Regression Models PA

Gerard NIGNON

16 août 2015

Synopsis

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (mpg) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Visual analysis

From the visual analysis of the boxplot in **Figure 1** we can assert that automatic cars have lower miles per gallon, and therefore a lower fuel efficiency, than manual cars do. However, it is possible that this apparent pattern happened by random chance– that is, that we just happened to pick a group of automatic cars with low efficiency and a group of manual cars with higher efficiency. We may have to push the analysis further than just visual.

Correlation analysis

We create nice pairwise scatter plots **Figure 2**. This is a good way to investigate the relationship between all the variables in this data set. For example, `mpg` has a strong and negative correlation (`-0.852`) with `cyl`.

Individual regression

We ran a simple linear regression model with mpg as outcome and the transmission (Automatic and Manual) as categorical precitor variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
factor(am)Manual	7.244939	1.764422	4.106127	0.000285

All the estimates provided here are in comparison with automatic transmission. The intercept of `17.14` is simply the mean MPG of automatic transmission. The slope of `7.24` is the change in the mean between manual transmission and automatic transmission. The p-value of `0.000285` for the mean MPG difference between manual and automatic transmission is significant. Therefore, we conclude from this model that manual transmission is more fuel efficient.

Group regression

To make our model more realistic, we add more independent variables without overfitting. The ideal is for all of the predictor variables to be correlated with the outcome variable, but not with others, to minimise the risk of multicollinearity. Therefore, in this part of the analysis, we use multivariable linear regression to develop a model that includes effects of other variables.

Before fitting the model, we want to perform a statistical test to determine which predictors are significant. To determine the ideal formula for prediction, we select the best regression model.

The All Subsets Regression is performed using the **regsubsets()** and **glmulti()** to select the variables with the highest adjusted R^2 as criterion,. The outcome is shown in **Figure 3.a**

The model with 3 variables has the highest adjusted R^2 . Variables marked with TRUE (**Figure 3.b.1 and Figure 3.b.2**) are the ones chosen: `mpg`(Intercept), `wt`, `qsec`, and `am`(factor).

Now let's fit the multiple regression model with the **lm()** function (**Figure 3.c**)

$\hat{mpg} = 9.6178 - 3.9165\text{wp} + 2.9358\text{factor (am) Manual} + 1.2259\text{qsec}$

The regression coefficients indicate the increase in the dependent variable for a unit change in a predictor variable, holding all other predictor variables constant. On average manual transmission cars have 2.94 miles per gallon more than automatic transmission cars. Our model plains 84.97% of the variance in miles per gallon.

Like most other statistical tests, regression analysis require that a set of assumptions about the data are met.

Statistical assumptions

Normality Figure 5 all the points fall close to the line and are within the confidence envelope, suggesting that we've met the normality assumption fairly well

Linearity Figure 6 The component plus residual plots confirm that you've met the linearity assumption. The form of the linear model seems to be appropriate for this dataset

Homoscedasticity

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.55815    Df = 1     p = 0.2119363
```

The score test is nonsignificant (p = 0.21), suggesting that we've met the constant variance assumption

Multicollinearity Multicollinearity can be detected using a statistic called the variance inflation factor (VIF).

