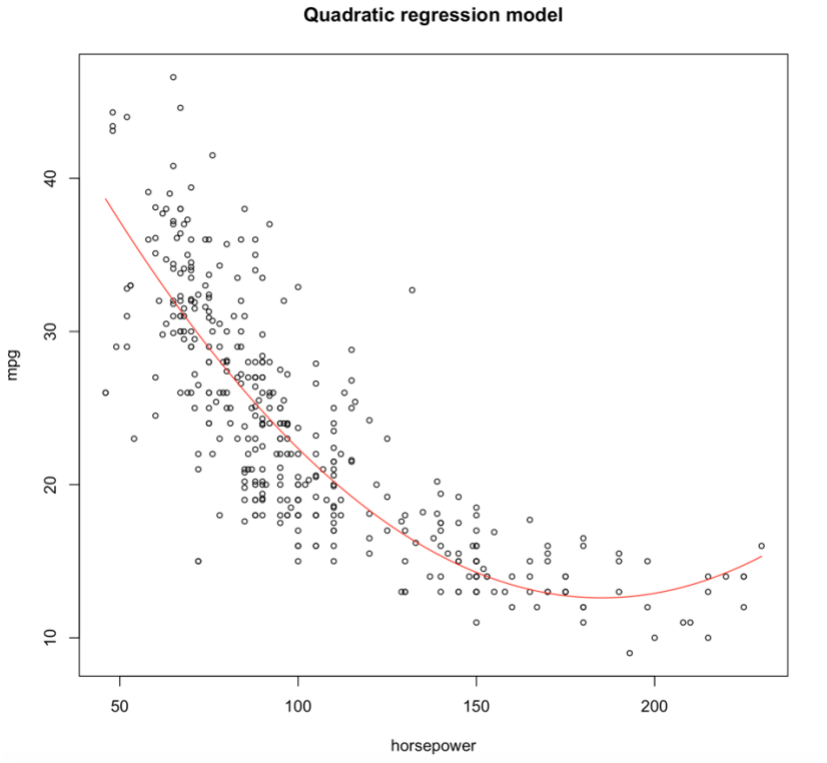
Shuting Chen

HW3

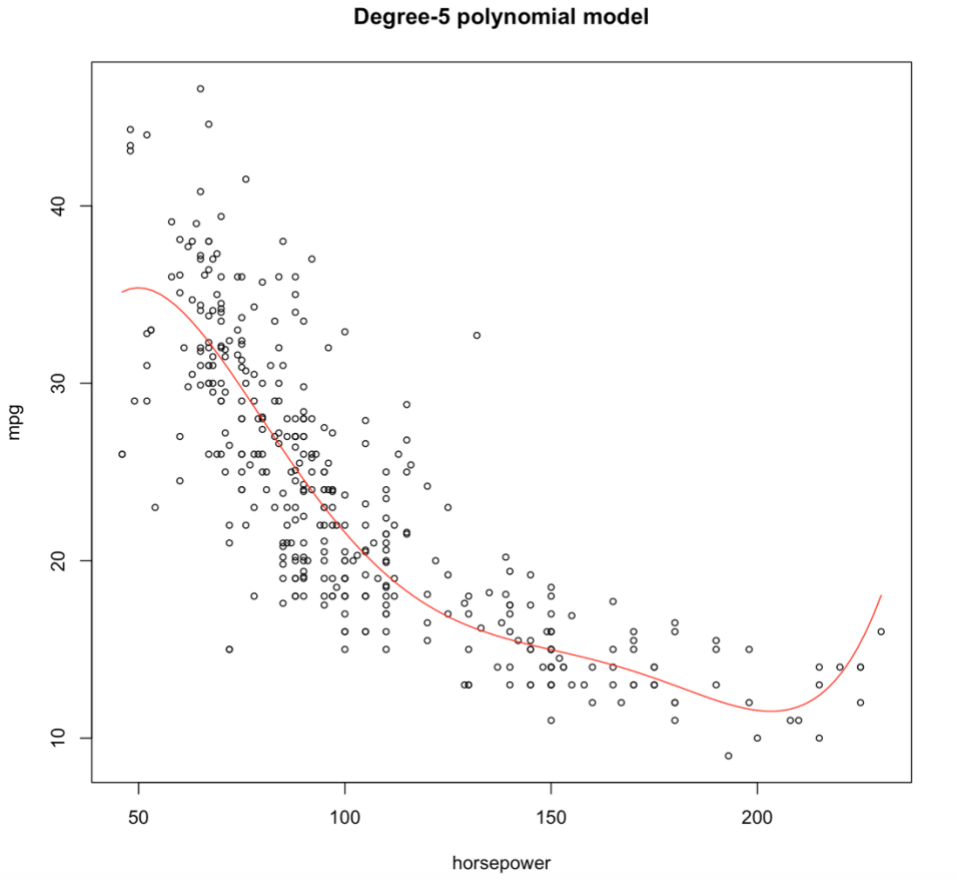
```
> library(ISLR)
> library(boot)
> library(leaps)
```

