Homework5

Shan Zhong

2018/11/06

1 Describe the dataset

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

```
##       id region GENDER agemons WEIGHT HEIGHT measure oedema HEAD MUAC  TRI
##   1:   1   West      2   22.45   10.4   84.8       l      n   NA 14.9  6.8
##   2:   2   West      2   21.47   11.7   83.5       l      n   NA 15.4  8.2
##   3:   3   West      2    7.70    7.2   67.0       l      n   NA 14.4 12.2
##   4:   4   East      1   60.83   21.3  118.0       h      n   NA 17.0  6.6
##   5:   5   West      2   17.01   11.0   80.6       l      n   NA 15.1  7.5
##  ---
## 494: 494   West      2   11.78    8.9   74.3       l      n   NA 14.6  8.6
## 495: 495  North      2   34.13   17.6   95.5       h      n   NA 18.9 11.2
## 496: 496   East      1    2.62    6.8   60.1       l      n 39.8   NA   NA
## 497: 497  South      2   50.27   17.2  105.1       h      n   NA 17.2  7.4
## 498: 498   East      2   50.81   13.5   95.5       h      n   NA 16.1  7.6
##      SUB        SW
##   1: 5.2  5598.031
##   2: 8.4 28113.578
##   3: 7.8  2865.472
##   4: 4.5 53687.994
##   5: 7.8  5988.650
##  ---
## 494: 6.0  5573.943
## 495: 7.2 13626.472
## 496:  NA 17962.687
## 497: 6.2 15941.096
## 498: 4.5 54077.649
```

```
summary(data)
```

```
##        id          region        GENDER        agemons
##  Min.   :  1.0   East :128   Min.   :1.00   Min.   : 0.000
##  1st Qu.:125.2   North:125   1st Qu.:1.00   1st Qu.: 9.338
##  Median :249.5   South:103   Median :1.00   Median :21.480
##  Mean   :249.5   West :142   Mean   :1.48   Mean   :24.597
##  3rd Qu.:373.8               3rd Qu.:2.00   3rd Qu.:38.657
##  Max.   :498.0               Max.   :2.00   Max.   :60.830
##
##      WEIGHT          HEIGHT       measure oedema       HEAD
##  Min.   : 3.30   Min.   : 49.20    : 23   n:492   Min.   :35.40
##  1st Qu.: 8.90   1st Qu.: 71.35   h:162   y:  6   1st Qu.:39.42
##  Median :11.70   Median : 83.50   l:313           Median :41.35
##  Mean   :12.07   Mean   : 83.45                   Mean   :41.15
##  3rd Qu.:15.05   3rd Qu.: 96.25                   3rd Qu.:43.27
##  Max.   :28.10   Max.   :120.70                   Max.   :46.50
```

