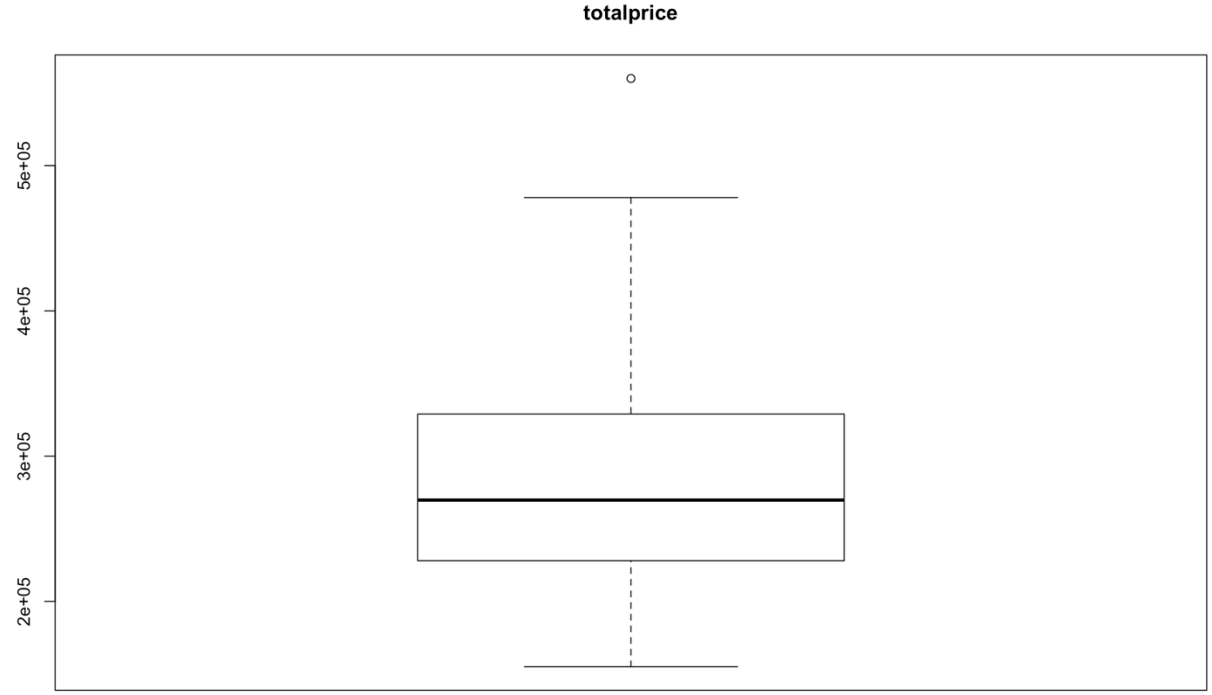
Shuting chen

HW1

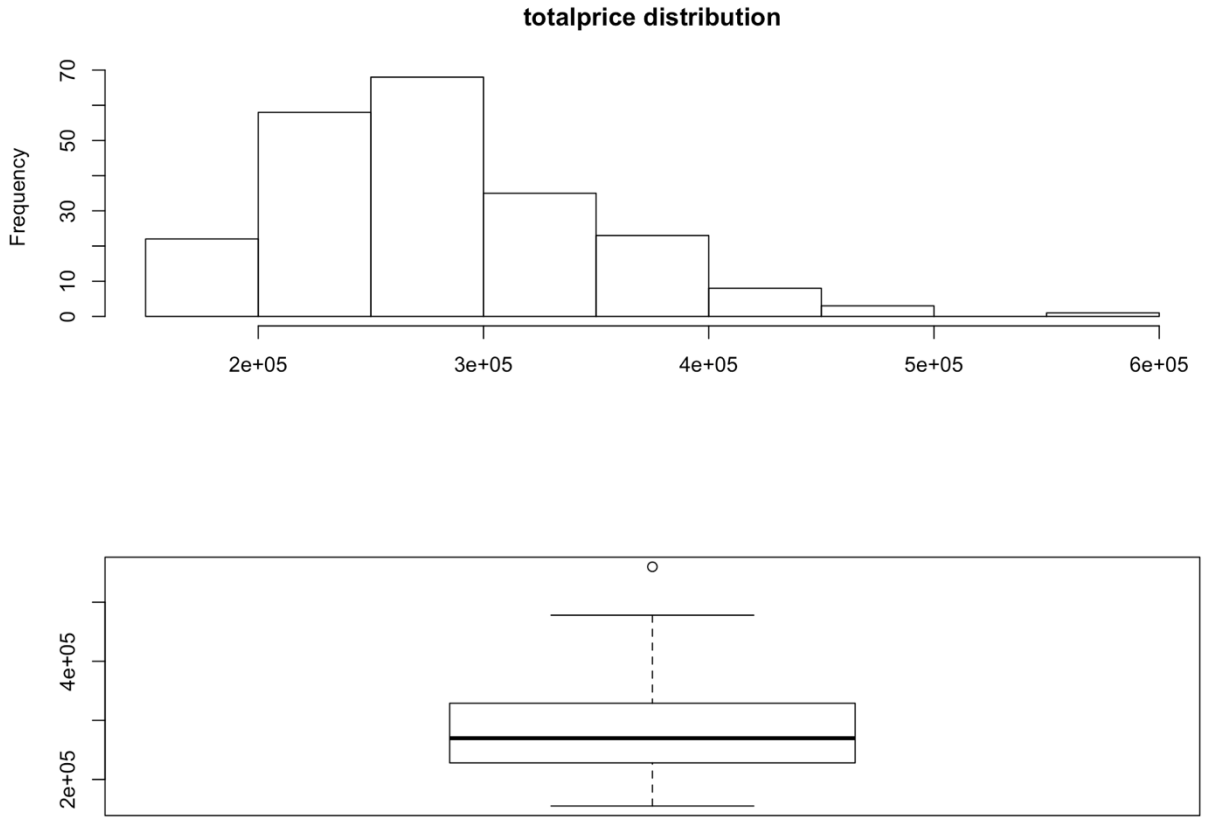
```
> install.packages("PASWR2")
> library(PASWR2)
>
> #1
a)
> d1=VIT2005
> mu=mean(d1$totalprice)
> mu
[1] 280741.5
> stdev=sd(d1$totalprice)
> stdev
[1] 69298.46
>
> boxplot(d1$totalprice,main="totalprice")
```



