605 Discussion 12

Jose Mawyin

11/16/2019

Linear Regression Fit to Boston Housing Data

1. Introduction

This analysis will use housing data for 506 census tracts of Boston from the 1970 census collected by Harrison and Rubinfeld (1979) to test the validity of using linear regression models to predict the median value of owner-occupied homes.

2. Data

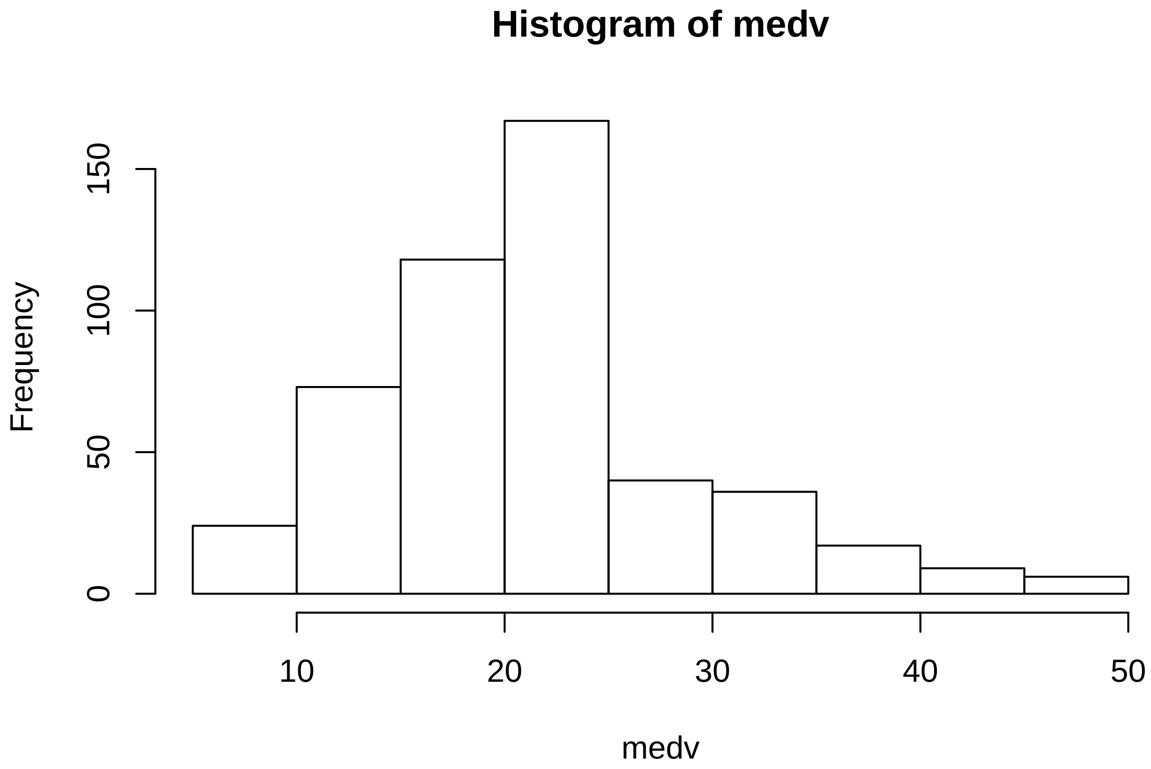
The original data are 506 observations on 14 variables, medv being the target variable:

1. *crim* per capita crime rate by town
2. *zn* proportion of residential land zoned for lots over 25,000 sq.ft
3. *indus* proportion of non-retail business acres per town
4. chas. Charles River dummy **dichotomous** variable (= 1 if tract bounds river; 0 otherwise)
5. *nox* nitric oxides concentration (parts per 10 million)
6. *rm* average number of rooms per dwelling
7. *age* proportion of owner-occupied units built prior to 1940
8. *dis* weighted distances to five Boston employment centres
9. *rad* index of accessibility to radial highways
10. *tax* full-value property-tax rate per USD 10,000
11. *ptratio* pupil-teacher ratio by town
12. *b* 1000(B-0.63)^2 where B is the proportion of blacks by town
13. *lstat* percentage of lower status of the population

The corrected data set has the following additional columns:

14. *cmedv* corrected median value of owner-occupied homes in USD 1000's
15. *town* name of town
16. *tract* census tract
17. *lon* longitude of census tract
18. *lat* latitude of census tract

Below is a histogram showing a distribution of the median value of owner-occupied homes in USD 1000's.