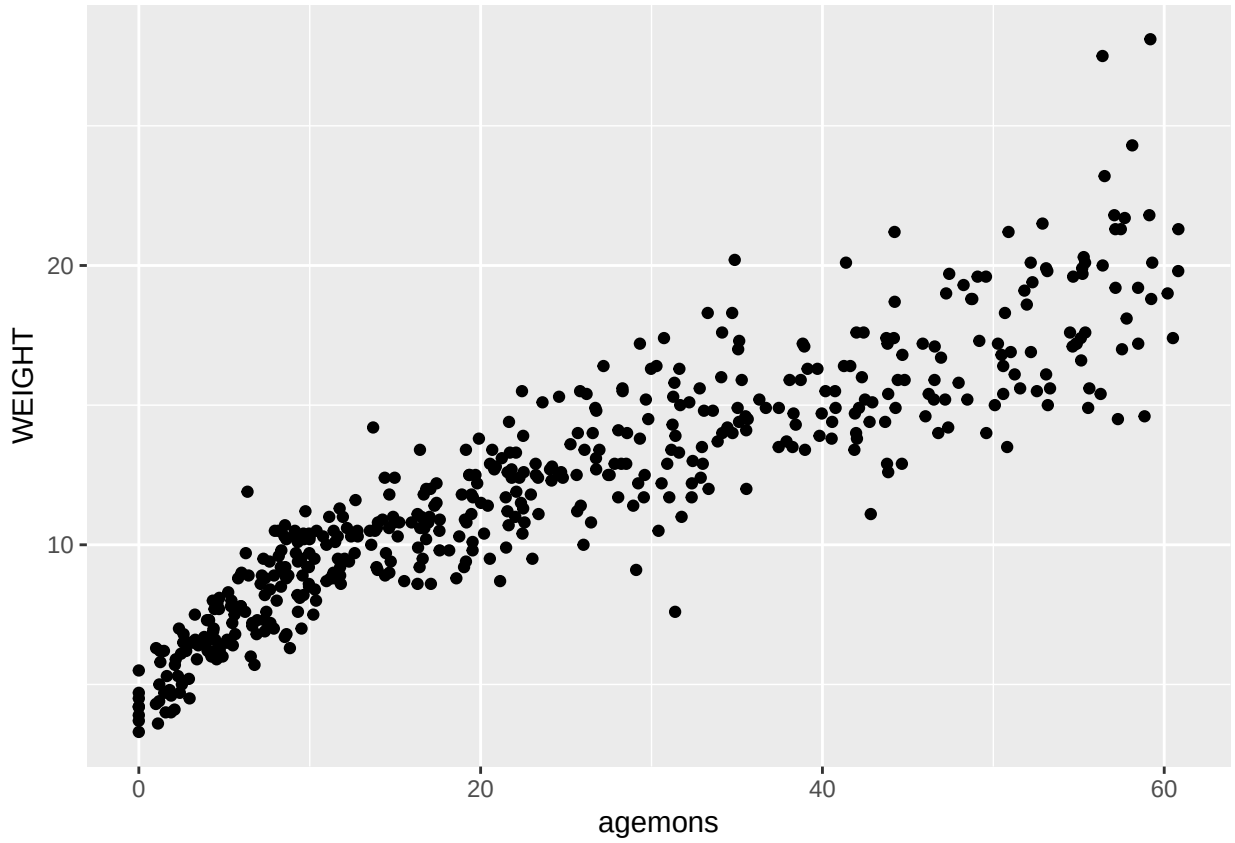
```
##  NA's   :3       NA's   :23                      NA's   :404
##       MUAC            TRI             SUB              SW
##  Min.   :10.80   Min.   : 4.00   Min.   : 3.000   Min.   : 1533
##  1st Qu.:14.70   1st Qu.: 8.20   1st Qu.: 5.800   1st Qu.: 6006
##  Median :15.70   Median : 9.90   Median : 7.000   Median : 8359
##  Mean   :15.82   Mean   :10.13   Mean   : 7.208   Mean   :13729
##  3rd Qu.:16.80   3rd Qu.:11.80   3rd Qu.: 8.200   3rd Qu.:17273
##  Max.   :22.60   Max.   :24.80   Max.   :18.200   Max.   :68578
##  NA's   :45      NA's   :48      NA's   :55
```

```
str(data)
```

```
## Classes 'data.table' and 'data.frame':   498 obs. of  13 variables:
##  $ id     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ region : Factor w/ 4 levels "East","North",..: 4 4 4 1 4 2 1 2 2 2 ...
