HW 6

Issac Li (zl368)

3/9/2017

```
bone <- read.csv("~/Documents/STAT665/HW6/bone.csv",header = T,stringsAsFactors = FALSE)
```

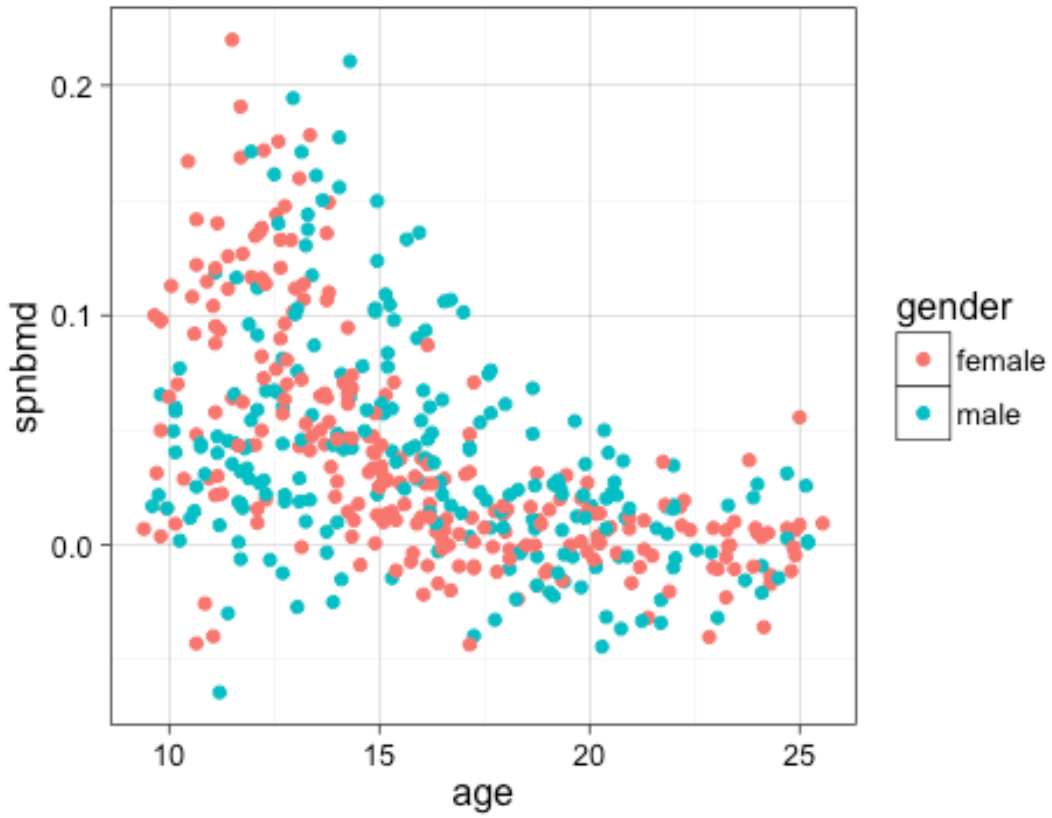
Part I

Begin with a plot of spnbmd against age, colorcoded by gender.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
p1<-qplot(data = bone,x = age, y=spnbmd,color=gender,group=idnum,geom = "point")+theme_linedraw()+
  ggtitle("Plot of `spnbmd` against `age`, color-coded by gender")