As we can see from boxplot the totalprice is not fairly symmetric.

```
> par(mfrow=c(2,1))
> hist(d1$totalprice,xlab="",main="totalprice distribution")
```

> boxplot(d1$totalprice)
> par(mfrow=c(1,1))



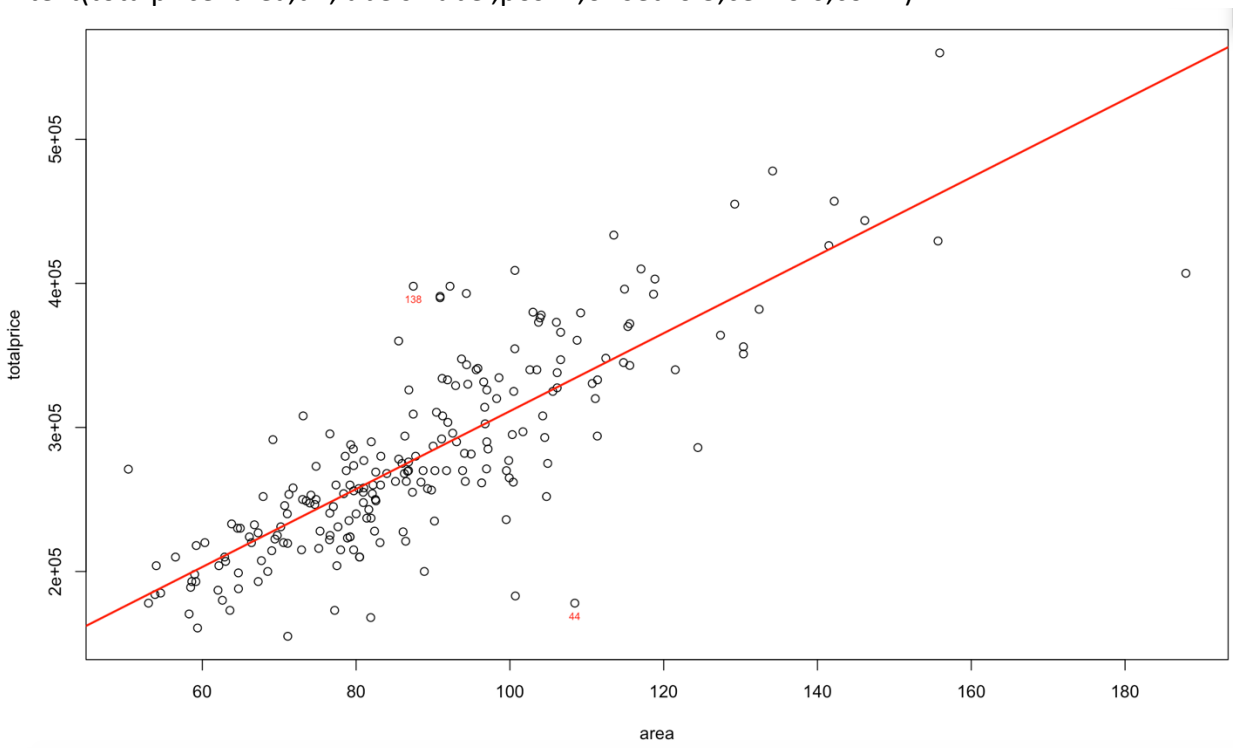
> outlier=boxplot.stats(d1$totalprice)$out
> outlier
[1] 560000
> which(grepl(outlier,d1$totalprice))
[1] 13
>