```
sqrt(vif(fit)) > 2 # problem?
```

```
##         wt factor(am)       qsec
##      FALSE      FALSE      FALSE
```

Unusual observations

Now we screen for unusual observation: outliers, high-leverage observations, and influential observations. **Figure 7**, identifies `Chrysler Imperial` as an unusual observation. Deleting this car will have a notable impact on the values of the intercept and slopes in the regression model.

Conclusions

The model seems robust and met all the underlying statistical assumptions . It will be more accure if we remove the `Chrysler Imperial`.

Appendix

All the figures

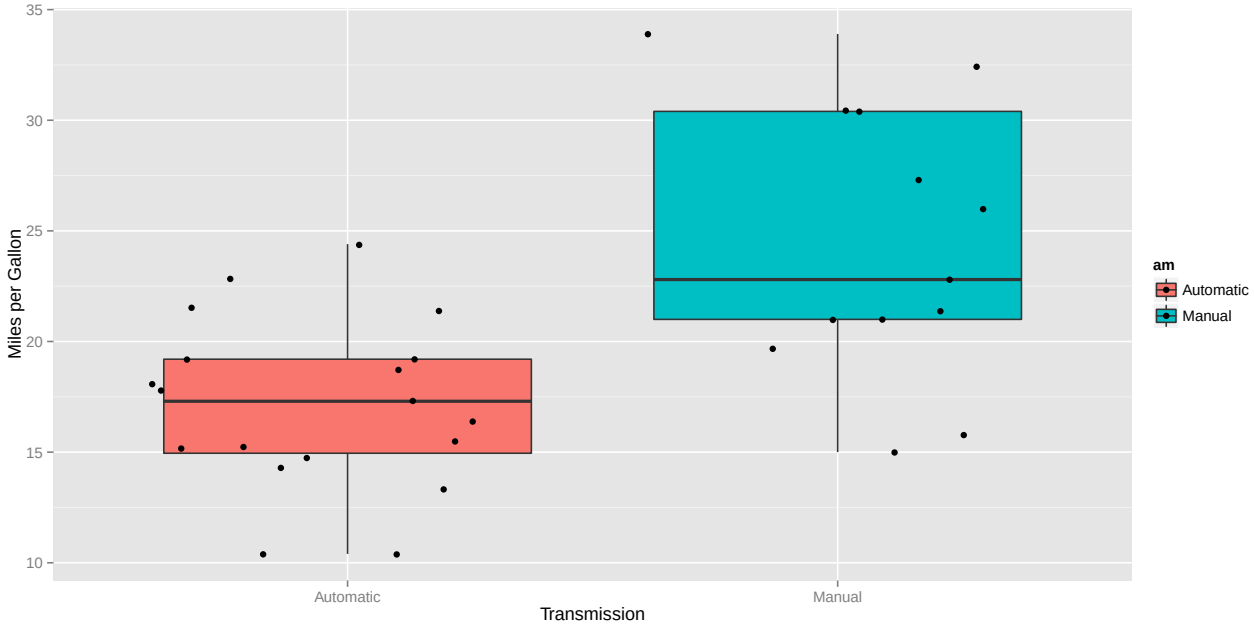


Figure 1: Boxplot of Mile per Gallon and Transmission

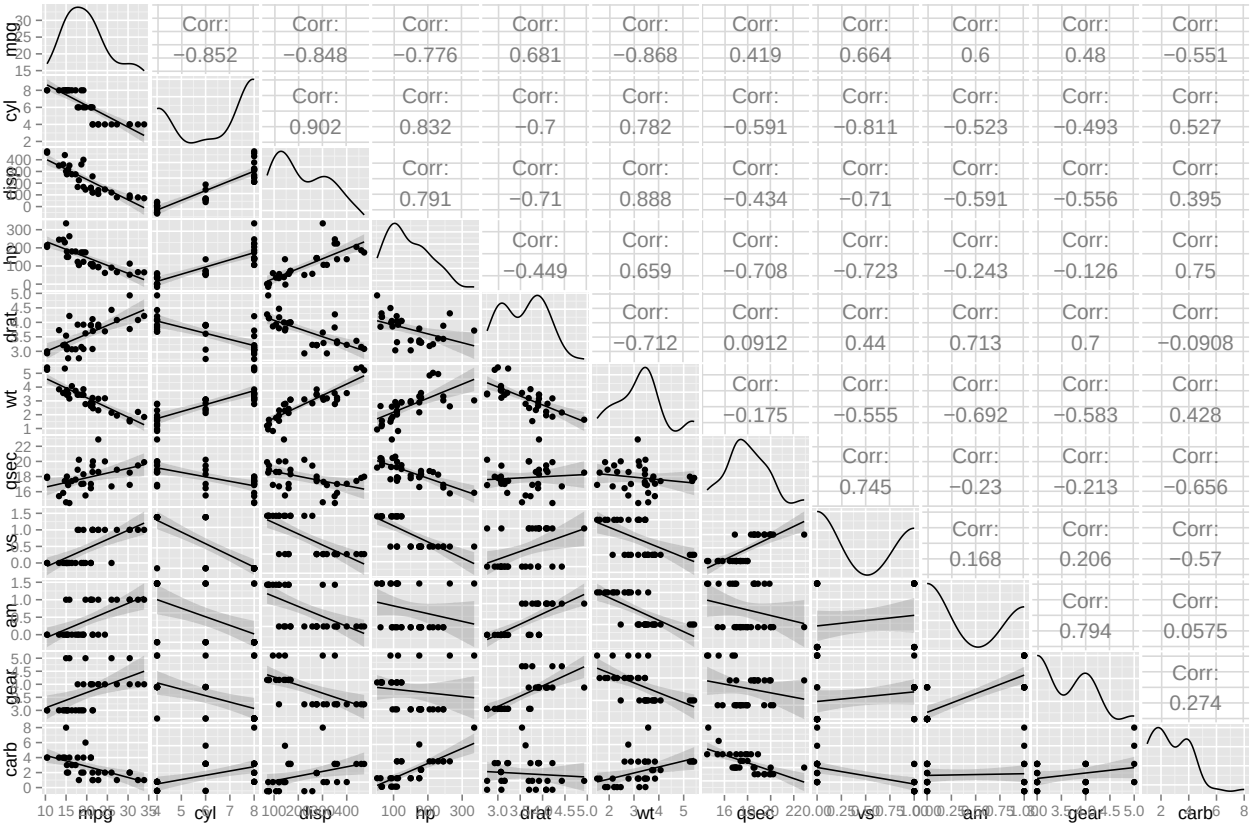


Figure 2: Scatter plot matrix of dependent and independent variables

Variables

