Homework6

Shan Zhong

November 13, 2018

```
library("data.table")
library("ggplot2")
data<-read.csv("http://people.stat.sc.edu/hoyen/Stat704/Data/survey.csv",sep=",")

data<-data.table(data)
```

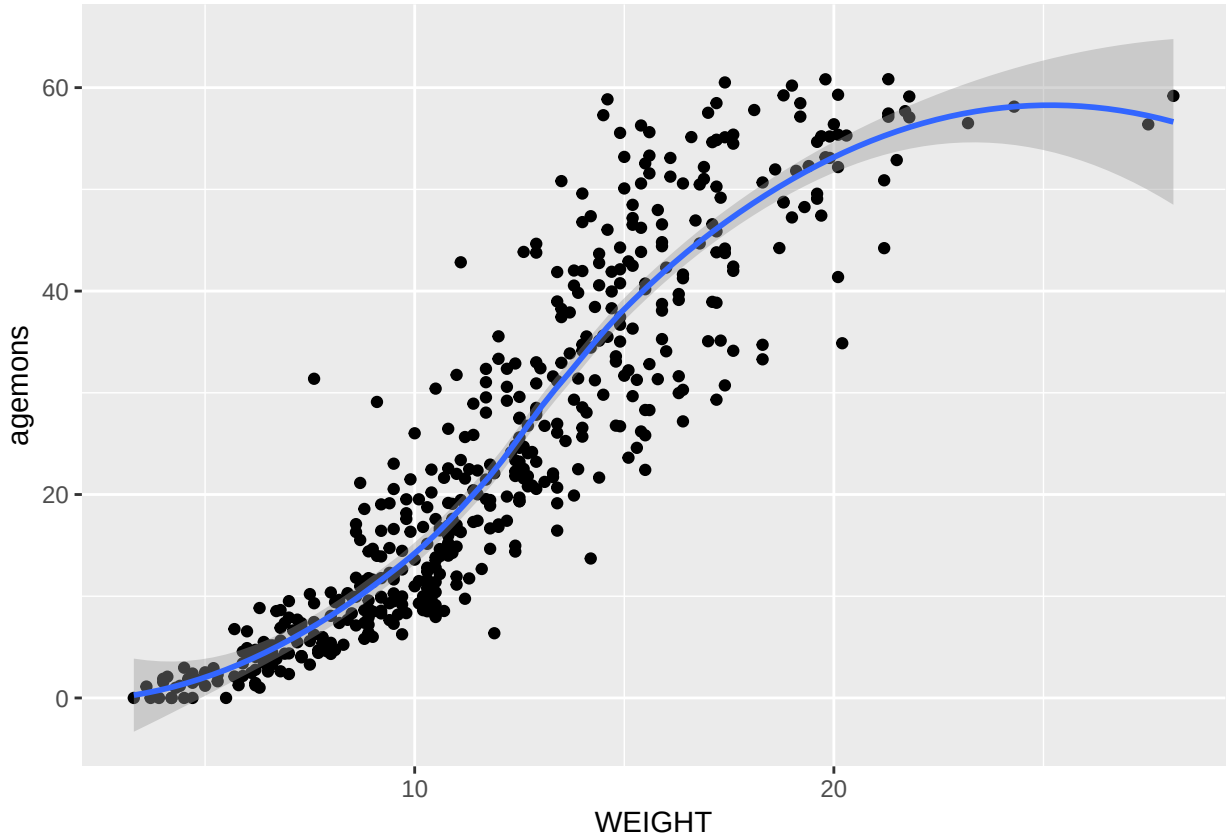
Question1

Problem 1

There seems linear relation between age and weight

```
data1<-na.omit(data[,c("WEIGHT","agemons")])
ggplot(data1,aes(x=WEIGHT,y=agemons))+geom_point()+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Problem 2

The Wight are more correlated with height, thus cause age to be insignificant

```
reg1<-lm(WEIGHT~agemons,data=data)
reg2<-lm(WEIGHT~agemons+HEIGHT,data=data)
summary(reg1)
```

```
##
## Call:
## lm(formula = WEIGHT ~ agemons, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0020 -1.1863 -0.0459  1.1028  8.2498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.515221   0.142211   45.81   <2e-16 ***
## agemons     0.225838   0.004727   47.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.823 on 493 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.8224, Adjusted R-squared:  0.822
## F-statistic:  2283 on 1 and 493 DF,  p-value: < 2.2e-16
```

