HW7

David Schultheiss

10/26/2020

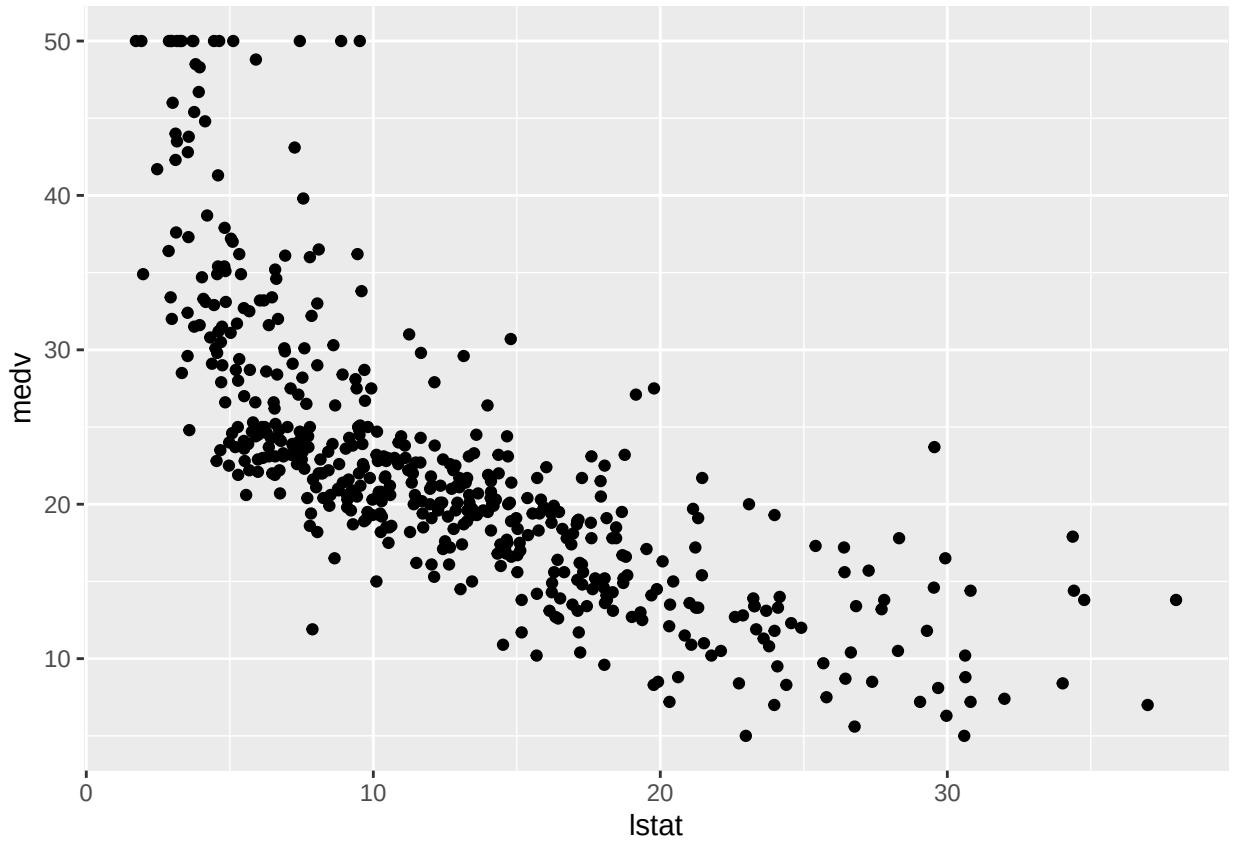
Problem 1

a)

```
library(MASS)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
Boston = Boston
ggplot(data = Boston, mapping = aes(x= lstat, y= medv)) +
  geom_point()
```



The relationship is non-linear.

b)