3. Linear Regresssion

Three different linear regression models were generated. The first one "Boston_cmedv.lm" that took into account all independent variables available in the dataset.

```
Boston_cmedv.lm <- lm(medv ~ crim + zn + indus + chas + nox +
    rm + dis + age + rad + ptratio + b + lstat, data = BostonHousing2)
```

A second model that used Backward Elimination of Factors to remove those independent variables with a low p value indicating that did not contribute much to the response variable "medv".

```
Boston_cmedv.lmV2 <- lm(medv ~ nox + rm + dis + ptratio + b +
    lstat, data = BostonHousing2)
```

A third model that tried to better fit independent variables dis and lstat into the model by considering them as **non-linear** variables.

```
Boston_cmedv.lmV3 <- lm(medv ~ nox + rm + log2(dis) + ptratio +
    b + log(lstat, base = 1/2), data = BostonHousing2)
```

4. Analysis

The first model using all 12 independent variables had a R^2 of *0.7678*. The second model that used only 6 independent variables had a R^2 of *0.7513*. The third model that took into account non-linearities in some of the 6 independent variables from model 2 had a R^2 of *0.7837*.

```r
summary(Boston_cmedv.lm)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + indus + chas + nox + rm + dis +
##     age + rad + ptratio + b + lstat, data = BostonHousing2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0286  -2.3752  -0.4924   1.7792  16.1796
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.250023   4.186502   7.226 1.99e-12 ***
## crim         -0.104842   0.026649  -3.934 9.59e-05 ***
## zn            0.023348   0.011194   2.086 0.037522 *
## indus        -0.138728   0.046277  -2.998 0.002862 **
## chas1         0.919597   0.749998   1.226 0.220753
## nox         -13.525418   3.110508  -4.348 1.68e-05 ***
## rm            3.863980   0.364073  10.613  < 2e-16 ***
## dis          -1.216163   0.163996  -7.416 5.57e-13 ***
## age          -0.024510   0.010872  -2.254 0.024624 *
## rad           0.060368   0.033714   1.791 0.073996 .
## ptratio      -0.876499   0.107224  -8.175 2.70e-15 ***
## b             0.008126   0.002174   3.737 0.000209 ***
## lstat        -0.350952   0.043404  -8.086 5.14e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.837 on 477 degrees of freedom
## Multiple R-squared:  0.7678, Adjusted R-squared:  0.762
## F-statistic: 131.4 on 12 and 477 DF,  p-value: < 2.2e-16
```

```r
summary(Boston_cmedv.lmV2)
```