b)
> d2 = subset(d1,age,subset = garage ==0)
> d3=apply(d2,2,mean)
> d3
   age
21.5988
the average age of apartments with no garage=21.5988
>

c)
> table(d1$garage,d1$storage)

```
    0   1   2
 0  39 128   0
 1   3  45   1
 2   1   1   0
```
the number apartments with no garage and a single storage unit=128
>

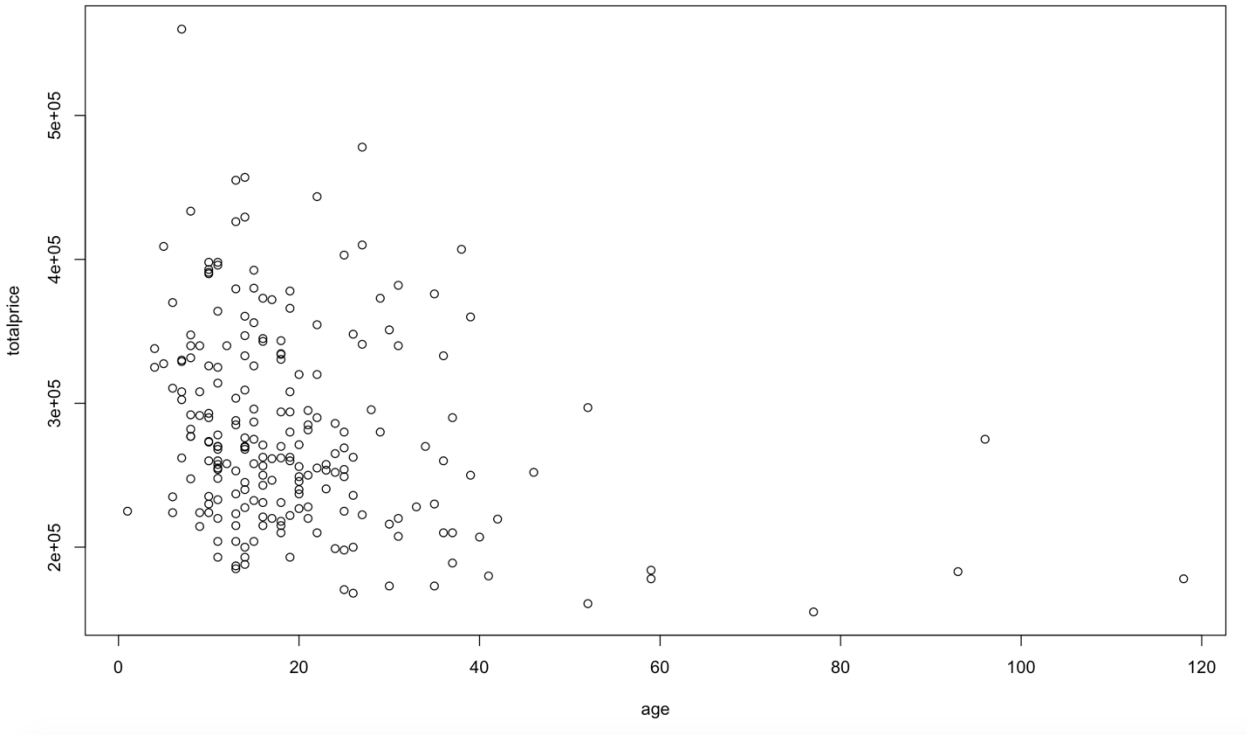
d)
```
> plot(totalprice~area,d1)
> m1=lm(totalprice~area,d1)
> abline(m1,col="red",lwd=2)
>
> label = rep("",392)
> res = resid(m1)
> idx = which(res==max(res))
> label[idx]=idx
> text(totalprice~area,d1,labels=label,pos=1,offset=0.5,cex=0.6,col=2)
>
> label = rep("",392)
> res = resid(m1)
> idx = which(res==min(res))
> label[idx]=idx
> text(totalprice~area,d1,labels=label,pos=1,offset=0.5,cex=0.6,col=2)
```



row number of outliers=138 and 44

e)
> newval1=data.frame(area=100)
> predict(m1,newval1)
      1
311297.5
Predict price of an apartment with 100 square meters=311297.5
>

f)
> plot(totalprice~age,d1)

