Code

- Show All Code
- Hide All Code

Final Project

Eugeniah Arthur

May 10, 2019

Title

A predictive model to ascertain the relationship between Spine bone mineral density and other variables of North American Adolescents in the Bone Data.

Introduction

It is a well known fact that there exist a relationship between the features of bone and the age of an object. This is what acheological studies is based on. For years, the age and possibly gender of ancient objects have been determine through information gotten from bones.As Andrea Waters-Rist writes in her course description on coursera:" Archaeology and anthropology enables us gain unique insights into the past and the present through the study of human skeletal remains. Therefore, for this study, I took the bone data on ElemStatLearn in R to investigate the relationship between age, bone mineral density and gender. The data consist of measurements in bone mineral density of 261 north american adolescents who made two consecutive visits to the hospital. The data has the average ages over the two visits , gender of the adolescents(factor: Female=1, Male=2) and the Id number of the individuals to show the repeated measurements.
According to an article by NIH Osteoporosis and related bone diseases national research center titled Bone Mass Measurement: What the Numbers Mean, the relative bone density test measures an individuals bone density and compares it to that of a healthy individual(an established norm) to give a score.The score is like a standard deviation. A zero score shows that your bone mineral density is as good as a healthy individual. However, a score below 0 shows a weaker bone density and hence a higher risk of fracture.
The original source of the data is from Bachrach et al who first gathered this data in 1999 for their support grant project. I would like to explore the relationship between these variable.

Objectives of the study

1. To investigate the relationship among the variables
2. To find the best model to predict bone mineral density using other variables as predictors

Methods and results

In order for me to execute my first objective, I explored and cleaned my data to make sure that it was suitable for the various models I wanted to use.

Exploratory Data Analysis

```
library(ElemStatLearn)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
b=na.omit(bone)
str(b)
```

```
## 'data.frame':    485 obs. of  4 variables:
##  $ idnum : int  1 1 1 2 2 2 3 3 3 4 ...
##  $ age   : num  11.7 12.7 13.8 13.2 14.3 ...
##  $ gender: Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ spnbmd: num  0.01808 0.06011 0.00586 0.01026 0.21053 ...
```

```
head(b)
```

```
##   idnum   age gender      spnbmd
## 1     1 11.70   male 0.018080670
## 2     1 12.70   male 0.060109290
## 3     1 13.75   male 0.005857545
## 4     2 13.25   male 0.010263930
## 5     2 14.30   male 0.210526300
## 6     2 15.30   male 0.040843210
```

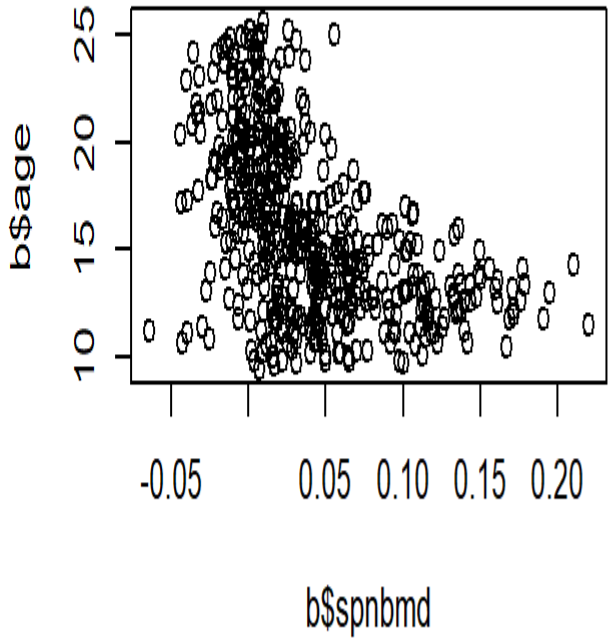
```
summary(b)
```

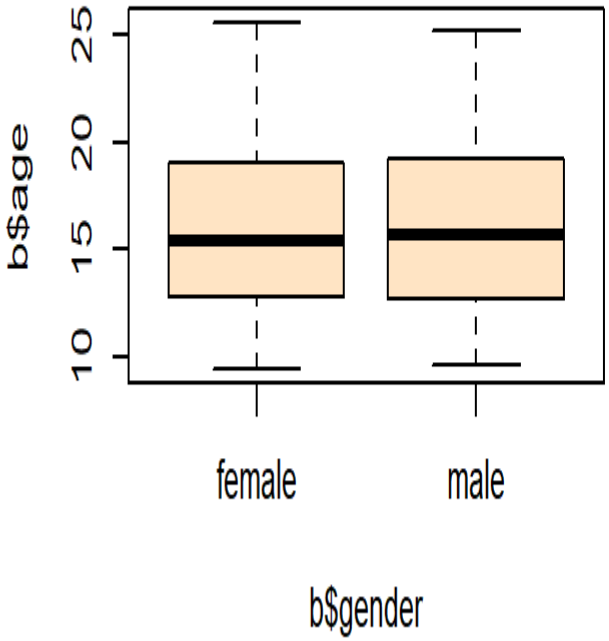
```
##      idnum            age           gender        spnbmd        
##  Min.   :  1.0   Min.   : 9.40   female:259   Min.   :-0.064103  
##  1st Qu.: 60.0   1st Qu.:12.70   male  :226   1st Qu.: 0.005858  
##  Median :124.0   Median :15.40                Median : 0.026591  
##  Mean   :151.5   Mean   :16.10                Mean   : 0.039252  
##  3rd Qu.:240.0   3rd Qu.:19.15                3rd Qu.: 0.064127  
##  Max.   :384.0   Max.   :25.55                Max.   : 0.219913
```

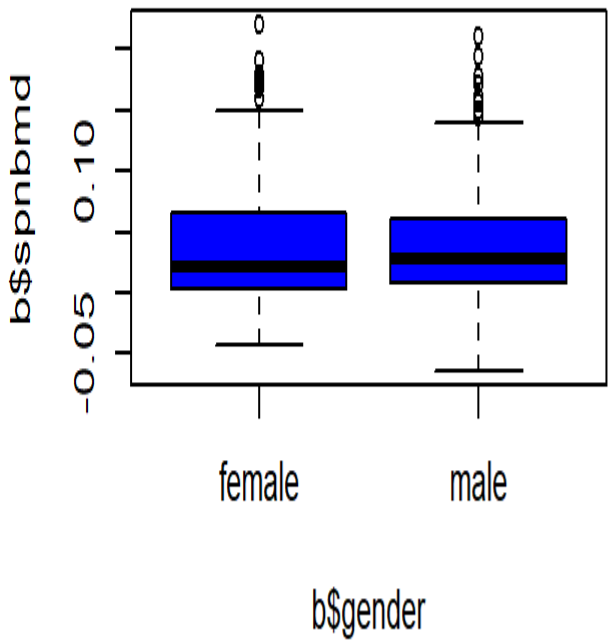
The summary of the data shows that some data were repeated with Idnum showing the repetition. The minimum age is 9.4 years and the maximum age is 25.55 years. With all the repeated measures, we had 259 males and 226 females . Also, the bone mineral density of the adolescents ranged from -0.064 and 0.2199.