```
summary(reg2)
```

```
##
## Call:
## lm(formula = WEIGHT ~ agemons + HEIGHT, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5021 -0.8057 -0.0388  0.6226  8.3958
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.62320    0.78296 -14.845   <2e-16 ***
## agemons      -0.02598    0.01132  -2.296   0.0221 *
## HEIGHT        0.29160    0.01250  23.330   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.225 on 471 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.9218, Adjusted R-squared:  0.9214
## F-statistic:  2775 on 2 and 471 DF,  p-value: < 2.2e-16
```

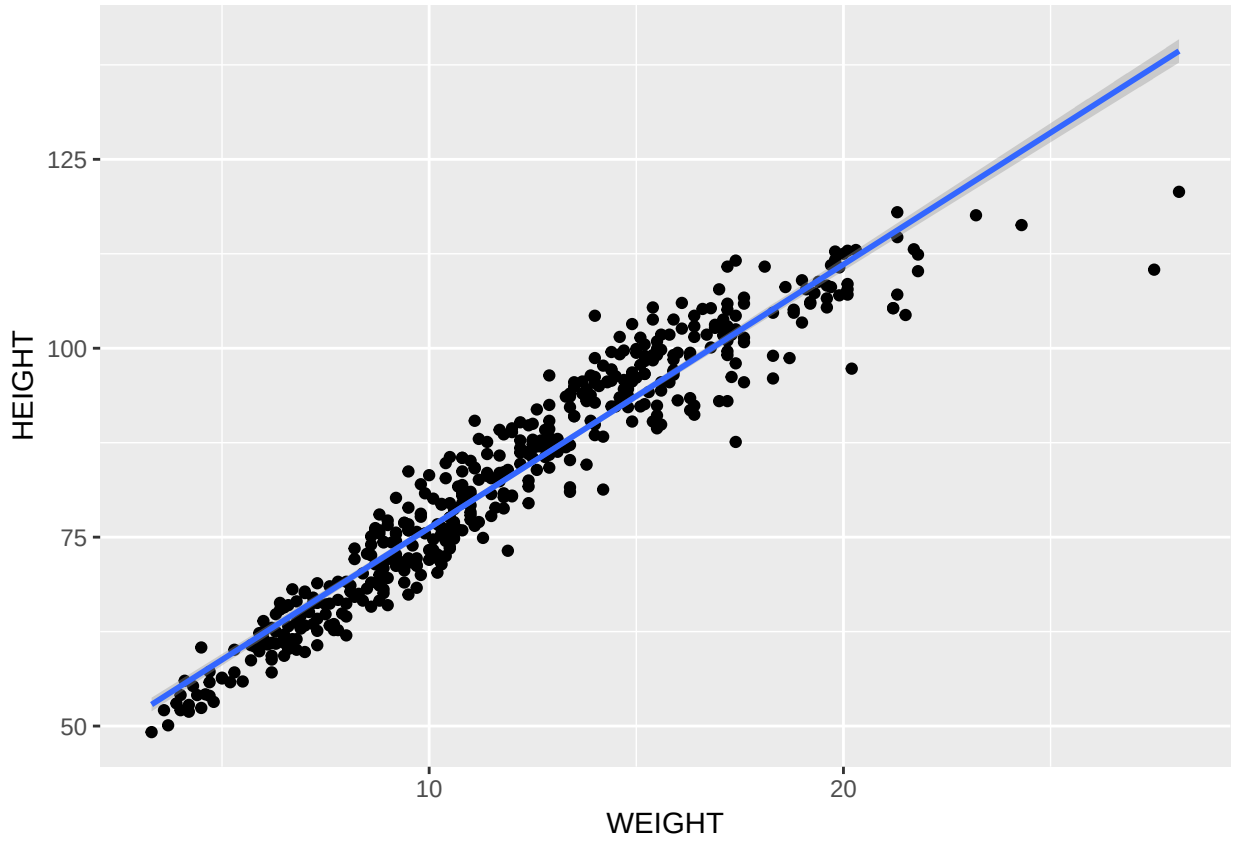
Problem3

Using single Height is a much better way to predict Weight, if we can know the height

age can be used if we do not know about height, using age and height together may cause confusion as age is not significant

10 piece from homework5

