15. This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

```
boston = read.csv("/Volumes/work/MTH522/data/Boston.csv")
head(boston)
```

```
##   X    crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
## 1 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
## 2 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
## 3 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
## 4 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
## 5 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
## 6 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
```

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

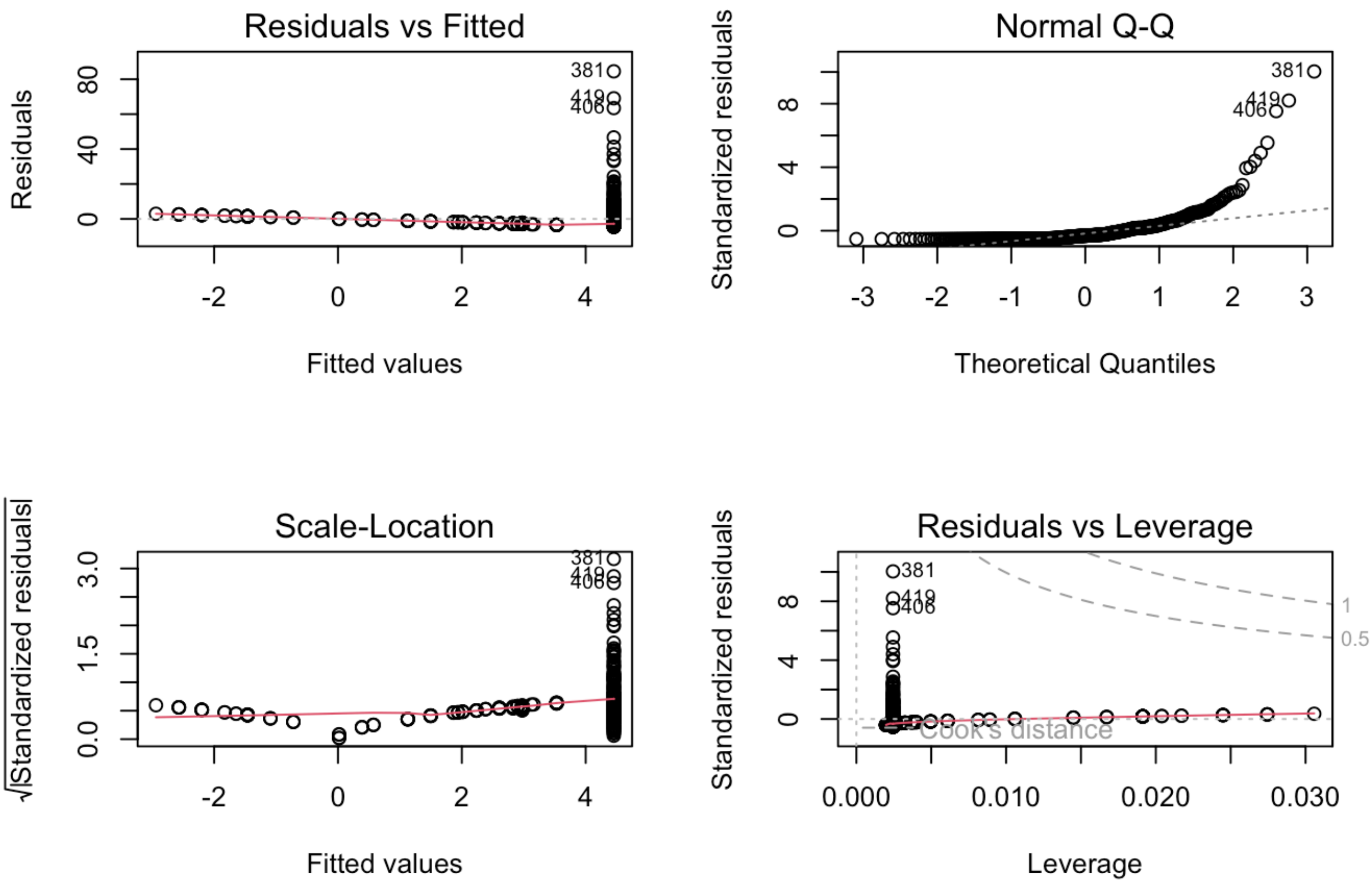
Crim (per capita crime rate) and zn (proportion of residential land zoned for lots over 25,000 sq.ft)

```
boston.zn <- lm(crim ~ zn, data=boston)
summary(boston.zn)
```

```
##
## Call:
## lm(formula = crim ~ zn, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.429 -4.222 -2.620  1.250 84.523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
## zn          -0.07393    0.01609  -4.594 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
## F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

Observation: 1. From the above summary, we can see that F-statistic is 21.1 and p-value is < 5.506e-06, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and zn.

```
par(mfrow = c(2, 2))
plot(boston.zn)
```

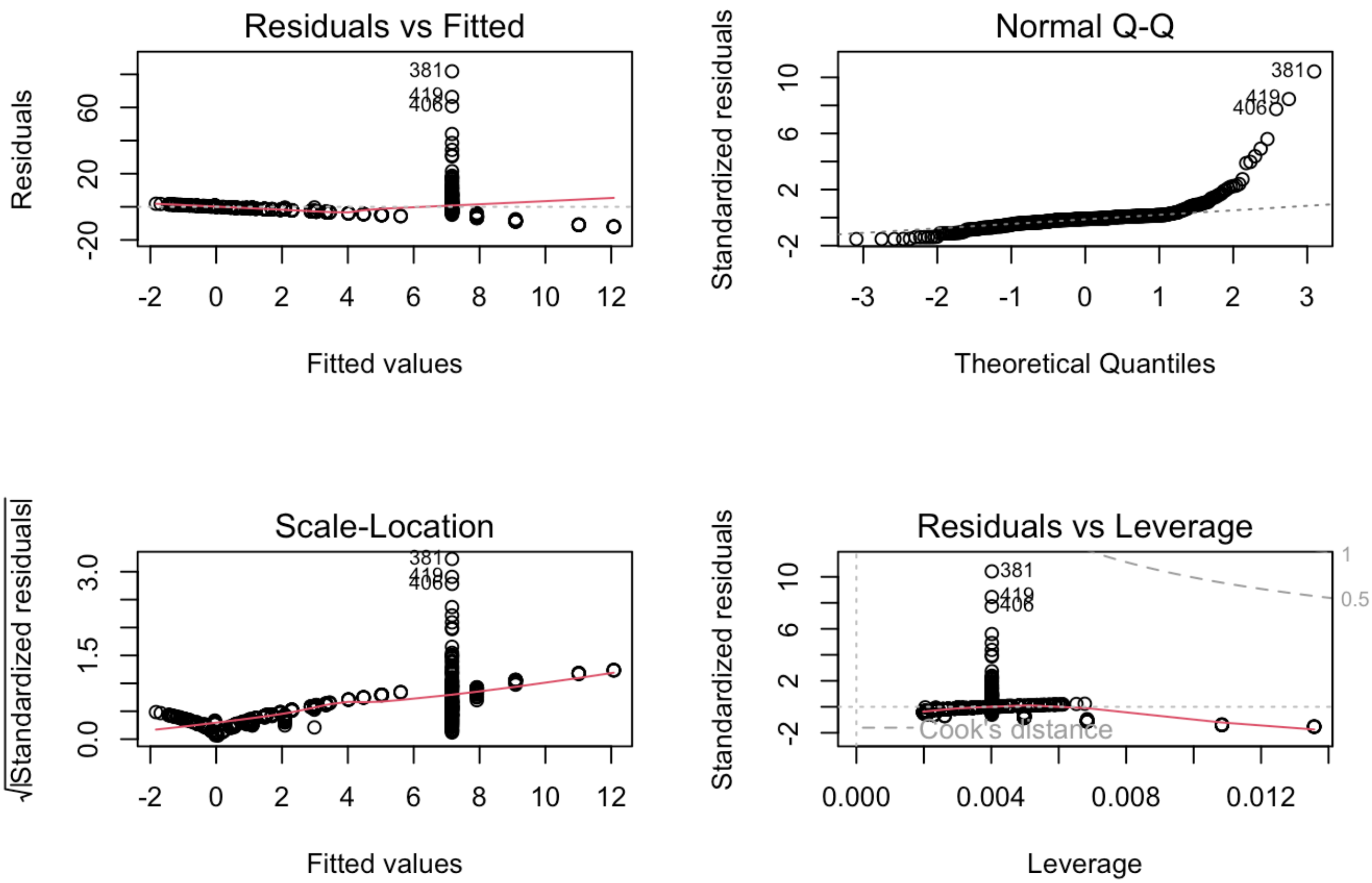


Per capita crime rate(crim) and Indus (proportion of non-retail business acres per town).

