

Net hourly wages across multiple countries Mc Donald's

[Code ▼](#)

Nishanth

01 August, 2018, 06:10

- 1 Importing data into R
 - 1.1 Viewing top rows in CSV
 - 1.2 viewing the structure of the dataset
 - 1.3 Basic Summary stats of dataset
 - 1.4 separating numerical and categorical variables from the dataset
- 2 Grapical representaion of the data
 - 2.1 scatter plot for net hourly wages and Big Mac price
 - 2.2 Univariate (Box plot) analysis for Outlier analysis
- 3 preprocessing
 - 3.1 extracting the outlier points, rows and removing them
 - 3.2 correlation between Blg Mac and Net Hourly Wage
- 4 Building the regression model
 - 4.1 applying linear regression to the Mac Donald's data
 - 4.2 viewing diagnostic plots of linear regression
 - 4.3 reengineering model
 - 4.3.1 Multivariate model approach for outliers (using cook's distance) and removal of outliers
 - 4.4 Again applying linear regression after removal of outliers
 - 4.5 viewing diagnostic plots of linear regression

1 Importing data into R

1.1 Viewing top rows in CSV

[Code](#)

```
##      Country Big.Mac.Price.... Net.Hourly.Wage....
## 1 Argentina          1.78              3.3
## 2 Australia          3.84             14.0
## 3  Brazil           4.91              4.3
## 4  Britain           3.48             13.9
## 5   Canada           4.00             12.8
```

1.2 viewing the structure of the dataset

[Code](#)

```
## 'data.frame':   27 obs. of  3 variables:
## $ Country      : Factor w/ 27 levels "Argentina","Australia",...: 1 2 3 4
## $ Big.Mac.Price.... : num  1.78 3.84 4.91 3.48 4 3.34 1.95 3.43 4.9 3.33 ...
## $ Net.Hourly.Wage....: num  3.3 14 4.3 13.9 12.8 3.1 3 5.1 17.7 3 ...
```

1.3 Basic Summary stats of dataset

[Code](#)

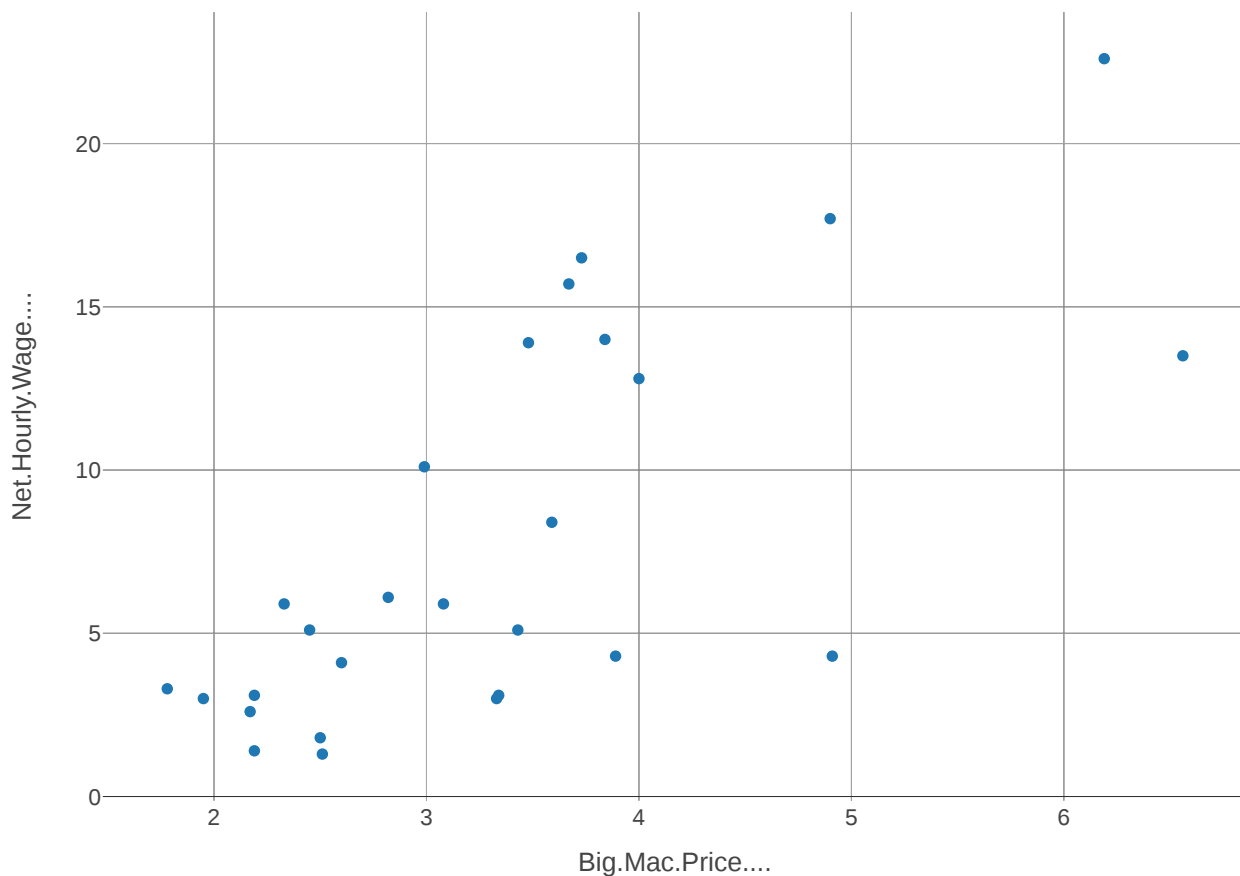
```
##      Country  Big.Mac.Price.... Net.Hourly.Wage....  
## Argentina: 1   Min.    :1.780      Min.    : 1.300  
## Australia: 1   1st Qu.:2.475      1st Qu.: 3.100  
## Brazil      : 1   Median :3.330      Median : 5.100  
## Britain    : 1   Mean    :3.349      Mean    : 7.726  
## Canada     : 1   3rd Qu.:3.785      3rd Qu.:13.150  
## Chile      : 1   Max.    :6.560      Max.    :22.600  
## (Other)    :21
```