p1
```



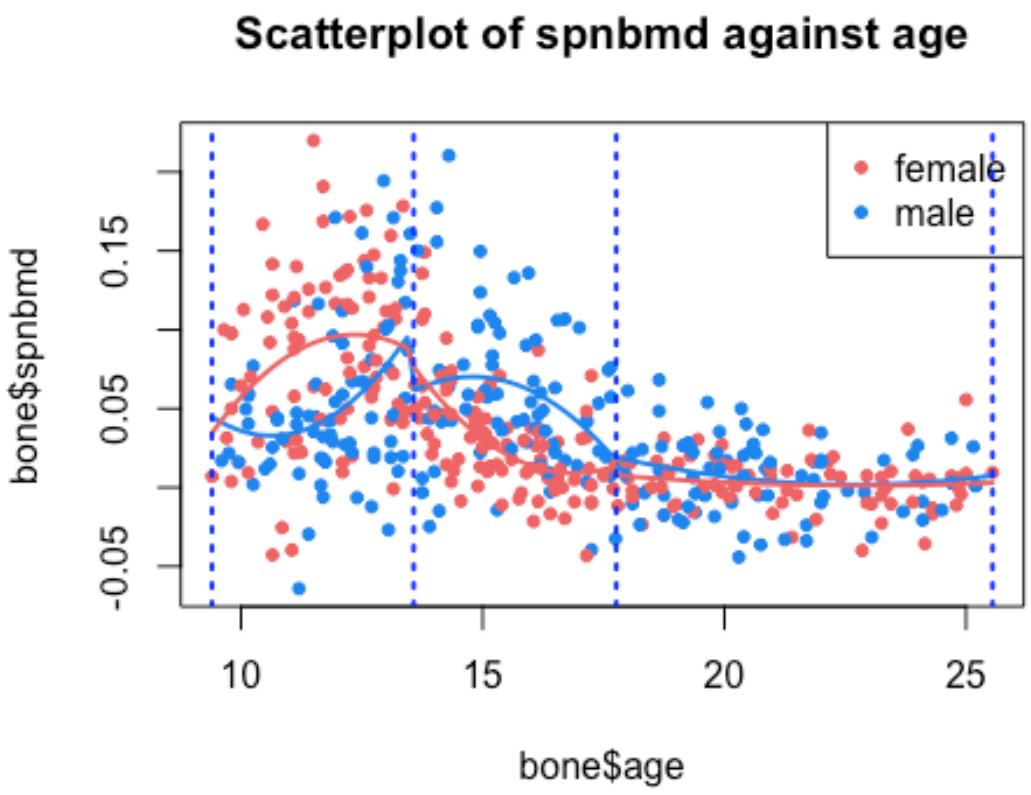
There is a little bit difference in the trend of bone mineral density between Male and Female.

Part II

For the plots in this section, I excluded confidence intervals (but I have the codes here) for the clarity of the plots.

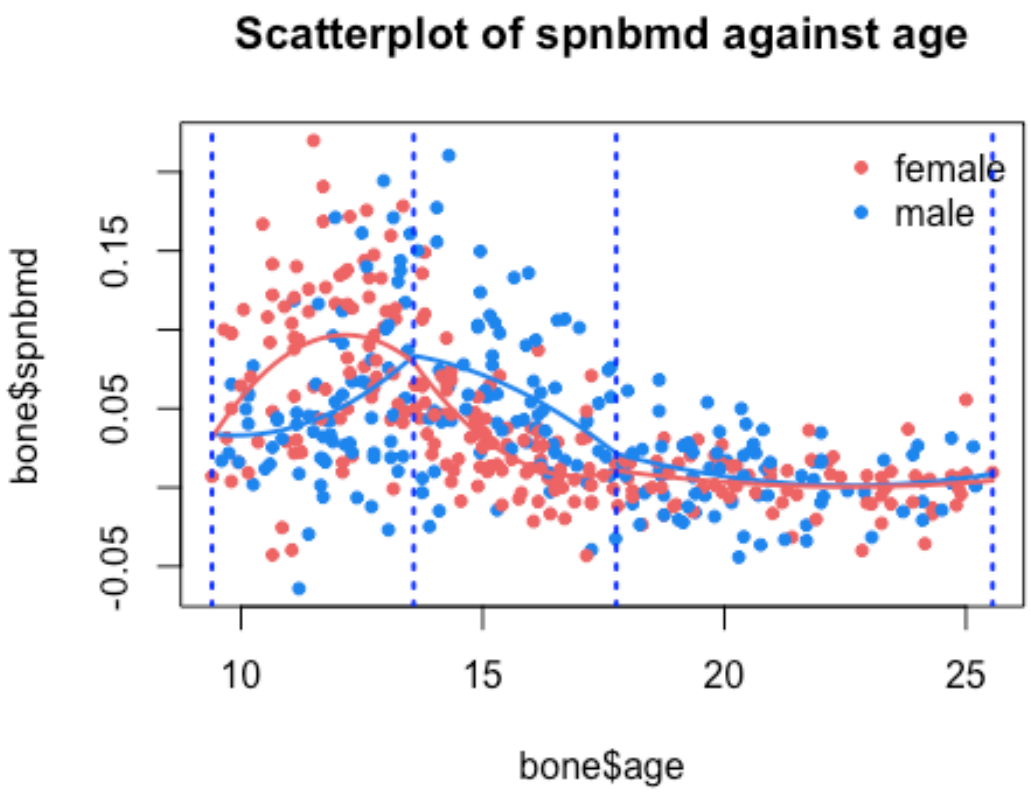
```
bone$age_group=cut(bone$age,quantile(bone$age, c(0, 0.33, 0.67, 1)), include.
lowest = TRUE)