```
data2<-na.omit(data[,c("WEIGHT","HEIGHT")])
ggplot(data2,aes(x=WEIGHT,y=HEIGHT))+geom_point()+geom_smooth(method=lm)
```



Problem4

Not reasonable as the slope is changing for each decile

Question2

```
data1<-data1[agemons<60,]
cut<-seq(0,60,by=2)
cut[1]<--0.01

bin<-paste("age-bin",1:30, sep="")
```

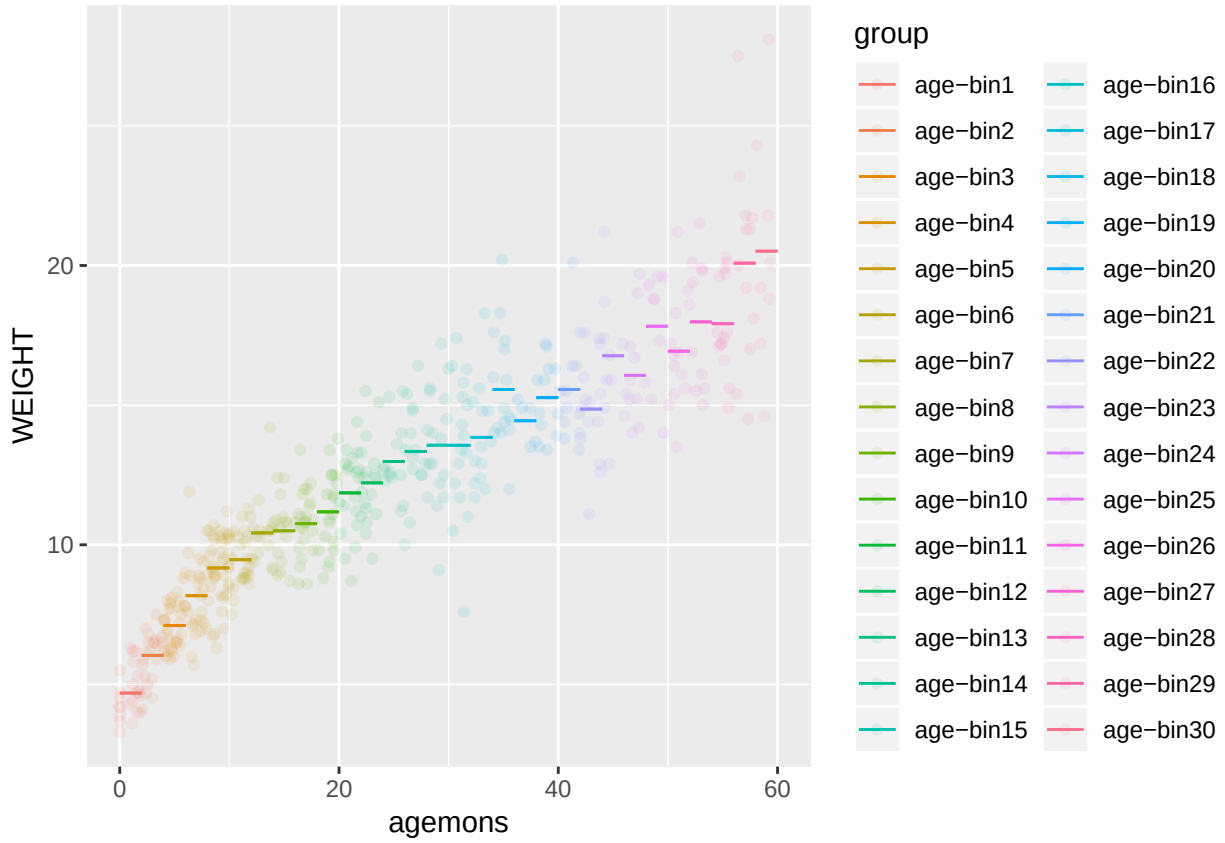
```r
summary(data1$agemons)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.31   21.24   24.30   38.16   59.30
```

