# Data Visualisation 2

Code

*Kapil Paranjape*

*08/03/2018*

# Multi-variate data

Typically, multi-variate data is provided as a table with rows corresponding to “individual” samples and columns giving the values of various measurements for those samples. For example, `R` has a data table called `cars` .

```
dim(cars)
```

```
[1] 50  2
```

We can look 4 randomly chosen measurements:

```
cars[sort(sample(c(1:50),4)),]
```

| | **speed**<br><dbl> | **dist**<br><dbl> |
|---|---|---|
| 4 | 7 | 22 |
| 15 | 12 | 28 |
| 24 | 15 | 20 |
| 50 | 25 | 85 |

4 rows

This table contains data for 50 cars about the “stopping distance” measured when a car is travelling at a certain speed. As usual, we can get a summary of individual columns:

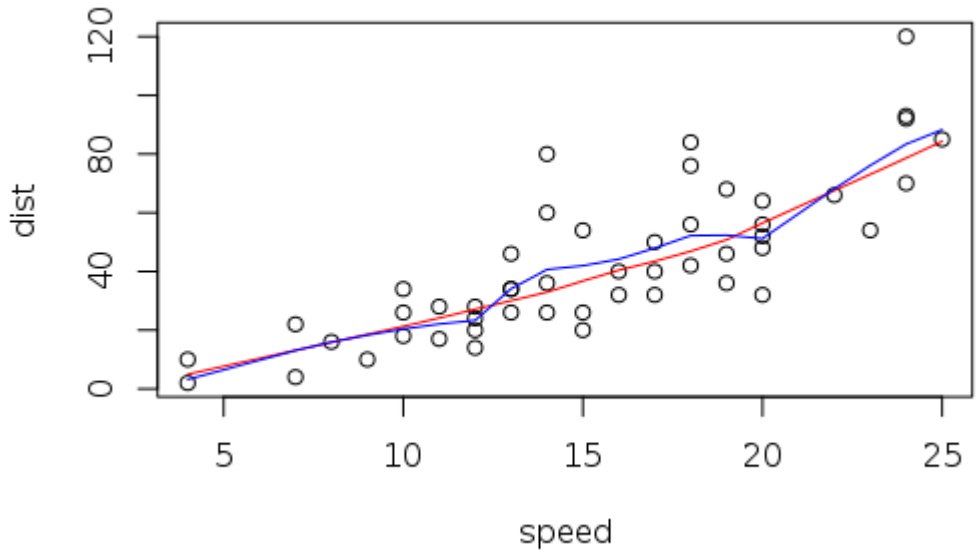
```
summary(cars)
```

```
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```

However, in most cases of multi-variate data, we are interested in the relation *between* the columns. In fact, in this particular table, the data in individual columns is not of interest at all! The relation between columns can be checked visually:

```
plot(cars)
lines(lowess(cars$speed,cars$dist),col="red")
```

```
lines(supsmu(cars$speed,cars$dist),col="blue")
```

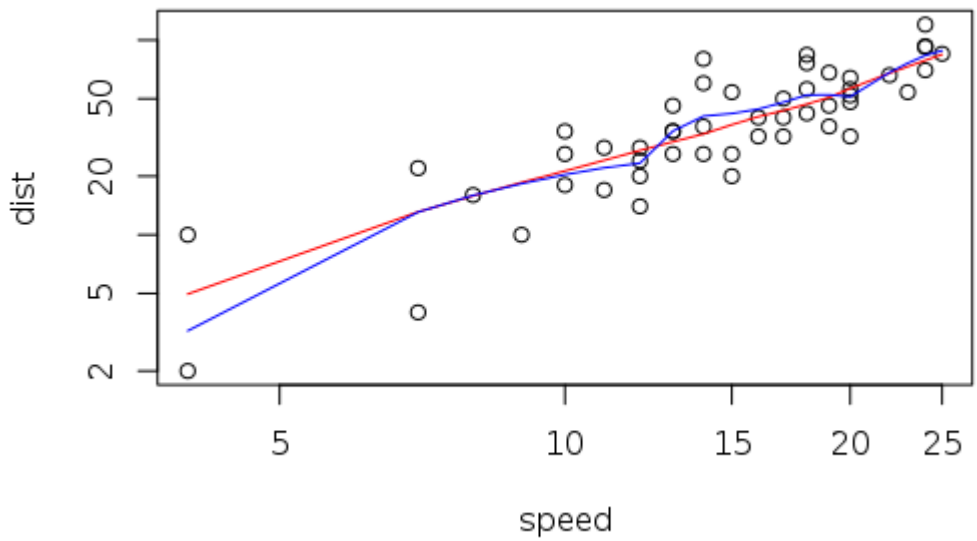



We see that there *appear* to be a relation between speed and stopping-distance which loosely speaking is: “The greater is the speed, the greater is the stopping distance”. This is already a lesson worth learning!

However, as scientists we would like to learn if there is a more precise relation. If the relation is a “power law”, then we expect that taking logarithms on both sides should “linearise” it. (Note that putting `log="xy"` in the parameter for the plot only changes the display, so the data being plotted in the previous picture and the current one is the same.)