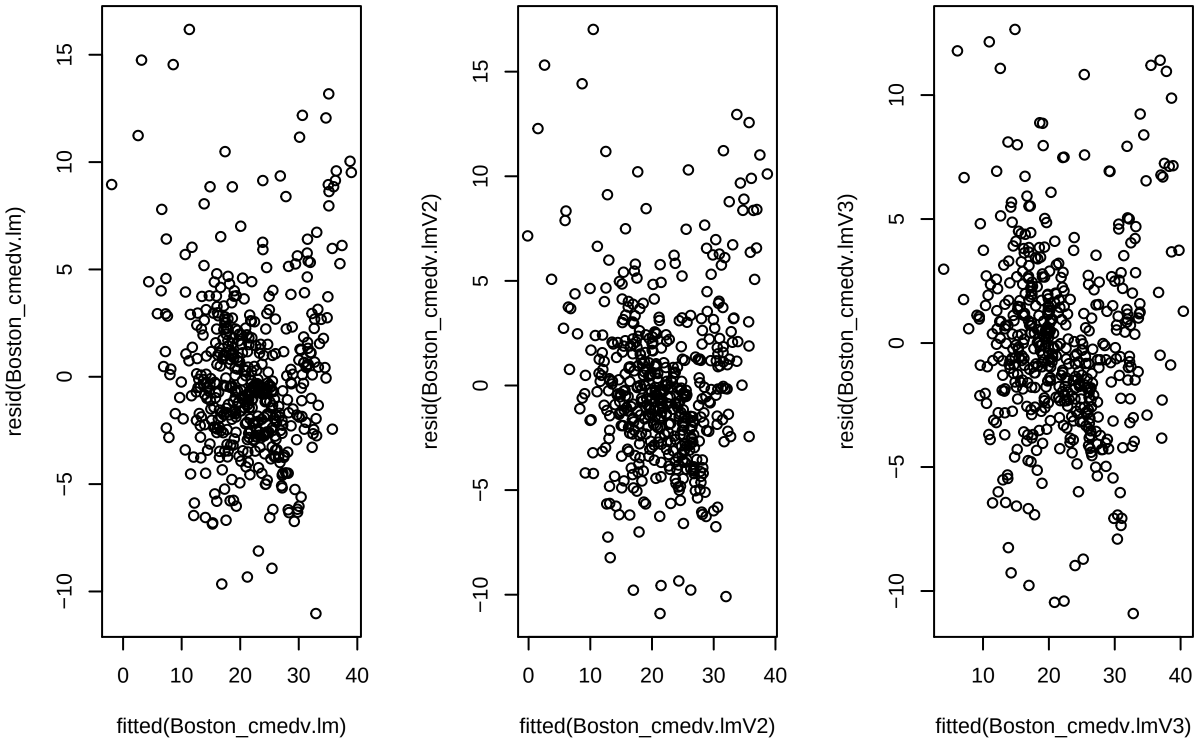
```
##
## Call:
## lm(formula = medv ~ nox + rm + dis + ptratio + b + lstat, data = BostonHousing2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.901  -2.350  -0.459   1.875  17.018
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.070615   4.020631   7.479 3.56e-13 ***
## nox         -17.293138   2.634768  -6.563 1.36e-10 ***
## rm            4.018509   0.352975  11.385  < 2e-16 ***
## dis          -0.786212   0.135207  -5.815 1.10e-08 ***
## ptratio      -1.011101   0.092263 -10.959  < 2e-16 ***
## b             0.008985   0.002143   4.192 3.29e-05 ***
## lstat        -0.418092   0.040271 -10.382  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.947 on 483 degrees of freedom
## Multiple R-squared:  0.7513, Adjusted R-squared:  0.7482
## F-statistic: 243.2 on 6 and 483 DF,  p-value: < 2.2e-16
```

```
summary(Boston_cmedv.lmV3)
```

```
##
## Call:
## lm(formula = medv ~ nox + rm + log2(dis) + ptratio + b + log(lstat,
##     base = 1/2), data = BostonHousing2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9084  -2.2245  -0.2441   1.7753  12.6464
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              45.240143   4.229140  10.697  < 2e-16 ***
## nox                     -15.887415   2.742186  -5.794 1.24e-08 ***
## rm                        3.147687   0.344231   9.144  < 2e-16 ***
## log2(dis)                -2.483261   0.403942  -6.148 1.65e-09 ***
## ptratio                  -0.871282   0.086901 -10.026  < 2e-16 ***
## b                         0.008895   0.001987   4.478 9.43e-06 ***
## log(lstat, base = 1/2)    4.954405   0.345263  14.350  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.68 on 483 degrees of freedom
## Multiple R-squared:  0.7837, Adjusted R-squared:  0.781
## F-statistic: 291.7 on 6 and 483 DF,  p-value: < 2.2e-16
```

