Regression Models: Peer Assessments

Created by H.Wang on December 23, 2015

Executive Summary

The purpose of this report is to analyze the relationship between MPG vs a set of variables in `mtcars` data set. The data was from the 1974 *Motor Trend* US magazine that comprises fuel consumption and 10 aspects of automobile design and performance for 32 cars (1973-1974 models). The regression models below is mainly used to explore how **transimission** (**automatic** (am = 0) and **manual** (am = 1)) features affect the **MPG**. Based on boxplot I firstly use t-test to evaluate the performance difference between cars with different transmission system. Then, I fit several linear regression models using different variable combinations and choose the one with highest Adjusted R-squared, and it yields to the fact that when cars are lighter in weight, manual transmitted cars usually have higher MPG than automatic transmitted cars on average; while as weight goes up, this figure tends to be higher in cars with automatic transmission than manual transmission.

Basic Settings

```
echo= TRUE  # make scripts visible to others
```

Load Data and Perform Basic Exploratory Data Analysis

```
# load the dataset
data(mtcars)
head(mtcars) #shows excerpt of dataset
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
# factorize variables
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am) #0 = automatic, 1 = manual
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
attach(mtcars)
```

Assumption

From boxplot in **Appendix: Figures**, I firstly make null hypothesis as the MPG of the automatic and manual transmission are from the same population (assuming MPG has a normal distribution). We use two sample T-test to test it.

```r
result <- t.test(mtcars$mpg ~ mtcars$am)
result$p.value
```

```
## [1] 0.001373638
```

```r
result$estimate #means
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Since p-value is 0.0013736 we reject our null hypothesis, which means the automatic and manual transmissions are from different populations. The mean of manual transmitted cars is 7.2449393 more than that of automatic transmitted cars.

Regression Analysis

I start to fit the full model use "mpg ~ ."

```r
# Full Model:
fullModel <- lm(mpg ~ ., data = mtcars)
summary(fullModel) #summary is hidden
```

This model has residual standard error as 2.833 on 15 degress of freedom, adjusted R-squared is 0.779 which reflects 78% of the variance of the MPG variables can be explained. However, none of coefficients are significant at 0.05 significant level.

```r
# StepModel: step back based on some significant variables:
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel) #summary is hidden
```

This model is "mpg ~ wt + qsec + am". It has residual standard error as 2.459 on 28 degrees of freedom, adjusted R-squared is 0.8336 slightly higher than that of full model, and all coefficients are significant at 0.05 significant level. A pair graph of these four listed variables is in **Appendix: Figures**.

Based on scatter plot from **Appendix: Figures** it tends to be a interaction between 'wt' and 'am' variable, since from pair graph automatic cars weights more than manual transmission cars, we add this interaction term to above model:

```r
# StepModel_2: add wt:am to model
stepModel_2 <- lm(mpg ~ wt + qsec + am + wt:am, data = mtcars)
summary(stepModel_2) #summary is hidden
```

This model has residual standard error 2.084 on 27 degrees of freedom, adjusted R-Squared 0.8804, and all coefficients are at 0.05 significant level.

Next analysis will be on simple model with MPG as the outcome variable and "am" as the predictor variable.

```r
#Simple Linear Regression Model
simpleModel <- lm(mpg ~ am, data = mtcars)
summary(simpleModel) #summary is hidden
```

It shows that automatic cars has 17.147 mpg on average but will increased by 7.245 if they are manual transmission ones. The model has residual standard error 4.902 on 30 degrees of freedom and adjusted R-square 0.3385 which means this model can explain about 34% of the variance of the MPG variable. This low value also indicates that we need to add other variables to the model.

```
#Compute analysis of variance tables for selected models
anova(simpleModel,stepModel,stepModel_2,fullModel)
confint(stepModel_2)
```

```
summary(stepModel_2)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

We end up choosing the model with highest adjusted R-Squared value "mpg ~ wt + qsec + am + wt:am". The results shows that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant, cars with manual transmission add 14.079 + (-4.141)*wt more MPG on average compares to cars with automatic transmission. That is, a manual transmitted car that weighs 2000lbs have 5.797 more MPG than an automatic transmitted car with same weight and 1/4 mile time.