Plots

```
par(mfrow=c(2,2))
plot(b$spnbmd,b$age)
boxplot(b$age~b$gender, col="bisque ")
boxplot(b$spnbmd~b$gender, col="blue")
```







The plot of the age with spnmbd shows a curved shape which gives a hint of a nonlinear relationship. Also, there was not much difference between the distribution of age in both genders.The difference in the distribution of bone mineral density in both genders was very little.

Since the data set had unequal number of revists of the adolescents, I decided to find the mean of the variables of each individual who revisited. And use those new variables for my model.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
b$gender=as.numeric(as.factor(b$gender))
dd=group_by(b, idnum) %>% summarize(meanspnbmd = mean(spnbmd),meanage=mean(age),
ngender=max(gender))
```

```
summary(dd)
```

```
##      idnum         meanspnbmd           meanage         ngender
##  Min.   :  1.0   Min.   :-0.064103   Min.   : 9.40   Min.   :1.000
##  1st Qu.: 74.0   1st Qu.: 0.006193   1st Qu.:12.82   1st Qu.:1.000
##  Median :170.0   Median : 0.024778   Median :15.90   Median :1.000
##  Mean   :178.7   Mean   : 0.036708   Mean   :16.34   Mean   :1.444
##  3rd Qu.:276.0   3rd Qu.: 0.062420   3rd Qu.:19.80   3rd Qu.:2.000
##  Max.   :384.0   Max.   : 0.194464   Max.   :25.15   Max.   :2.000
```

```
str(dd)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    261 obs. of  4 variables:
##  $ idnum     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ meanspnbmd: num  0.02802 0.08721 0.00685 0.13487 0.06257 ...
##  $ meanage   : num  12.7 14.3 12.4 11.5 13.7 ...
##  $ ngender   : num  2 2 2 1 1 1 1 2 1 1 ...
```

```
head(dd)
```

```
## # A tibble: 6 x 4
##   idnum meanspnbmd meanage ngender
##   <int>      <dbl>   <dbl>   <dbl>
## 1     1    0.0280     12.7       2
## 2     2    0.0872     14.3       2
## 3     3    0.00685    12.4       2
## 4     4    0.135      11.5       1
## 5     5    0.0626     13.7       1
## 6     6    0.00584    18         1
```

```
newdata=dd
```

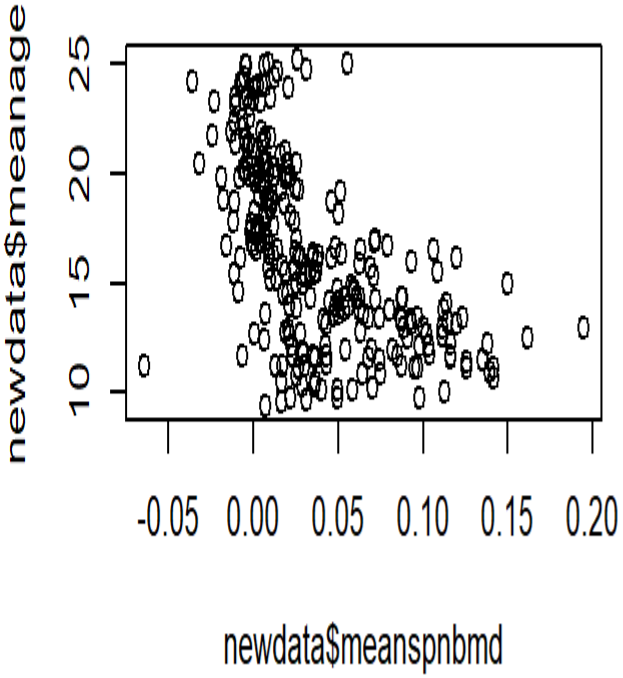
```
head(newdata$ngender)
```

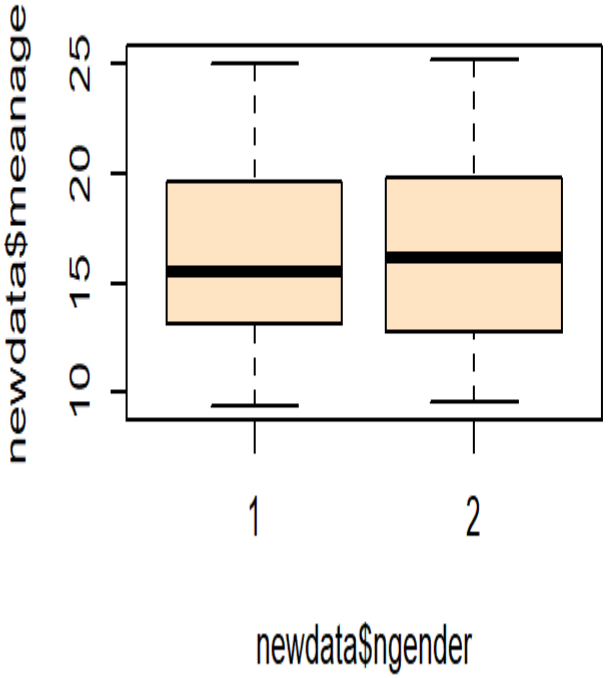
```
## [1] 2 2 2 1 1 1
```

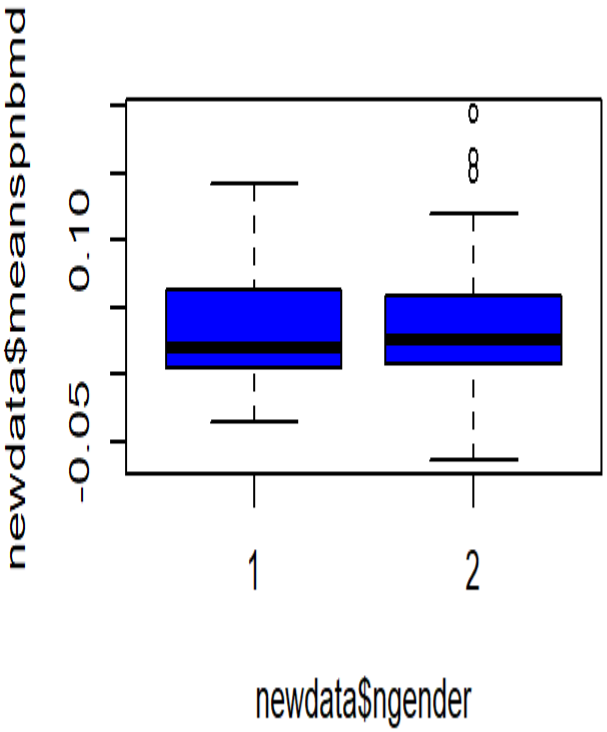
```
newdata$ngender=as.factor(as.numeric(newdata$ngender))
str(newdata)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    261 obs. of  4 variables:
##  $ idnum     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ meanspnbmd: num  0.02802 0.08721 0.00685 0.13487 0.06257 ...
##  $ meanage   : num  12.7 14.3 12.4 11.5 13.7 ...
##  $ ngender   : Factor w/ 2 levels "1","2": 2 2 2 1 1 1 1 2 1 1 ...
```