```
##           cyl disp hp drat wt qsec vs am1 gear carb
## 1  ( 1 )                   *
## 2  ( 1 )    *              *
## 3  ( 1 )                   *    *      *
## 4  ( 1 )           *       *    *      *
## 5  ( 1 )         *  *      *    *      *
## 6  ( 1 )         *  *    *  *    *      *
## 7  ( 1 )         *  *    *  *    *      *    *
## 8  ( 1 )         *  *    *  *    *      *    *    *
## 9  ( 1 )         *  *    *  *    *  *   *    *    *
## 10  ( 1 )   *    *  *    *  *    *  *   *    *    *
```

Figure 3.a: Variables choice

```
summary.out$which[3,]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
##        TRUE       FALSE       FALSE       FALSE       FALSE        TRUE
##        qsec          vs    amManual        gear        carb
##        TRUE       FALSE        TRUE       FALSE       FALSE
```

Figure 3.b.1: Variables choice

```
## [[1]]
## mpg ~ 1 + factor(am) + wt + qsec
## <environment: 0x7ffc9ee6f780>
##
## [[2]]
## mpg ~ 1 + factor(am) + hp + wt + qsec
## <environment: 0x7ffc9ee6f780>
##
## [[3]]
## mpg ~ 1 + factor(am) + wt + qsec + carb
## <environment: 0x7ffc9ee6f780>
```

Figure 3.b.2: The 3 best model

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
factor(am)1	2.935837	1.4109045	2.080819	0.0467155
qsec	1.225886	0.2886696	4.246676	0.0002162

R^2