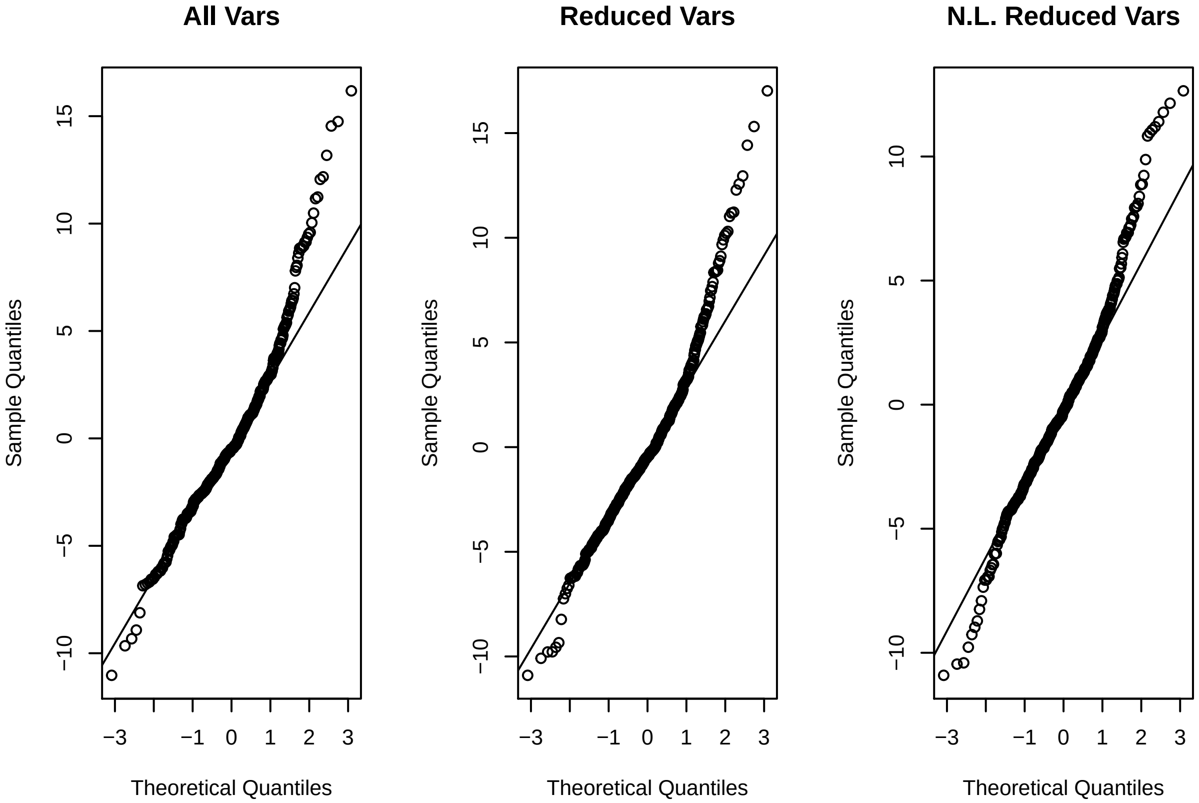
We notice an improvement in the residuals when we take into account non-linearities in some of the 6 independent variables from model 2. Comparing the rightmost residual plot before we see a more even spread around the zeroth line as compared to the middle plot that use the same independent variables as is.

```
par(mfrow = c(1, 3))
plot(fitted(Boston_cmedv.lm), resid(Boston_cmedv.lm))
plot(fitted(Boston_cmedv.lmV2), resid(Boston_cmedv.lmV2))
plot(fitted(Boston_cmedv.lmV3), resid(Boston_cmedv.lmV3))
```



Improvements in the fit are not obvious in the Q-Q plots below. Most changes happen in a reduction in the spread of the Sample Quartiles from -10:+15 to -10:+10 when taking into account non-linearities in the independent variables. Despite some effort, it was not possible to identify the causes for the tails to diverge from the normal distribution fit line.

```r
par(mfrow = c(1, 3))
qqnorm(resid(Boston_cmedv.lm), main = "All Vars")
qqline(resid(Boston_cmedv.lm))
qqnorm(resid(Boston_cmedv.lmV2), main = "Reduced Vars")
qqline(resid(Boston_cmedv.lmV2))
qqnorm(resid(Boston_cmedv.lmV3), main = "N.L. Reduced Vars")
qqline(resid(Boston_cmedv.lmV3))
```



5. Comments

We have seen how it was possible to reduce by half the number of independent variables in a linear regression model that still produced a similar R^2 value that describes the measured data. We also saw how taking into account non-linearities in independent variables was used to generate a better fit to the linear regression model.