Part 1: Simple Linear Regression and Diagnostics

Dominique Tanner

Background: ** Outliers are observations for which the response y_i is unusual for given predictor value x_i. In contrast, observations with high leverage have unusual value for x_i. ** How do we judge certain x is unusual? We use the following leverage statistic: h_i = 1/n + ((x_i- (x)) ¯^2)/(∑_(j=1)^n(x_j- x¯)^2). This expression shows that if x_i is far away from the mean x¯, then h_i will be large. ** Note: h_i is always between 1/n and 1. The average of all leverage values is given by 2/n. If an h_i far above the mean value 2/n, then suspect that the corresponding point has a high leverage.

Goal: * Using Boston data and its analysis to set a stage for analyzing housing prices locally and nationally. * Reviewing the data to analyze and interpret various response variables and predictors * download the Boston data from the MASS package

```
library(MASS)
data(Boston)
dim(Boston)
```

```
## [1] 506  14
```

```
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

```
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
summary(Boston)
```

```
##       crim                zn              indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   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
```

Comments: * review the documentation of the Boston data

```
?Boston
```

```
## starting httpd help server ... done
```

Comments: * The variables in the data folder can be recalled independently without referring to the primary folder ‘Boston’. To do this, use the ‘attach’ command.

```
attach(Boston)
```

Comments: * The response variable is medv.

Step 1: Perform a simple linear regression of medv on lstat.

```
Reglstat <- lm(medv ~ lstat)
Reglstat
```

```
##
## Call:
## lm(formula = medv ~ lstat)
##
## Coefficients:
## (Intercept)        lstat
##       34.55        -0.95
```

```
summary(Reglstat)
```

```
##
## Call:
## lm(formula = medv ~ lstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## 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:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

Comments: * the predictor 'lstat' is very, very significant (p-value < 0.001). * R^2 is significant. 54% of variation present in medv is accounted by the predictor lstat
* (p-value = 2.2*10^-16).

Step 2: look at the other information that is avaiable in the folder 'Reglstat'

```
names(Reglstat)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
coef(Reglstat)
```

```
## (Intercept)       lstat
##  34.5538409  -0.9500494
```

Step 3: getting the confidence intervals for the regression parameters of the model.

```
confint(Reglstat)
```

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

Comments: * For each of the predictor lstat = 5, 10 , and 15.

Step 4: Need to predict E(medv) along with a 95% confidence interval.

```
predict(Reglstat, data.frame(lstat = (c(5, 10, 15))), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 29.80359 29.00741 30.59978
## 2 25.05335 24.47413 25.63256
## 3 20.30310 19.73159 20.87461
```

Comments: * For each of the predictor lstat = 5, 10 , and 15.

Step 5: Need to predict medv along with a 95% confidence interval.