1.4 separating numerical and categorical variables from the dataset

[Code](#)

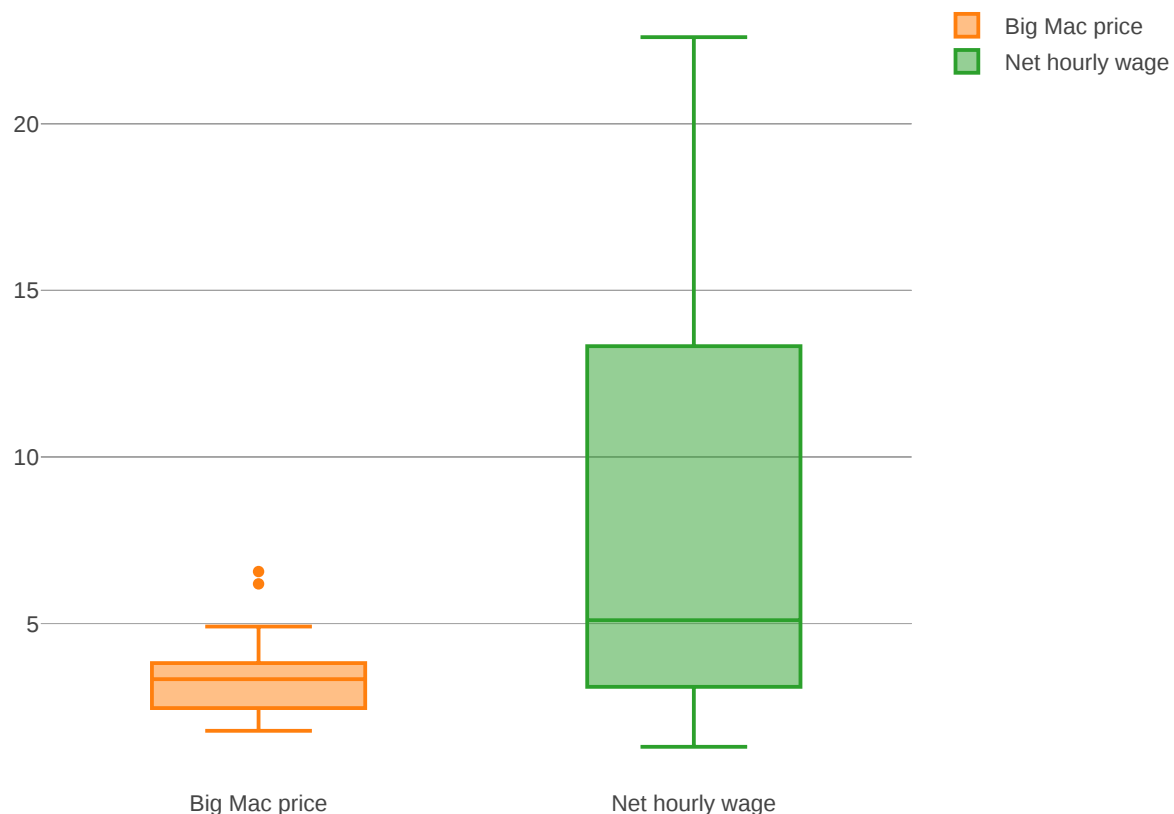
2 Grapical representaion of the data

2.1 scatter plot for net hourly wages and Big Mac price

[Code](#)

As price of the Big Mac increases net hourly wages also increase, there is a positive relationship between Big Mac prices and Net hourly wages

2.2 Univariate (Box plot) analysis for Outlier analysis

[Code](#)

we see couple of outliers in the data for Big Mac

3 preprocessing

3.1 extracting the outlier points, rows and removing them

[Code](#)

```
## $Big.Mac.Price....  
## [1] 6.56 6.19  
##  
## $Net.Hourly.Wage....  
## numeric(0)
```

[Code](#)

```
## [1] 22 23
```

[Code](#)

There are couple of outliers in Big Mac price and no outliers in net hourly wages

3.2 correlation between Blg Mac and Net Hourly Wage

[Code](#)

```
## [1] 0.717055
```

