3.2) Variance Inflation Factor (VIF) and Outliers

Vitor Kamada

December 2018

Reference

Tables, Graphics, and Figures from

An Introduction to Statistical Learning

James et al. (2017): Chapter 3

Boston Data Set from library(ISLR)

medv (median house value)

lstat (percent of households with low socioeconomic status)

Statistic	N	Mean	St. Dev.	Min	Max
crim	506	3.614	8.602	0.006	88.976
zn	506	11.364	23.322	0.000	100.000
indus	506	11.137	6.860	0.460	27.740
chas	506	0.069	0.254	0	1
nox	506	0.555	0.116	0.385	0.871
rm	506	6.285	0.703	3.561	8.780
age	506	68.575	28.149	2.900	100.000
dis	506	3.795	2.106	1.130	12.127
rad	506	9.549	8.707	1	24
tax	506	408.237	168.537	187	711
ptratio	506	18.456	2.165	12.600	22.000
black	506	356.674	91.295	0.320	396.900
lstat	506	12.653	7.141	1.730	37.970
medv	506	22.533	9.197	5.000	50.000

Variance Inflation Factor (VIF)

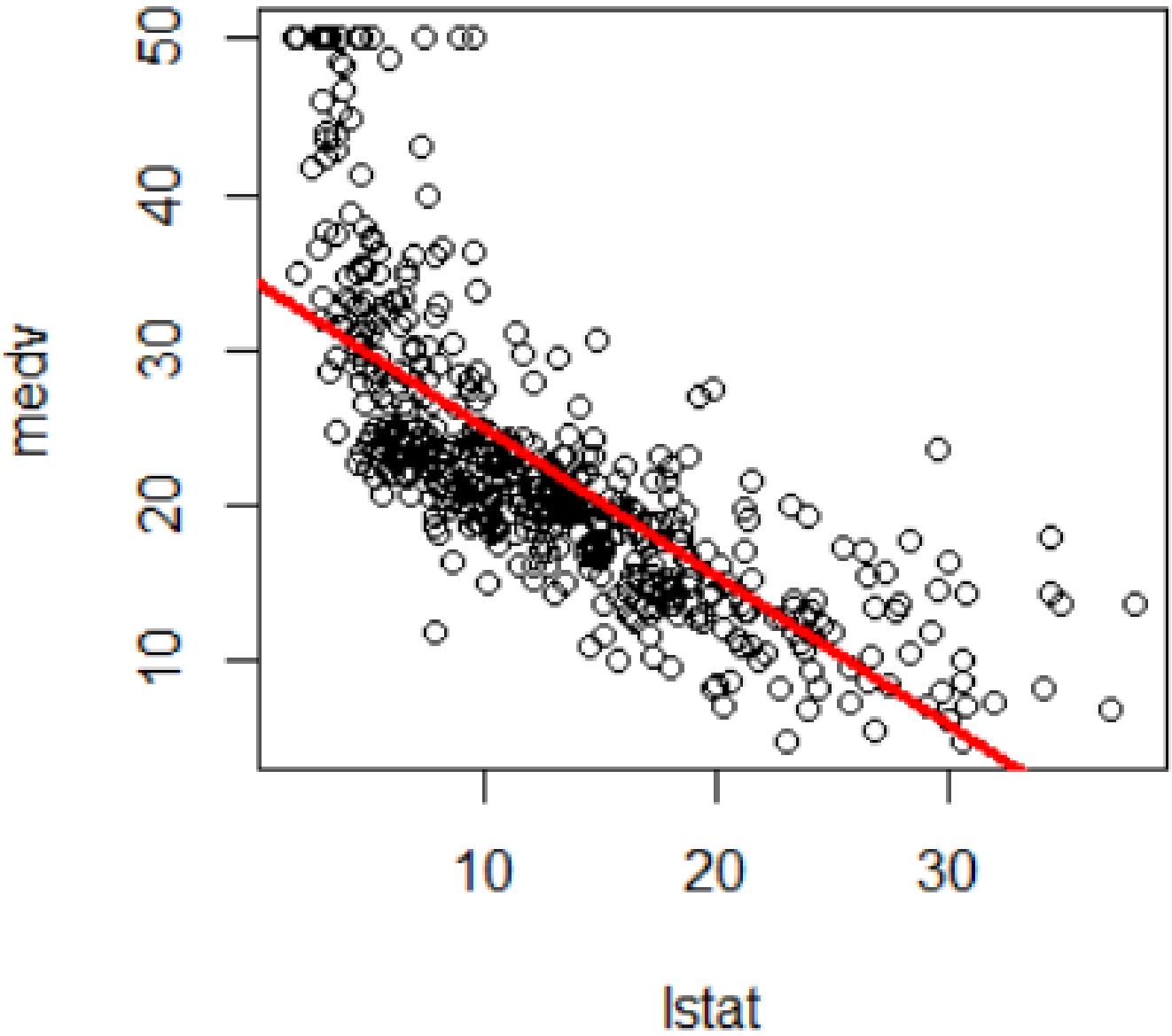
$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