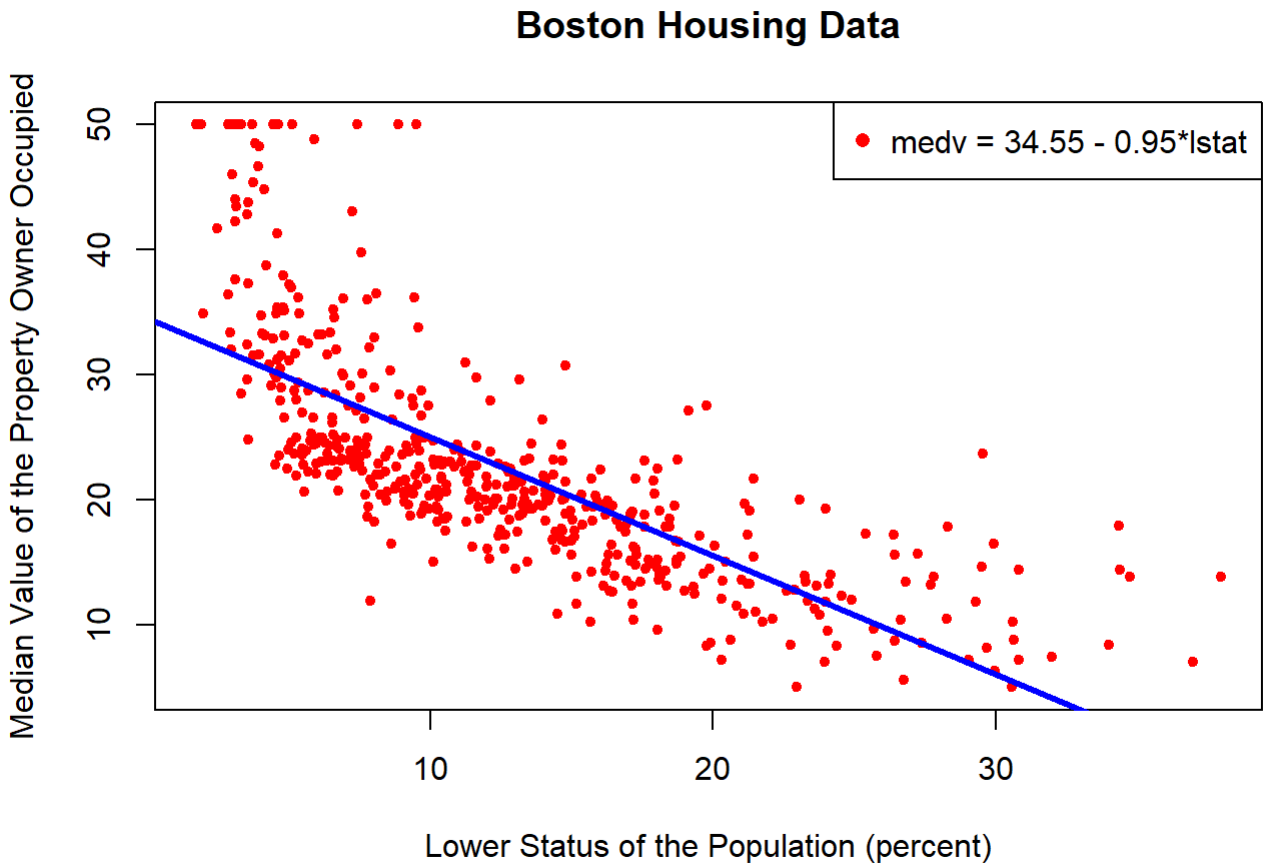
```
predict(Reglstat, data.frame(lstat = (c(5, 10, 15))), interval = "prediction")
```

```
##        fit       lwr      upr
## 1 29.80359 17.565675 42.04151
## 2 25.05335 12.827626 37.27907
## 3 20.30310  8.077742 32.52846
```

Comments: * observe the differences between the confidence interval and prediction interval.

Step 6: Generate a scatter plot and draw the regression line on the scatter plot.