```
par(mfrow=c(2,2))
plot(newdata$meanspnbmd,newdata$meanage)
boxplot(newdata$meanage~newdata$ngender, col="bisque ")
boxplot(newdata$meanspnbmd~newdata$ngender, col="blue")
```







A plot of the newdata shows the same results as before. However, it seems the distribution of the relative spinal bone mineral density in Males is skewed with some outliers. Though, the median is the same for both males and females.After reorganising, I got 261 data points with new variable names.

```
# Removing the idnum
newdata=newdata[,-1]
str(newdata)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    261 obs. of  3 variables:
##  $ meanspnbmd: num  0.02802 0.08721 0.00685 0.13487 0.06257 ...
##  $ meanage   : num  12.7 14.3 12.4 11.5 13.7 ...
##  $ ngender   : Factor w/ 2 levels "1","2": 2 2 2 1 1 1 1 2 1 1 ...
```

```
n=newdata[,c(1,2)]
r=cor(n)
r
```

```
##             meanspnbmd    meanage
## meanspnbmd  1.0000000 -0.5835802
## meanage    -0.5835802  1.0000000
```

The scatterplot shows that there exist a negative correlation between meanage and the mean spnbmd.Thus, as age increases, the mean spinal bone density decreases. However, the relationship looks nonlinear. The correlation plot shows the relationship is around -0.58 which is not very strong.

Since one of my main objective is to find a good predictive model to measure the bone mineral density of an individual. I am splitting my data into train and test data so that I am able to compare the test errors of the various models.

train and test data

```
set.seed(1)
```

```
train = sample(1:nrow(newdata),132)
btrain = newdata[train, ]
btest = newdata[-train, ]
nrow(btrain)
```

```
## [1] 132
```

```
nrow(btest)
```

```
## [1] 129
```

Linear Model

Since there seems to be a negative correlation, I am going to check if a linear regression model is a good model for predicting mean age

```
siml=lm( meanspnbmd~.,data= btrain)
summary(siml)
```

```
##
## Call:
## lm(formula = meanspnbmd ~ ., data = btrain)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.126463 -0.024352 -0.004232  0.020074  0.108130
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.130739   0.012504  10.456  < 2e-16 ***
## meanage     -0.005546   0.000727  -7.628 4.61e-12 ***
## ngender2    -0.006266   0.006229  -1.006    0.316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03552 on 129 degrees of freedom
## Multiple R-squared:  0.3146, Adjusted R-squared:  0.304
## F-statistic:  29.6 on 2 and 129 DF,  p-value: 2.624e-11
```

The linear regression model shows that the intercept and meanage is significant to the model. However, gender was not significant at all. The adjusted r square value , 0.361 which is very low and shows the model may not be a good fit.

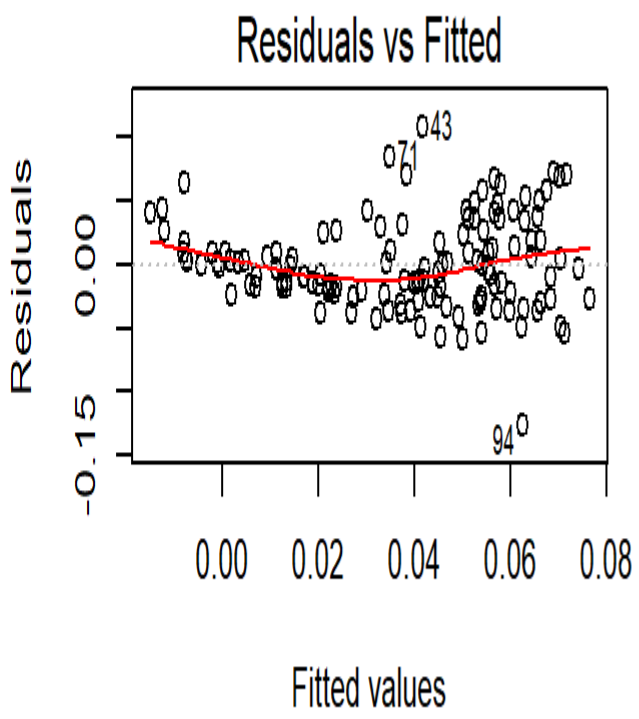
** Test error**

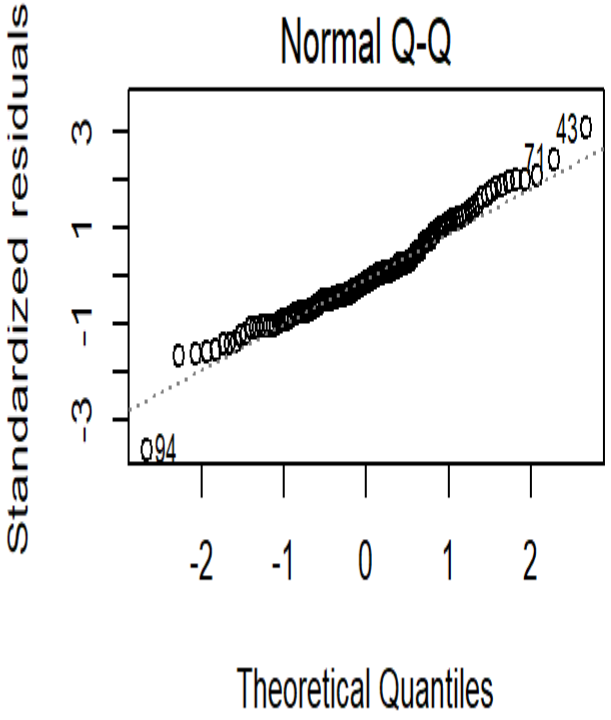
```
cpred = predict(siml, btest)
lmtesterr=mean((btest$meanspnbmd - cpred)^2)
lmtesterr
```

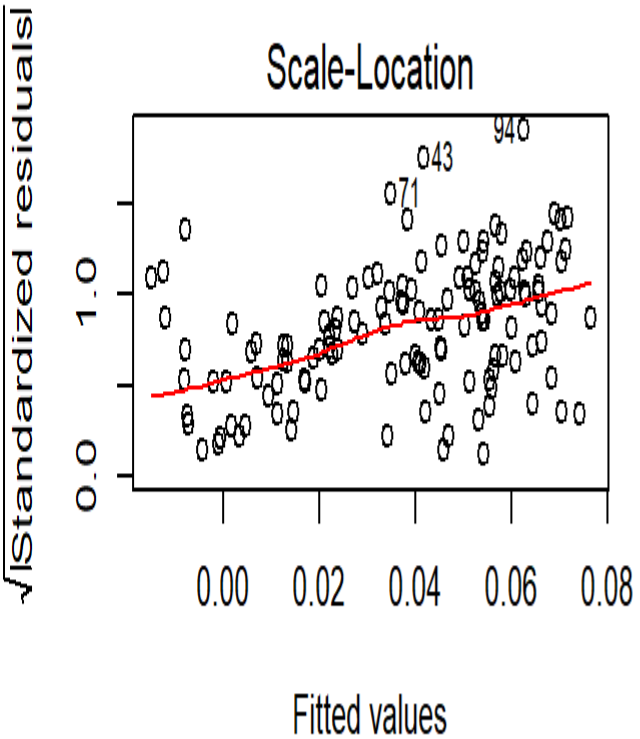
```
## [1] 0.001155216
```

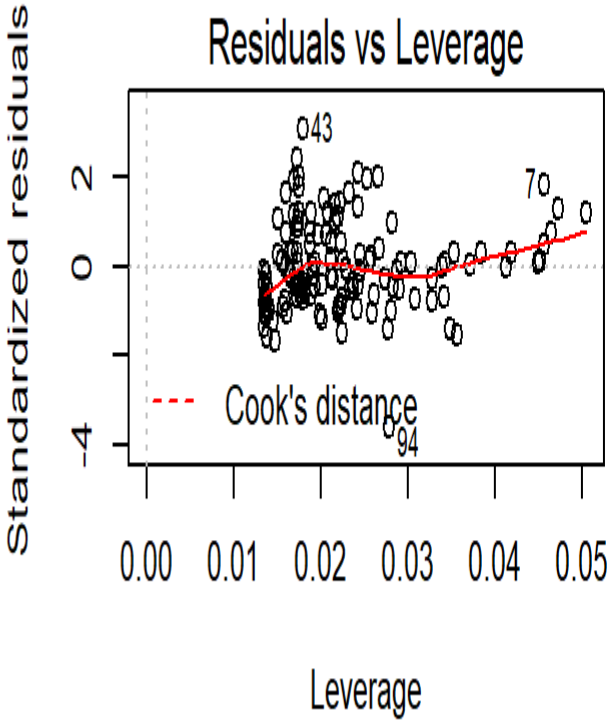
** Linear Model Diagonistic**