library(car); vif(lm.fit)

```
    crim       zn    indus     chas      nox       rm
1.792192 2.298758 3.991596 1.073995 4.393720 1.933744
     age      dis      rad      tax  ptratio    black
3.100826 3.955945 7.484496 9.008554 1.799084 1.348521
   lstat
2.941491
```

$>$ 5 or 10 indicates a problematic collinearity

plot(lstat,medv); abline(lm.fit,lwd=3,col="red")



lm.fit=lm(medv~lstat,data=Boston)

summary(lm.fit)

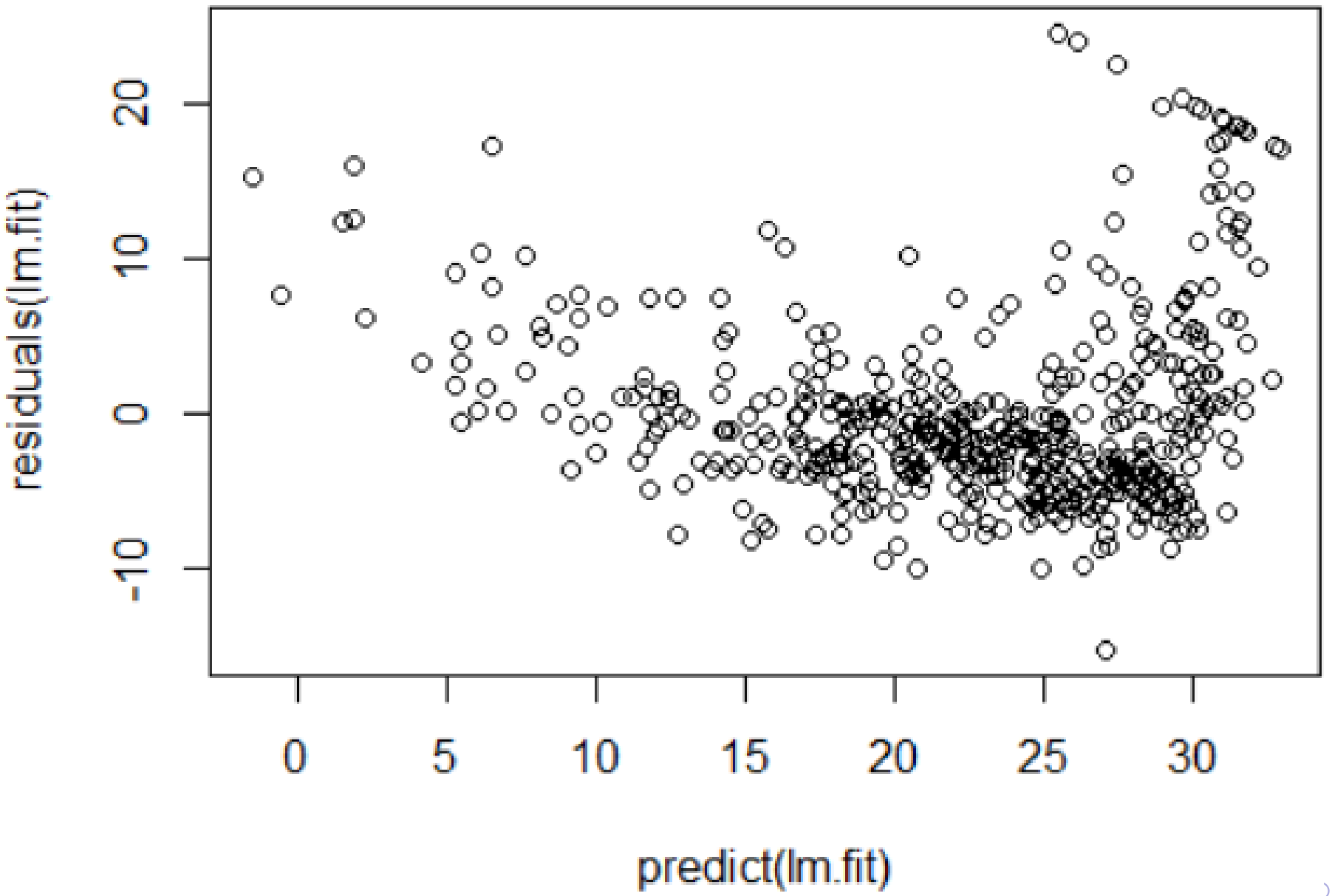
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5
432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