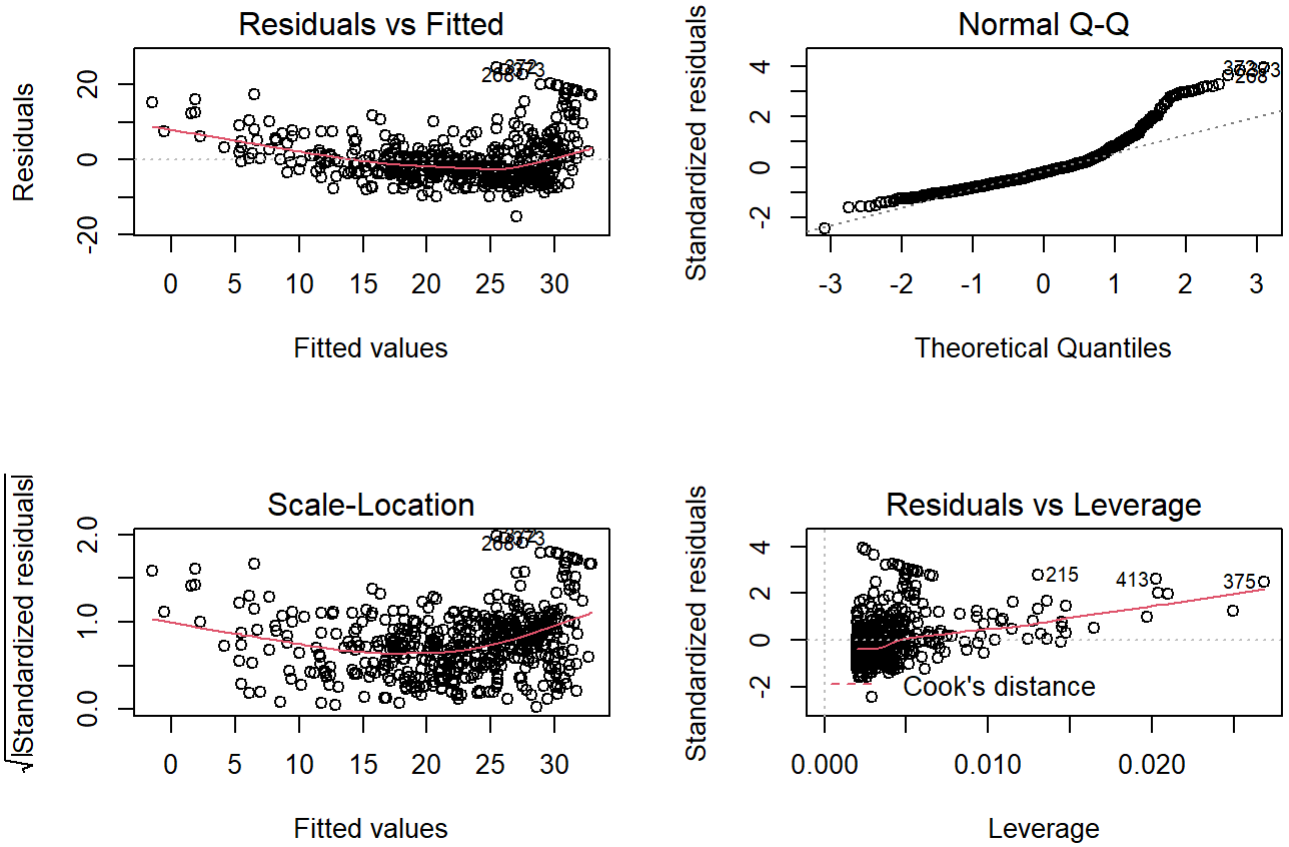
```
plot(lstat, medv, pch = 20, col = "red", xlab = "Lower Status of the Population (percent)",
     ylab = "Median Value of the Property Owner Occupied", main = "Boston Housing Data")

abline(Reglstat, lwd = 3, col = "blue")

legend("topright", legend = "medv = 34.55 - 0.95*lstat", pch = 16, col = "red")
```



Comments: * can discern non-linear relationship between lstat and medv. * The residual analysis graphs confirm this. * will try a quadratic regression model.

Step 7: Perform a residual analysis to identify outliers and points with high leverage. There will be four graphs; need all the graphs in a single frame. To do this, create a blank graph with room for four graphs arranged in the form of a 2 by 2 grid.

```
par(mfrow = c(2,2))
plot(Reglstat)
```