Residual Analysis and Diagnostics

Please refer to **Appendix: Figures** for residual analysis, according to the plot we can verify the underlying assumptions:

1. The Residuals vs. Fitted Plot shows no consistent pattern, supporting the accuracy of the independent assumption.
2. The Normal Q-Q plot indicates that residuals are normally distributed, most of points are distributed linearlly.
3. Scale-Location plot confirms the constant variance assumption as points are randomly distributed.
4. Residuals vs. Leverage points out no outliers are present all values are within 0.5 bands.

As for Dfbetas, the measure of how much an observation has affected the estimate of a regression coefficient, the result is:

```
sum((abs(dfbetas(stepModel_2)))>1)
```

```
## [1] 0
```

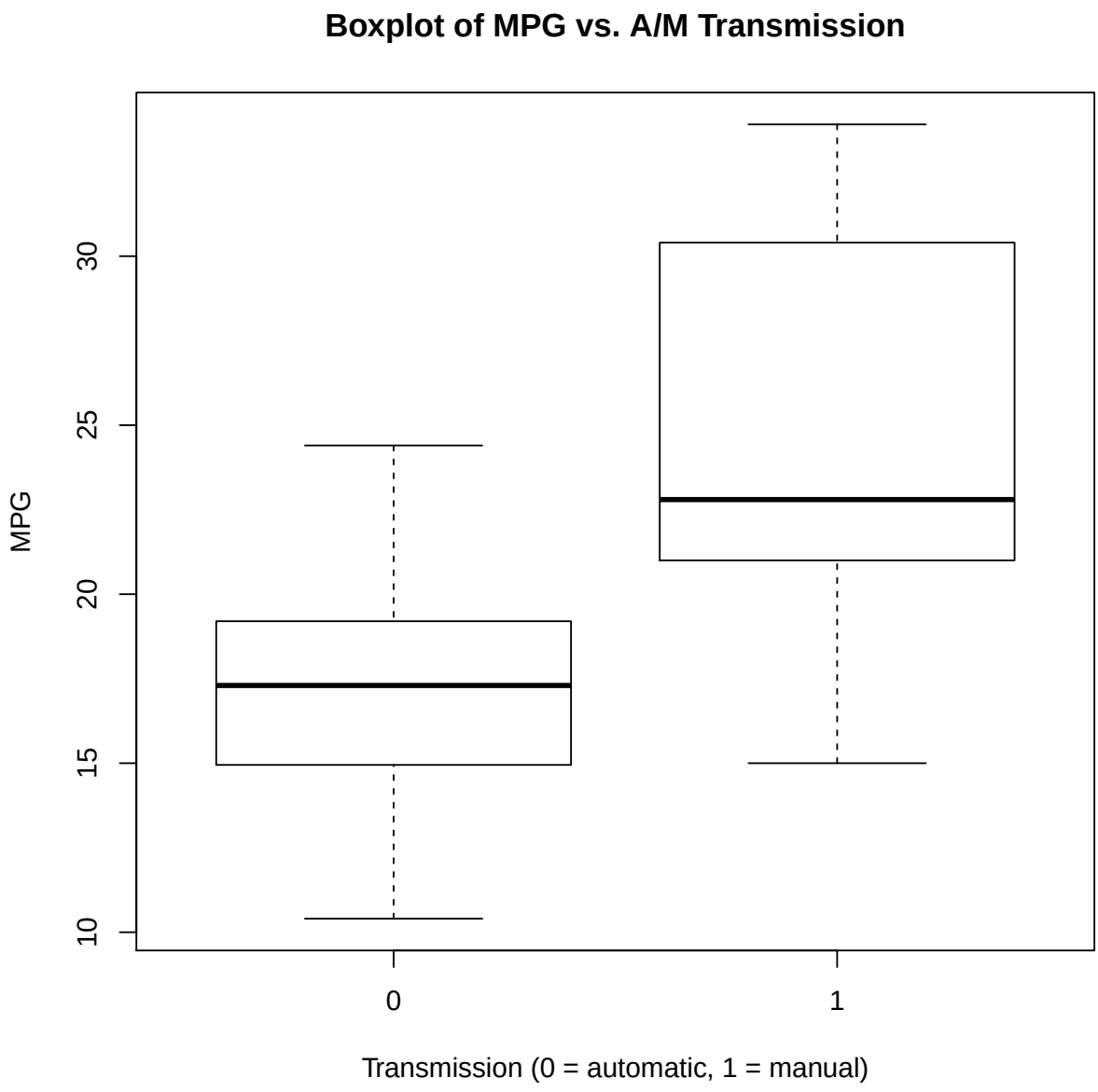
Thus, the above analyses meet all basic assumptions of linear regression.

```
## \newpage
```