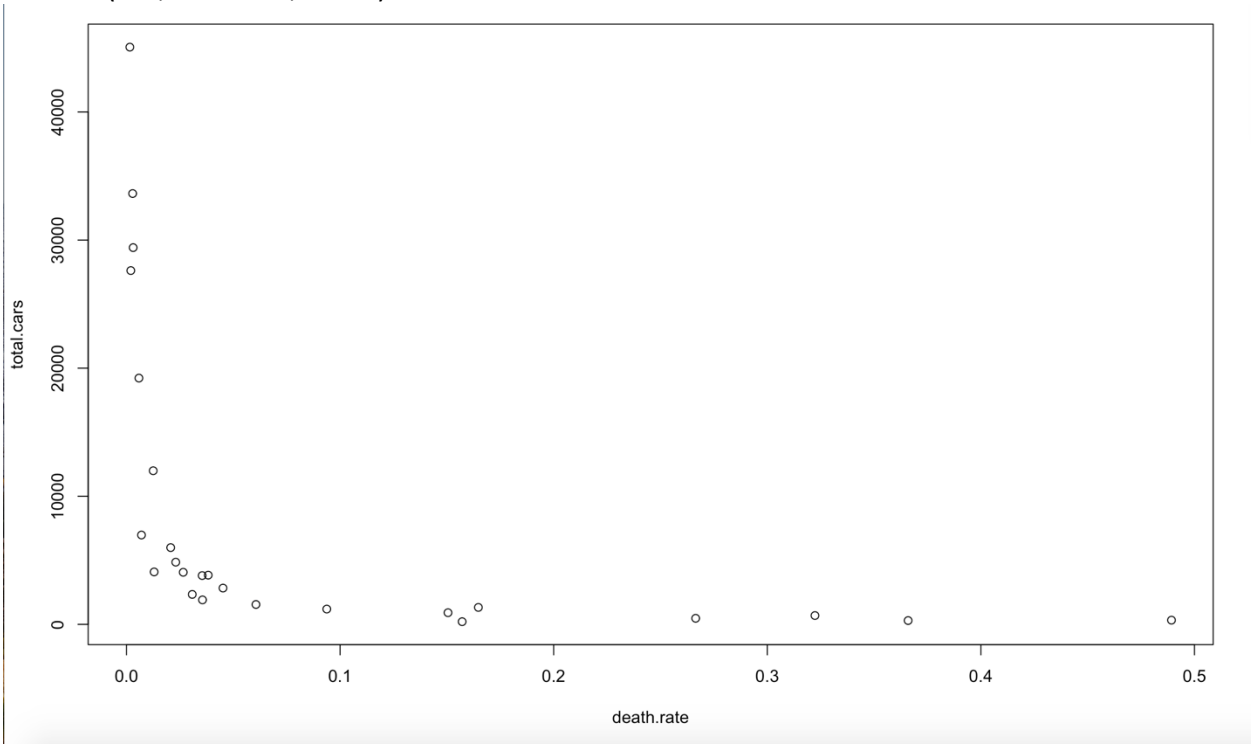


>

> d4=subset(d1,select = c(totalprice,area,age))
> trainingset=d4[1:109,1:3]
> F1=lm(totalprice~area,trainingset)
> F2=lm(totalprice~age,trainingset)
> testset=d4[110:218,1:3]
> newval2=data.frame(area=testset$area)
> r1=predict(F1,newval2)
> newval3=data.frame(age=testset$age)
> r2=predict(F2,newval3)
> e1=(sum(testset$totalprice-r1)^2)/109
> e1

[1] 1192937863
> e2=(sum(testset$totalprice-r2)^2)/109
> e2
[1] 6490303797
>
Area is better predictor. As we can see from picture the relation of age and totalprice is weak and hard to find the rules, but area have strong influence to the totalprice.
As we can see from the picture the relation of age and totalprice is weak, and we can divide the sample into two part one of them use for training the model, rest of them use for test and calculate the variance of expected number and actual number. We can see that the e1<e2, so the area is better predictor in this case.