```r
data1$group<-cut(data1$agemons, breaks = cut , label=bin)

cata<-model.matrix( ~ group - 1, data=data1 )

data1<-cbind(data1,cata)

table1<-data1[,mean(.SD$WEIGHT),by=c("group")]
table1<-table1[order(group)]
colnames(table1)<-c("group","WEIGHT")
table1
```

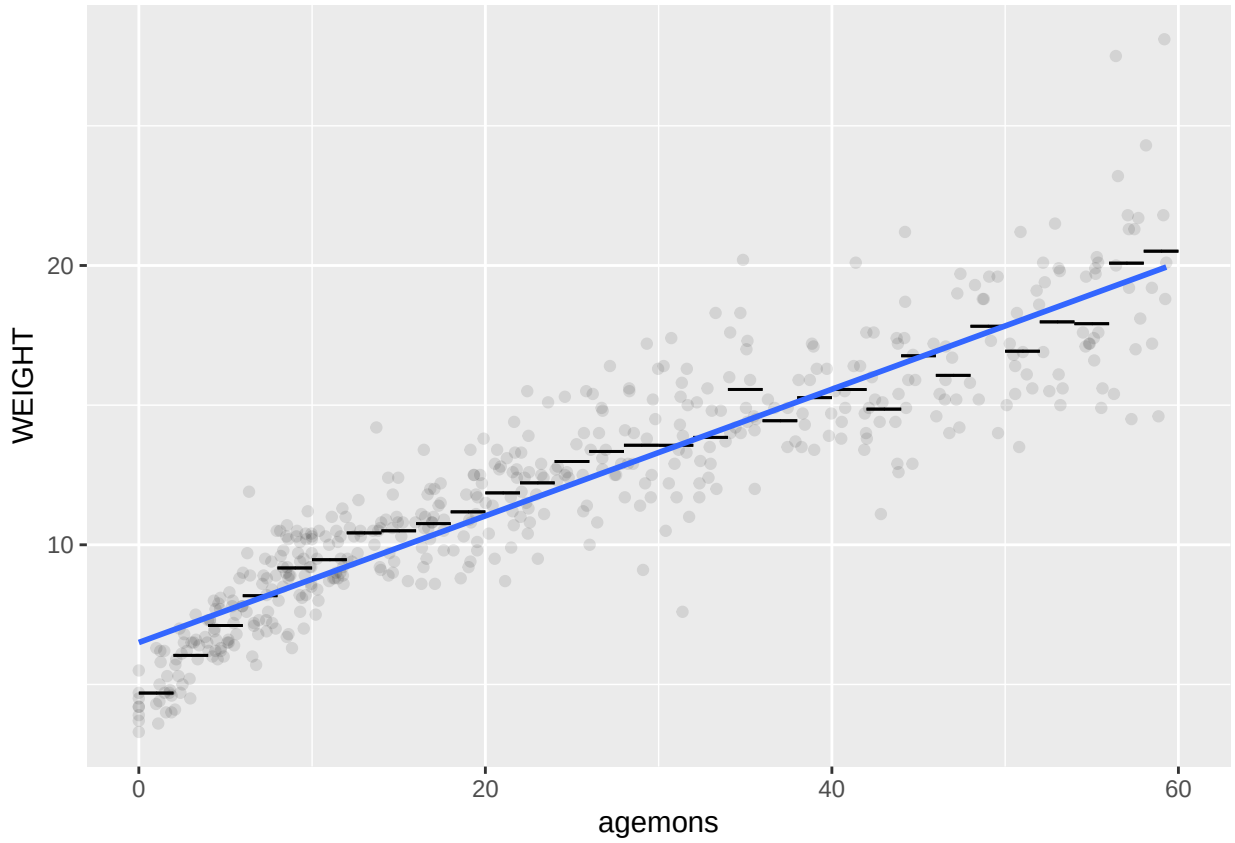
```
##         group    WEIGHT
##  1:  age-bin1  4.691304
##  2:  age-bin2  6.040909
##  3:  age-bin3  7.110000
##  4:  age-bin4  8.176000
##  5:  age-bin5  9.170000
##  6:  age-bin6  9.465217
##  7:  age-bin7 10.425000
##  8:  age-bin8 10.500000
##  9:  age-bin9 10.756522
## 10: age-bin10 11.178947
## 11: age-bin11 11.857895
## 12: age-bin12 12.216667
## 13: age-bin13 12.984615
## 14: age-bin14 13.342857
## 15: age-bin15 13.564706
## 16: age-bin16 13.562500
## 17: age-bin17 13.846154
## 18: age-bin18 15.562500
## 19: age-bin19 14.440000
## 20: age-bin20 15.266666
## 21: age-bin21 15.558333
## 22: age-bin22 14.857143
## 23: age-bin23 16.766667
## 24: age-bin24 16.066667
## 25: age-bin25 17.825000
## 26: age-bin26 16.930769
## 27: age-bin27 17.980000
## 28: age-bin28 17.914286
## 29: age-bin29 20.083333
## 30: age-bin30 20.512500
##         group    WEIGHT
```