correlation between Big Mac and Net Hourly wage is strong and positively correlated

4 Building the regression model

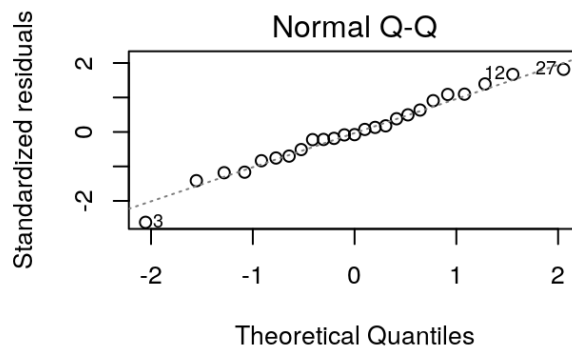
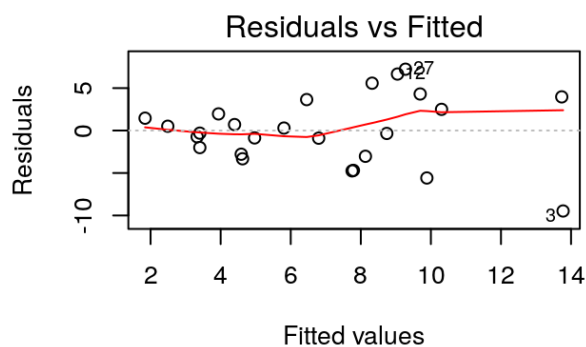
4.1 applying linear regression to the Mac Donald's data

[Code](#)

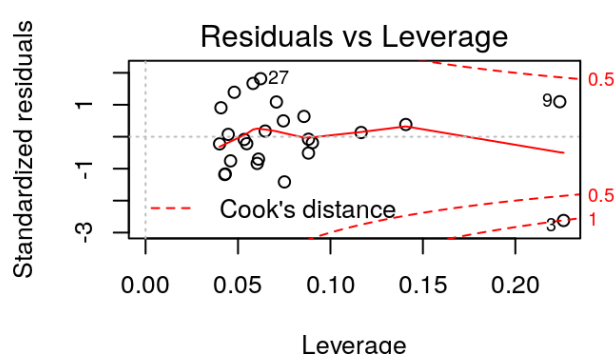
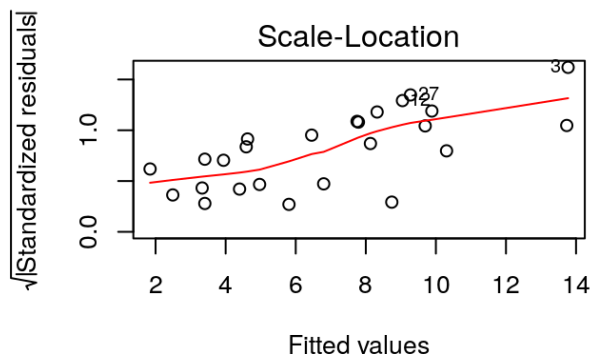
```
##
## Call:
## lm(formula = Net.Hourly.Wage.... ~ Big.Mac.Price...., data = data_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.4727 -2.7873 -0.3057  2.4957  7.2248
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -4.9411     3.1612  -1.563  0.131697
## Big.Mac.Price....  3.8114     0.9826   3.879  0.000759 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.107 on 23 degrees of freedom
## Multiple R-squared:  0.3955, Adjusted R-squared:  0.3692
## F-statistic: 15.05 on 1 and 23 DF, p-value: 0.0007594
```