```
## [1] 0.8496636
```

Figure 3.c: Multiple linear regression

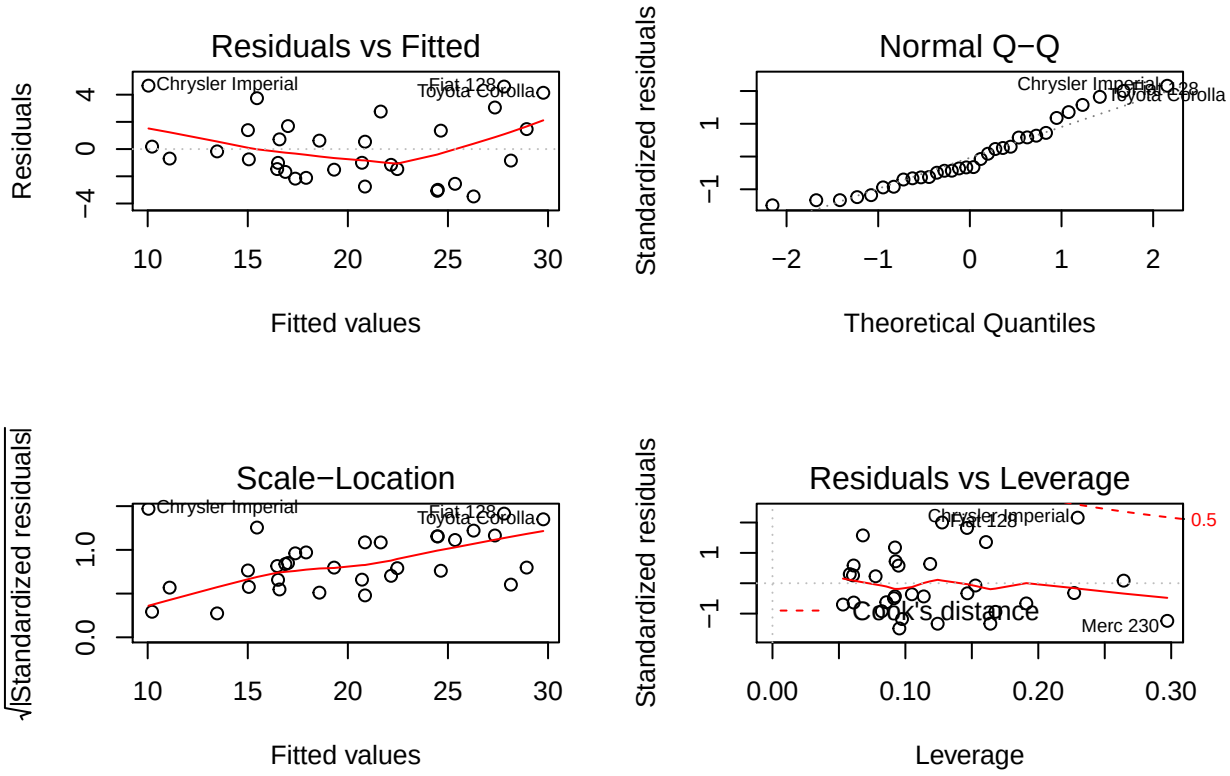


Figure 4:Diagnostic plots for the regression

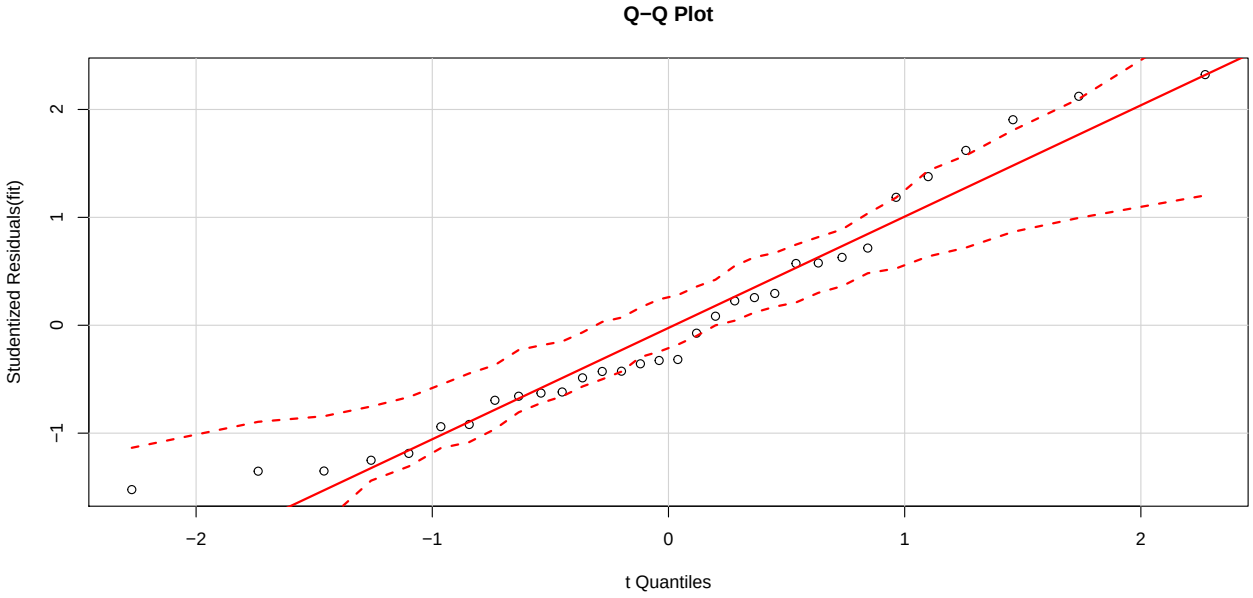


Figure 5: Q-Q plot for studentized residuals

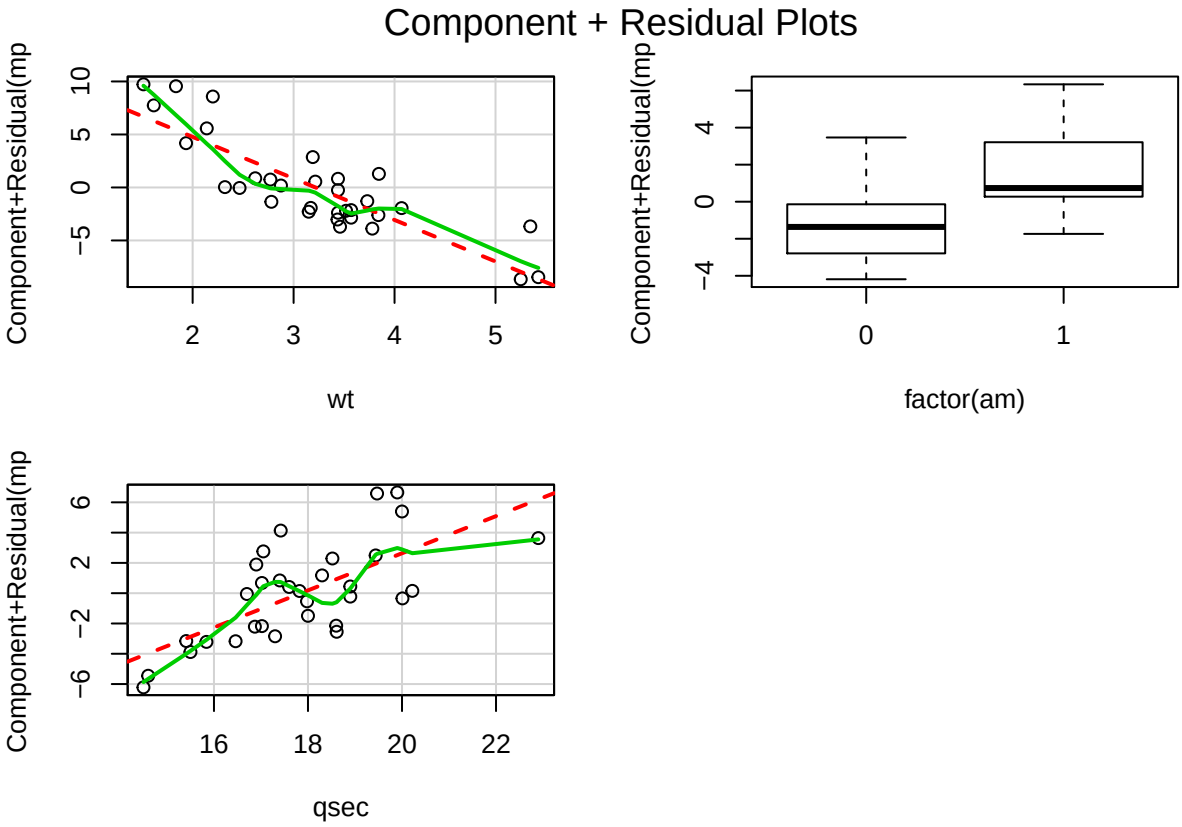