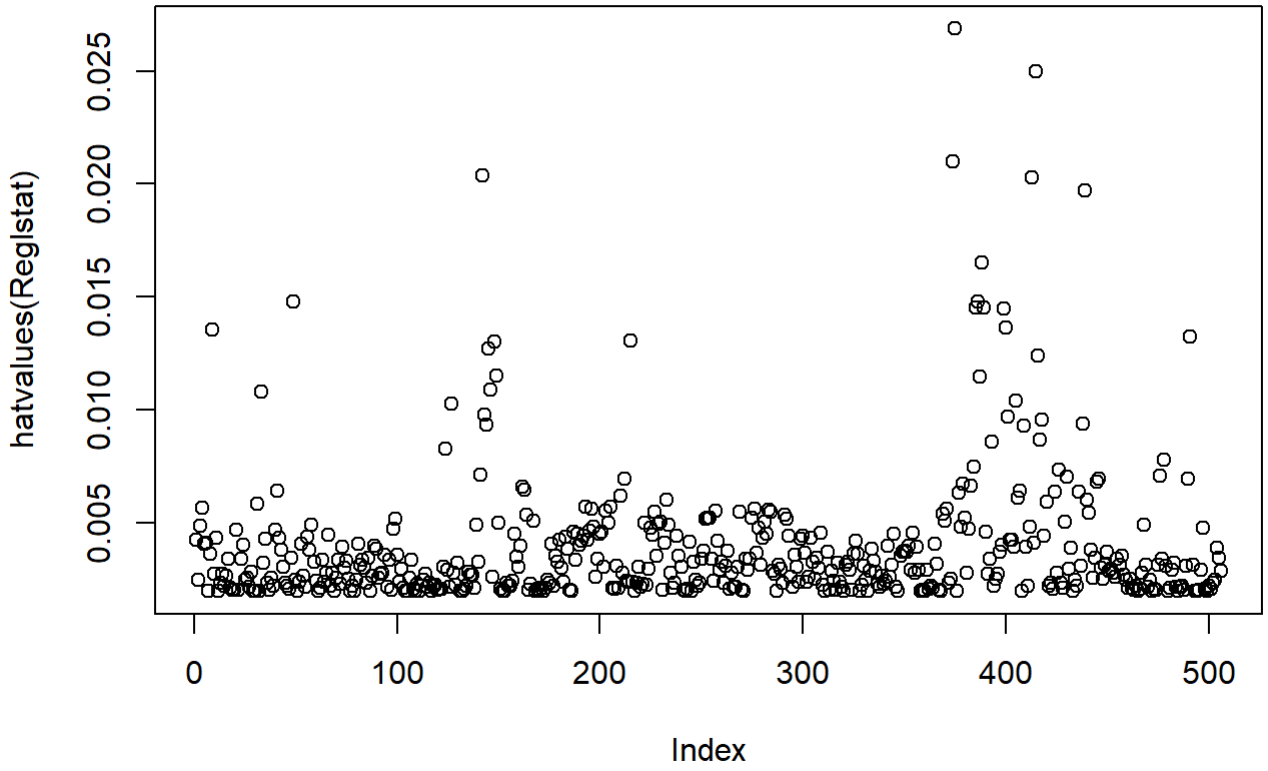
Comments: * observe the graph at the top left hand corner. * The plot is that of (ŷ_i,ϵˆ_i), where yˆ_i = (β_0)ˆ+ (β_1)ˆ* x_i and (ε_i)ˆ= y_i- (β_0)ˆ- (β_1)ˆ* x_i. * The ŷ_i s are the predicted values or fitted values as per the model. This graph is used to check on homoscedasticity (constant standard deviation). * If homoscedasticity holds, it should be expected that the residuals will be distributed evenly on either side of the x-axis. * it should also be expected that the red curve (LOWESS curve) to be more or less coincide with the x-axis. * A LOWESS curve is fitted with residuals being the response and fitted values being the predictor. * Homoscedasticity is doubtful. The graph indicates a quadratic model.

- Look at the top right hand graph. This graph is used for checking normality.
- If the points lies more or less on a straight line, normality is validated.

- Since the normality appears to be suspicious, some formal tests of normality will need to be conducted.
- Look at the bottom left hand graph. This could be used to detect outliers.
- The y-axis is the square root of the absolute value of the standardized residuals.
- Taking square root is a nuisance. If we are looking for the cut-off 3, then the observations whose y-value is greater than sqrt(3) = 1.732 needs to be observed
- There are some outliers whose indices are identified. We could look at the outliers directly.
- Look at the bottom right hand graph.
- Outliers and high leverage points can be identified.
- a separate graph with leverage values is present

Step 8: identify the observation with the highest leverage value.

```
plot(hatvalues(Reglstat))
```




```
which.max(hatvalues(Reglstat))
```

```
## 375
## 375
```

```
Boston[375, ]
```

```
##        crim zn indus chas   nox    rm age   dis rad tax ptratio black lstat
## 375 18.4982  0  18.1    0 0.668 4.138 100 1.137  24 666    20.2 396.9 37.97
##     medv
## 375 13.8
```

Continue with part 2 of Simple Linear Regression and Diagnostics…