```r
table1$agemons<-seq(1,59,by=2)
ggplot(data1,aes(x=agemons,y=WEIGHT,color=group))+geom_point(alpha=0.1)+
  geom_errorbar(data=table1,aes(ymin=WEIGHT,ymax=WEIGHT),width=2)
```

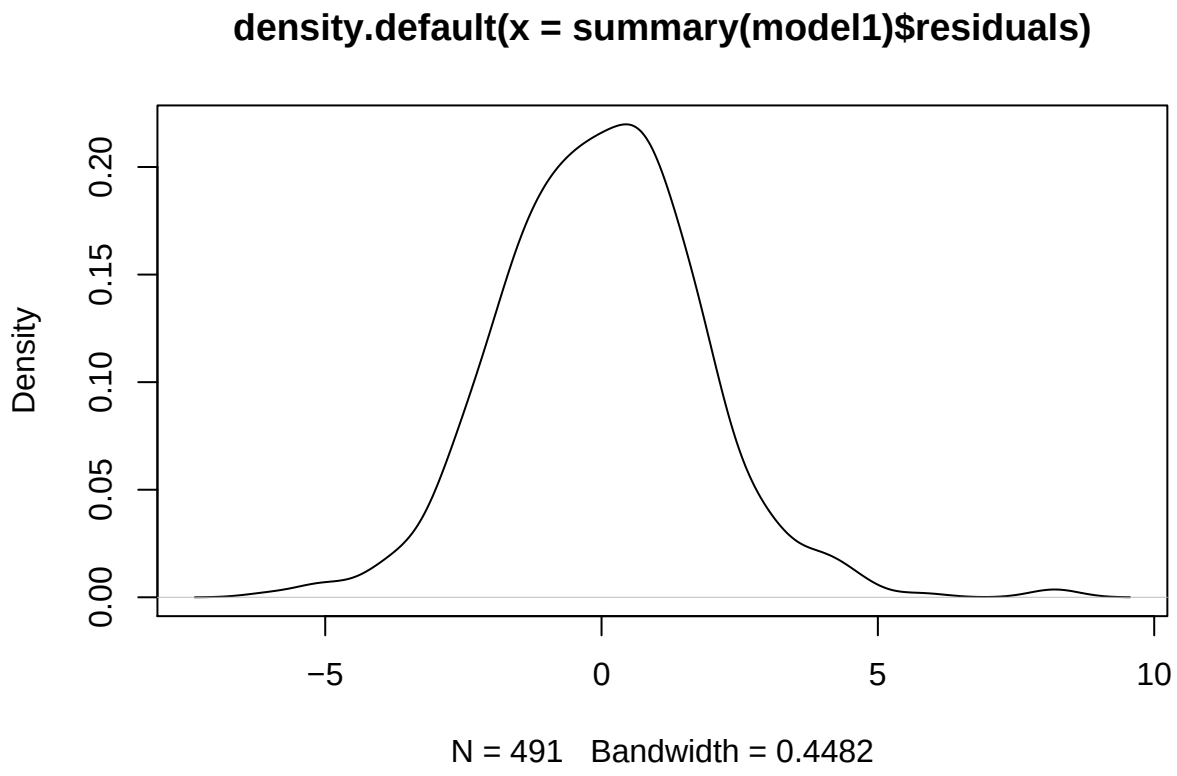


question2

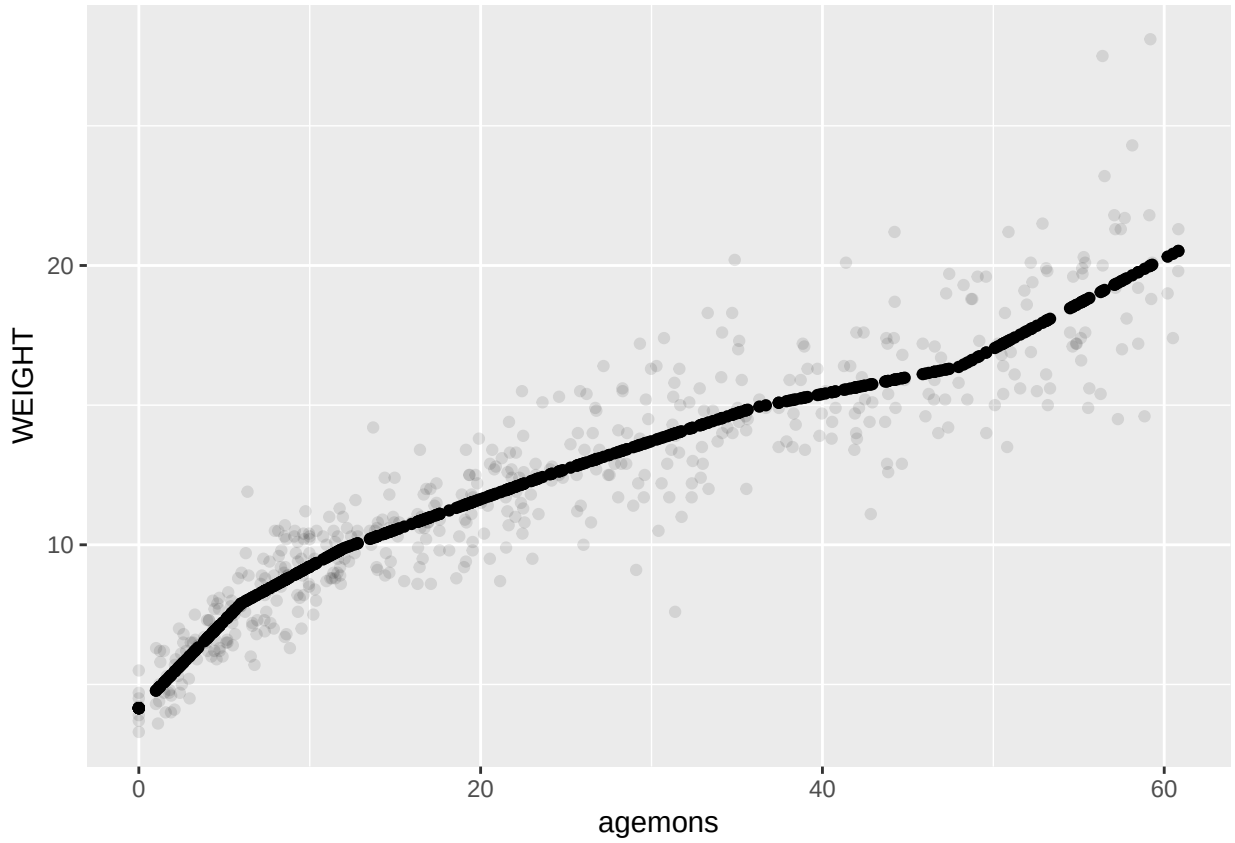
```
ggplot(data1,aes(x=agemons,y=WEIGHT))+geom_point(alpha=0.1)+
  geom_errorbar(data=table1,aes(ymin=WEIGHT,ymax=WEIGHT),width=2)+
  geom_smooth(method=lm,formula=y~x, se=FALSE)
```