Appendix: Figures

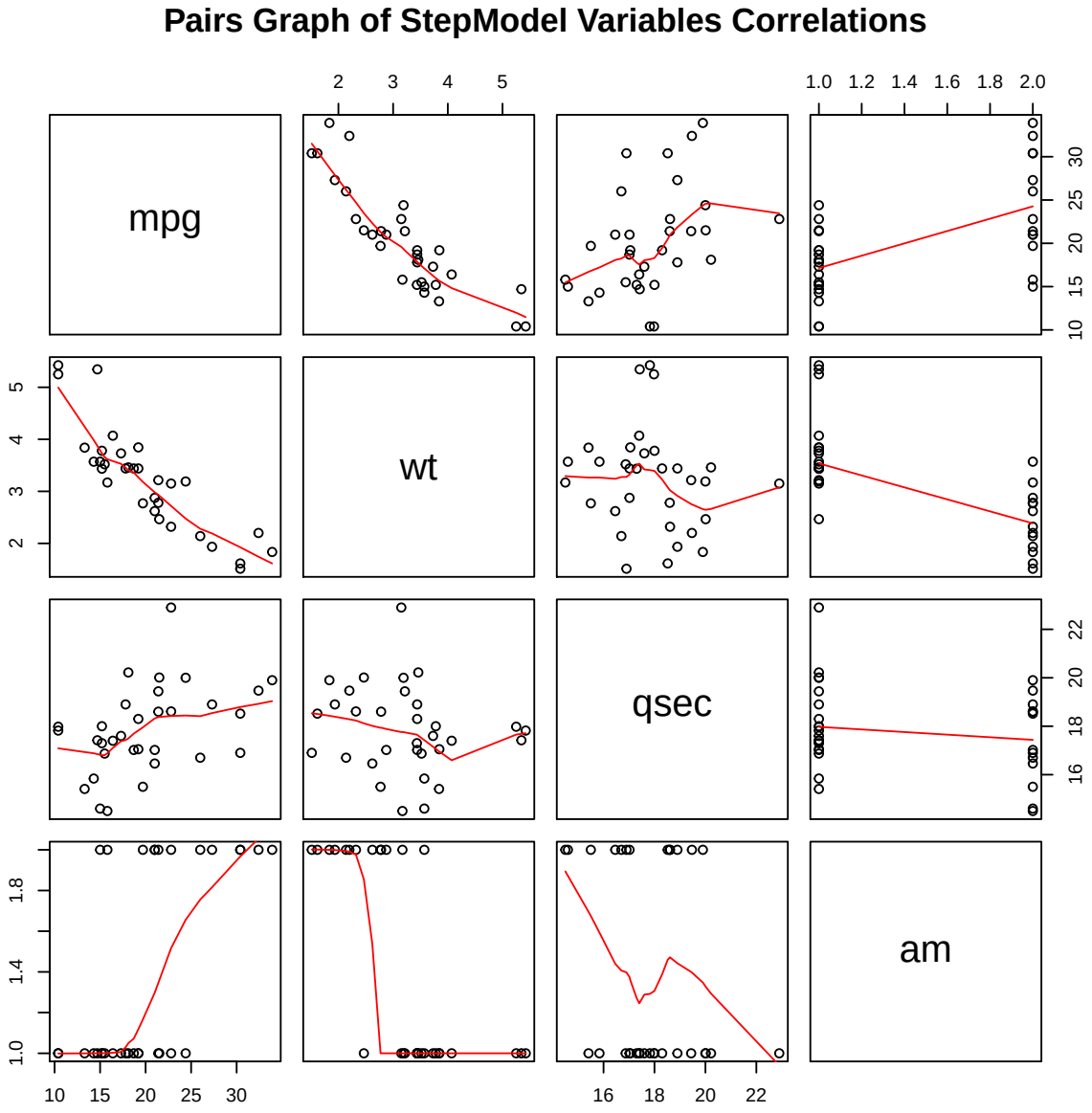
1. Boxplot of MPG vs. A/M Transmission

```
boxplot(mtcars$mpg ~ mtcars$am,xlab = "Transmission (0 = automatic, 1 = manual)",ylab = "MPG",
        main = "Boxplot of MPG vs. A/M Transmission")
```



