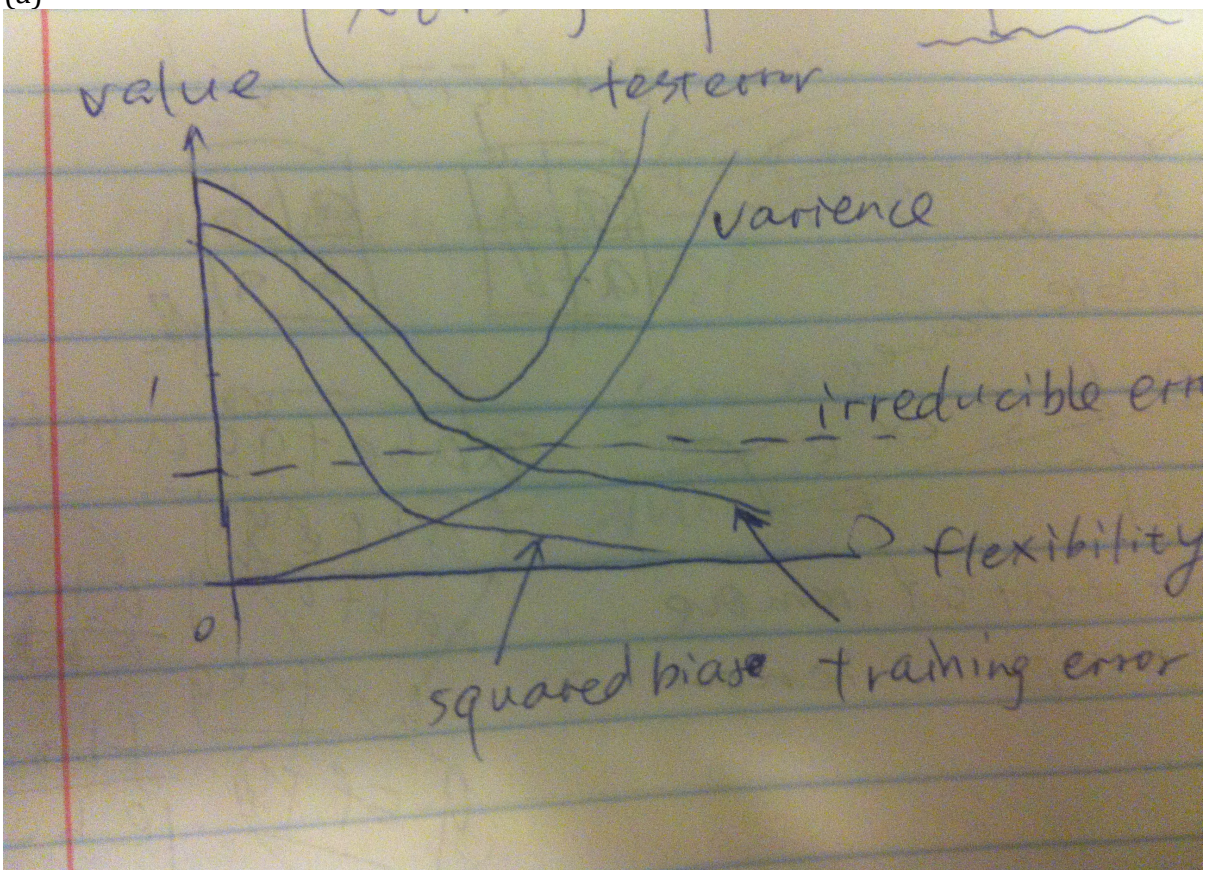
2.
(a) regression, inference, n= 500 firms in the US, p= profit, number of employees, industry
(b) classification, prediction, n= 20 similar products previously launched, p= price charged, marketing budget, comp. price, ten other variables
(c) regression, prediction, n= 52 weeks of 2012 weekly data, p= % change in US market, % change in British market, % change in German market

3.
(a)



(b)
As flexibility increases, and at one point the model starts over-fitting, squared bias and training error decreases, while variance increases all the time.

Test error first decreases and then starts to increase since flexibility yields an overfit.

As for irreducible error, which is more than zero and less than one, it is the lower bound of test error as we could see from the formula. After training error is lower than the irreducible error, overfitting starts

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

5.
For regression or classification:
Advantages: to obtain a better fit for non-linear models, and to reduce bias.
Disadvantages: to increase variance as well as test error, and require estimating more parameters.