```
boston.indus <- lm(crim ~ indus, data=boston)
summary(boston.indus)
```

```
##
## Call:
## lm(formula = crim ~ indus, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.972  -2.698  -0.736   0.712  81.813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.06374    0.66723  -3.093  0.00209 **
## indus        0.50978    0.05102   9.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.866 on 504 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637
## F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(boston.indus)
```



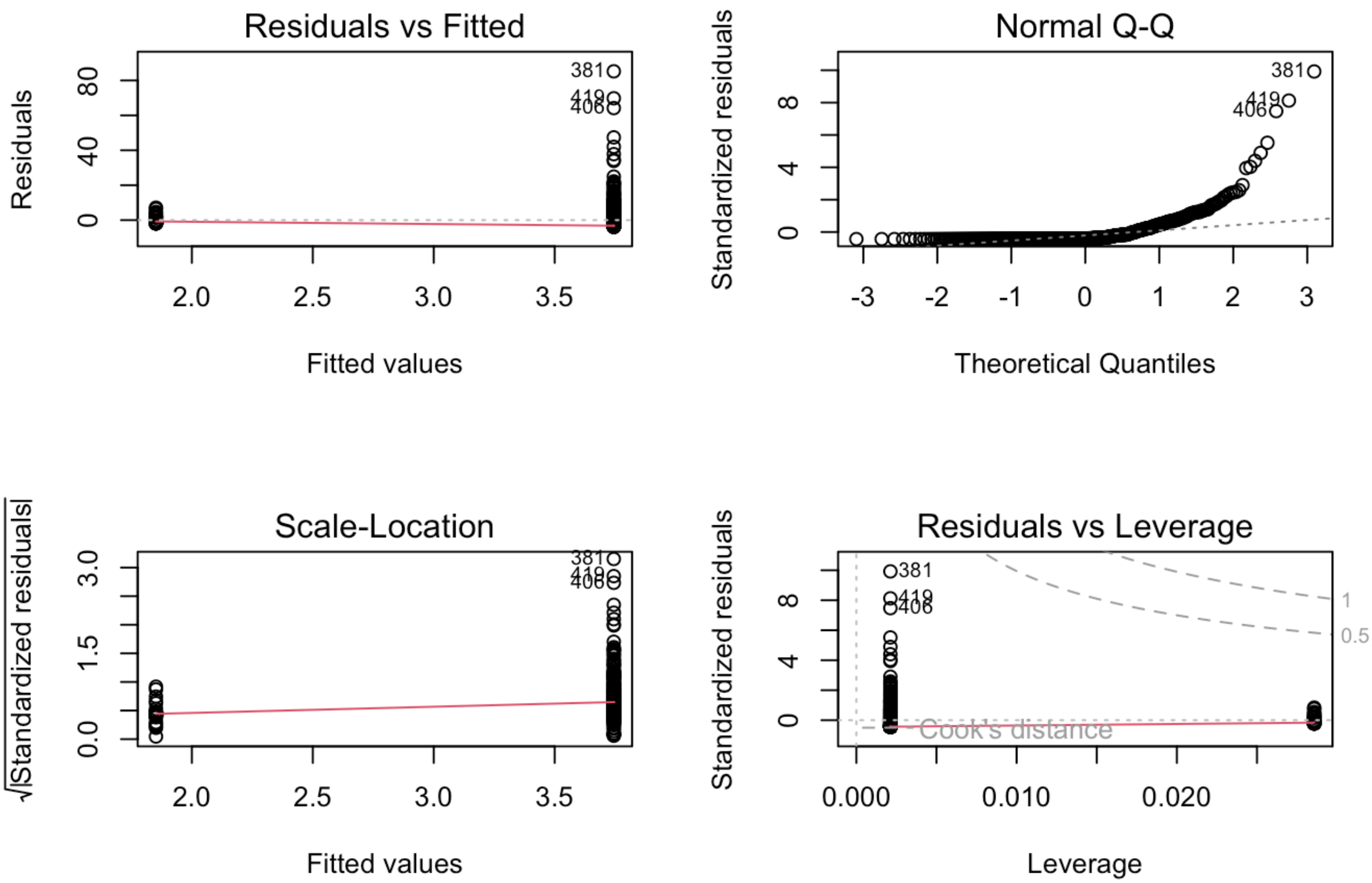
Observation: 1. From the above summary, we can see that F-statistic is 99.82 and p-value is < 2.2e-16, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and indus.

Per capita crime rate(crim) and chas (Charles River dummy variable)

```
boston.chas <- lm(crim ~ chas, data=boston)
summary(boston.chas)
```

```
##
## Call:
## lm(formula = crim ~ chas, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7444     0.3961   9.453   <2e-16 ***
## chas         -1.8928     1.5061  -1.257    0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.597 on 504 degrees of freedom
## Multiple R-squared:  0.003124,   Adjusted R-squared:  0.001146
## F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

```
par(mfrow = c(2, 2))
plot(boston.chas)
```



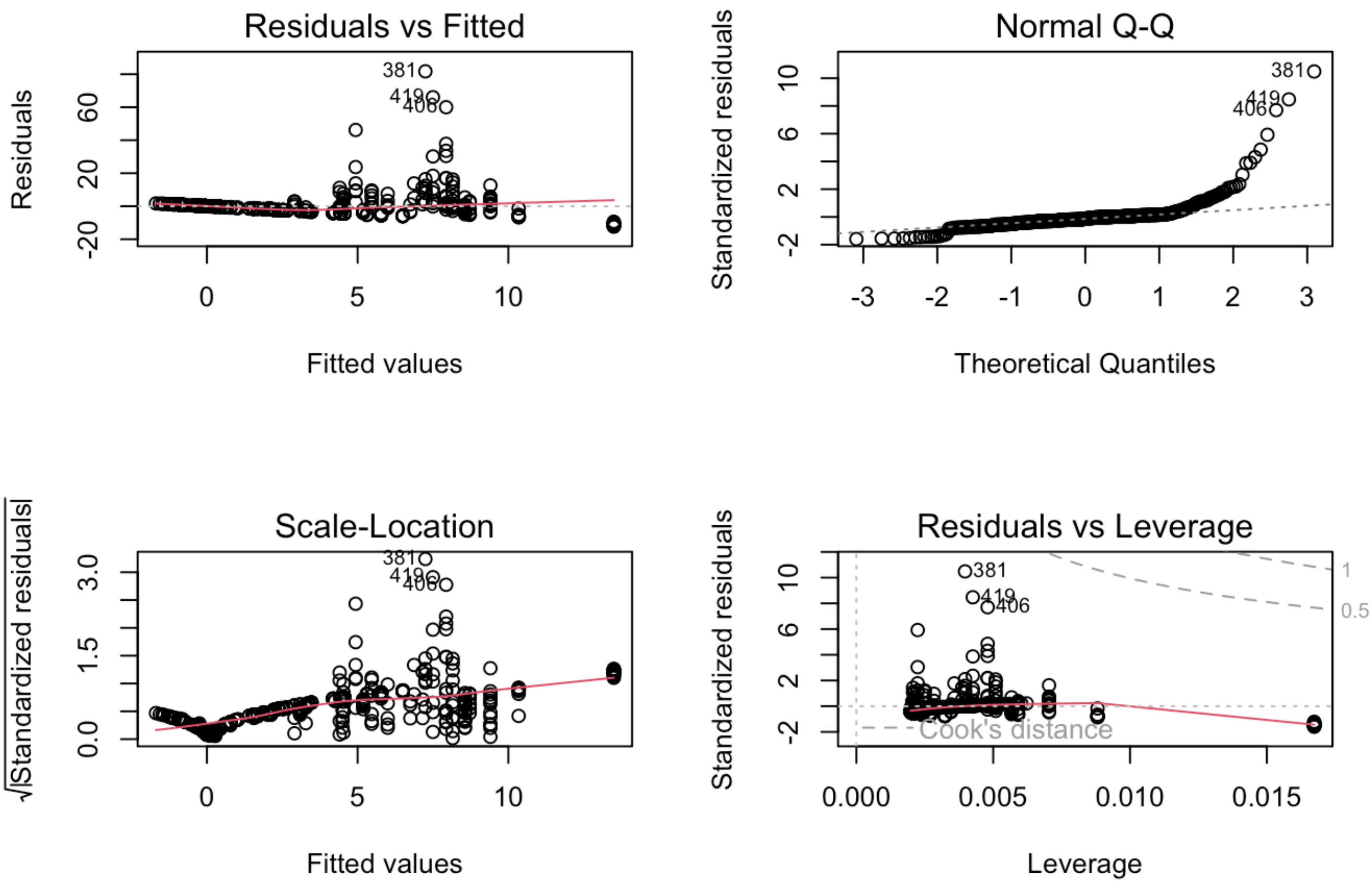
Observation: 1. The p-value of the model is 0.2094 which is greater than 0.05 and this means that the chances of having a null hypothesis are high and therefore chas is not statistically significant. 2. We can also see from the plot that, increase in the per capita crime rate is not effecting the change in chas. we can conclude that there is no relationship between chas and crim.