bds <- cbind(lower = quantile(bone$age, c(0, 0.33, 0.67, 1))[1:3],
                   upper = quantile(bone$age, c(0, 0.33, 0.67, 1))[2:4])

table1<-cbind(bds,table(bone$age_group))
colnames(table1)[3]<-"count"

bds <- c(bds[,1],bds[nrow(bds),2])

# Piecewise quadratic

model.a1<-lm(spnbmd ~ age_group*poly(age,2,raw=T),data = bone,subset = bone$g
ender=="male")
model.a2<-lm(spnbmd ~ age_group*poly(age,2),data = bone,subset = bone$gender=
="female")

# no of coefficients
length(coefficients(model.a1))

## [1] 9

plot(bone$age, bone$spnbmd, col = ifelse(bone$gender == "female", "indianred2
", "dodgerblue2"),
main = "Scatterplot of spnbmd against age",pch=20)
legend("topright", legend = c("female", "male"), col = c("indianred2", "dodge
rblue2"), pch = c(20, 20))

# For Male
curve(predict(model.a1,data.frame(age_group = cut(x, breaks = bds, include.lo
west = T),
                                          age = x)), lwd = 2, col = "dodgerbl
ue2", add=TRUE)

# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                           age = x),
#               interval = "confidence")[,3], lwd = 2,lty = 2, col = "deepsky
blue3", add = TRUE)
#
# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
```

```r
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,2], lwd = 2,lty = 2, col = "deepsky
blue3", add = TRUE)


# For Female
curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.lo
west = T),
                                                age = x)), lwd = 2, col = "indianre
d2", add=TRUE)

# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,3], lwd = 2,lty = 2, col = "dark or
ange", add = TRUE)
#
# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,2], lwd = 2,lty = 2, col = "dark or
ange", add = TRUE)