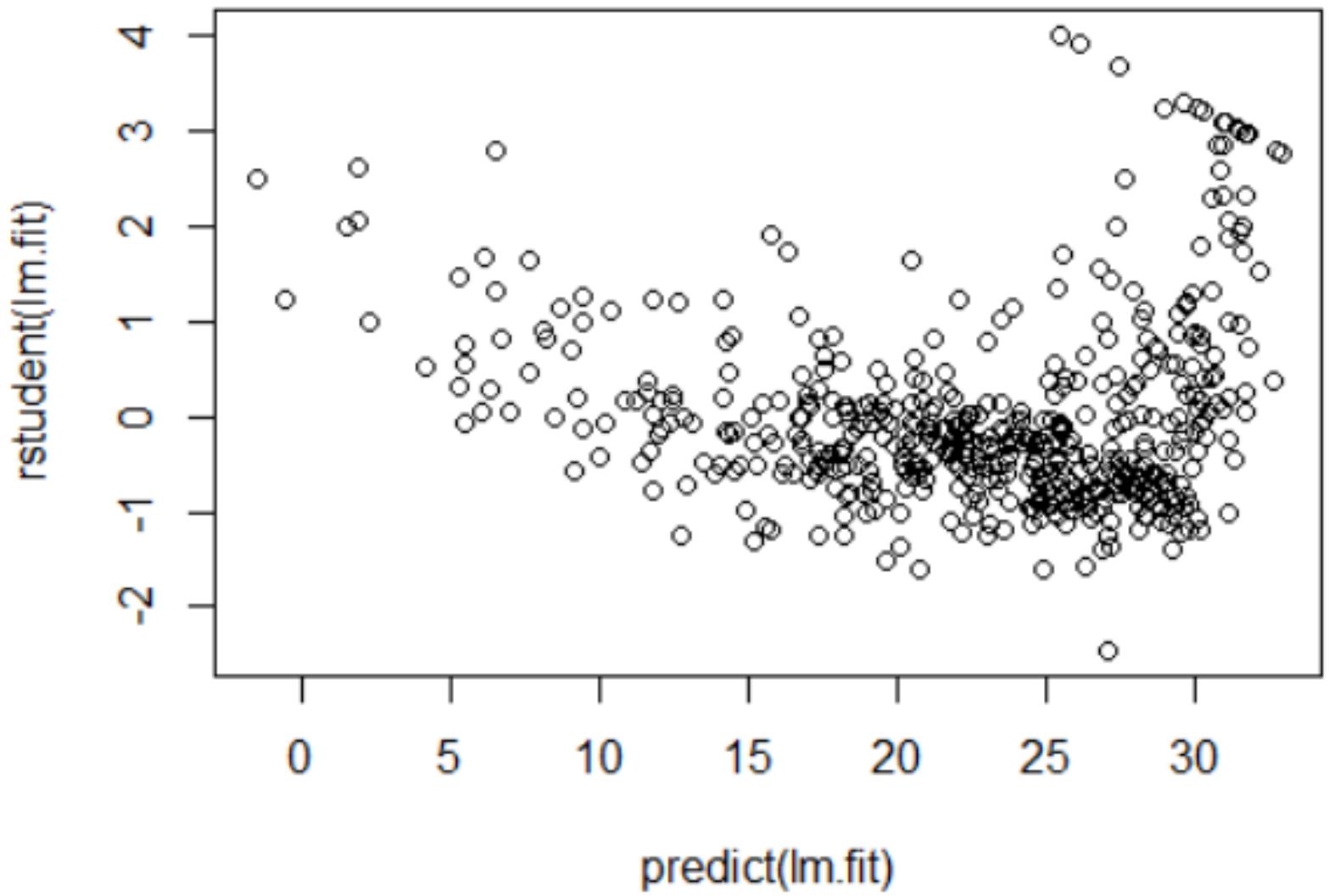
confint(lm.fit)

```
                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
```