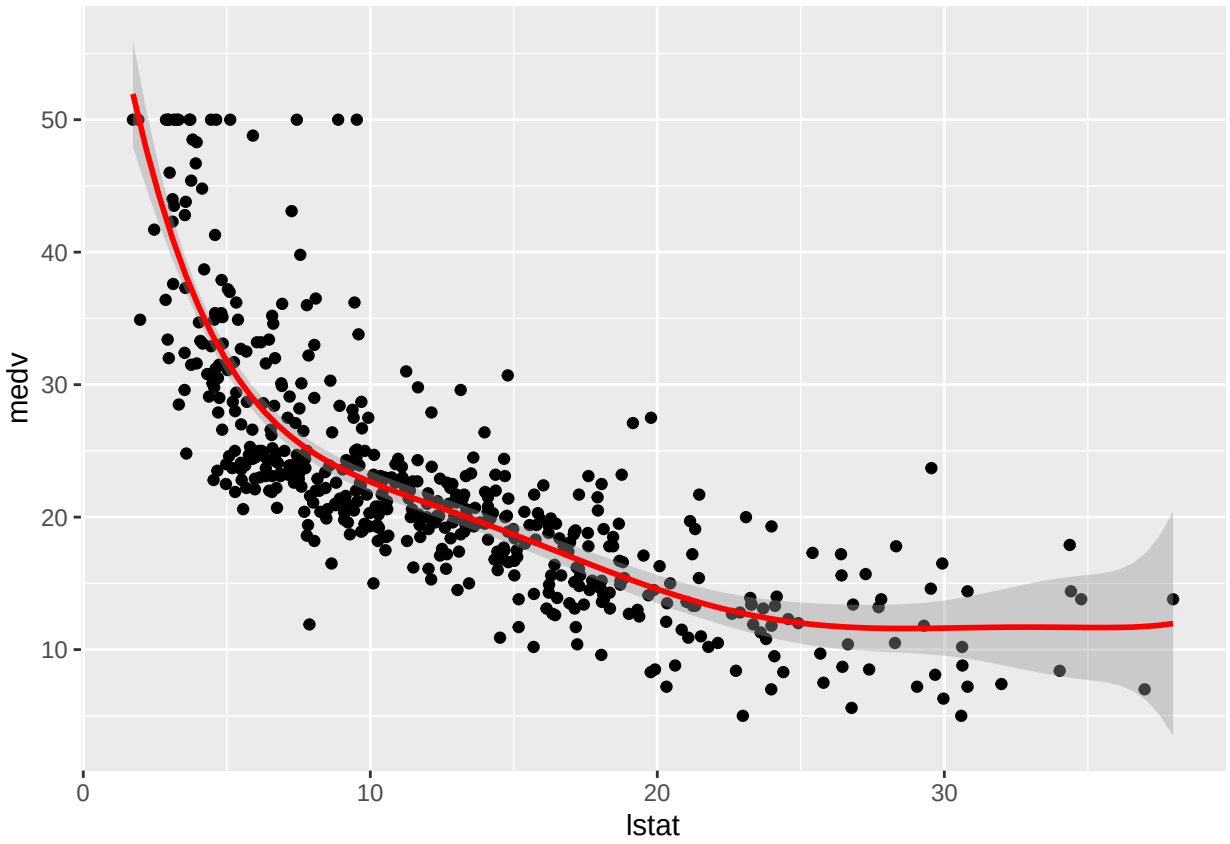
```
library(boot)
cv.error = rep(0,10)
for(i in 1:10) {
  glm.fit = glm(data= Boston, medv ~ poly(lstat, i, raw= T))
  cv.error[i] = cv.glm(Boston, glm.fit, K= 5)$delta[1]
}

cv.error
```

```
##  [1] 38.95867 30.56585 29.63028 27.97419 27.69826 27.96951 27.76969 27.57908
##  [9] 28.28069 33.86763
```

```
which.min(cv.error)
```

```
## [1] 8
```

```
ggplot(data = Boston, mapping = aes(x= lstat, y= medv)) +
  geom_point() +
  stat_smooth(method= 'glm', formula = y ~ poly(x, 6, raw=T), colour= 'red')
```