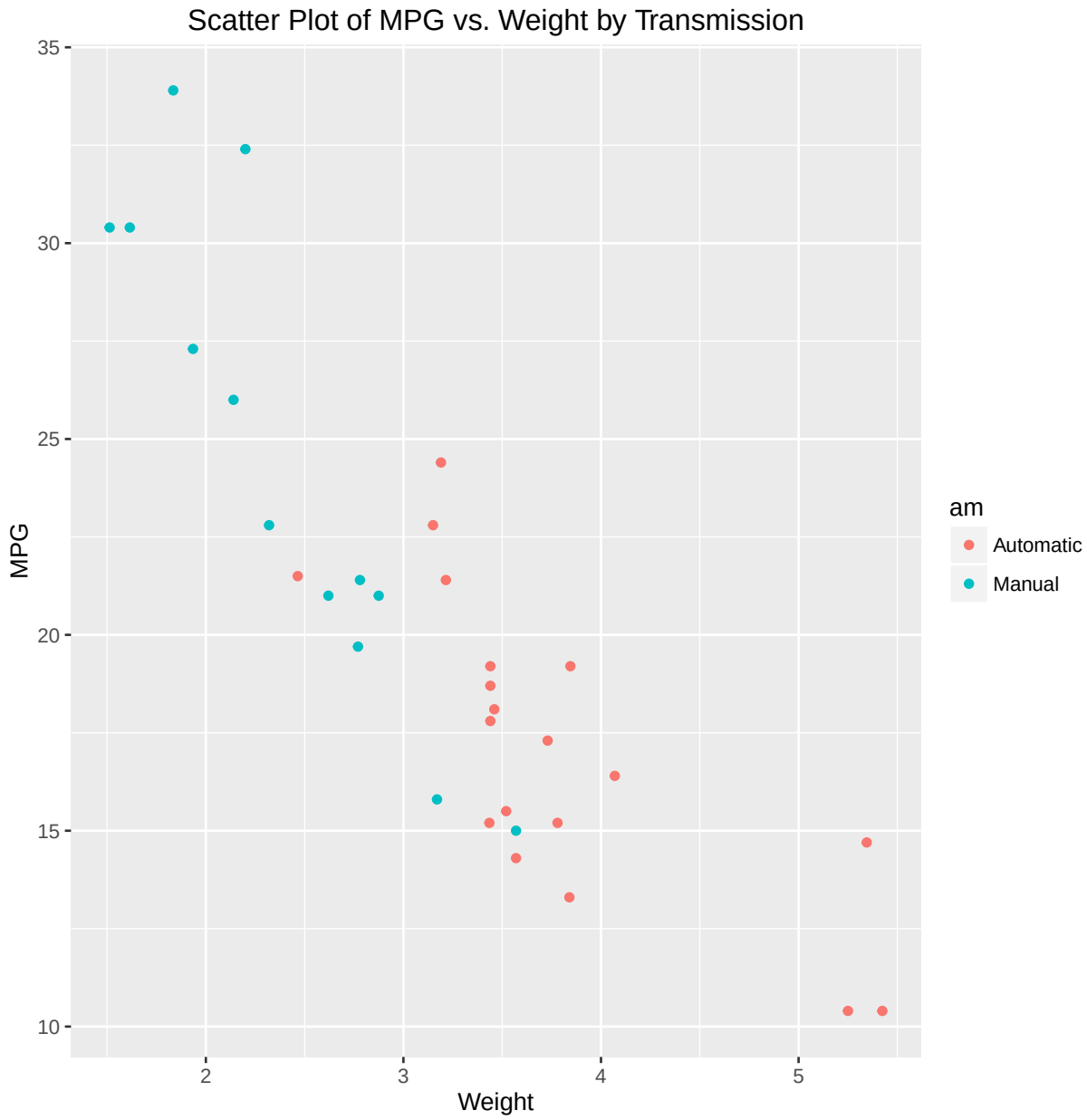
2. Pairs Graph of StepModel Variables Correlations

```
pairs(mtcars[,c(1,6,7,9)], panel=panel.smooth, main="Pairs Graph of StepModel Variables Correlations")
```