abline(v = bds, lwd = 2, lty = 3, col = "blue")
```



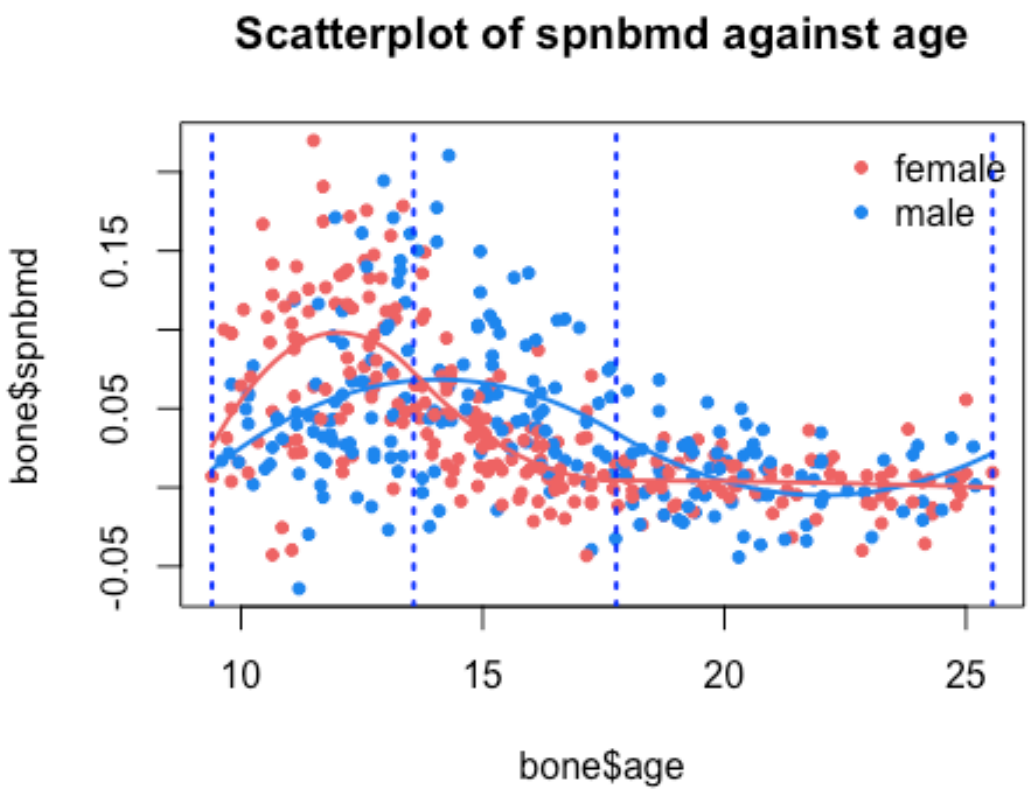

```
model.b1 <- lm(spnbmd ~ poly(age, 2, raw = TRUE) +
                           poly(pmax(I(age - bds[2]), 0), 2)
                         + poly(pmax(I(age - bds[3]), 0), 2),
                         data = bone, subset=bone$gender == 'male')

model.b2 <- lm(spnbmd ~ poly(age, 2, raw = TRUE) +
                           poly(pmax(I(age - bds[2]), 0), 2)+
                           poly(pmax(I(age - bds[3]), 0), 2),
                         data = bone, subset=bone$gender == 'female')

# Numer of coefficints
length(coef(model.b1))

## [1] 7

plot(bone$age, bone$spnbmd, col = ifelse(bone$gender == "female", "indianred2
", "dodgerblue2"),
main = "Scatterplot of spnbmd against age",pch=20)
legend("topright", legend = c("female", "male"), bty="n",col = c("indianred2"
, "dodgerblue2"), pch = c(20, 20))

# For Male
curve(predict(model.b1,data.frame(age_group = cut(x, breaks = bds, include.lo
```

```r
west = T),
                                                age = x)), lwd = 2, col = "dodgerbl
ue2", add=TRUE)

# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,3], lwd = 2,lty = 2, col = "deepsky
blue3", add = TRUE)
#
# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,2], lwd = 2,lty = 2, col = "deepsky
blue3", add = TRUE)



# For Female
curve(predict(model.b2,data.frame(age_group = cut(x, breaks = bds, include.lo
west = T),
                                                age = x)), lwd = 2, col = "indianre
d2", add=TRUE)

# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,3], lwd = 2,lty = 2, col = "dark or
ange", add = TRUE)
#
# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,2], lwd = 2,lty = 2, col = "dark or
ange", add = TRUE)