```
par(mfrow=c(2,2))
plot(siml)
```









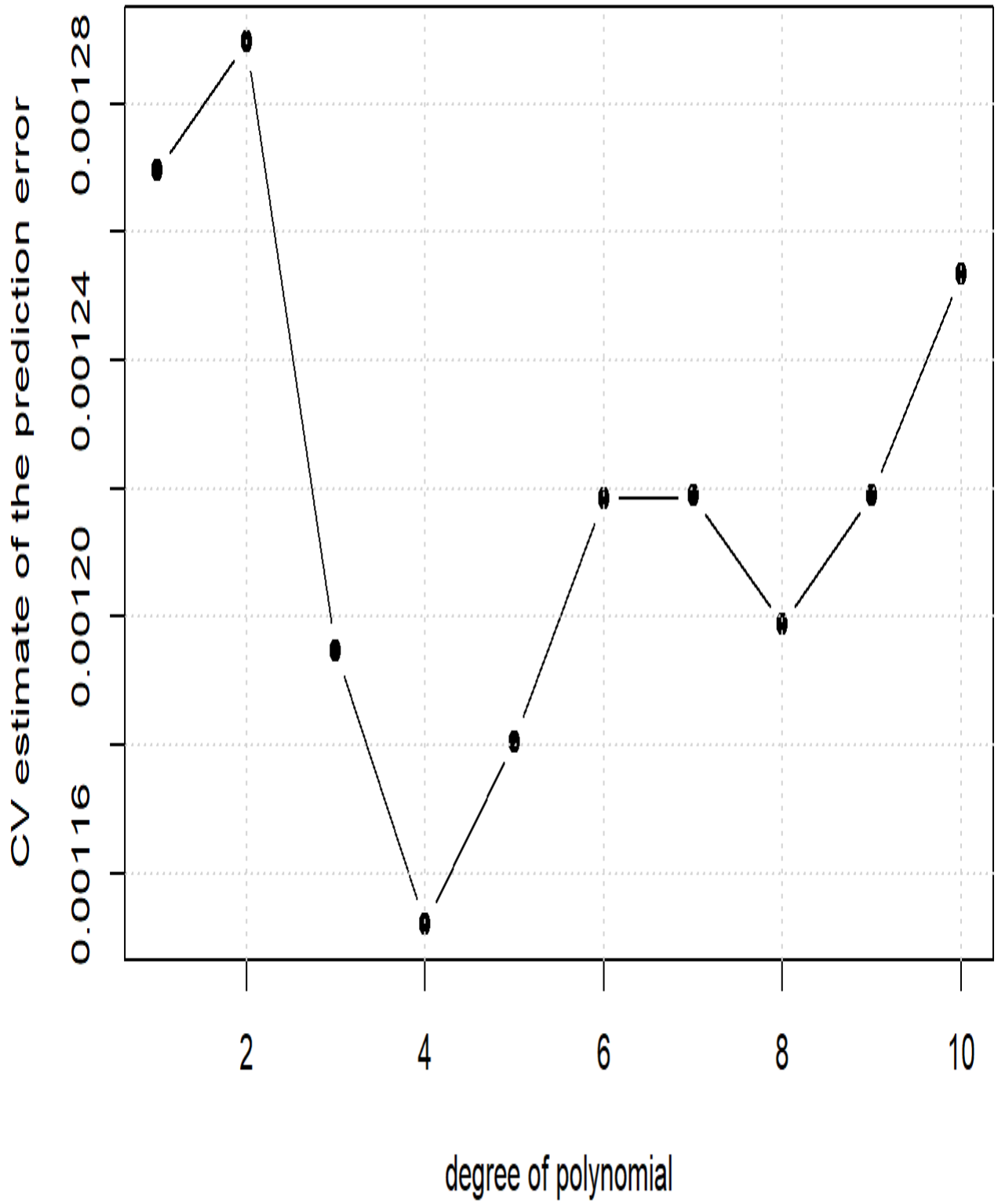
From the model diagonistic, the residual verses fitted model shows that the erorrs are correlated and have a pattern.Hence, there is a problem of mutlicollinearity.The normal qq plot shows the residual errors deviates from the normal distribution. The standardised verses leverage shows that there exist influential

points .The scale location is also not horizontal and hence the covariance of the error terms are not constant (problem of heteroscedascticity). From the model diagnostics, it is realised that the model does not fit the assumptions of the linear regression model . Hence, we will try some polynomial models.

Polynomial Regression

From the initial plot there seemed to exit a non linear relationship between the meanage and meanspmbn. We would use the cross validation method to fit a polynomial model for the data.