plot(predict(lm.fit), residuals(lm.fit))

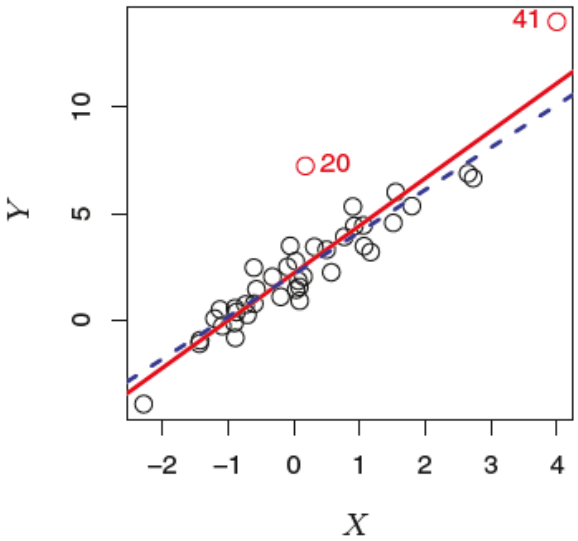


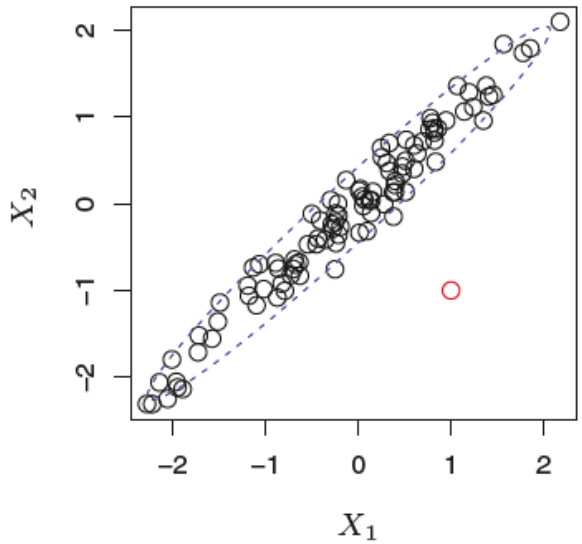
plot(predict(lm.fit), rstudent(lm.fit))

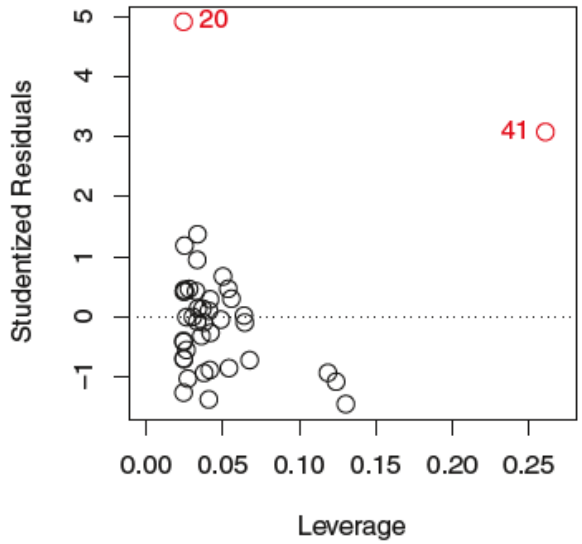


$> |3|$ are possible outliers

High Leverage Points (Unusual Value for x_i)