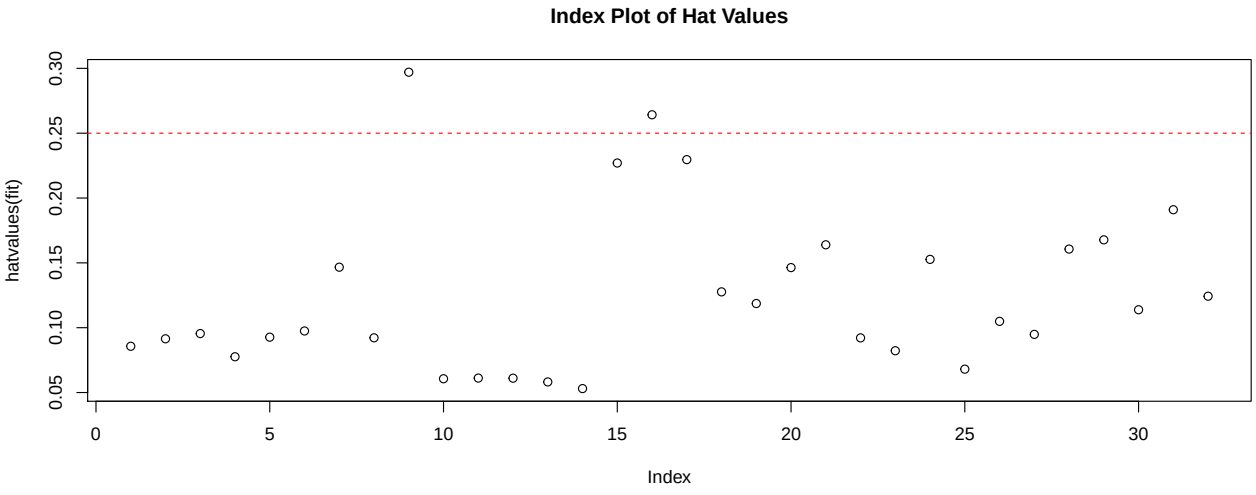


Figure 6: Component plus residual plots for the regression

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##                   rstudent unadjusted p-value Bonferonni p
## Chrysler Imperial 2.323119           0.027949      0.89437
```