Per capita crime rate(crim) and nox (nitrogen oxides concentration)

```
boston.nox <- lm(crim ~ nox, data=boston)
summary(boston.nox)
```

```
##
## Call:
## lm(formula = crim ~ nox, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.371  -2.738  -0.974   0.559  81.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
## nox           31.249      2.999  10.419  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(boston.nox)
```



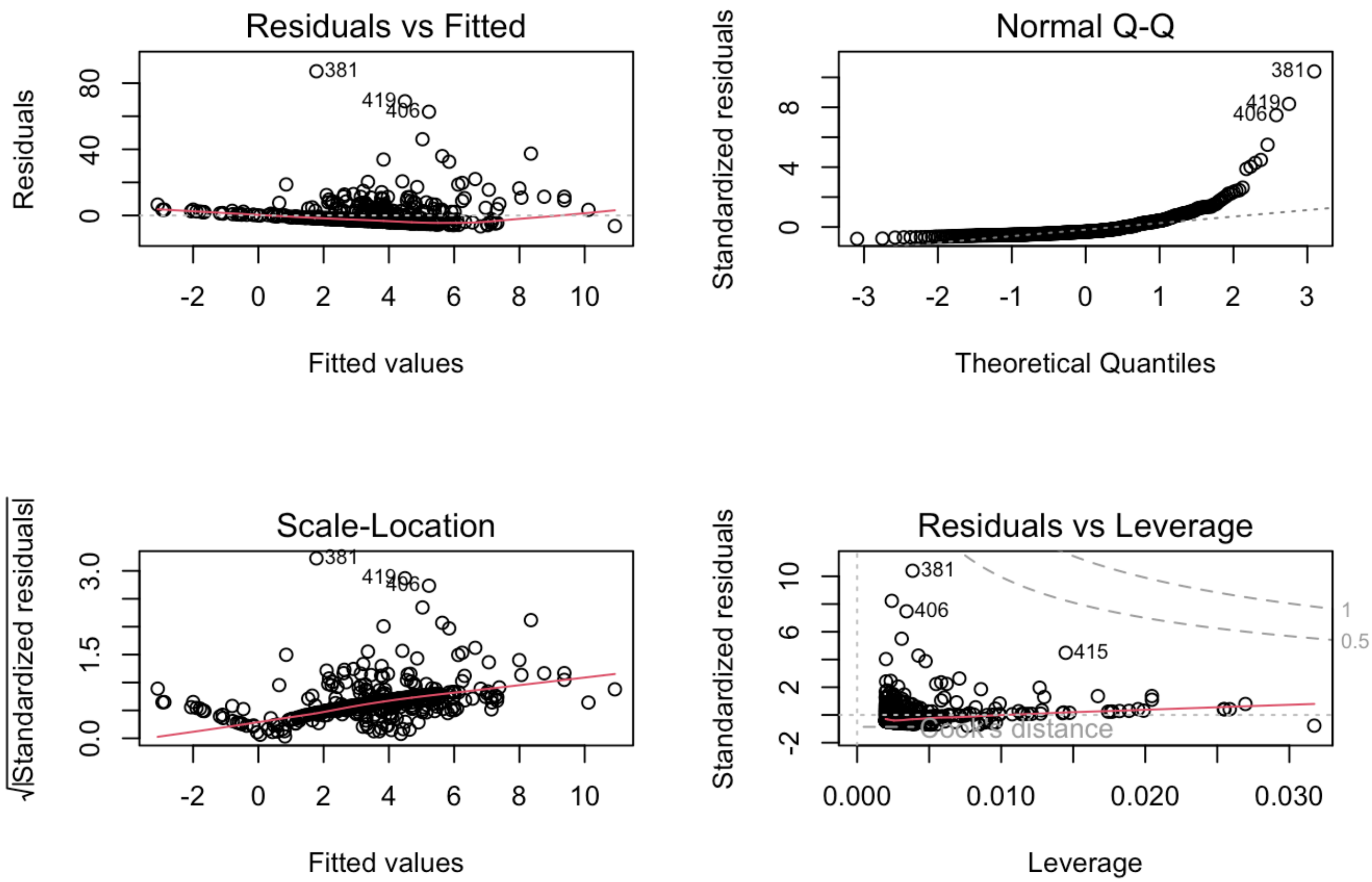
Observations: 1. From the above summary, we can see that F-statistic is 108.6 and p-value is < 2.2e-16, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and nox.

Per capita crime rate(crim) and rm (average number of rooms per dwelling)

```
boston.rm <- lm(crim ~ rm, data=boston)
summary(boston.rm)
```

```
##
## Call:
## lm(formula = crim ~ rm, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.604 -3.952 -2.654  0.989 87.197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.482      3.365   6.088 2.27e-09 ***
## rm            -2.684      0.532  -5.045 6.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.401 on 504 degrees of freedom
## Multiple R-squared:  0.04807,    Adjusted R-squared:  0.04618
## F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