4.2 viewing diagnostic plots of linear regression

[Code](#)



we can



observe there is little bit of upword treand in residuals of linear regression model(heteroscedasticity) in 3rd diagnostic graph

4.3 reengineering model

4.3.1 Multivariate model approach for outliers (using cook's distance) and removal of outliers

Cook's distance is a measure computed with respect to a given regression model and therefore is impacted only by the X variables included in the model. But, what does cook's distance mean? It computes the influence exerted by each data point (row) on the predicted outcome.

The cook's distance for each observation i measures the change in \hat{Y} (fitted Y) for all observations with and without the presence of observation i, so we know how much the observation i impacted the fitted values. Mathematically, cook's distance D_i for observation i is computed as

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \times MSE}$$

Cook's distance formula

where,

\hat{Y}_j is the value of jth fitted response when all the observations are included.

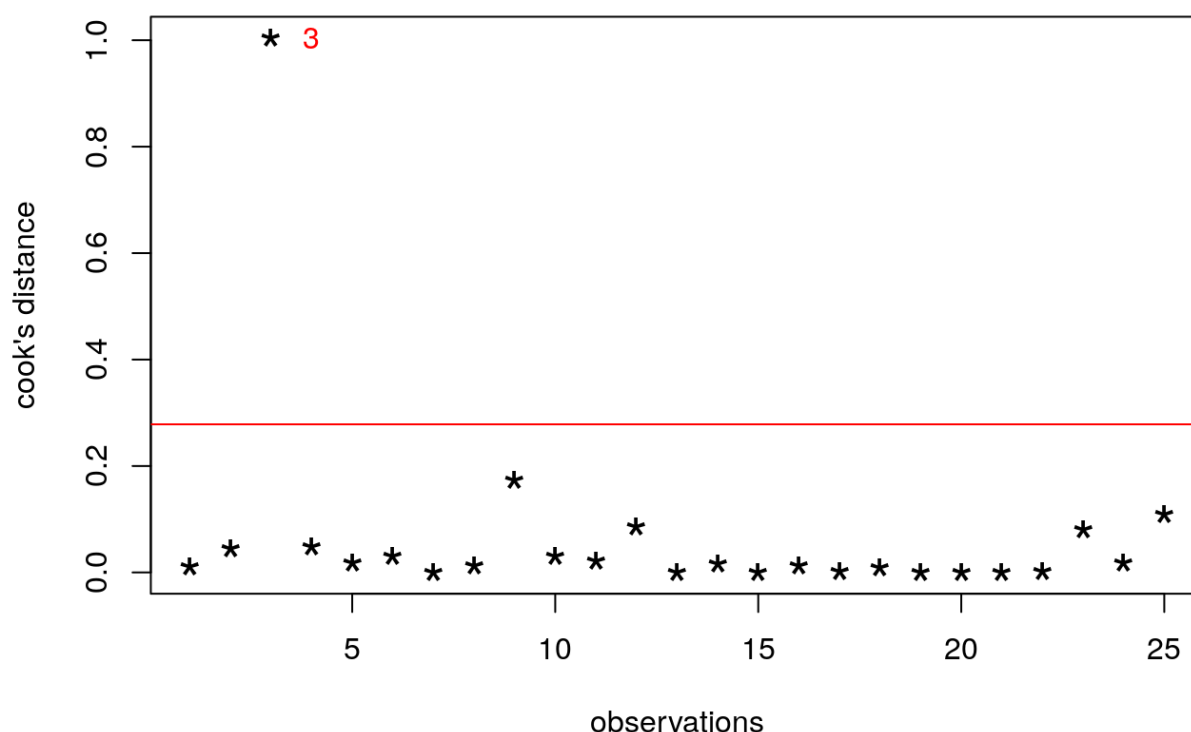
$\hat{Y}_{j(i)}$ is the value of jth fitted response, where the fit does not include observation i.