Figure 7.a : Identifying unusual observations - Outliers




```
## integer(0)
```

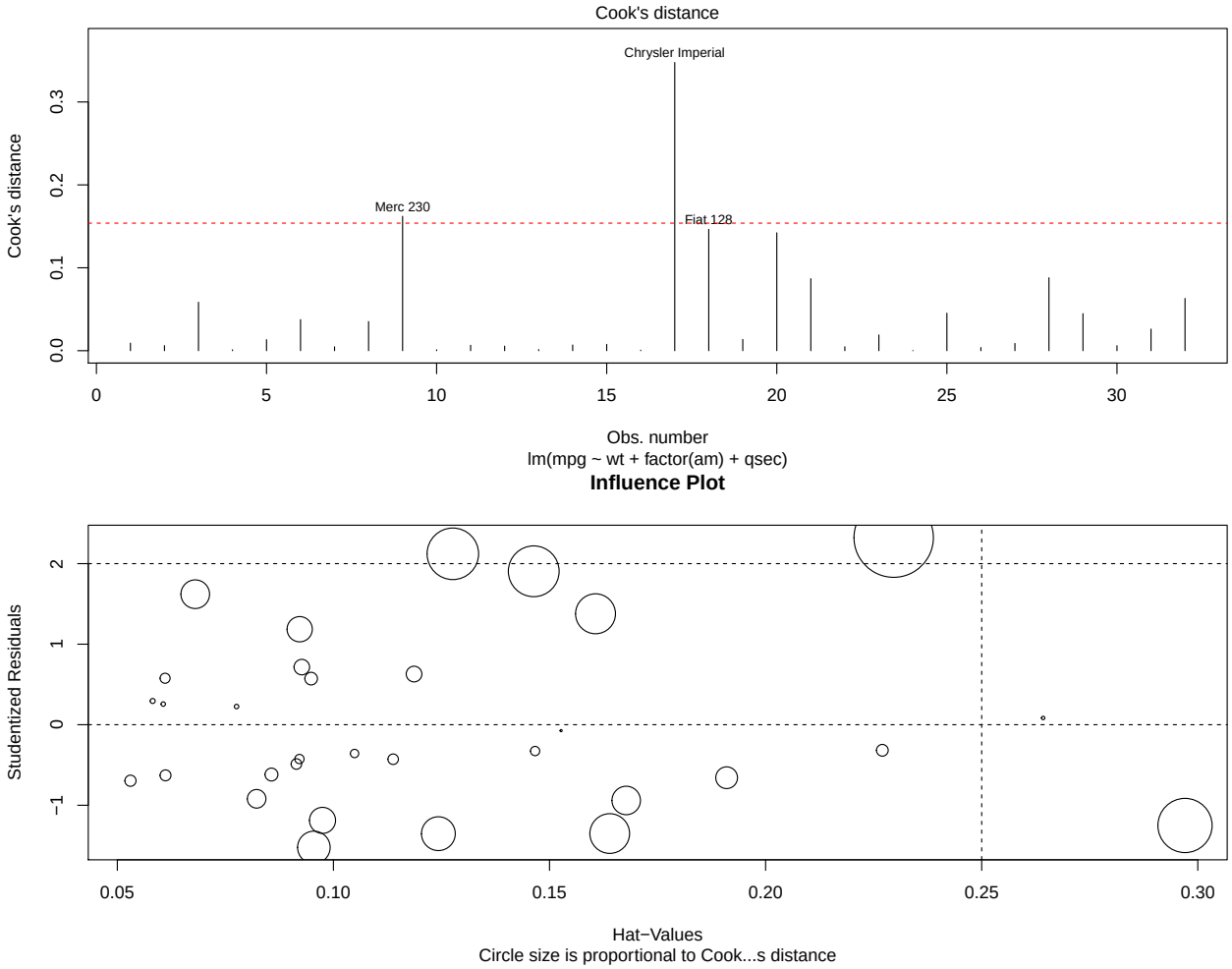


Figure 7.b : Identifying unusual observations - High leverage points and Influential observations