##  $ GENDER : int  2 2 2 1 2 1 1 2 2 1 ...
##  $ agemons: num  22.4 21.5 7.7 60.8 17 ...
##  $ WEIGHT : num  10.4 11.7 7.2 21.3 11 ...
##  $ HEIGHT : num  84.8 83.5 67 118 80.6 ...
##  $ measure: Factor w/ 3 levels "","h","l": 3 3 3 2 3 2 3 3 2 3 ...
##  $ oedema : Factor w/ 2 levels "n","y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ HEAD   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MUAC   : num  14.9 15.4 14.4 17 15.1 ...
##  $ TRI    : num  6.8 8.2 12.2 6.6 7.5 ...
##  $ SUB    : num  5.2 8.4 7.8 4.5 7.8 ...
##  $ SW     : num  5598 28114 2865 53688 5989 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

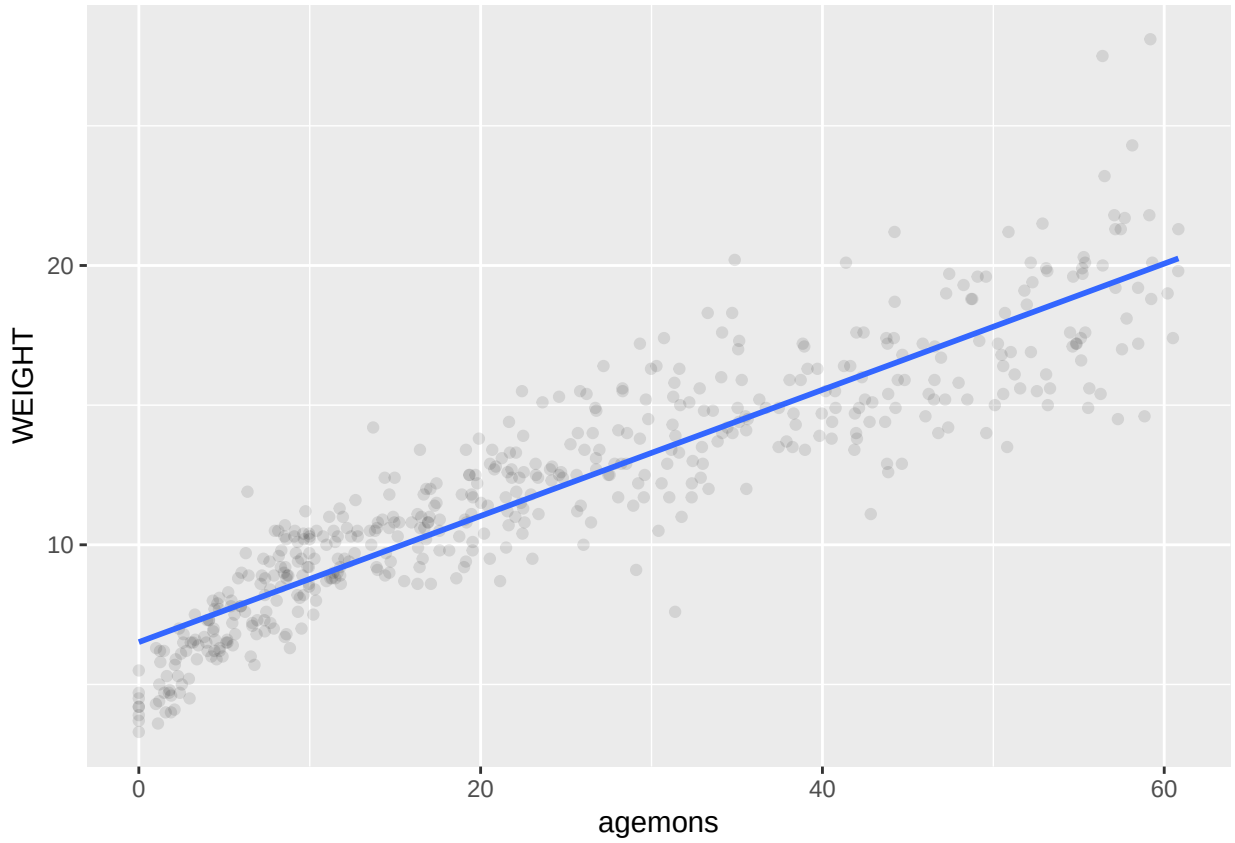
2 Plot wight against age

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



3 Regression Line