c)

```
library(splines)
cv.error = rep(0,20)
for(i in 1:20) {
  glm.fit = glm(data= Boston, medv ~ bs(lstat, df= i))
  cv.error[i] = cv.glm(Boston, glm.fit, K= 5)$delta[1]
}
```

```
## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.73, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.73, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.92, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.92, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases
```

```
## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.73, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.73, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.92, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.92, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, df = i): 'df' was too small; have used 3

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.92, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.92, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.73, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = numeric(0), Boundary.knots = c(1.73, :
## some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(`50%` = 11.235), Boundary.knots =
## c(1.73, : some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(`50%` = 11.235), Boundary.knots =
## c(1.73, : some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(`50%` = 11.365), Boundary.knots =
## c(1.92, : some 'x' values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(`50%` = 11.365), Boundary.knots =
## c(1.92, : some 'x' values beyond boundary knots may cause ill-conditioned bases
```

```
## Warning in bs(lstat, degree = 3L, knots = c(‘33.33333%‘ = 8.48333333333333, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘33.33333%‘ = 8.48333333333333, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘33.33333%‘ = 8.14, ‘66.66667%‘
## = 14.6633333333333: some ’x’ values beyond boundary knots may cause ill-
## conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘33.33333%‘ = 8.14, ‘66.66667%‘
## = 14.6633333333333: some ’x’ values beyond boundary knots may cause ill-
## conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘25%‘ = 7.14, ‘50%‘ = 11.32, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘25%‘ = 7.14, ‘50%‘ = 11.32, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘25%‘ = 6.75, ‘50%‘ = 11.28, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘25%‘ = 6.75, ‘50%‘ = 11.28, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘20%‘ = 6.1, ‘40%‘ = 9.476, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘20%‘ = 6.1, ‘40%‘ = 9.476, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘20%‘ = 6.48, ‘40%‘ = 9.68, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘20%‘ = 6.48, ‘40%‘ = 9.68, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘16.66667%‘ = 5.68333333333333, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘16.66667%‘ = 5.68333333333333, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘16.66667%‘ = 5.53666666666667, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘16.66667%‘ = 5.53666666666667, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘14.28571%‘ = 5.23285714285714, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases
```

```
## Warning in bs(lstat, degree = 3L, knots = c(‘14.28571%‘ = 5.23285714285714, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘14.28571%‘ = 5.31857142857143, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘14.28571%‘ = 5.31857142857143, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘12.5%‘ = 5.1975, ‘25%‘ = 7.21, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘12.5%‘ = 5.1975, ‘25%‘ = 7.21, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘12.5%‘ = 5.03375, ‘25%‘ = 6.84, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘12.5%‘ = 5.03375, ‘25%‘ = 6.84, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘11.11111%‘ = 4.85777777777778, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘11.11111%‘ = 4.85777777777778, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘11.11111%‘ = 5.02444444444444, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘11.11111%‘ = 5.02444444444444, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘10%‘ = 4.676, ‘20%‘ = 6.282, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘10%‘ = 4.676, ‘20%‘ = 6.282, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘10%‘ = 4.694, ‘20%‘ = 6.36, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘10%‘ = 4.694, ‘20%‘ = 6.36, : some
## ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘9.090909%‘ = 4.52909090909091, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘9.090909%‘ = 4.52909090909091, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases
```

```
## Warning in bs(lstat, degree = 3L, knots = c(‘9.090909%‘ = 4.62272727272727, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘9.090909%‘ = 4.62272727272727, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘8.333333%‘ = 4.36, ‘16.66667%‘
## = 5.49333333333333, : some ’x’ values beyond boundary knots may cause ill-
## conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘8.333333%‘ = 4.36, ‘16.66667%‘
## = 5.49333333333333, : some ’x’ values beyond boundary knots may cause ill-
## conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘8.333333%‘ = 4.55333333333333, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘8.333333%‘ = 4.55333333333333, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘7.692308%‘ = 4.5, ‘15.38462%‘ =
## 5.52, : some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘7.692308%‘ = 4.5, ‘15.38462%‘ =
## 5.52, : some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘7.692308%‘ = 4.45769230769231, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘7.692308%‘ = 4.45769230769231, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘7.142857%‘ = 4.20285714285714, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘7.142857%‘ = 4.20285714285714, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘7.142857%‘ = 4.07285714285714, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘7.142857%‘ = 4.07285714285714, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘6.666667%‘ = 4.20666666666667, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘6.666667%‘ = 4.20666666666667, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘6.666667%‘ = 4.01933333333333, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases
```

```
## Warning in bs(lstat, degree = 3L, knots = c(‘6.666667%‘ = 4.01933333333333, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘6.25%‘ = 3.975, ‘12.5%‘ = 5.235, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘6.25%‘ = 3.975, ‘12.5%‘ = 5.235, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘6.25%‘ = 3.925625, ‘12.5%‘ =
## 4.85375, : some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘6.25%‘ = 3.925625, ‘12.5%‘ =
## 4.85375, : some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘5.882353%‘ = 4.06823529411765, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘5.882353%‘ = 4.06823529411765, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘5.882353%‘ = 3.94294117647059, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘5.882353%‘ = 3.94294117647059, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘5.555556%‘ = 3.78222222222222, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases

## Warning in bs(lstat, degree = 3L, knots = c(‘5.555556%‘ = 3.78222222222222, :
## some ’x’ values beyond boundary knots may cause ill-conditioned bases
```