```
par(mfrow = c(2, 2))
plot(boston.rm)
```

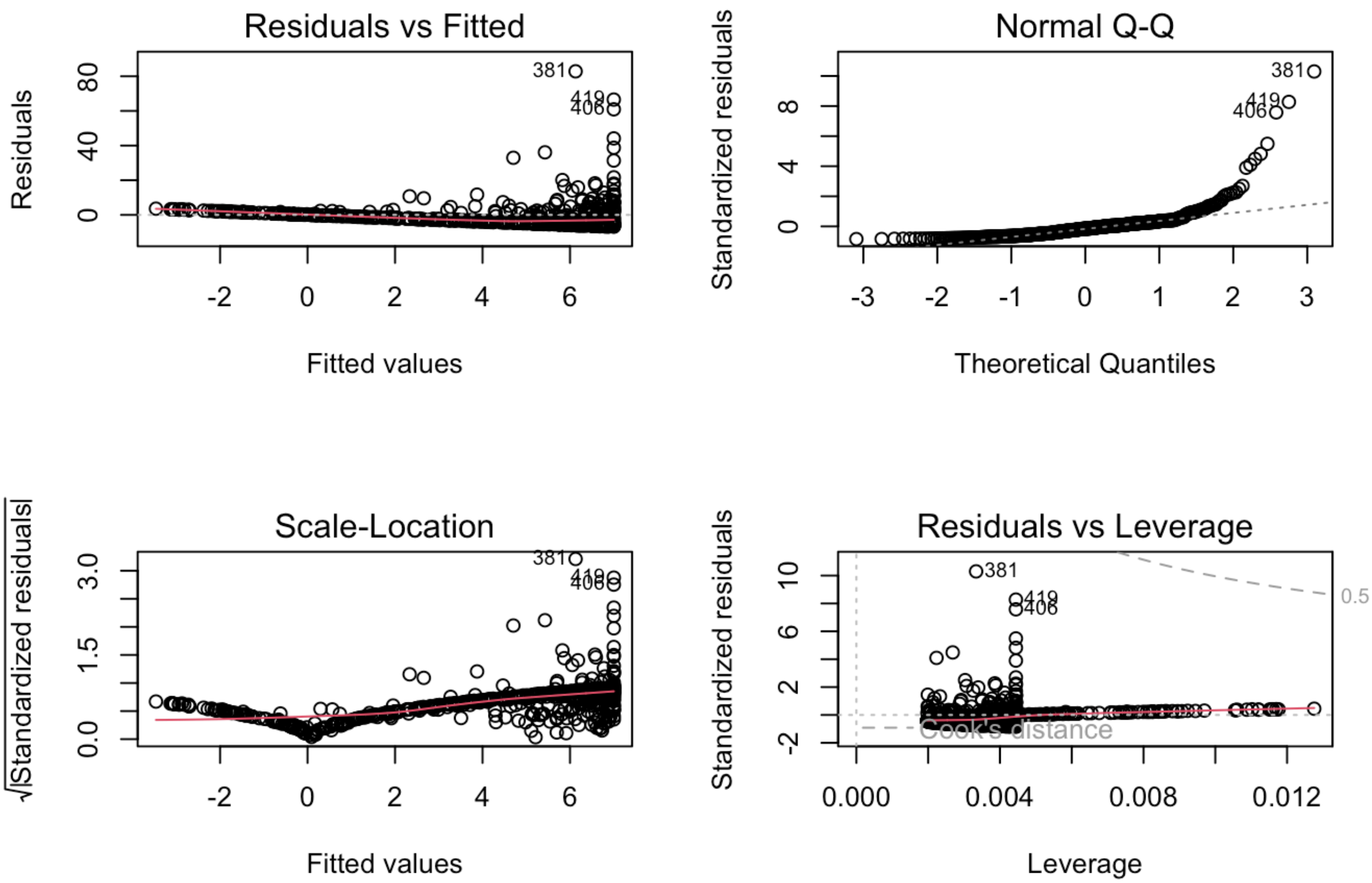


Observation: 1. From the above summary, we can see that F-statistic is 25.45 and p-value is < 6.347e-07, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and rm. 2. But this influence is low because of the low R squared value of 0.04807 and Adjusted R squared value of 0.04618.

```
boston.age <- lm(crim ~ age, data=boston)
summary(boston.age)
```

```
##
## Call:
## lm(formula = crim ~ age, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.789 -4.257 -1.230  1.527 82.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
## age          0.10779    0.01274   8.463 2.85e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
## F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

```
par(mfrow = c(2, 2))
plot(boston.age)
```



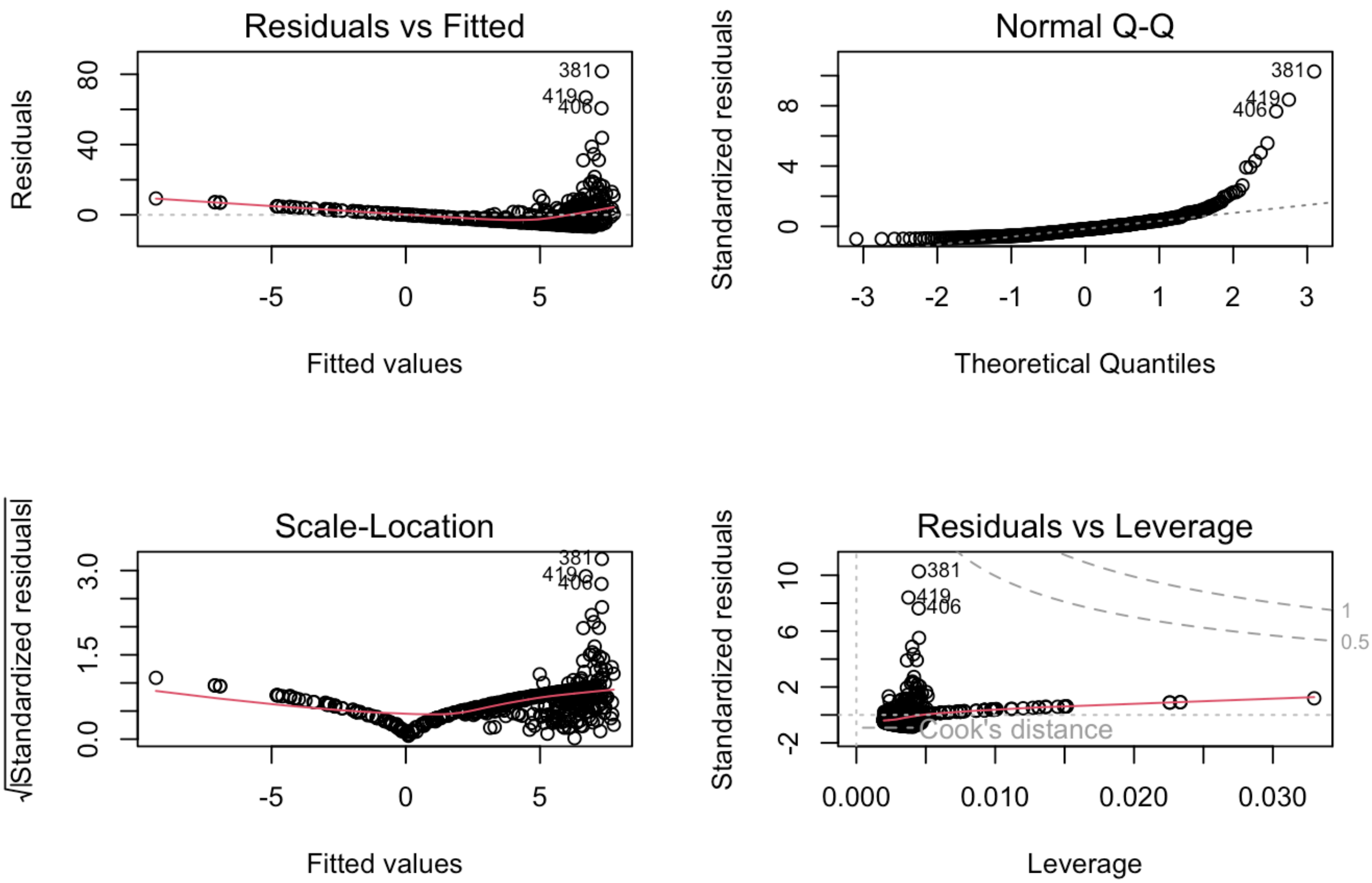
Observation: 1. From the above summary, we can see that F-statistic is 71.62 and p-value is < 2.855e-16, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and age. 2. But influence is quite small due to the low values of the R squared value of 0.1244 and Adjusted R squared value of 0.1227.