```
ggplot(data1, aes(x=agemons,y=WEIGHT))+
  geom_point(alpha=0.1)+geom_smooth(method='lm',formula=y~x, se=FALSE)
```



4 Regression analysis

```
output<-lm(WEIGHT~agemons,data=data1)
summary(output)
```

```
##
## Call:
## lm(formula = WEIGHT ~ agemons, data = data1)
##
## 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
## Multiple R-squared:  0.8224, Adjusted R-squared:  0.822
## F-statistic:  2283 on 1 and 493 DF,  p-value: < 2.2e-16
```

5 Assumptions

Weight = 6.51+agemons*0.23 ### Linearity,Constant variance , Independence ,Weak exogeneity ,Lack of perfect multicollinearity (from wiki)

6 it seems variance is not constant, as age goes larger , variance of weight become larger

there is not perfect multicollinearity, agemons seems to be random variable

linear model still seems(to me) a nice fit

7(separate analysis)

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

oneyear<-data1[agemons>=0&agemons<12]
two_to_six<-data1[agemons>=12&agemons<60]

output1<-lm(WEIGHT~agemons,data=oneyear)
output2<-lm(WEIGHT~agemons,data=two_to_six)
summary(output1)
```

```
##
## Call:
## lm(formula = WEIGHT ~ agemons, data = oneyear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5230 -0.8014  0.0774  0.8026  4.2619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.59942    0.17477   26.32   <2e-16 ***
## agemons      0.47778    0.02425   19.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.064 on 161 degrees of freedom
## Multiple R-squared:  0.7068, Adjusted R-squared:  0.7049
## F-statistic:   388 on 1 and 161 DF,  p-value: < 2.2e-16
```

```
summary(output2)
```

```
##
## Call:
## lm(formula = WEIGHT ~ agemons, data = two_to_six)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2240 -1.2652 -0.0951  1.1193  8.7921
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.636011   0.280807   27.19   <2e-16 ***
## agemons     0.197194   0.007802   25.27   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.955 on 326 degrees of freedom
## Multiple R-squared:  0.6621, Adjusted R-squared:  0.6611
## F-statistic: 638.8 on 1 and 326 DF,  p-value: < 2.2e-16
```