```
plot(cars,log="xy")
lines(lowess(cars$speed,cars$dist),col="red")
```

```
lines(supsmu(cars$speed,cars$dist),col="blue")
```




In both cases, we have tried to fit a smooth curve to the data to get a sense of the relation between the two columns.

Let us assume that we sense that there *is* a power law relation. In other words, we feel that there is a linear model that fits the logarithms of the colums.

```
model1 <- lm(log(dist)~log(speed),data=cars)
model1
```

```

Call:
lm(formula = log(dist) ~ log(speed), data = cars)

Coefficients:
(Intercept)   log(speed)
    -0.7297       1.6024
```

This tells us that if we fit a linear model to the logarithms of the columns we will get the formula

$$\log(\text{dist}) = -0.7297 + 1.6024 \cdot \log(\text{speed})$$

We will learn later how this linear model is calculated.

We can calculate

```
exp(-0.7297)
```

```
[1] 0.4820536
```

So roughly speaking the formula becomes:

$$\text{stopping distance} = 0.5 \cdot (\text{speed})^{1.6}$$

This gives us a second life lesson! The relation is *not* linear. The stopping-distance increases *faster* than the speed. (The constant $0.5$ is just something to do with the chosen units.)

There is more to the linear model! We would like to know how well this model fits the data. Here is how `R` summarises the model.

```
summary(model1)
```

```
Call:
lm(formula = log(dist) ~ log(speed), data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00215 -0.24578 -0.02898  0.20717  0.88289