```
model1<-lm(WEIGHT~agemons,data=data1)
plot(density(summary(model1)$residuals))
```




```
data1<-na.omit(data[,c("WEIGHT","agemons")])

groupbin1<-ifelse(data1$agemons > 6, data1$agemons-6, 0)
groupbin2<-ifelse(data1$agemons > 12, data1$agemons-12, 0)
groupbin3<-ifelse(data1$agemons > 24, data1$agemons-24, 0)
groupbin4<-ifelse(data1$agemons > 36, data1$agemons-36, 0)
groupbin5<-ifelse(data1$agemons > 48, data1$agemons-48, 0)
data2<-cbind(data1,groupbin1,groupbin2,groupbin3,groupbin4,groupbin5)
data2$agecen<-data2$agemons-mean(data2$agemons)

fit1<-lm(WEIGHT ~ agecen + groupbin1+groupbin2+groupbin3+groupbin4+groupbin5,data=data2)

data2$WEIGHThat<-predict(fit1,data2)

ggplot(data2,aes(x=agemons,y=WEIGHT))+geom_point(alpha=0.1)+geom_point(aes(x=agemons,y=WEIGHThat))
```



Problem4 for each additional months in age, the increase rate of WEIGHT change by the cofficient

e The spline model fit the data better

rest: professor I am so dumb I can't figure the rest out

```
library(knitr)
purl("Homework5.Rmd")
```