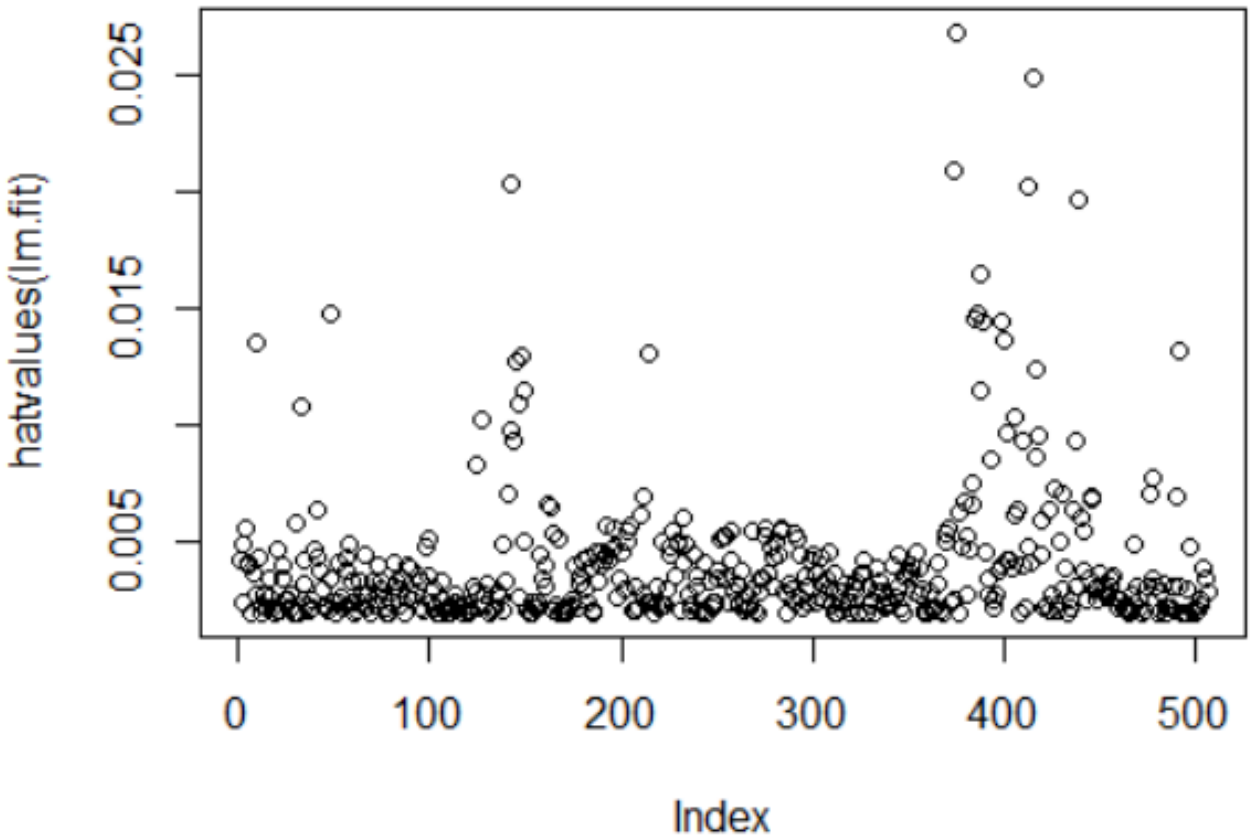




$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$$

lm.fit=lm(medv~lstat,data=Boston)

plot(hatvalues(lm.fit))



which.max(hatvalues(lm.fit)) **375**