Per capita crime rate(crim) and dis (weighted mean of distances to five Boston employment centers).

```
boston.dis <- lm(crim ~ dis, data=boston)
summary(boston.dis)
```

```
##
## Call:
## lm(formula = crim ~ dis, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.708 -4.134 -1.527  1.516 81.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4993     0.7304  13.006   <2e-16 ***
## dis          -1.5509     0.1683  -9.213   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.965 on 504 degrees of freedom
## Multiple R-squared:  0.1441, Adjusted R-squared:  0.1425
## F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(boston.dis)
```



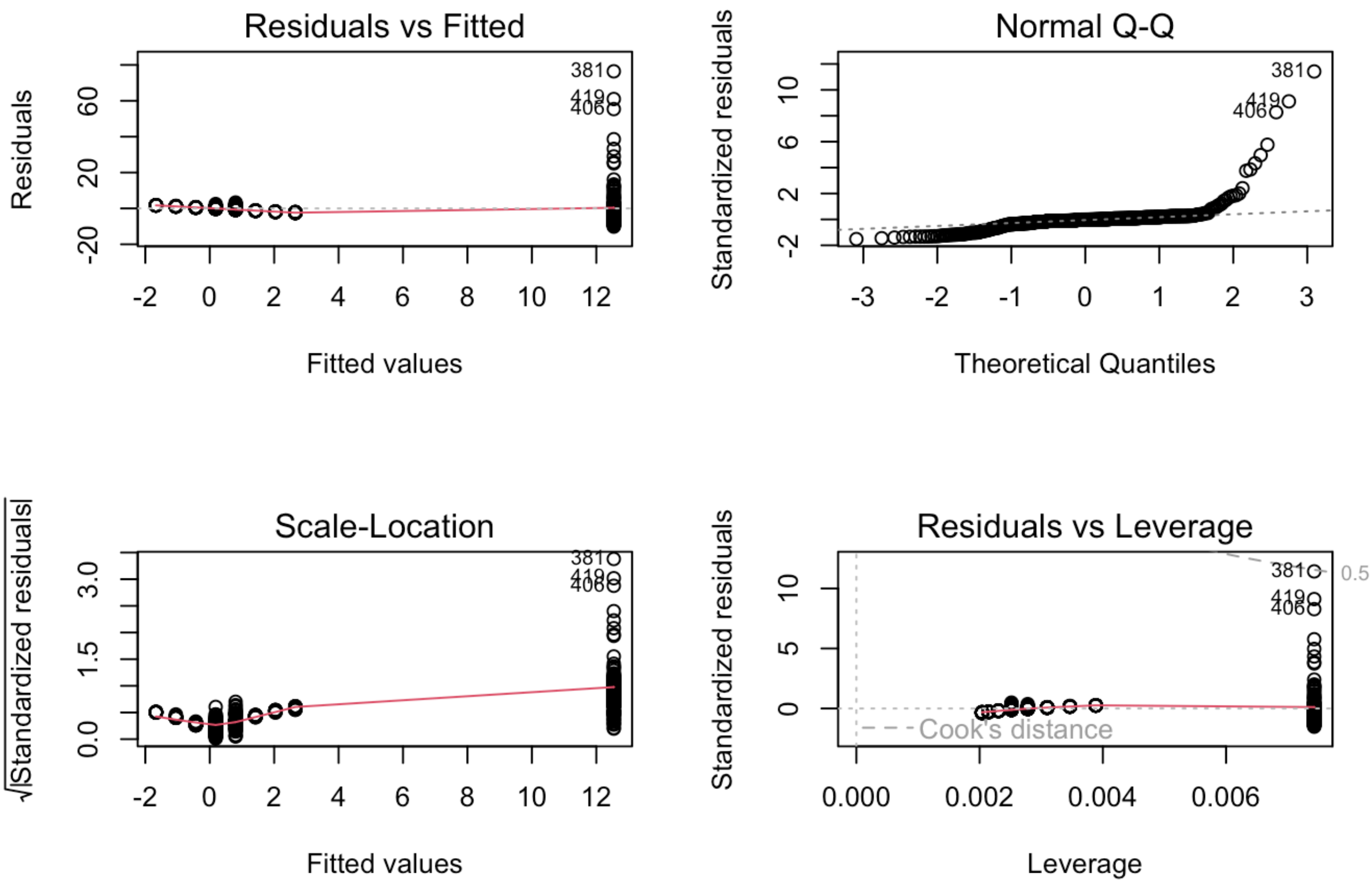
Observation: 1. From the above summary, we can see that F-statistic is 84.89 and p-value is < 2.2e-16, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and dis. 2. But influence is quite small due to the low values of the R squared value of 0.1441 and Adjusted R squared value of 0.1425.

Per capita crime rate(crim) and rad (index of accessibility to radial highways).

```
boston.rad <- lm(crim ~ rad, data=boston)
summary(boston.rad)
```

```
##
## Call:
## lm(formula = crim ~ rad, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.164  -1.381  -0.141   0.660  76.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
## rad          0.61791    0.03433  17.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.718 on 504 degrees of freedom
## Multiple R-squared:  0.3913, Adjusted R-squared:   0.39
## F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(boston.rad)
```



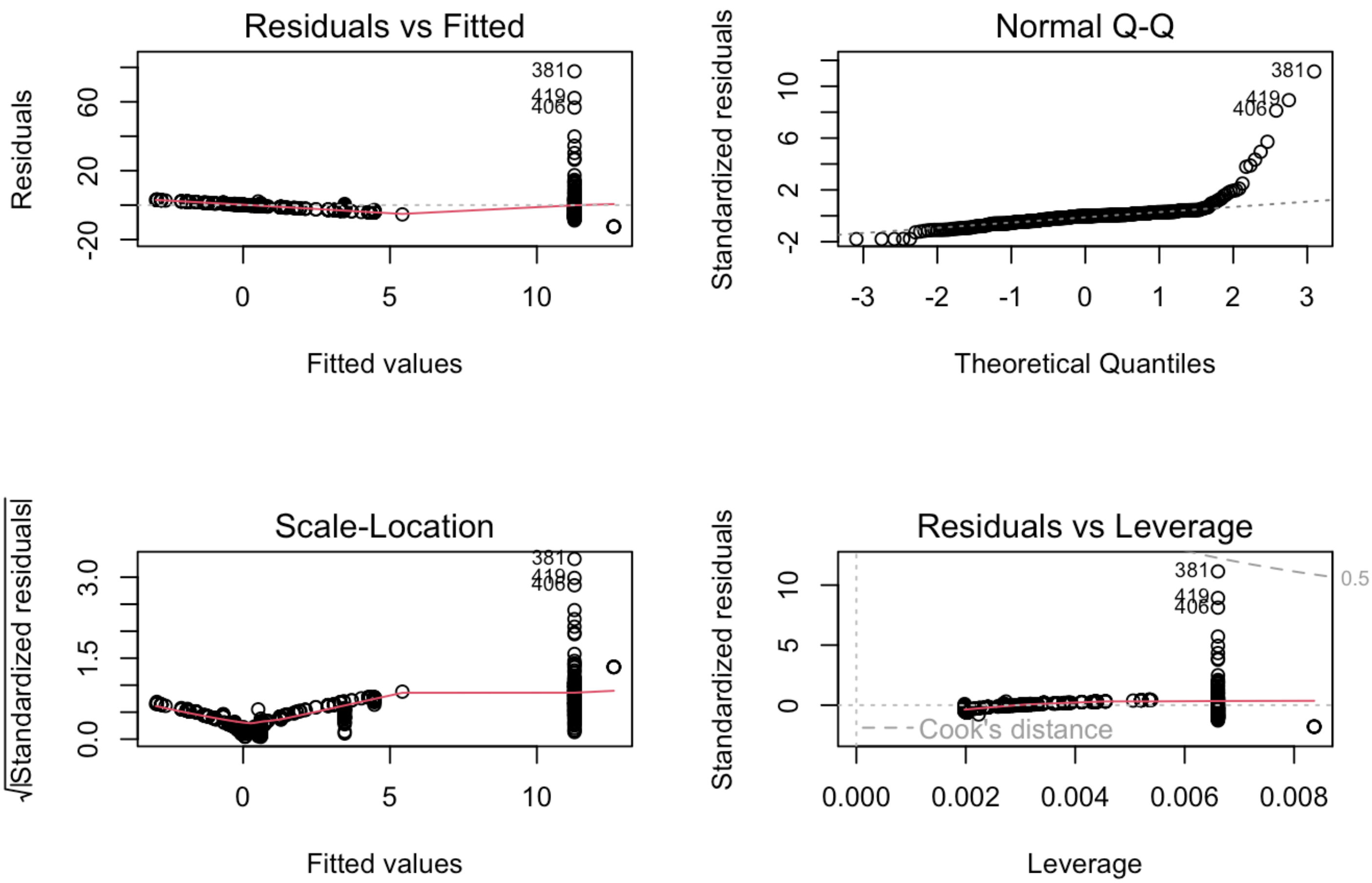
Observation: 1. From the above summary, we can see that F-statistic is 323.9 and p-value is < 2.2e-16, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and rad. 2. But influence is quite small due to the low values of the R squared value of 0.3913 and Adjusted R squared value of 0.39.