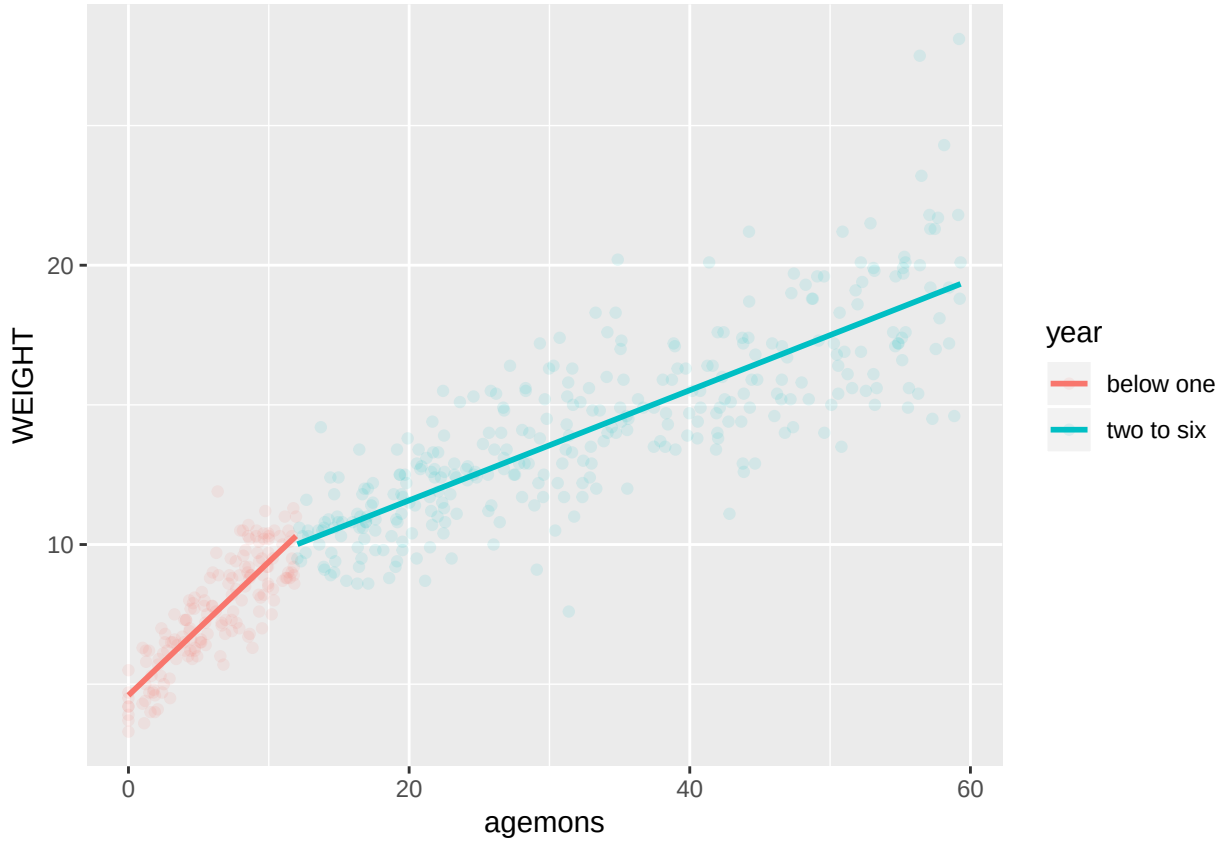
8 significant difference

the sploe and intercept are all different

so baye of difffernt age should be treat differently

```
data2<-data1[agemons>=0&agemons<=60]
data2[,year:=(ifelse(agemons < 12, "below one", "two to six"))]

ggplot(data2, aes(x=agemons,y=WEIGHT,color=year))+
  geom_point(alpha=0.1)+geom_smooth(method='lm',formula=y~x, se=FALSE)
```