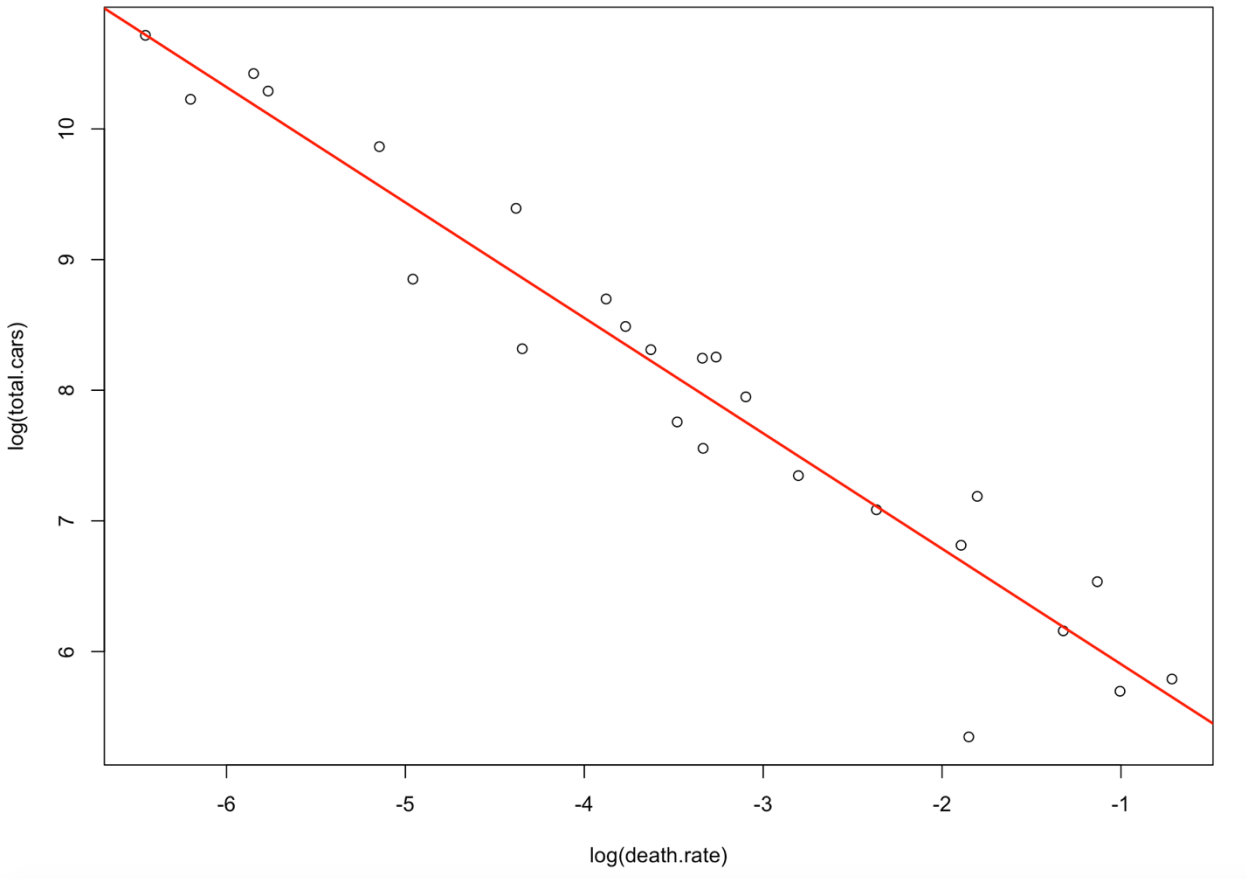
> #2
a)
> D1=CARS2004
> D1$total.cars=D1$cars*D1$population/1000
> D1$death.rate=D1$deaths/D1$total.cars
> plot(total.cars~death.rate,D1)
> M2=lm(total.cars~death.rate,D1)
> abline(M2,col="red",lwd=2)



it is like power function f(x)=x^a (a<0)
>

b)c)
> D1$ltc=log(D1$total.cars)

```
> D1$ldr=log(D1$death.rate)
> plot(ltc~ldr,D1,xlab="log(death.rate)",ylab="log(total.cars)")
> M1=lm(ltc~ldr,D1)
> abline(M1,col="red",lwd=2)
```



The relationship between log(death.rate) and log(total.cars) is like linear funtion f(x) = kx + b (k<0,b>0)

>

d)

```
> x=log(0.02305206,base=exp(1))
> newval4=data.frame(ldr=x)
> A=predict(M1,newval2)
> n=exp(A)
> n
      1
4233.815
>
```

total number of cars does this model predict is 4233.815