Per capita crime rate(crim) and tax (full-value property-tax rate per $10,000).

```
boston.tax <- lm(crim ~ tax, data=boston)
summary(boston.tax)
```

```
##
## Call:
## lm(formula = crim ~ tax, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.513  -2.738  -0.194   1.065  77.696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
## tax          0.029742   0.001847   16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.997 on 504 degrees of freedom
## Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
## F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(boston.tax)
```



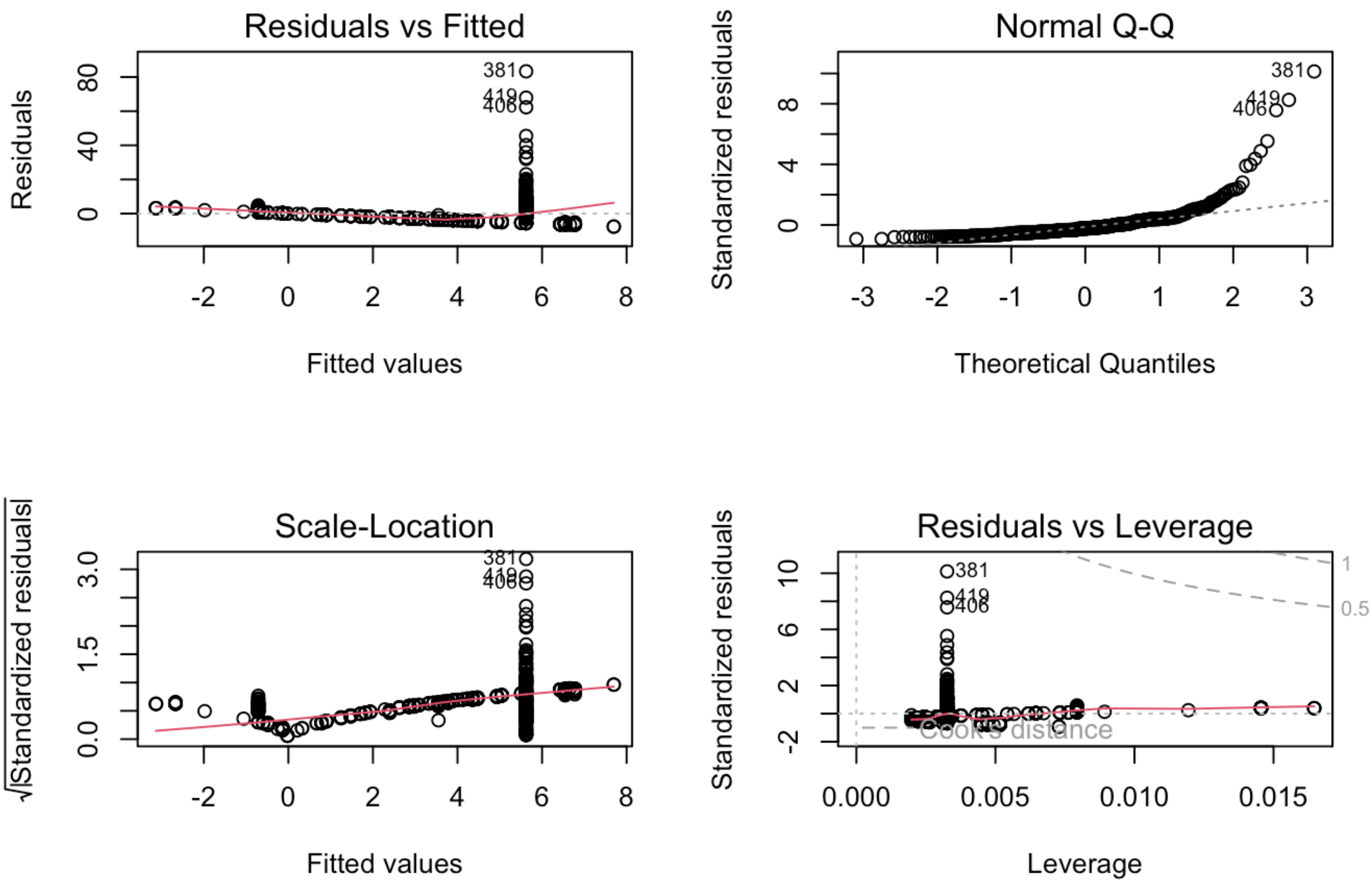
Observation: 1. From the above summary, we can see that F-statistic is 259.2 and p-value is < 2.2e-16, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and rad. 2. But influence is quite small due to the low values of the R squared value of 0.3396 and Adjusted R squared value of 0.3383.

Per capita crime rate(crim) and ptratio (pupil-teacher ratio by town)

```
boston.ptratio <- lm(crim ~ ptratio, data=boston)
summary(boston.ptratio)
```

```
##
## Call:
## lm(formula = crim ~ ptratio, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.654 -3.985 -1.912  1.825 83.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
## ptratio       1.1520     0.1694   6.801 2.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.24 on 504 degrees of freedom
## Multiple R-squared:  0.08407,    Adjusted R-squared:  0.08225
## F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

```
par(mfrow = c(2,2))
plot(boston.ptratio)
```



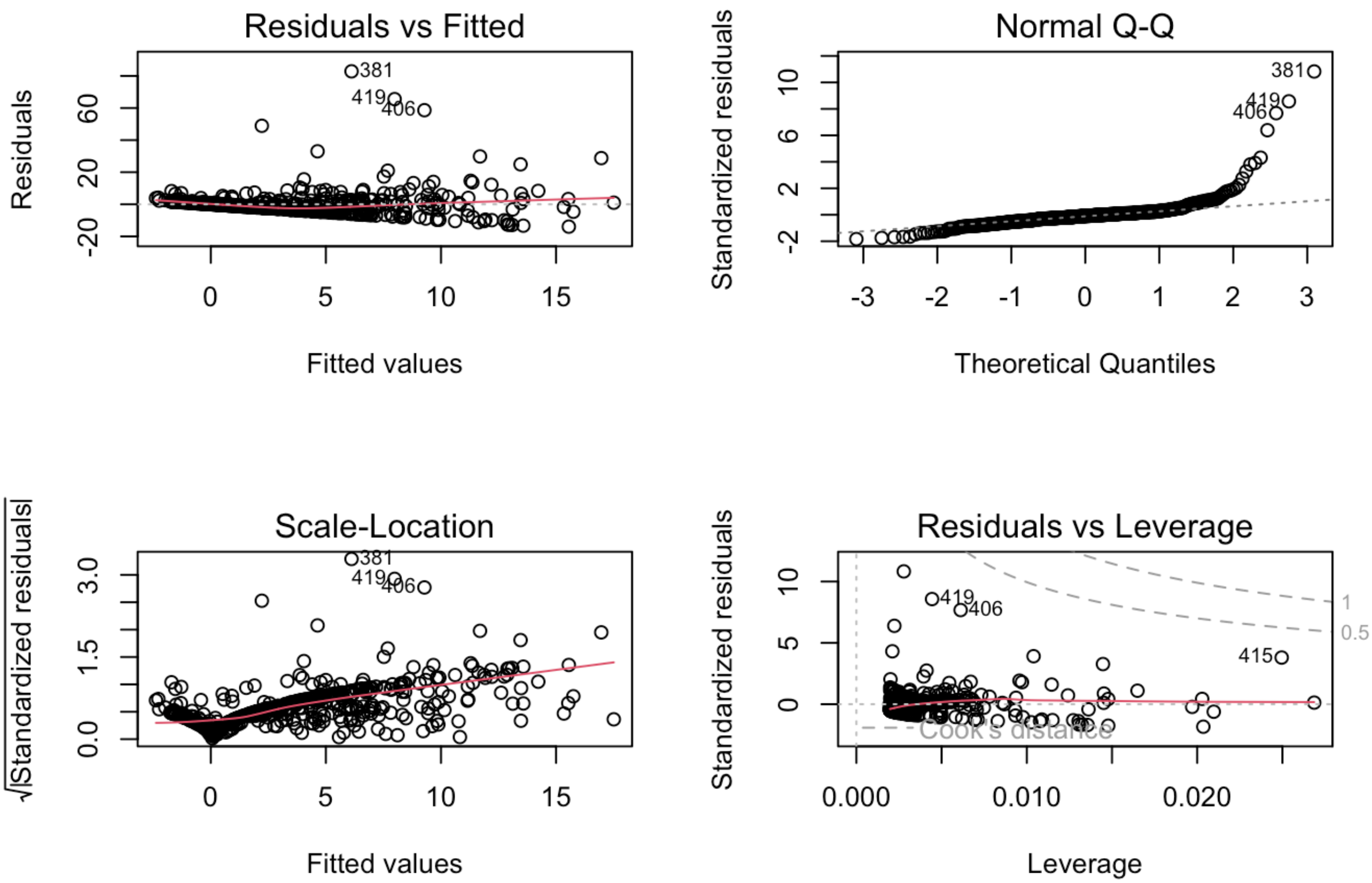
Observation: 1. From the above summary, we can see that F-statistic is 46.25 and p-value is < 2.943e-11, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and ptratio. 2. But influence is quite small due to the low values of the R squared value of 0.08407 and Adjusted R squared value of 0.08225.

Per capita crime rate(crim) and lstat (lower status of the population (percent)).