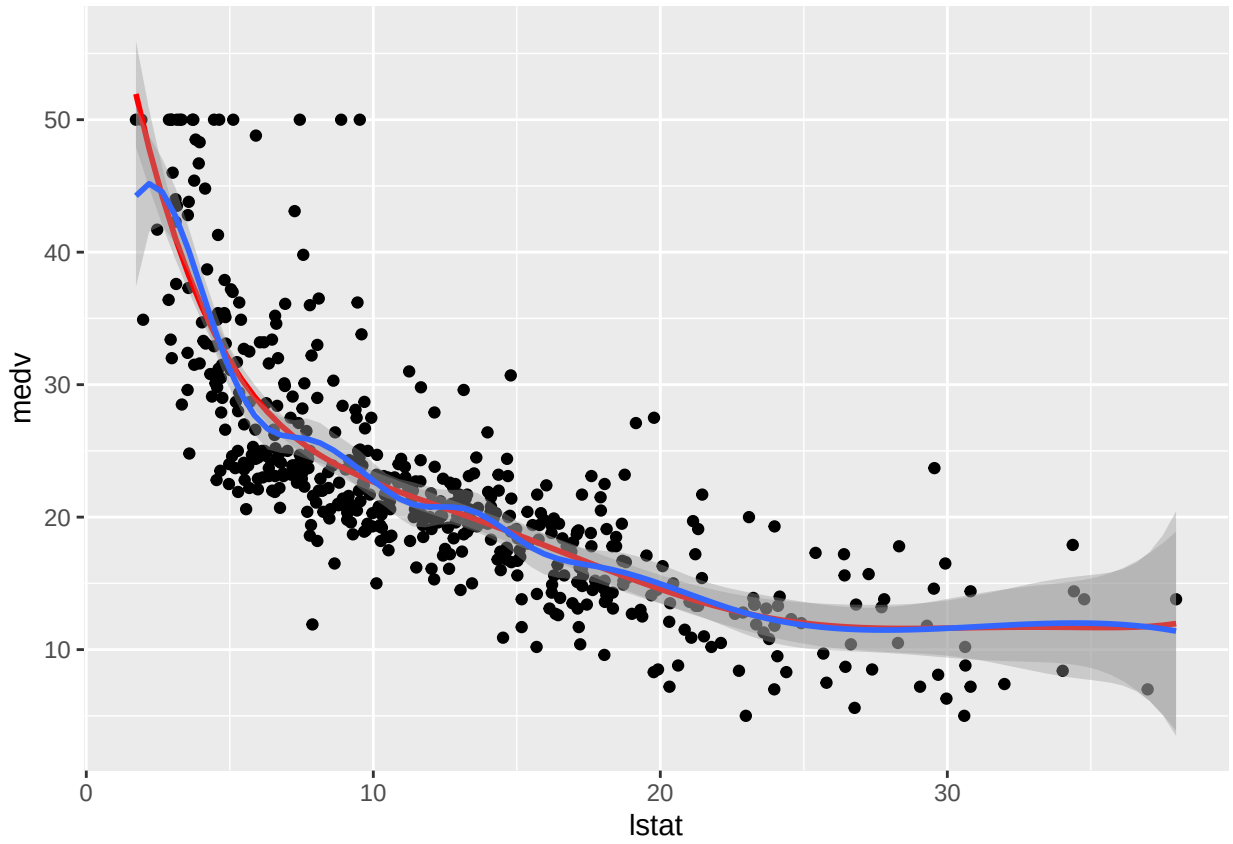
```
cv.error
```

```
##  [1] 29.40564 29.34908 29.84852 27.66809 27.54198 27.29312 27.37795 27.62532
##  [9] 27.45441 27.19180 27.44698 27.85406 27.50446 27.78776 27.79638 27.90734
## [17] 27.49285 27.97460 27.34896 28.73574
```

```
which.min(cv.error)
```

```
## [1] 10
```

```
ggplot(data = Boston, mapping = aes(x= lstat, y= medv)) +
  geom_point() +
  stat_smooth(method= ’glm’, formula = y ~ poly(x, 6, raw=T), colour= ’red’)+
  stat_smooth(method= ’glm’, formula = y ~ bs(x, df= 12))
```