```
library(boot)
set.seed(1)
cv = rep(0,10)
for (i in 1:10) {
    glm.fit = glm(meanspnbmd ~ poly(meanage, i)+ ngender, data = btrain)
    cv[i]=cv.glm(btrain, glm.fit, K = 10)$delta[1]
}

plot(1:10, cv, pch = 19, type = "b", xlab = "degree of polynomial", ylab = "CV
estimate of the prediction error")
grid()
```




```
min.point = which.min(cv)
min.point
```

```
## [1] 4
```

```
fit.cv= glm(meanspnbmd ~ poly(meanage,min.point)+ngender, data= btrain)

summary(fit.cv)
```

```
##
## Call:
## glm(formula = meanspnbmd ~ poly(meanage, min.point) + ngender,
##     data = btrain)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.131448  -0.019551  -0.002488   0.014842   0.103039
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.038973   0.003959   9.844  < 2e-16 ***
## poly(meanage, min.point)1 -0.270944   0.033834  -8.008 6.66e-13 ***
## poly(meanage, min.point)2  0.052679   0.033836   1.557 0.122006
## poly(meanage, min.point)3  0.117918   0.034301   3.438 0.000795 ***
## poly(meanage, min.point)4 -0.048198   0.033872  -1.423 0.157221
## ngender2                  -0.002354   0.006022  -0.391 0.696525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001144707)
##
##     Null deviance: 0.23743  on 131  degrees of freedom
## Residual deviance: 0.14423  on 126  degrees of freedom
## AIC: -511.52
##
## Number of Fisher Scoring iterations: 2
```

The cross validation method selected a polynomial of degree 4 to give the minimum error rate. Also, gender was not significant in the prediction of mean spinal bone density

```
#Polynomial Test Error Rate
lm.pred = predict(fit.cv, btest)
testerr=mean((btest$meanspnbmd-lm.pred)^2)

testerr
```

```
## [1] 0.0009967901
```

The test MSE is relatively smaller than the linear model.

Ridge Regression Model

The ridge regrssion uses a criterion to shrink the variables that are not significant. However, it does not do variable selection as the lasso model that forces all the insignificnt variables to zero.

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```
xtrain=model.matrix(meanspnbmd~ngender+ poly(meanage,5), btrain)
xtest=model.matrix(meanspnbmd~ngender+poly(meanage,5), btest)
ytrain=btrain$meanspnbmd
ytest=btest$meanspnbmd
set.seed(1)
```

```
rm=cv.glmnet(xtrain,ytrain, alpha=0)
blam=rm$lambda.min
blam
```

```
## [1] 0.002588286
```

```
# The test error is given as
rp=predict(rm, s=blam, newx=xtest)
rtestmse=mean((rp-ytest)^2)
rtestmse
```

```
## [1] 0.001025186
```

For the ridge regression, the lambda value that gave the minimum MSE was 0.0026. The test error of the ridge regression was 0.00121. Which is relatively low.

Lasso model

```
set.seed(1)
lassomod=cv.glmnet(xtrain,ytrain, alpha=1)
blasm=lassomod$lambda.min
blasm
```

```
## [1] 0.00029759
```

```
lassopr=predict(lassomod,s=blasm,newx=xtest)
lassterr=mean((lassopr-ytest)^2)
 lassterr
```

```
## [1] 0.001014677
```

```
#lasso model
predict(lassomod, s=blasm, type="coefficients")
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                               1
## (Intercept)        0.038898292
## (Intercept)        .
## ngender2          -0.002183273
## poly(meanage, 5)1 -0.267525883
## poly(meanage, 5)2  0.049273022
## poly(meanage, 5)3  0.114659000
## poly(meanage, 5)4 -0.044824298
## poly(meanage, 5)5  0.031145035
```

```
#
```

The lasso model chose the best shrinkage method lambda as 0.0009333. During its variable selection, it forced the coefficient of the intercept to zero. The lasso model had a test error rate of 0.001014 which is relatively small.