> #1

```
> d0=Auto
> set.seed(9)
> n=nrow(d0)
> train = sample(1:n,n/2)
> glm1=glm(mpg~poly(horsepower,2),data=d0,subset = train)
> mpg = d0$mpg
> res1 = (mpg-predict(glm1,d0))[-train]^2
> mspe1 = mean(res1)
> #MSPE1=17.47851
> r=range(d0$horsepower)
> xlim1=seq(r[1],r[2],len=100)
> ylim1=predict(glm1,newdata = data.frame(horsepower=xlim1))
> plot(mpg~horsepower,d0,cex=0.7,main="Quadratic regression model")
> lines(xlim1,ylim1,col="red")
>
```



```
> glm2=glm(mpg~poly(horsepower,5),data=d0,subset = train)
> res2 = (mpg-predict(glm2,d0))[-train]^2
> mspe2=mean(res2)
> mspe2
[1] 17.13565
> #MSPE2=17.13565
> xlim2=seq(r[1],r[2],len=100)
> ylim2=predict(glm2,newdata = data.frame(horsepower=xlim2))
> plot(mpg~horsepower,d0,cex=0.7,main="Degree-5 polynomial model")
> lines(xlim2,ylim2,col="red")
>
```



> #2

```
> #a
> setwd("/Users/shutingchen/Desktop/ISE 529          Data Analytics/HW")
> set.seed(1)
> d2=read.csv("homes.csv",header = T)
> d3=d2[,c(1,8,2,3,4,6,5)]
> n=nrow(d3)
> k=n
> mspe=matrix(0,k,6)
```

```
> predict.regsubsets <- function(object, newdata, id, ...)
+ {
+   form <- as.formula(object$call[[2]])
+   mat <- model.matrix(form, newdata)
+   coefi = coef(object, id = id)
+   xvars <- names(coefi)
+   mat[, xvars]%*%coefi
+ }
> for (j in 1:k)
+ {
+   y=d3$price[j]
+   d4=d3[-j,]
+   models = regsubsets(price~.,d4)
+   for (i in 1:6)
+   {
+     newdata = d3[j,] #test set
+     yhat <- predict.regsubsets(models,newdata,id=i)#predict jth vector
+     mspe[j,i]<-mean((y-yhat)^2)
+   }
+ }
> CV1= colMeans(mspe)
> CV1
[1] 6309808439 5442770459 5110873304 4981528600 4920519950 4923110359
> aux=which.min(CV1)
> regfit.best<-regsubsets(price~.,d3)
> coef(regfit.best,aux)
  (Intercept)       lotsize          area          beds          year       garage
-3.774167e+06  1.594788e+00  1.317622e+02 -1.116391e+04  1.883998e+03  2.307114e+04
> #Leave-One-Out cross validation
> #y=-3.774167e+06+1.594788e+00*lotsize+1.317622e+02*area-
1.116391e+04*beds+1.883998e+03*year+2.307114e+04*garage
>
> #b
> k=10
> set.seed(1)
> x = rep(1:10,each=52)
> x = sample(x)
> x2=sample(1:10,2)
> folds=c(x,x2)
> mspe=matrix(0,10,6)
> predict.regsubsets <- function(object, newdata, id, ...)
+ {
+   form <- as.formula(object$call[[2]])
+   mat <- model.matrix(form, newdata)
```

```
+   coefi = coef(object, id = id)
+   xvars <- names(coefi)
+   mat[, xvars]%*%coefi
+ }
> for (j in 1:k)
+ {
+   y=d3$price[folds==j] #y-value in the jth fold
+   d4=d3[folds!=j,]   #training set ignores jth fold
+   cvmodels = regsubsets(price~.,d4)
+   for (i in 1:6)
+   {
+     newdata = d3[folds == j,] #test set
+     yhat = predict.regsubsets(cvmodels,newdata,id=i)#predict jth fold(vector)
+     mspe[j,i]=mean((y-yhat)^2)
+   }
+ }
> CV2= colMeans(mspe)
> aux1=which.min(CV2)
> best.fit=regsubsets(price~.,d3)
> coef(best.fit,aux1)
  (Intercept)      lotsize         area         beds        baths         year       garage
-3.567709e+06  1.554990e+00  1.257386e+02 -1.304139e+04  7.987552e+03  1.779611e+03
2.253038e+04
>
> #10-folds cross validation
> #y=-3.567709e+06+1.554990e+00*lotsize+1.257386e+02*area-
1.304139e+04*beds+7.987552e+03*baths+1.779611e+03*year+2.253038e+04*garage
>
```