1 Regression Diagnostics with R

An excellent review of regression diagnostics is provided in John Fox's aptly named *Overview of Regression Diagnostics*. Dr. Fox's car package provides advanced utilities for regression modeling.

(1) Fox, John. (1991). Regression Diagnostics: An Introduction. Sage Publications.

```
# Assume that we are fitting a multiple linear regression
# on the MTCARS data
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
```

1.1 Outliers

Assessment of Outliers can be carried out using the `outlierTest` function.

```
outlierTest(fit) # Bonferonni p-value for most extreme obs
qqPlot(fit, main="QQ Plot") #qq plot for studentized resid
leveragePlots(fit) # leverage plots
```

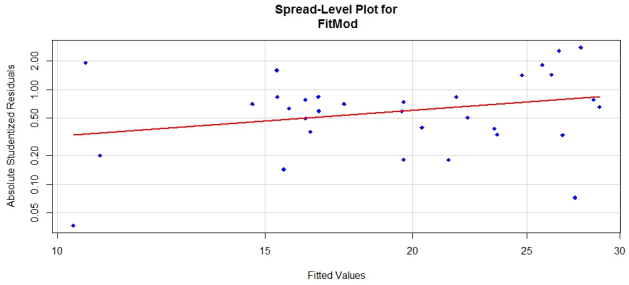
1.2 Added Variable Plots

```
# added variable plots
av.Plots(fit)
```

1.3 Non-constant Error Variance

```
# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(FitMod)
# plot studentized residuals vs. fitted values
spreadLevelPlot(FitMod)
```

```
> ncvTest(FitMod)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.330027    Df = 1     p = 0.06802577
```




```
Suggested power transformation:  0.08866484
```

1.4 Influential Observations

```
# Influential Observations

# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(fit, id.method="identify", main="Influence Plot", sub="Circle size is propor
```

2 Diagnostic Plots for Linear Models with `R`

Plot Diagnostics for an `lm` Object

2.1 Description

Six plots (selectable by `which`) are currently available:

1. a plot of residuals against fitted values,
2. a Scale-Location plot of *sqrt(— residuals —)* against fitted values,
3. a Normal Q-Q plot,
4. a plot of Cook's distances versus row labels,
5. a plot of residuals against leverages,
6. a plot of Cook's distances against leverage/(1-leverage).

By default, the first three and 5 are provided.

I explained the assumption of homoscedasticity and the plots that can help you assess it (including scale-location plots [2]) on CV here: What does having constant variance in a linear regression model mean? I have discussed qq-plots [3] on CV here: QQ plot does not match histogram. So, what's left is primarily just understanding [5], the residual-leverage plot.

To understand this, we need to understand three things:

- leverage,
- standardized residuals, and
- Cook's distance.

2.1.1 Leverage

To understand leverage, recognize that *Ordinary Least Squares* regression fits a line that will pass through the centre of your data, $(\bar{x}, \bar{y})$. The line can be shallowly or steeply sloped, but it will pivot around that point like a lever on a fulcrum. We can take this analogy fairly literally: because OLS seeks to minimize the vertical distances between the data and the line, the data points that are further out towards the extremes of X will push / pull harder on the lever (i.e., the regression line); they have more leverage. One result of this could be that the results you get are driven by a few data points; that's what this plot is intended to help you determine.

Another result of the fact that points further out on X have more leverage is that they tend to be closer to the regression line (or more accurately: the regression line is fit so as to be closer to them) than points that are near $\bar{x}$. In other words, the residual standard deviation can differ at different points on X (even if the error standard deviation is constant). To correct for this, residuals are often standardized so that they have constant variance (assuming the underlying data generating process is homoscedastic, of course).

One way to think about whether or not the results you have were driven by a given data point is to calculate how far the predicted values for your data would move if your model were fit without the data point in question. This calculated total distance is called **Cook's distance**. Fortunately, you don't have to rerun your regression model N times to find out how far the predicted values will move, Cook's D is a function of the leverage and standardized residual associated with each data point.

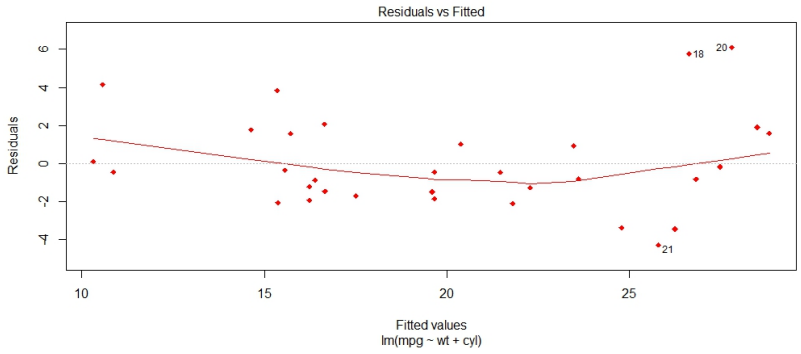
With these facts in mind, consider the plots associated with four different situations:

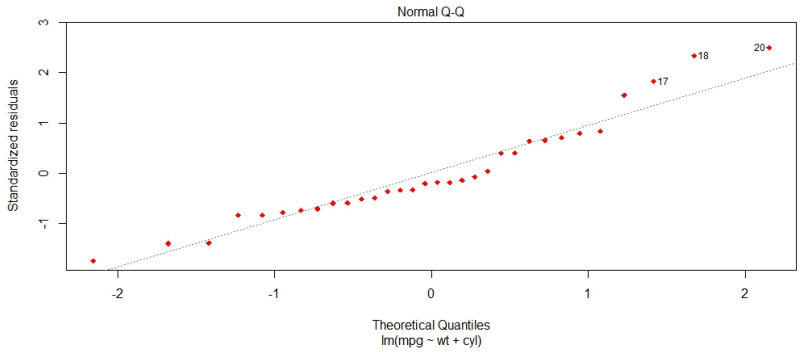
1. a dataset where everything is fine
2. a dataset with a high-leverage, but low-standardized residual point
3. a dataset with a low-leverage, but high-standardized residual point
4. a dataset with a high-leverage, high-standardized residual point