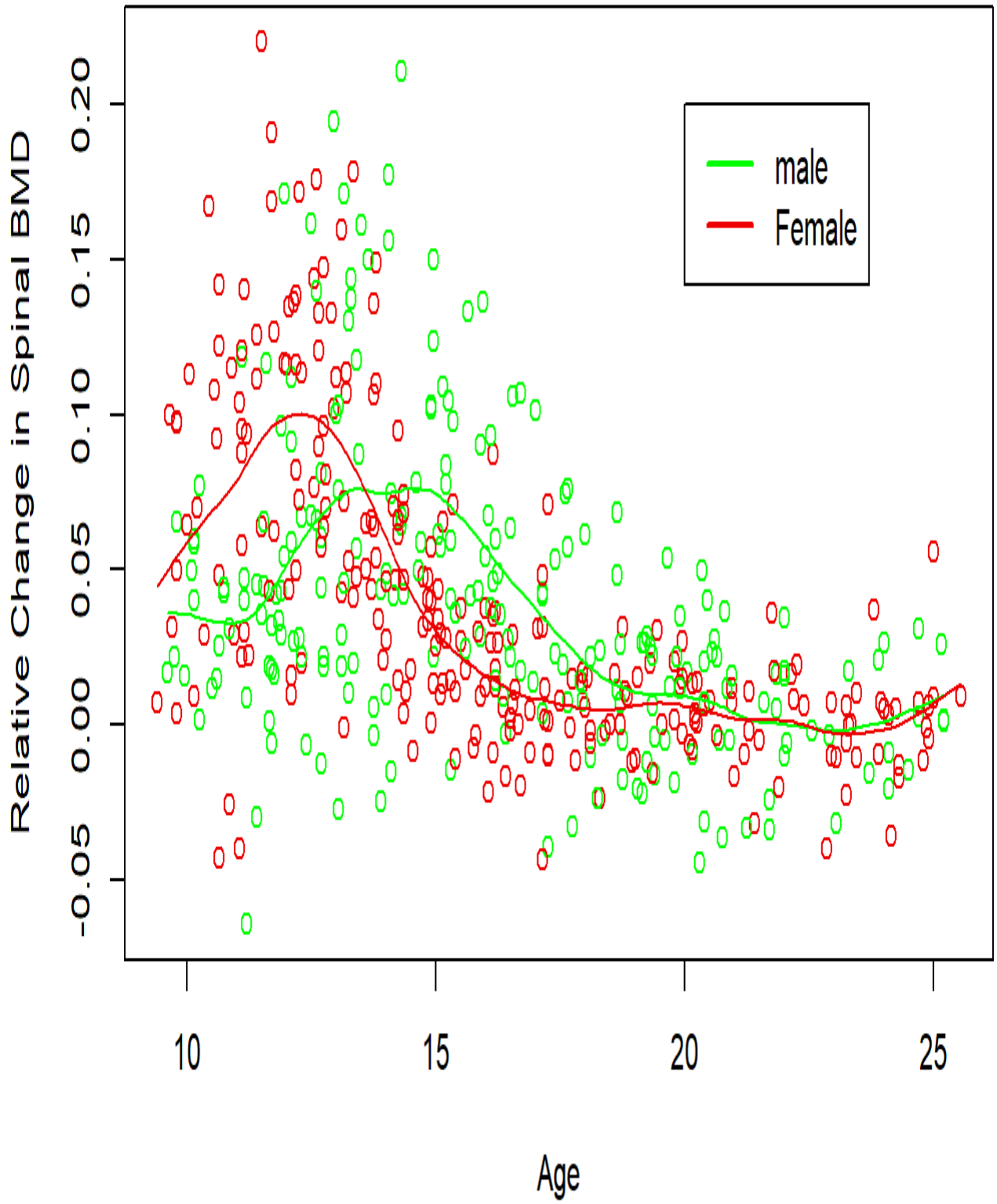
Spline Models

Since there seems to be some nonlinear relationship between meanage and meanspnbmd, I explored some other non linear plots.

This is a plot predicting the relative spinal bone density using age for the two genders. This is a plot seen in the book Elements of Statistical learning.

```
plot(spnbmd ~ age, data=bone, col =
```

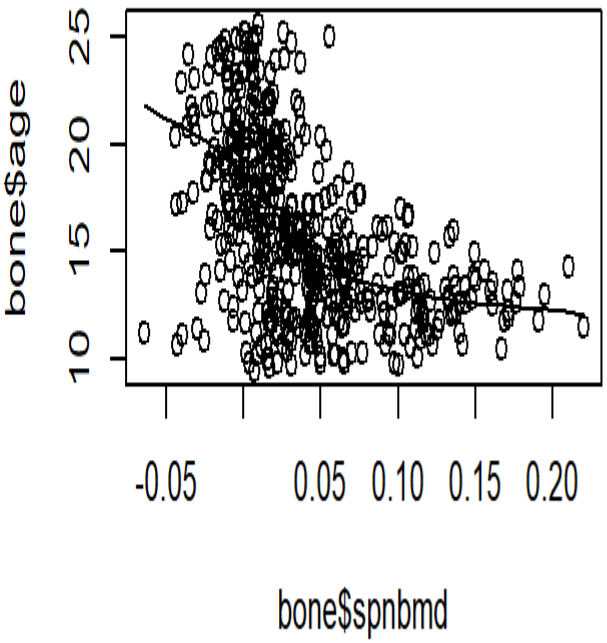
```
     ifelse(gender=="male", "green", "red2"),
     xlab="Age", ylab="Relative Change in Spinal BMD")
bone.spline.male <- with(subset(bone,gender=="male"),
                smooth.spline(age, spnbmd,df=12))
bone.spline.female <- with(subset(bone, gender=="female"),
                smooth.spline(age, spnbmd, df=12))
lines(bone.spline.male, col="green")
lines(bone.spline.female, col="red2")
legend(20,0.20, legend=c("male", "Female"), col=c("green", "red2"),
          lwd=2)
```

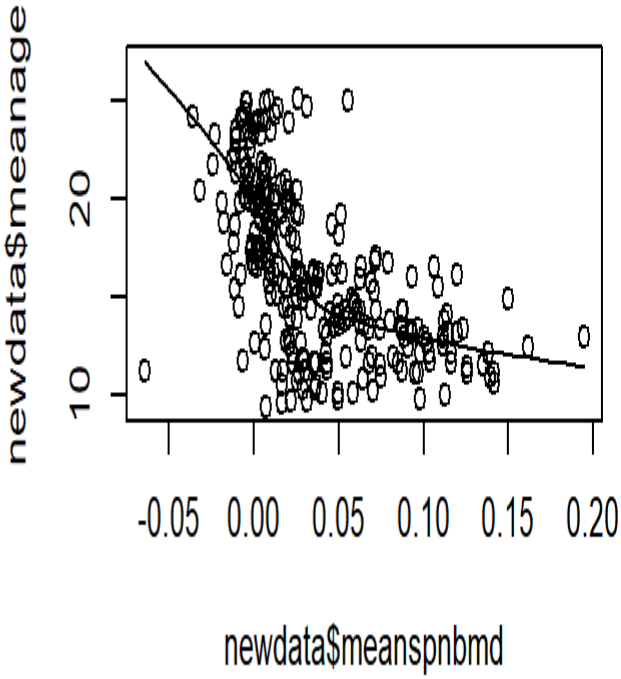


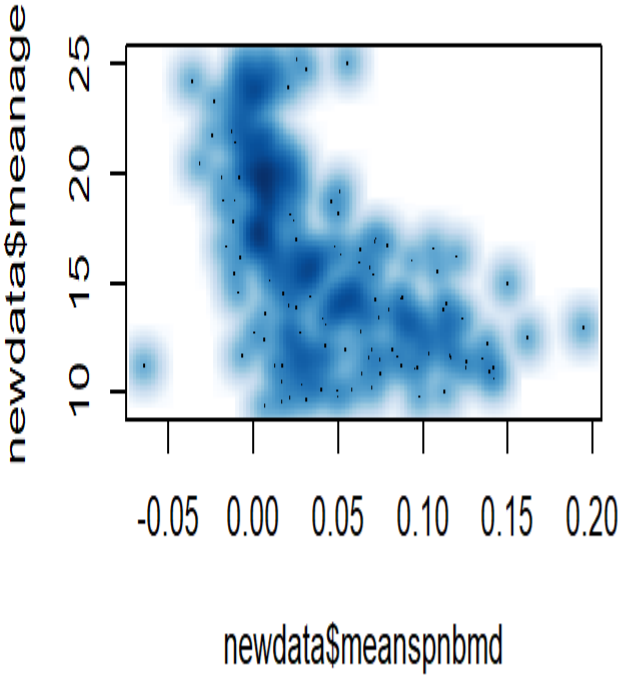
This plot showed that the relative change in spinal BMD was higher in younger females than males which reduces as age increase. However, after age 20, the relative change in BMD becomes the same in both genders.