d)

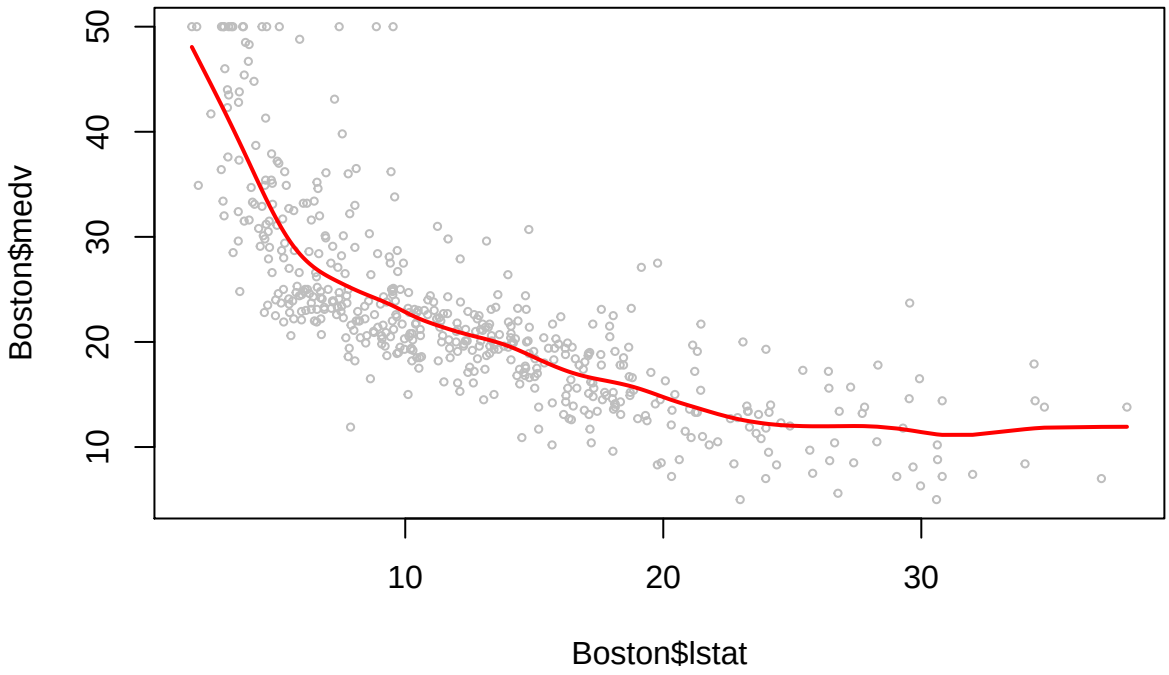
```
plot(Boston$lstat, Boston$medv, cex= .5, col= 'grey')
fit.ss = smooth.spline( x= Boston$lstat, y= Boston$medv, cv= T)
```

```
## Warning in smooth.spline(x = Boston$lstat, y = Boston$medv, cv = T): cross-
## validation with non-unique 'x' values seems doubtful
```

```
fit.ss$df
```

```
## [1] 11.3742
```

```
lines(fit.ss, col= 'red', lwd= 2)
```