abline(v = bds, lwd = 2, lty = 3, col = "blue")
```



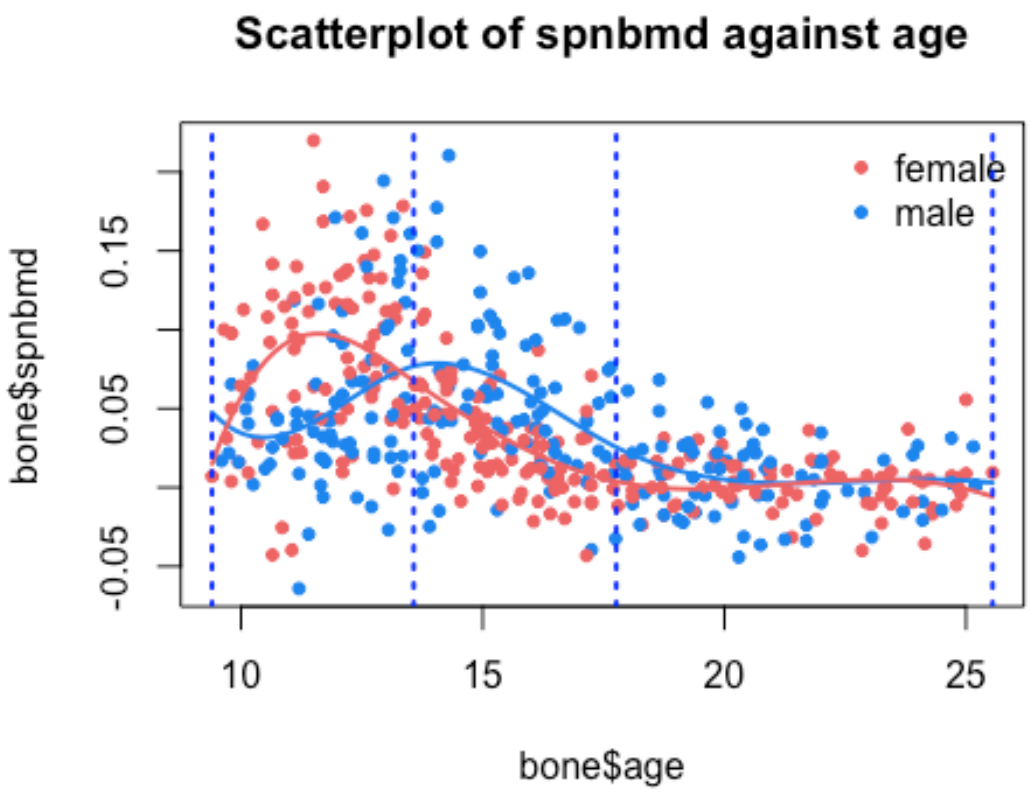

```
model.c1 <- lm(spnbmd ~ poly(age, 2, raw = TRUE) +
                                  I(pmax(age - bds[2], 0)^2) +
                                  I(pmax(age - bds[3], 0)^2),
                                data = bone, subset=bone$gender == 'male')

model.c2 <- lm(spnbmd ~ poly(age, 2, raw = TRUE) +
                                  I(pmax(age - bds[2], 0) ^2) +
                                  I(pmax(age - bds[3], 0) ^2),
                                data = bone, subset=bone$gender == 'female')

# Numer of coefficints
length(coef(model.c1))

## [1] 5

plot(bone$age, bone$spnbmd, col = ifelse(bone$gender == "female", "indianred2
", "dodgerblue2"),
main = "Scatterplot of spnbmd against age",pch=20)
legend("topright", legend = c("female", "male"), bty="n",col = c("indianred2"
, "dodgerblue2"), pch = c(20, 20))

# For Male
curve(predict(model.c1,data.frame(age_group = cut(x, breaks = bds, include.lo
```

```r
west = T),
                                                age = x)), lwd = 2, col = "dodgerbl
ue2", add=TRUE)

# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,3], lwd = 2,lty = 2, col = "deepsky
blue3", add = TRUE)
#
# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,2], lwd = 2,lty = 2, col = "deepsky
blue3", add = TRUE)



# For Female
curve(predict(model.c2,data.frame(age_group = cut(x, breaks = bds, include.lo
west = T),
                                                age = x)), lwd = 2, col = "indianre
d2", add=TRUE)

# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,3], lwd = 2,lty = 2, col = "dark or
ange", add = TRUE)
#
# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,2], lwd = 2,lty = 2, col = "dark or
ange", add = TRUE)


abline(v = bds, lwd = 2, lty = 3, col = "blue")
```




```
model.d1 <- lm(spnbmd ~ poly(age, 3, raw = TRUE) +
                                  pmax(I(age - bds[2]) ^3, 0) +
                                  pmax(I(age - bds[3]) ^3, 0),
                                data = bone, subset=bone$gender == 'male')

model.d2 <- lm(spnbmd ~ poly(age, 3, raw = TRUE) +
                                  pmax(I(age - bds[2]) ^3, 0) +
                                  pmax(I(age - bds[3]) ^3, 0),
                                data = bone, subset=bone$gender == 'female')