```
boston.lstat <- lm(crim ~ lstat,data=boston)
summary(boston.lstat)
```

```
##
## Call:
## lm(formula = crim ~ lstat, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
## lstat        0.54880    0.04776  11.491  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
## F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(boston.lstat)
```



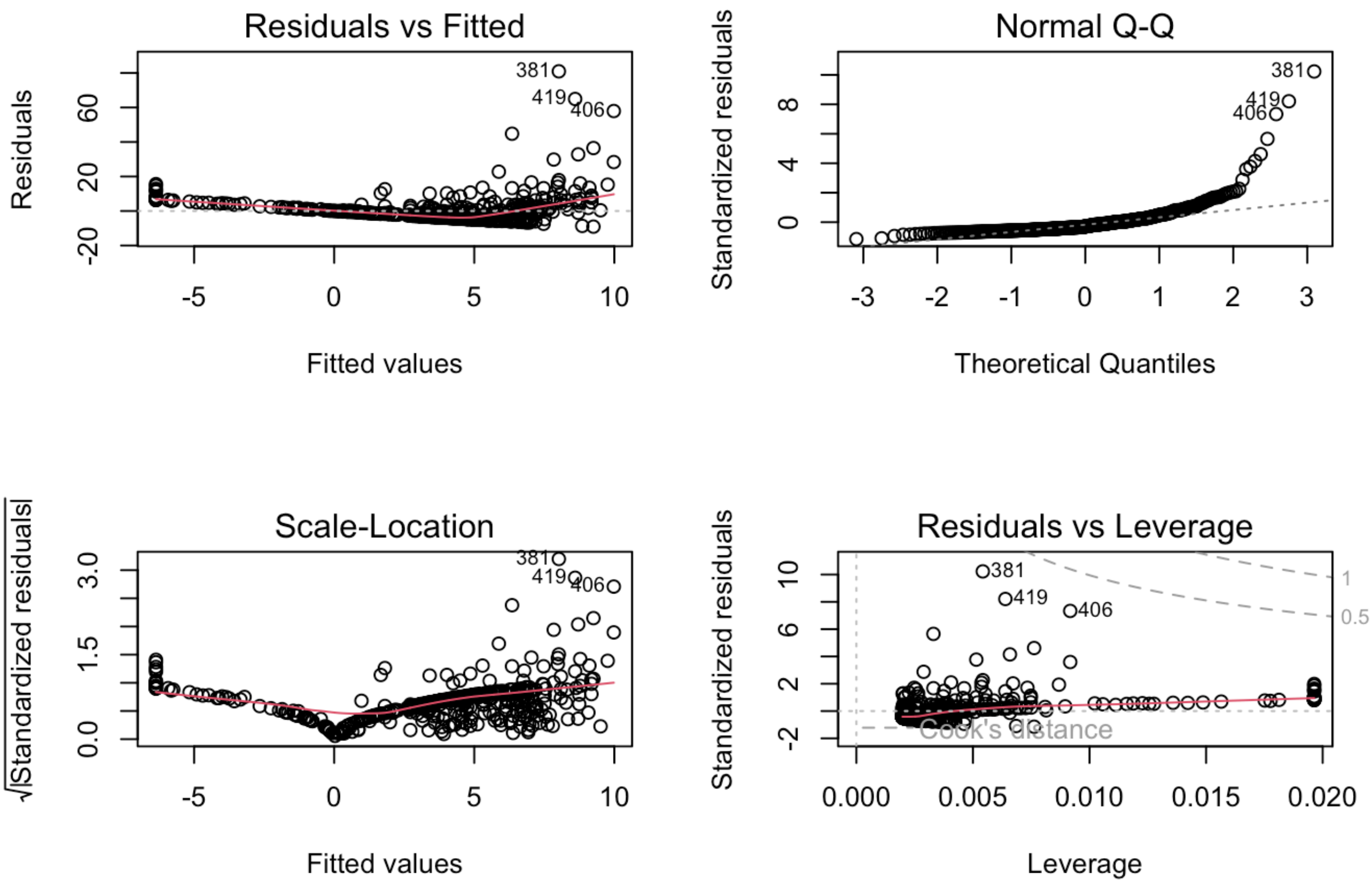
Observation: 1. From the above summary, we can see that F-statistic is 132 and p-value is < 2.2e-16, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and lstat. 2. But influence is quite small due to the low values of the R squared value of 0.2076 and Adjusted R squared value of 0.206.

Per capita crime rate(crim) and medv (median value of owner-occupied homes in $1000s).

```
boston.medv <- lm(crim ~ medv,data = boston)
summary(boston.medv)
```

```
##
## Call:
## lm(formula = crim ~ medv, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## medv        -0.36316    0.03839   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(boston.medv)
```



Observation: 1. From the above summary, we can see that F-statistic is 89.49 and p-value is < 2.2e-16, meaning the chance of having a null hypothesis (β0) is very low. So, there is a statistically significant association between crim and medv. 2. But influence is quite small due to the low values of the R squared value of 0.1508 and Adjusted R squared value of 0.1491

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : βj = 0?

```
boston.allvar <- lm(crim~.-X,data = boston)
summary(boston.allvar)
```

```
##
## Call:
## lm(formula = crim ~ . - X, data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.534 -2.248 -0.348  1.087 73.923
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.7783938  7.0818258   1.946 0.052271 .
## zn           0.0457100  0.0187903   2.433 0.015344 *
## indus       -0.0583501  0.0836351  -0.698 0.485709
## chas        -0.8253776  1.1833963  -0.697 0.485841
## nox         -9.9575865  5.2898242  -1.882 0.060370 .
## rm           0.6289107  0.6070924   1.036 0.300738
## age         -0.0008483  0.0179482  -0.047 0.962323
## dis         -1.0122467  0.2824676  -3.584 0.000373 ***
## rad          0.6124653  0.0875358   6.997 8.59e-12 ***
## tax         -0.0037756  0.0051723  -0.730 0.465757
## ptratio     -0.3040728  0.1863598  -1.632 0.103393
## lstat        0.1388006  0.0757213   1.833 0.067398 .
## medv        -0.2200564  0.0598240  -3.678 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.46 on 493 degrees of freedom
## Multiple R-squared:  0.4493, Adjusted R-squared:  0.4359
## F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