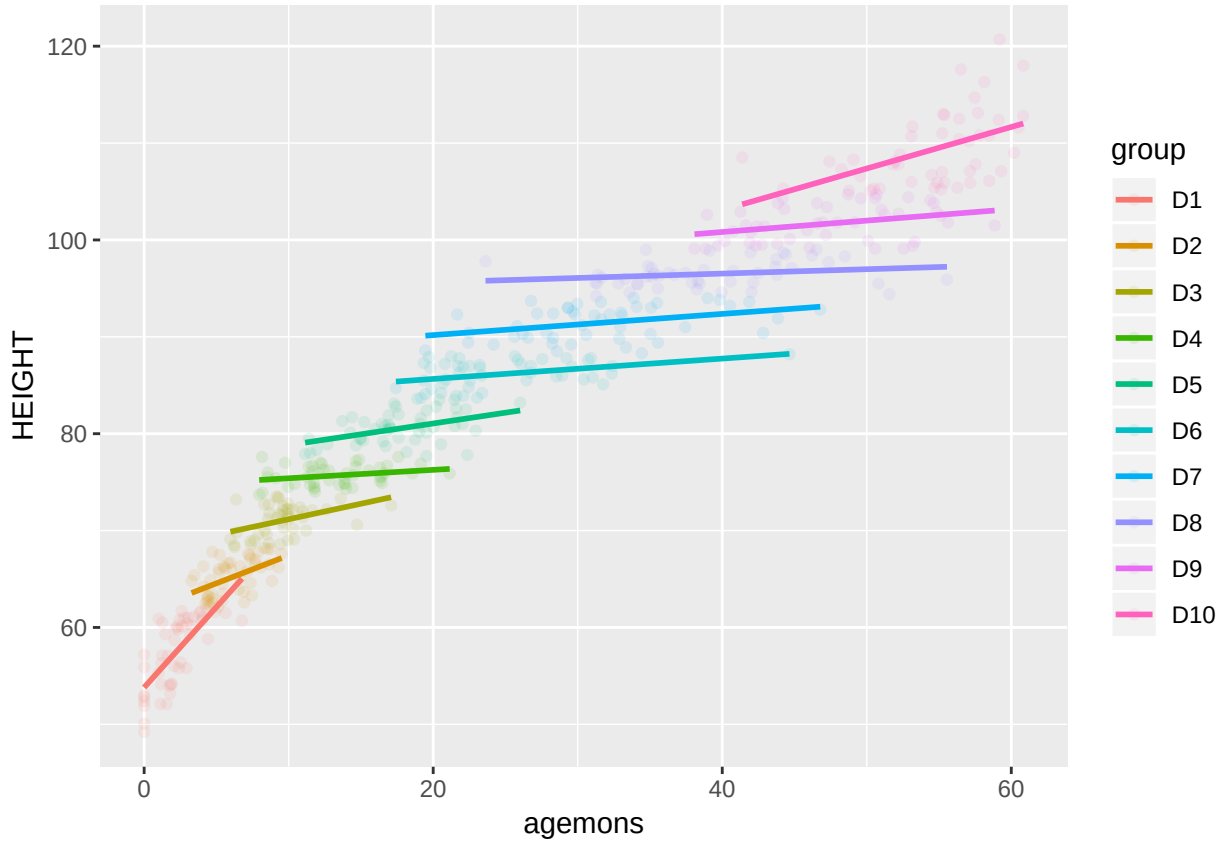


Problem 10 table

```
library("gmodels")
data1<-na.omit(data[,c("agemons","HEIGHT")])
probsvalue<-seq(0,1,by=0.1)
D<-paste("D",1:10, sep="")
cut<-quantile(data1$HEIGHT,probs = probsvalue)
cut[1]<-0
```

```
data1$group<-cut(data1$HEIGHT, breaks = cut , label=D)

ggplot(data1, aes(x=agemons,y=HEIGHT,color=group))+
  geom_point(alpha=0.1)+geom_smooth(method='lm',formula=y~x, se=FALSE)
```




```
table<-data1[,.("Intercept"=summary(lm(HEIGHT~agemons,data=.SD))$coefficients[1,1],
         "Age Slope(b1)"=summary(lm(HEIGHT~agemons,data=.SD))$coefficients[1,1],
         "se(b1)" = summary(lm(HEIGHT~agemons,data=.SD))$coefficients[2,2],
         "Residual std dev" = summary(lm(HEIGHT~agemons,data=.SD))$sigma
         ),by=c("group")][order(group)]

table
```

```
##     group Intercept Age Slope(b1)     se(b1) Residual std dev
##  1:    D1  53.80616      53.80616 0.23675045         2.554439
##  2:    D2  61.66605      61.66605 0.14171367         1.599902
##  3:    D3  67.97244      67.97244 0.10445421         1.496832
##  4:    D4  74.53460      74.53460 0.04830058         1.007346
##  5:    D5  76.59688      76.59688 0.06340201         1.456245
##  6:    D6  83.55069      83.55069 0.03680209         1.294713
##  7:    D7  87.99181      87.99181 0.03939028         1.600930
##  8:    D8  94.72020      94.72020 0.03034512         1.307400
##  9:    D9  96.11413      96.11413 0.04852296         1.753775
## 10:   D10  85.93438      85.93438 0.11539765         3.398505
```