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7297     0.3758  -1.941   0.0581 .
log(speed)    1.6024     0.1395  11.484 2.26e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4053 on 48 degrees of freedom
Multiple R-squared:  0.7331,    Adjusted R-squared:  0.7276
F-statistic: 131.9 on 1 and 48 DF,  p-value: 2.259e-15
```

That is a lot of output! Let us try to understand it piece by piece.

First of all, there is just the statement of what model we are examining.

The word “residuals” is what we can loosely called the “error” in the model; it is the difference between the actual value and the predicted value. The second part gives the distribution of these residuals for the linear model as expressed by the formula:

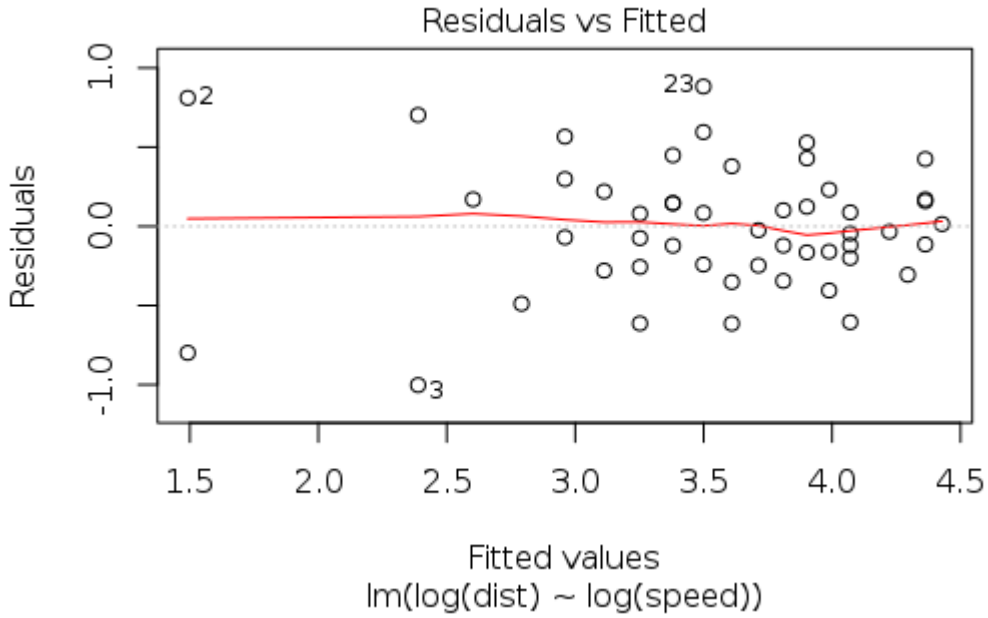
$$\log(\text{dist}) = A + B \cdot \log(\text{speed})$$

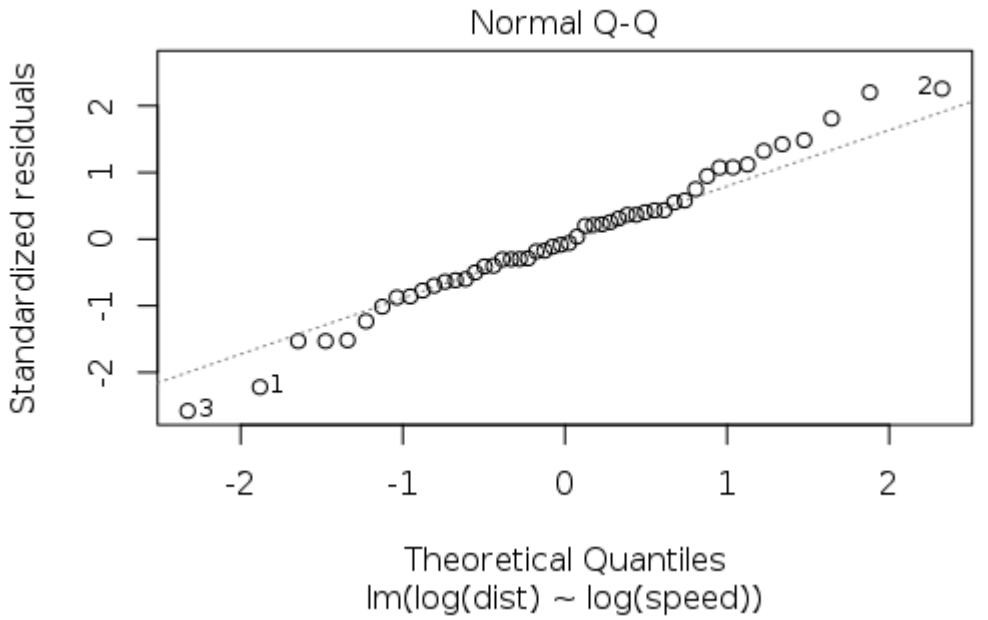
with $A = -0.7297$ and $B = 1.6024$

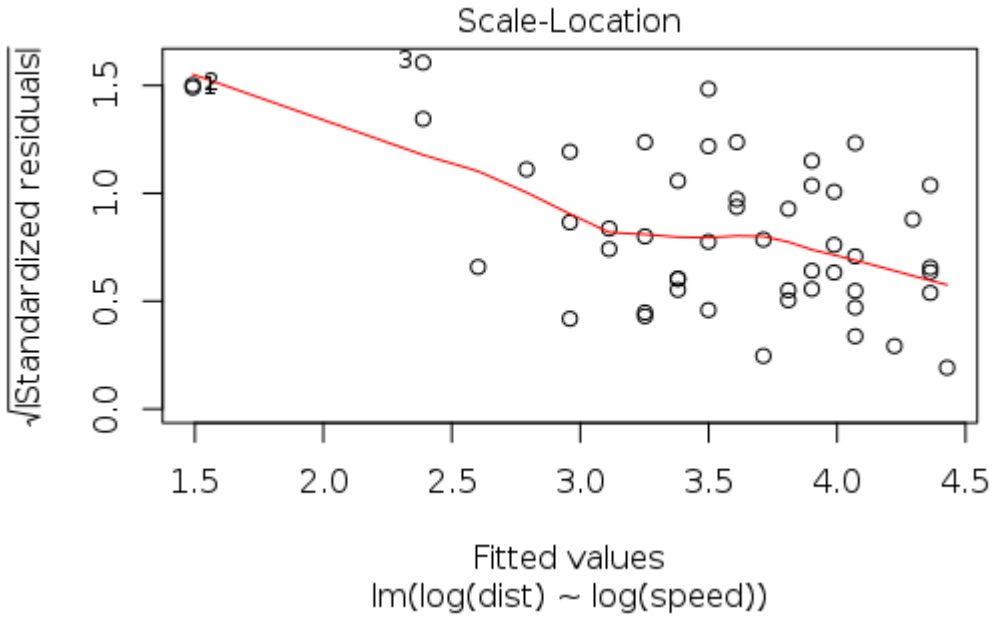
The *real* data is distributed around the model and $A$ and $B$ would themselves be random variables from certain distributions. The next part of the summary gives the estimated values of $A$ and $B$ (essentially the expected values of these distributions) and the “standard error” in these estimates (which is the “standard deviation” of these distributions). There are two more columns giving a $t$-value and a probability of $A$ and $B$ respectively going beyond this value. We will learn more about this later; suffices at this point to say that these columns give an idea of how erroneous the values of $A$ and $B$ might be and what is the probability that model does not really fit!

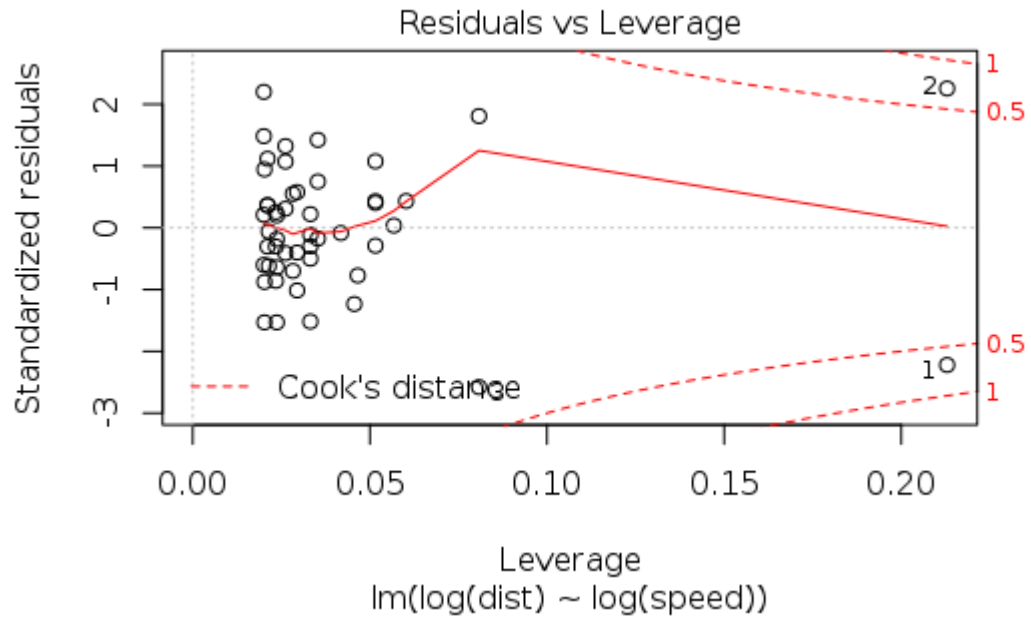