2.2 Diagnostic Plots for LMs

- The **Scale-Location** plot, also called Spread-Location (or S-L plot), takes the square root of the absolute residuals in order to diminish skewness (sqrt($|E|$)) is much less skewed than $|E|$ for Gaussian zero-mean E).
- The **Residual-Leverage** plot shows contours of equal Cook's distance, for values of `cook.levels` (by default 0.5 and 1) and omits cases with leverage one with a warning. If the leverages are constant (as is typically the case in a balanced aov situation) the plot uses factor level combinations instead of the leverages for the x-axis.
  (The factor levels are ordered by mean fitted value.)

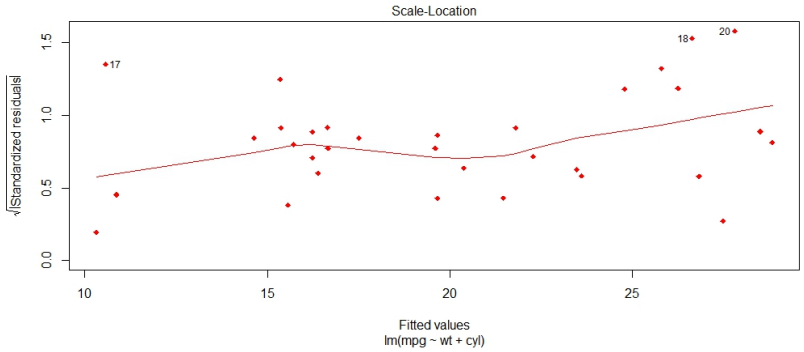
```
plot(lm(mpg~wt+cyl),which=c(1),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(2),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(3),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(4),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(5),pch=18,col="red")
plot(lm(mpg~wt+cyl),which=c(6),pch=18,col="red")
```



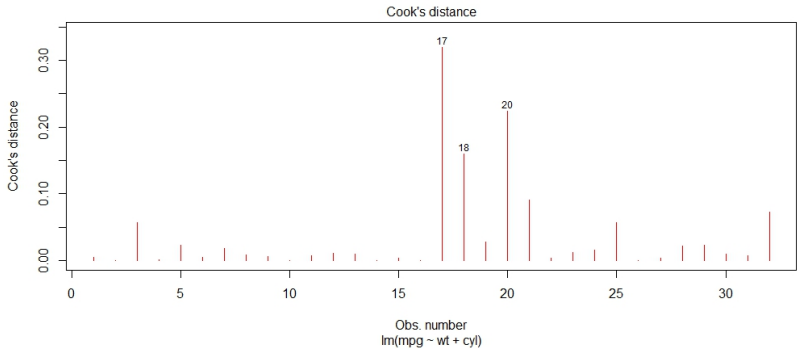


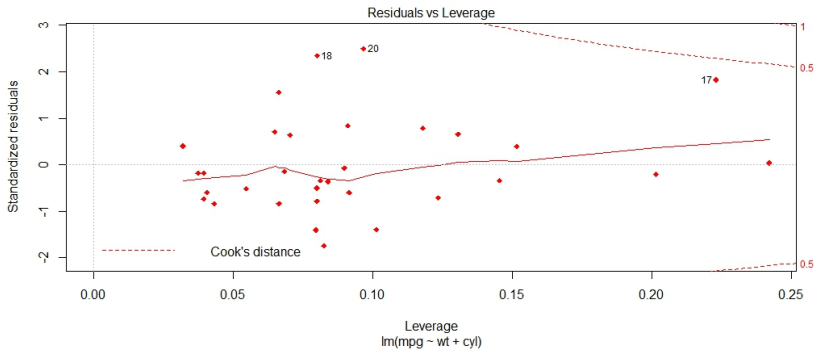
2.2.1 Plot 3 : Normal Probability Plot

This plot is used to assess the validity of the normality of the residuals.



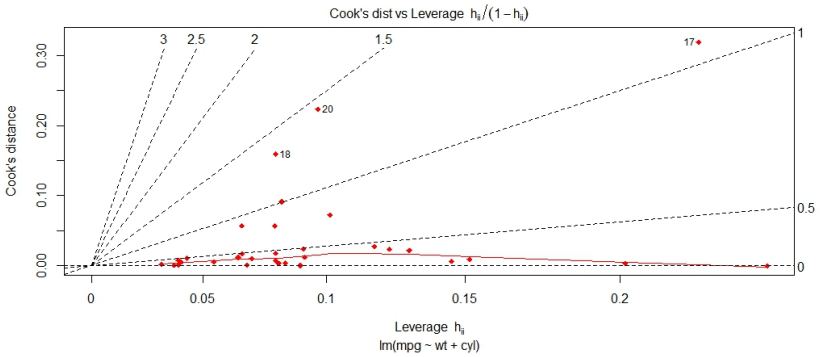
2.2.2 Plot 5 : Cook's Distance





2.2.3 Plot 6 : Cook's Distance vs Leverage

```
par(mfrow=c(4,1))
plot(fittedmodel)
par(opar)
```


Cook's dist vs Leverage $h_{ii}/(1-h_{ii})$
3
2.5
2
1.5
1
0.5
0
17
20
18
Cook's distance
0.00
0.10
0.20
0.30
0
0.05
0.1
0.15
0.2
Leverage h_{ii}
lm(mpg ~ wt + cyl)