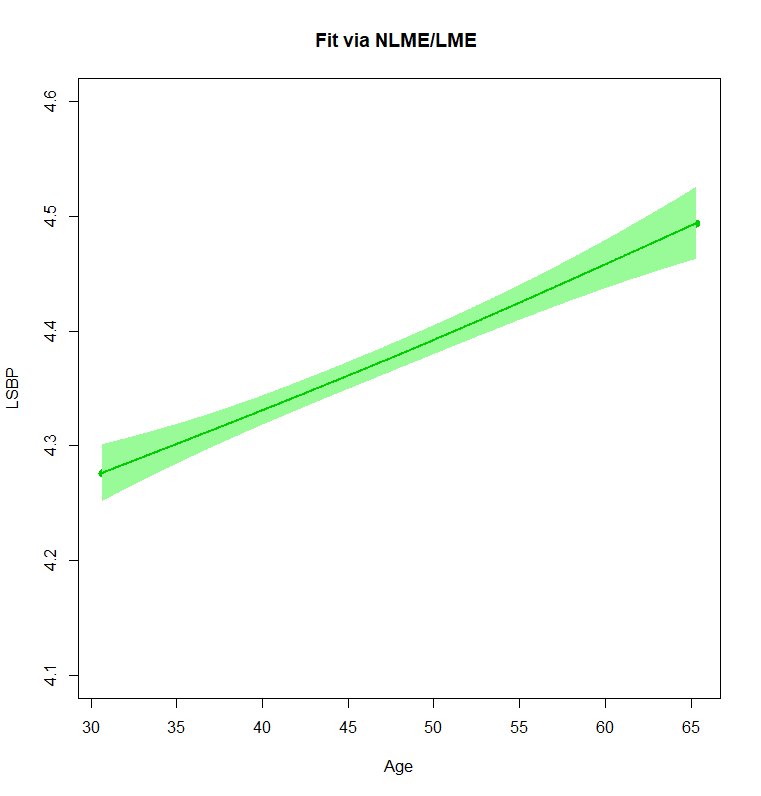
polygon(c(xg , rev(xg)) , c(CIlowg , rev(CIuppg)), col = "palegreen" , border = F)

lines(xg , fHatg , col = 3 , lwd = 2 , ylim = c(4.1, 4.6),

xlab = "Age" , ylab = "LSBP" ,

main = "Fit via NLME/LME ")

####Plot on the next page



**################Answer to Question 14 ####################**

summary(fit)

####Linear mixed-effects model fit by REML

####Data: NULL

####AIC BIC logLik

####-492.401 -470.8576 250.2005

####

####Random effects:

#### Formula: ~-1 + Z | dummyID

####Structure: Multiple of an Identity

####Z1 Z2 Z3 Z4 Z5 Z6 Z7 Z8 Z9

####StdDev: 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042

####Z10 Z11 Z12 Z13 Z14 Z15 Z16 Z17 Z18

####StdDev: 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042

####Z19 Z20 Z21 Z22 Z23 Z24 Z25 Residual

####StdDev: 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.0003722042 0.2059625

####

####Fixed effects: y ~ x

####Value Std.Error DF t-value p-value

####(Intercept) 4.080427 0.028263462 1613 144.37110 0

####x 0.006281 0.000608534 1613 10.32184 0

####Correlation:

#### (Intr)

####x -0.982

####

####Standardized Within-Group Residuals:

#### Min Q1 Med Q3 Max

####-3.70736855 -0.66630572 -0.05824728 0.58982495 4.40312810

####

####Number of Observations: 1615

####Number of Groups: 1

####We can see from the above summary that we have statistically significant coef (pvalue = 0)

**################Answer to Question 15 ####################**

fit.lin.age = lm (y ~ x)

summary(fit.lin.age)

fit.quad.age = lm (y ~ poly(x , 2))

summary(fit.quad.age)

fit.gam.age = gam(y ~ s(x))

anova(fit.lin.age , fit.quad.age , test = "F")

####Analysis of Variance Table

####

####Model 1: y ~ x

####Model 2: y ~ poly(x, 2)

#### Res.Df RSS Df Sum of Sq F Pr(>F)

####1 1613.0 68.439

####2 1612.0 68.366 1.00000 0.073051 1.7225 0.1896

anova(fit.lin.age , fit.gam.age)

####Analysis of Variance Table

####Model 1: y ~ x

####Model 2: y ~ s(x)

####Res.Df RSS Df Sum of Sq F Pr(>F)

####1 1613.0 68.439

####2 1612.7 68.412 0.28496 0.027155 2.2464 0.1247

####The results of above annoa (partial F tests) suggests that linear fit is approprioate as we dont see any significant gains in fit

#### from linear to quadratic ( p value = 0.1896)

####OR from linear to semiparamteric (p value = .1247)