e)

```
library(leaps)
reg.fit = regsubsets(medv~., data= Boston, nvmax= 3, method= 'forward')
summary(reg.fit)
```

```
## Subset selection object
## Call: regsubsets.formula(medv ~ ., data = Boston, nvmax = 3, method = "forward")
## 13 Variables  (and intercept)
##         Forced in Forced out
## crim        FALSE      FALSE
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas        FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## black       FALSE      FALSE
## lstat       FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: forward
##          crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
## 1  ( 1 ) " "  " " " "   " "  " " " " " " " " " " " "     " "   "*"
## 2  ( 1 ) " "  " " " "   " "  " " "*" " " " " " " " "     " "   "*"
## 3  ( 1 ) " "  " " " "   " "  " " "*" " " " " " " " " "*"     " "   "*"
```

1-variable model: lstat 2-variable model: lstat, rm 3-variable model: lstat, rm, ptratio

f)

```
library(gam)
```

```
## Loading required package: foreach
```

```
## Warning: package 'foreach' was built under R version 3.6.2
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
## Loaded gam 1.16.1
```

```
testsample = sample(1:nrow(Boston), .2*nrow(Boston))
test = Boston[testsample, ]
training = Boston[-testsample, ]

gam.fit = gam(data= training, medv~ s(lstat) + rm + ptratio)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored
```

```
summary(gam.fit)
```

```
##
## Call: gam(formula = medv ~ s(lstat) + rm + ptratio, data = training)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -10.398  -2.879  -0.518   2.052  29.063
##
## (Dispersion Parameter for gaussian family taken to be 22.5443)
##
##     Null Deviance: 35052.08 on 404 degrees of freedom
## Residual Deviance: 8972.632 on 398.0002 degrees of freedom
## AIC: 2420.049
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## s(lstat)    1 18970.6 18970.6 841.479 < 2.2e-16 ***
## rm          1  2366.1  2366.1 104.953 < 2.2e-16 ***
## ptratio     1   801.1   801.1  35.535 5.525e-09 ***
## Residuals 398  8972.6    22.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Anova for Nonparametric Effects
##             Npar Df Npar F     Pr(F)
## (Intercept)
## s(lstat)          3 33.341 < 2.2e-16 ***
## rm
## ptratio
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gam.pred = predict(gam.fit, test)
mean((test$medv - gam.pred)^2)
```

```
## [1] 21.05001
```

```
gam.fit = gam(data= training, medv~ s(lstat) + s(rm) + s(ptratio))
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored
```

```
summary(gam.fit)
```

```
##
## Call: gam(formula = medv ~ s(lstat) + s(rm) + s(ptratio), data = training)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6425  -2.3048  -0.5304   1.8584  28.5414
##
## (Dispersion Parameter for gaussian family taken to be 18.7979)
##
##     Null Deviance: 35052.08 on 404 degrees of freedom
## Residual Deviance: 7368.775 on 391.9999 degrees of freedom
## AIC: 2352.294
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## s(lstat)     1 20867.8 20867.8 1110.111 < 2.2e-16 ***
## s(rm)        1  2207.2  2207.2  117.416 < 2.2e-16 ***
## s(ptratio)   1   447.3   447.3   23.795 1.561e-06 ***
## Residuals  392  7368.8    18.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##             Npar Df Npar F     Pr(F)
## (Intercept)
## s(lstat)          3 15.640 1.259e-09 ***
## s(rm)             3 34.411 < 2.2e-16 ***
## s(ptratio)        3  1.217    0.3031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
gam.pred = predict(gam.fit, test)
mean((test$medv - gam.pred)^2)
```

```
## [1] 16.23631
```

Using the splines on all of our variables (with default df=4) gives us a lower MSE than only using spline on lstat, which we know has a non-linear relationship with medv from our graph. This means the other variables likely have non-linear relationships with medv.