MSE is the mean squared error.

p is the number of coefficients in the regression model



Influential Obs. by Cooks distance



Code

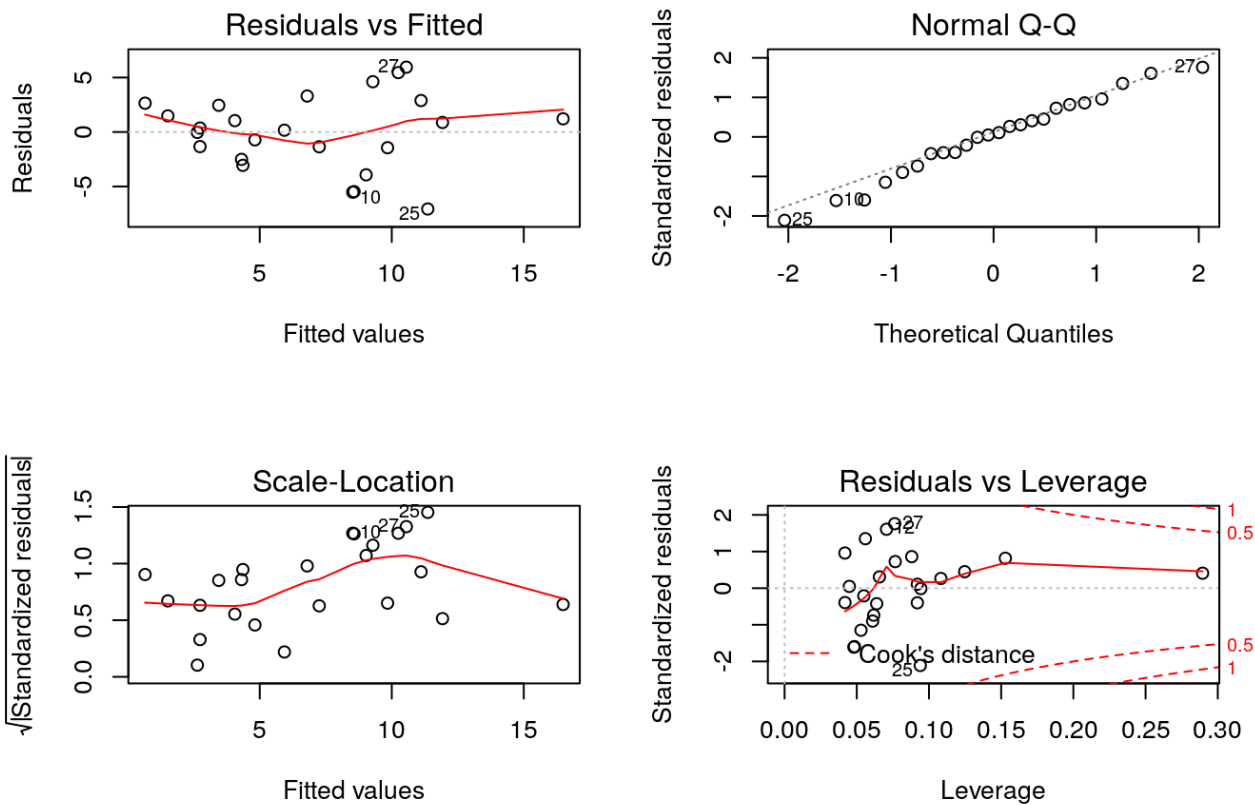
4.4 Again applying linear regression after removal of outliers

Code

```
##
## Call:
## lm(formula = Net.Hourly.Wage.... ~ Big.Mac.Price...., data = data_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.0640 -1.7088  0.2643  2.5001  5.9479
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -8.3760     2.9298  -2.859  0.00912 **
## Big.Mac.Price....  5.0745     0.9369   5.416 1.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.516 on 22 degrees of freedom
## Multiple R-squared:  0.5714, Adjusted R-squared:  0.552
## F-statistic: 29.33 on 1 and 22 DF, p-value: 1.936e-05
```

After removal of outliers we observe p value for big mac price is more significant, adjusted R-squared increased from 36.9 to 55.2 and Multiple R squared increased from 39.5 to 57.1, overall model significance F test p - value becomes more significant.

4.5 viewing diagnostic plots of linear regression

[Code](#)

Now we observe there is no upward trend in residuals (heteroscedasticity) in 3rd diagnostic graph.