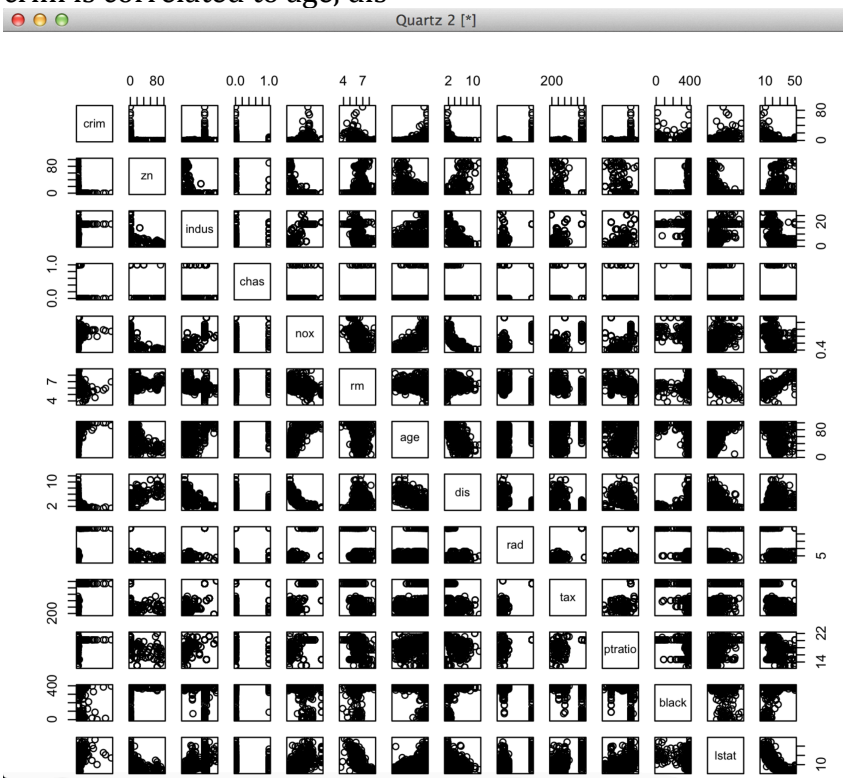
Prediction prefers more flexibility, while interpretability is interested in less flexible.

6.
A parametric approach only estimates a set of parameters, while a non-parametric approach requires a very large number of observations to accurately estimate function f.

Advantages: only require a set of parameters to estimate f.

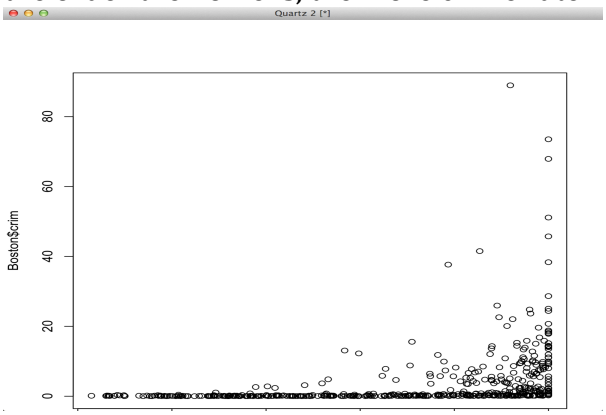
Disadvantages: requires a very large number of observations to accurately estimate function f compared to the parametric approach, overfits the observations if more flexible models are used

10.
(a) 506 rows and 14 columns, representing housing values in suburbs of Boston
(b) >pairs(Boston)
crim is correlated to age, dis

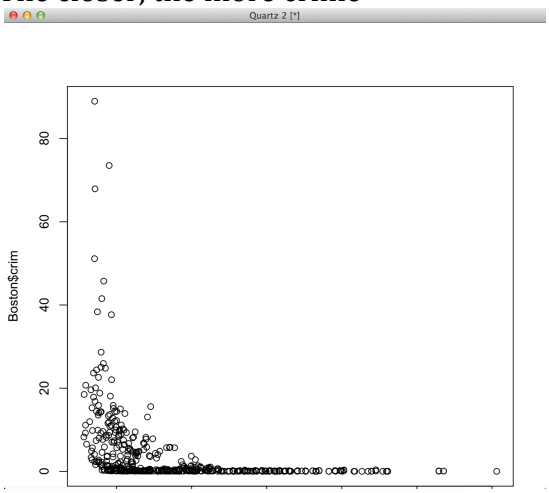


(c) >plot (Boston$age, Boston$crim)

the order the home is, the more crime rate.