Smooth scatter

```
par(mfrow=c(2,2))
scatter.smooth(bone$spnbmd,bone$age)
scatter.smooth(newdata$meanspnbmd,newdata$meanage)
smoothScatter(newdata$meanspnbmd,newdata$meanage)
```



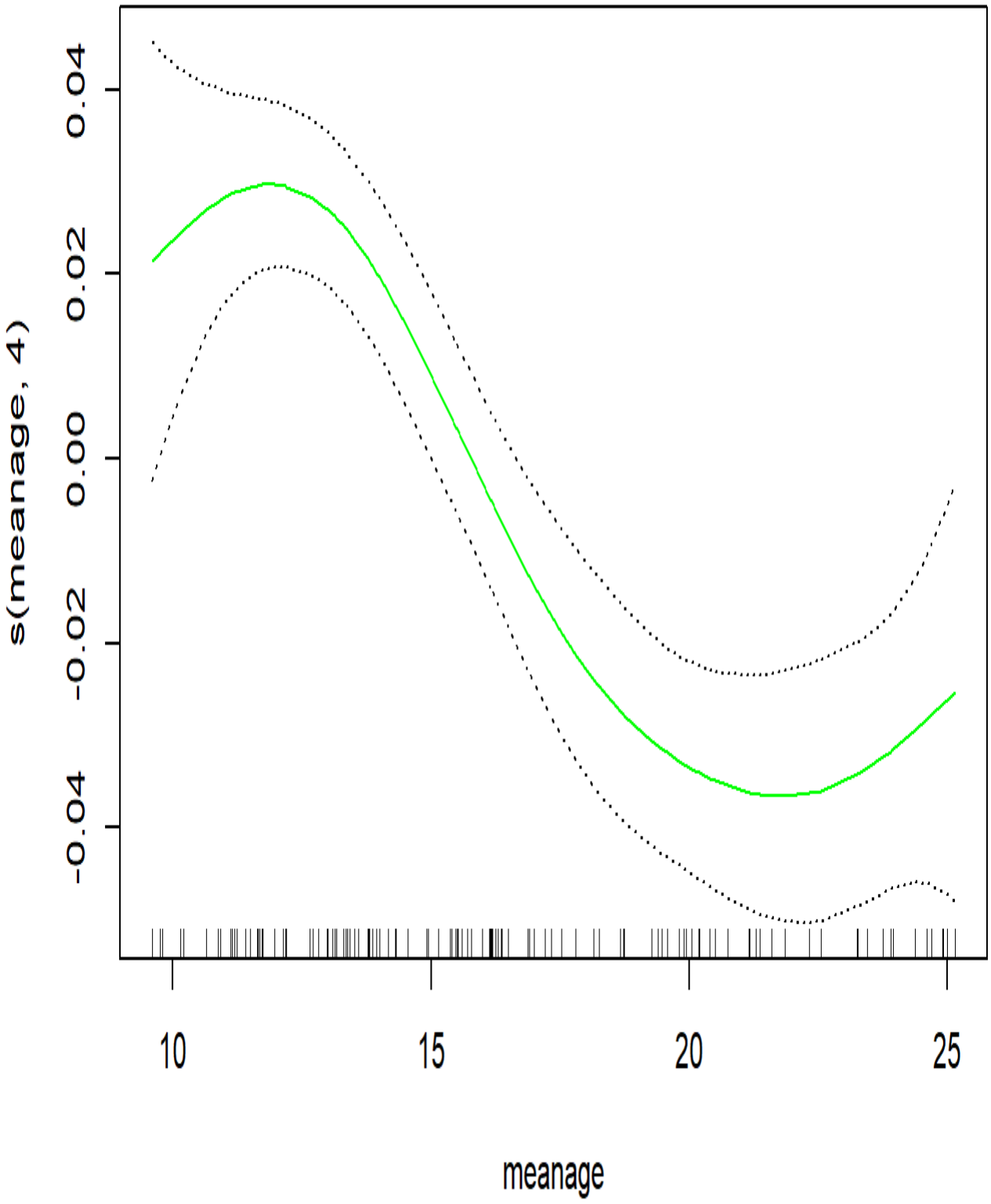




These plots show that there exist a non linear relationship between the meanspnbmd and the meanage.There seemed to be a change around a mean relative change in spinal bone at 0.025 and at age 15.