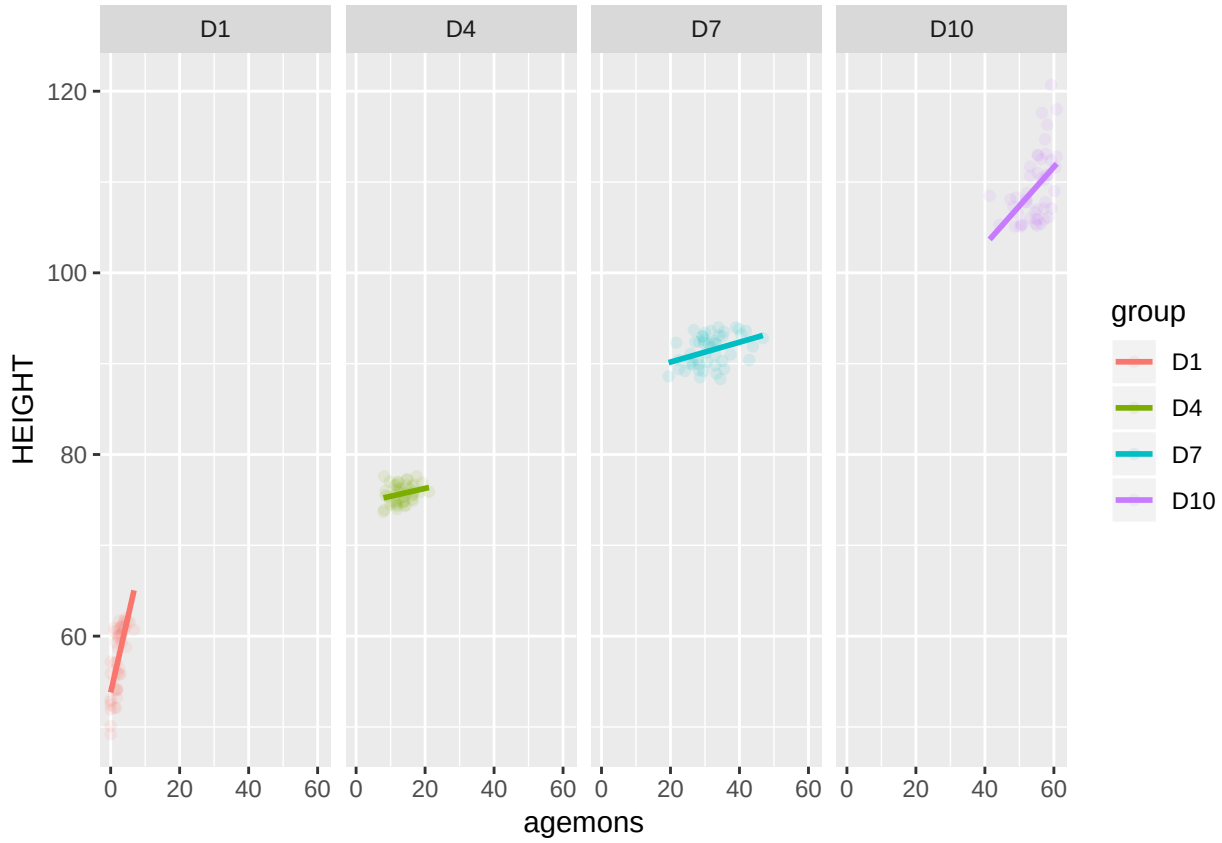
```
colMeans(table[,2:3])
```

```
##     Intercept Age Slope(b1)
##      78.28873      78.28873
```

Problem 9 plot

```
data2<-data1[data1$group %in% c("D1","D4","D7","D10")]

ggplot(data2, aes(x=agemons,y=HEIGHT,color=group))+
  geom_point(alpha=0.1)+geom_smooth(method='lm',formula=y~x, se=FALSE)+facet_grid(. ~ group)
```



11 H0: D1=D2...=D10 = 0

H1: NOt True

Model = Height = b0 + b1*X*+ *D1*b2*X*+*D2*b3*X+.....

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