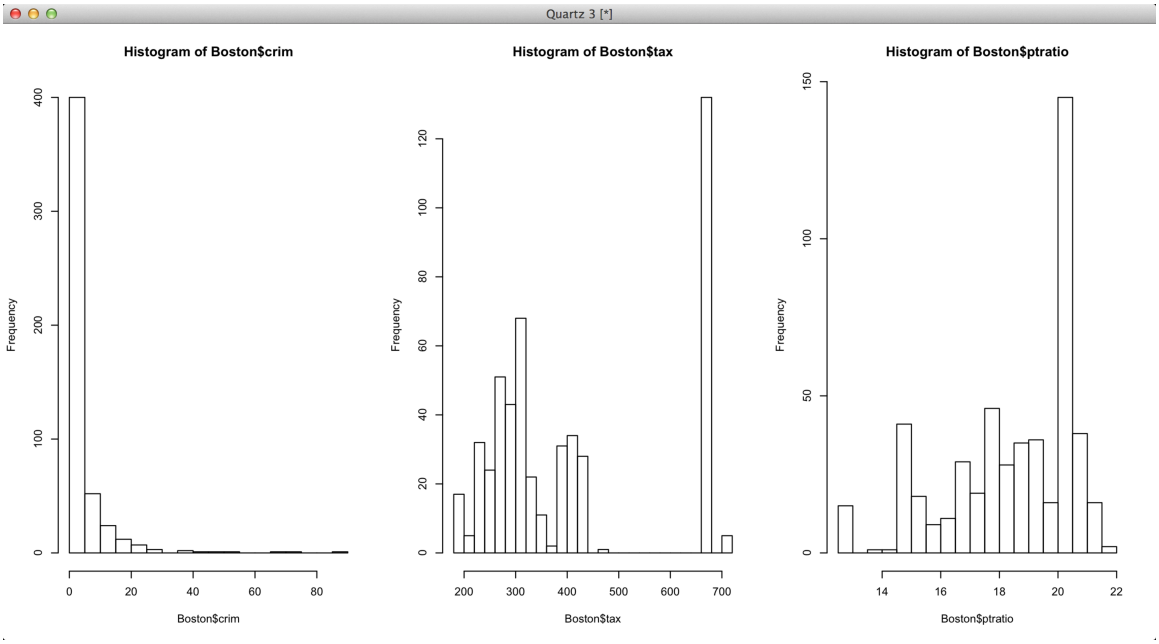


>plot(Boston$dis, Boston$crim)
The closer, the more crime



(d)
>par(mfrow=c(1,3))
#combine multiple plots into one overall graph using par()
>hist(Boston$crim[Boston$crim>1], breaks=25)
>hist(Boston$tax, breaks=25)
>hist(Boston$ptratio, breaks=25)

most cities have low crime rates(below 20), while a small amount of suburbs have a crime rate reaching above 80
tax rate focuses on two extremes
ratio peaks at 20-21, while others have an almost equal distribution



(e)
>dim(subset(Boston, chas == 1))

35
(f)
>median(Boston$ptratio)
19.95

(g) t(subset(Boston, medv == min(Boston$medv)))
#t() function is to transpose

	399	406
crim	38.3518	67.9208
zn	0.0000	0.0000
indus	18.1000	18.1000
chas	0.0000	0.0000
nox	0.6930	0.6930
rm	5.4530	5.6830
age	100.0000	100.0000
dis	1.4896	1.4254
rad	24.0000	24.0000
tax	666.0000	666.0000
ptratio	20.2000	20.2000
black	396.9000	384.9700
lstat	30.5900	22.9800
medv	5.0000	5.0000

>summary(Boston)

crim	zn	indus	chas
Min. : 0.00632	Min. : 0.00	Min. : 0.46	Min. :0.00000

```
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
     nox               rm             age              dis
 Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
 1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
 Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
 Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
 3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
 Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
     rad              tax            ptratio          black
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
    lstat            medv
 Min.   : 1.73   Min.   : 5.00
 1st Qu.: 6.95   1st Qu.:17.02
 Median :11.36   Median :21.20
 Mean   :12.65   Mean   :22.53
 3rd Qu.:16.95   3rd Qu.:25.00
 Max.   :37.97   Max.   :50.00
```

It's not the best, nor the worst place.

(h)

```
> dim(subset(Boston, rm > 7))
[1] 64 14
> dim(subset(Boston, rm > 8))
[1] 13 14
```

relatively low crime rate