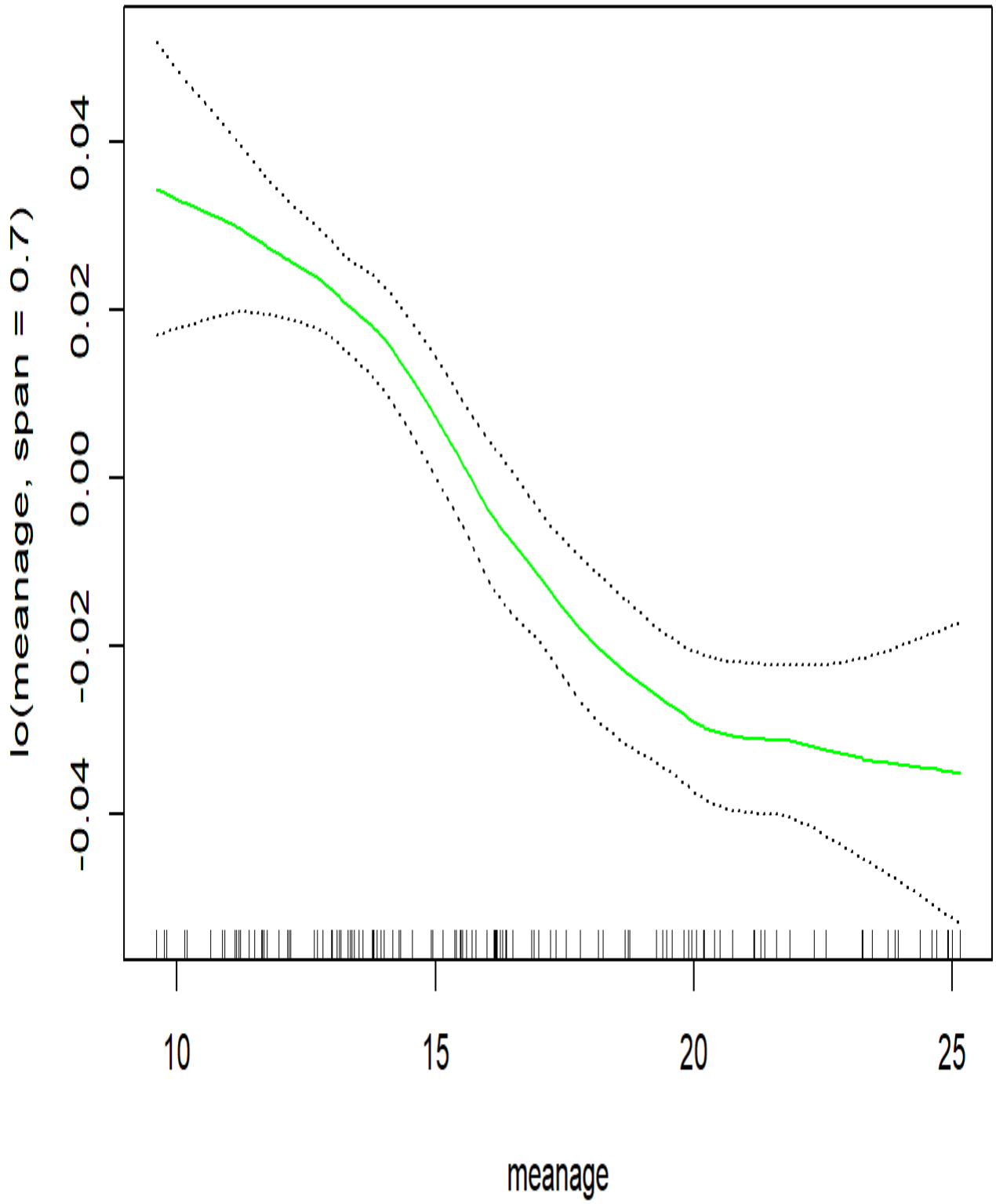
GAM model

```
library(gam)

## Loading required package: splines

## Loaded gam 1.16

library(akima)

gal = lm(meanspnbmd ~ ns(meanage, 4) , data = btrain)
gam1= gam(meanspnbmd ~ s(meanage, 4) , data = btrain)

## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored

anova(gal, gam1, test = "F")

## Analysis of Variance Table
##
## Model 1: meanspnbmd ~ ns(meanage, 4)
## Model 2: meanspnbmd ~ s(meanage, 4)
##   Res.Df     RSS         Df  Sum of Sq F Pr(>F)
## 1    127 0.14515
## 2    127 0.14597 6.3896e-05 -0.0008257

summary(gam1)

##
## Call: gam(formula = meanspnbmd ~ s(meanage, 4), data = btrain)
## Deviance Residuals:
##       Min        1Q    Median        3Q       Max
## -0.130941 -0.020879 -0.002483  0.013927  0.102054
##
## (Dispersion Parameter for gaussian family taken to be 0.0011)
##
##     Null Deviance: 0.2374 on 131 degrees of freedom
## Residual Deviance: 0.146 on 126.9999 degrees of freedom
## AIC: -511.942
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                Df   Sum Sq  Mean Sq F value    Pr(>F)
## s(meanage, 4)   1 0.073416 0.073416  63.874 6.992e-13 ***
## Residuals     127 0.145973 0.001149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##               Npar Df Npar F    Pr(F)
## (Intercept)
## s(meanage, 4)       3 5.2322 0.001943 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

preds = predict(gam1, newdata = btest)
gammse=mean((preds-ytest)^2)
gammse

## [1] 0.0009979

plot(gam1, se = TRUE, col = "green")
```



The output of the gam model showed that the coefficent of the polynomial of degree 4 was highly significant.

```
gamlo = gam(meanspnbmd~  lo(meanage, span = 0.7) , data = btrain)

## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored

plot(gamlo, se = TRUE, col = "green")
```




```
predslo = predict(gamlo, newdata = btest)
```

```
## Warning in gam.lo(data[["lo(meanage, span = 0.7)"]], z, w, span = 0.7,
## degree = 1, : eval 9.4
```

```
## Warning in gam.lo(data[["lo(meanage, span = 0.7)"]], z, w, span = 0.7,
## degree = 1, : lowerlimit 9.5222

## Warning in gam.lo(data[["lo(meanage, span = 0.7)"]], z, w, span = 0.7,
## degree = 1, : extrapolation not allowed with blending

gamlomse=mean((predslo-ytest)^2)
gamlomse

## [1] 0.001052628
```

Comparism of TEST MSE

	Linear Model	Polynomial Regression	Ridge Regression	Lasso Model	Generalised Additive Model	LOESS
Test Errors	0.0011552	0.0009968	0.0010252	0.0010147	0.0009979	0.0010526

Comparing the test MSE of each of the models used it is realised that all of them are relatively low with the heighest test mse being the linear model. This may be due to the fact that the assumptions of the linear model were violated. The polynomial regression gave the lowest test MSE. Therefore, The polynomial regression model is selected as the best model to predict the meanspnmb.

Summary

The aim of this study was to ascertain the relationship between the variables in the bone data and to find the best model to predict the spine bone mineral density.However, since the data was contained uneven visits by individuals. I decided to use the mean of the repeated measures.
After the data was reorganised into the mean of the variables for each individual, it was realised that there existed a relationship between the age and spinal bone mineral density of the adolescents. However, there was not much difference in the bone mineral density of each gender. Also, there was no difference in the distribution of age in both genders.

The data was then split into equal sized train and test data, 132.

After the analysis, it was realised there existed a nonlinear negative relationship between bone density and age.The polynomial regression model chose a polynomial of degree 4 as the best model.

The results showed that there was a change in the spinal BMD around age 15.

All the models showed that gender was not a significant factor in predicting bone mineral density. Thus, there was no significant difference in the bone mineral density of both genders.

Conclusion

With regards to the objective of this study.

1. Age was the only significant variable in the prediction of spinal bone mineral density.Age and spinal BMD had a negative nonlinear association.

2. The model with the lowest Test error rate and hence best for the prediction was the polynomial model with degree 4.

References

Andrea W., ://www.coursera.org/learn/truthinourbones-osteoarchaeology-archaeology

Kjetil B. Halvorsen (2019). ElemStatLearn: Data Sets, Functions and Examples from the Book: â The Elements of Statistical Learning, Data Mining, Inference, and Predictionâ by Trevor Hastie, Robert Tibshirani and Jerome Friedman. R package version 2015.6.26.1. https://CRAN.R-project.org/package=ElemStatLearn