# Numer of coefficints
length(coef(model.d1))

## [1] 6

plot(bone$age, bone$spnbmd, col = ifelse(bone$gender == "female", "indianred2
", "dodgerblue2"),
main = "Scatterplot of spnbmd against age",pch=20)
legend("topright", legend = c("female", "male"), bty="n",col = c("indianred2"
, "dodgerblue2"), pch = c(20, 20))

# For Male
curve(predict(model.d1,data.frame(age_group = cut(x, breaks = bds, include.lo
```

```r
west = T),
                                                age = x)), lwd = 2, col = "dodgerbl
ue2", add=TRUE)

# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,3], lwd = 2,lty = 2, col = "deepsky
blue3", add = TRUE)
#
# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,2], lwd = 2,lty = 2, col = "deepsky
blue3", add = TRUE)



# For Female
curve(predict(model.d2,data.frame(age_group = cut(x, breaks = bds, include.lo
west = T),
                                                age = x)), lwd = 2, col = "indianre
d2", add=TRUE)

# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,3], lwd = 2,lty = 2, col = "dark or
ange", add = TRUE)
#
# curve(predict(model.a2,data.frame(age_group = cut(x, breaks = bds, include.
lowest = TRUE),
#                                                 age = x),
#               interval = "confidence")[,2], lwd = 2,lty = 2, col = "dark or
ange", add = TRUE)



abline(v = bds, lwd = 2, lty = 3, col = "blue")
```



Part III

Based on visual inspection, I think model b is the best model because there is a tendency for spnbmd to increase after age = 25. Only model b depicts this trend while model c and d fail to do so for both gender groups. The reason that model b seems concerning is that the derivatives at the knots are not continuous.

```
# 5-fold cross-validation
k <- 5
bone.f <- bone[bone$gender == 'female', ]
bone.m <- bone[bone$gender == 'male', ]

dim(bone.f)

## [1] 259   5

dim(bone.m)

## [1] 226   5

set.seed(111)
bone.f$fold <- c(rep(sample(x = 5),51),sample(4))
bone.m$fold <- c(rep(sample(x = 5),45),sample(1))
```

```
mses.f <- NULL
mses.m <- NULL

for (knot in 2: 7) {
  bds.temp <- round(quantile(bone$age, seq(0, 1, length.out = knot + 2)))
  form <- "spnbmd ~ poly(age, 2, raw = TRUE)"
  for (i in 2: (knot + 1)){
  form <- paste(form, " + poly(pmax(I(age - bds.temp[", i, "]), 0), 2, raw =
T)",
  sep = "")
  }
  f.temp <- as.formula(form)
  ## Cross Validation Lreg
  mse_f <- 0
  mse_m <- 0
  for (fold in 1:k){
  train.f <- bone.f[bone.f$fold != fold, ]
  train.m <- bone.f[bone.m$fold != fold, ]
  valid.f <- bone.f[bone.f$fold == fold, ]
  valid.m <- bone.f[bone.m$fold == fold, ]
  fit.train.f <- lm(f.temp, data = train.f)
  fit.train.m <- lm(f.temp, data = train.m)
  pred.valid.f <- predict(fit.train.f, newdata = valid.f)
  pred.valid.m <- predict(fit.train.m, newdata = valid.m)
  mse_f <- mse_f + mean((pred.valid.f - valid.f$spnbmd) ** 2)
  mse_m <- mse_m + mean((pred.valid.m - valid.m$spnbmd) ** 2)
  }
  mses.f <- c(mses.f, mse_f)
  mses.m <- c(mses.m, mse_m)
}

seq(2,7)[which.min(mses.m)]

## [1] 2

seq(2,7)[which.min(mses.f)]

## [1] 2
```

For my model, the MSEs for both female and male increase from knots = 2 to knots= 7, and the the optimal value of knots is 2 for both cases.