All this is fine, but we would much prefer a *visual* representation as we have been doing so far.

```
#par(mfrow=c(2,2))
plot(model1)
```






This is again a lot of information!

First notice that in each case the vertical axis is labelled as “Standardized Residuals” (in the third one it is the square root). This value is the residual divided by the standard deviation of all the residuals.

The first plot merely shows how these values are distributed as compared with the “fitted” (or model) values. Note that the $x$-axis is now given by the model values of $\log(\text{dist})$ based on the tabulated value of $\log(\text{speed})$.

The next plot on the right compares the standardized residuals against the “standard” (essentially normal) distribution by using a Quantile-Quantile plot. This seems to be quite a good fit. In other words, it *does* seem as if the (Standardized) residuals are normally distributed. Such a fit is often considered an indication that the model is a good one. Note that there *are* a few points well outside

the “ideal” Q-Q plot.

The next plot is the scale and location plot. The expected distribution of the residuals has two parameters, the location is determined by the expectation and the scale is determined by the variance. This plot allows us to visualise around each fitted value, how far the actual values deviate from the model.

The last plot uses a number of terms which we cannot explain at this point of the course! “Leverage” denotes the extent to which a particular data point affects the values in our model. In other words, it is helpful in determining which data-points are “outliers” and could perhaps be excluded so as to make the model fit the remaining data better.