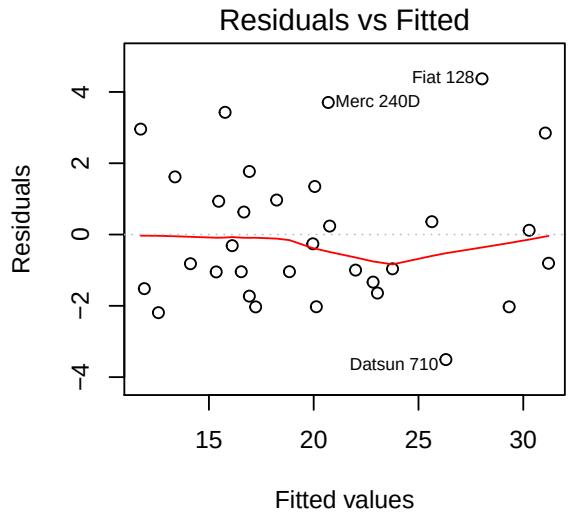
3. Scatter Plot of MPG vs. Weight by Transmission

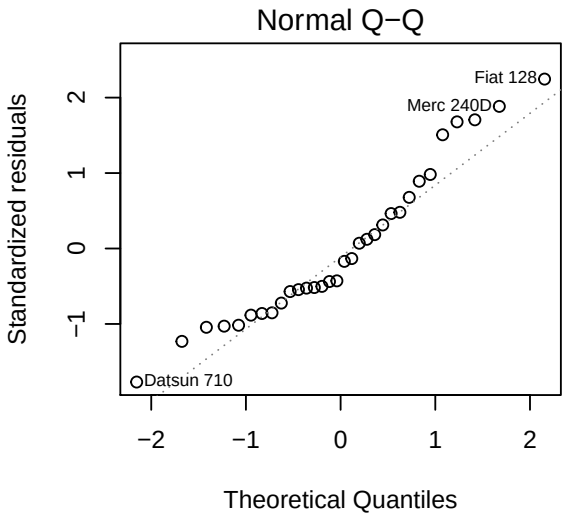
```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("Weight") + ylab("MPG") + ggtitle("Scatter Plot of MPG vs. Weight by Transmission")
```

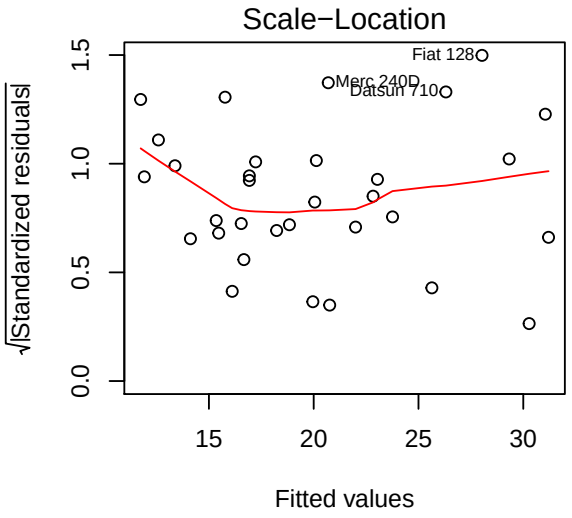


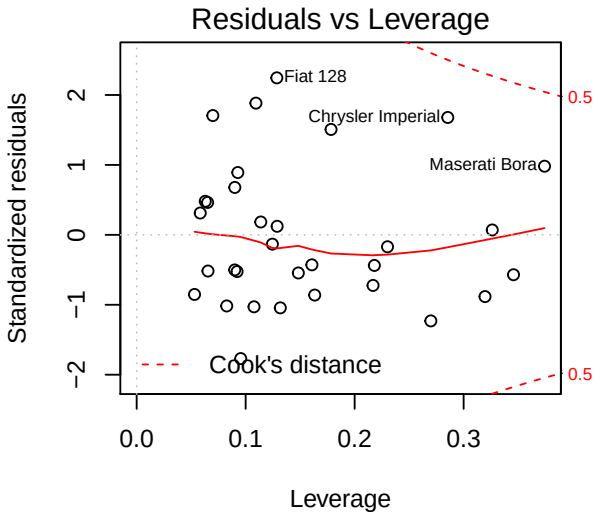
4.Residual Plots

```
par(mfrow=c(2,2))
plot(stepModel_2)
```


Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Merc 240D
Datsun 710


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710


Scale−Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Datsun 710


Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5