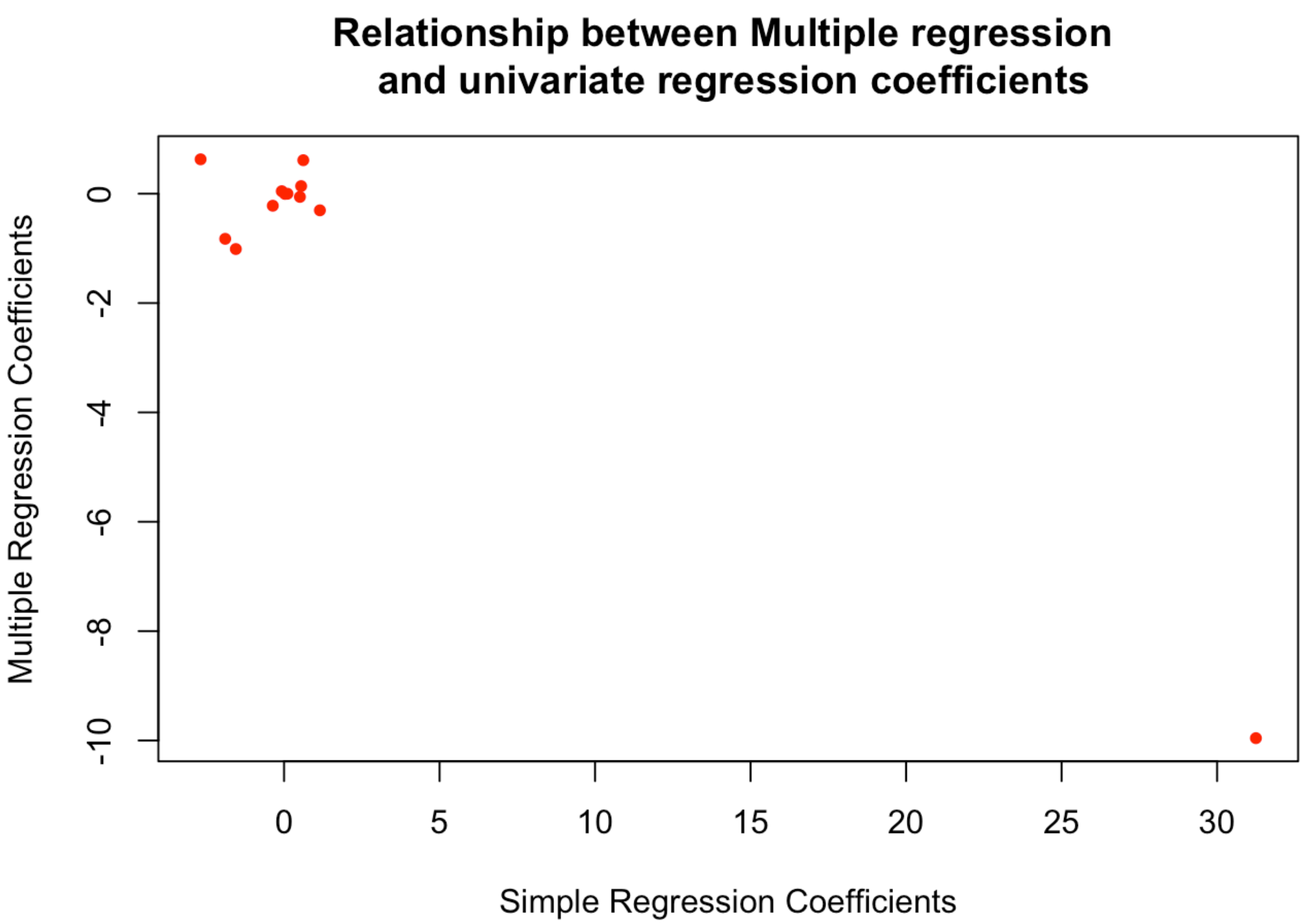
Observations: 1. we can only reject the null hypothesis for “zn”, ”dis”, ”rad”, ”black” and “medv” because these predictors are fitted multiple regression model are found to be statistically significant.

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regres- sion model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

```
df_coefs = data.frame("multi_coefs"=summary(boston.allvar)$coef[-1,1])
df_coefs$simple_coefs = NA
```

```
for(i in row.names(df_coefs)){
  reg_model = lm(crim~eval(str2lang(i)), data=boston)
  df_coefs[row.names(df_coefs)==i, "simple_coefs"] = coef(reg_model)[2]
}
```

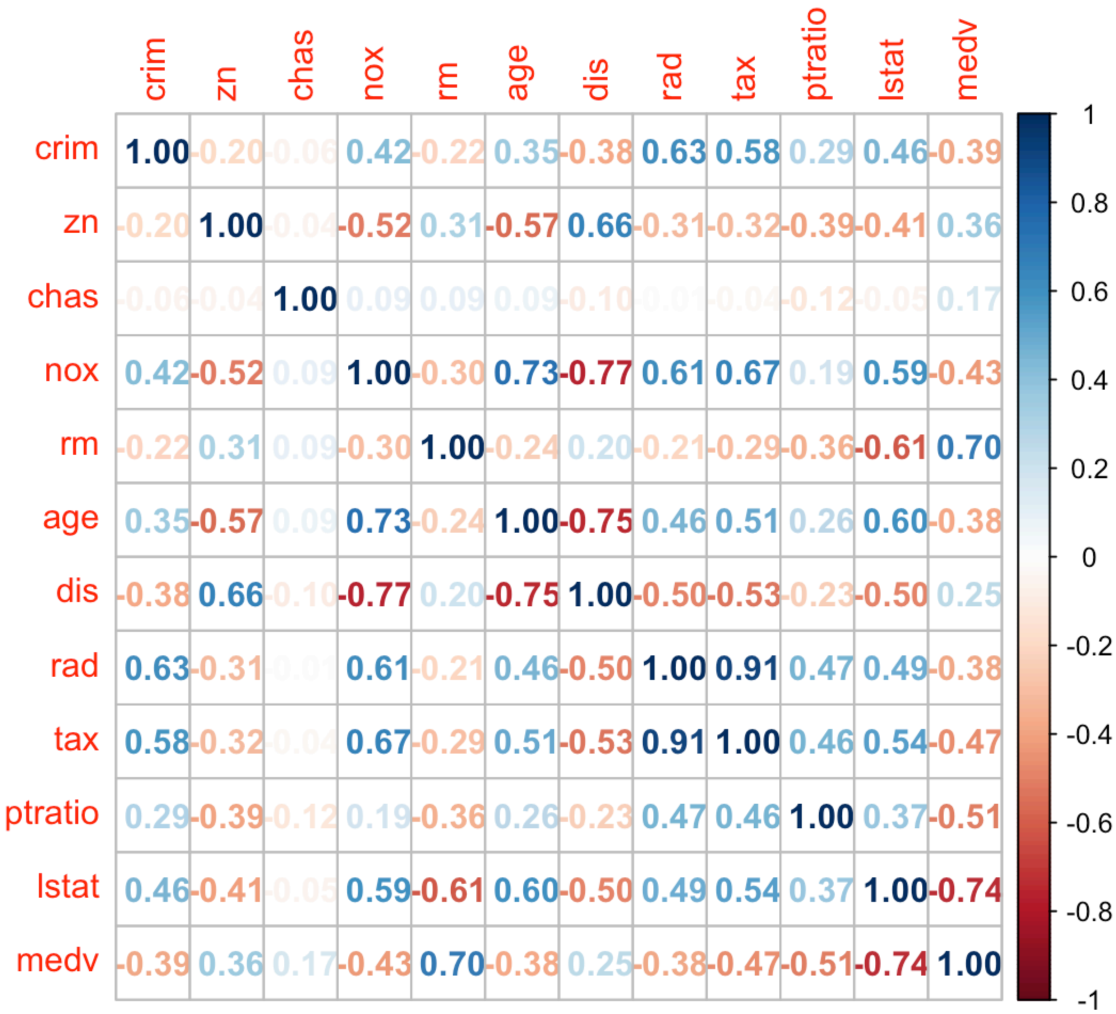
```
plot(df_coefs$simple_coefs, df_coefs$multi_coefs, col = "red", pch =20,
     xlab="Simple Regression Coefficients",
     ylab="Multiple Regression Coefficients",
     main = "Relationship between Multiple regression \n and univariate regression co
efficients")
```




```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corr <-round(cor(boston[-c(1,4)]),3)
corrplot(corr, method = "number")
```



note: The above figure is the Correlation between different variables of the Boston Dataset