R lecture notes 2022-11-29 (Linear models)

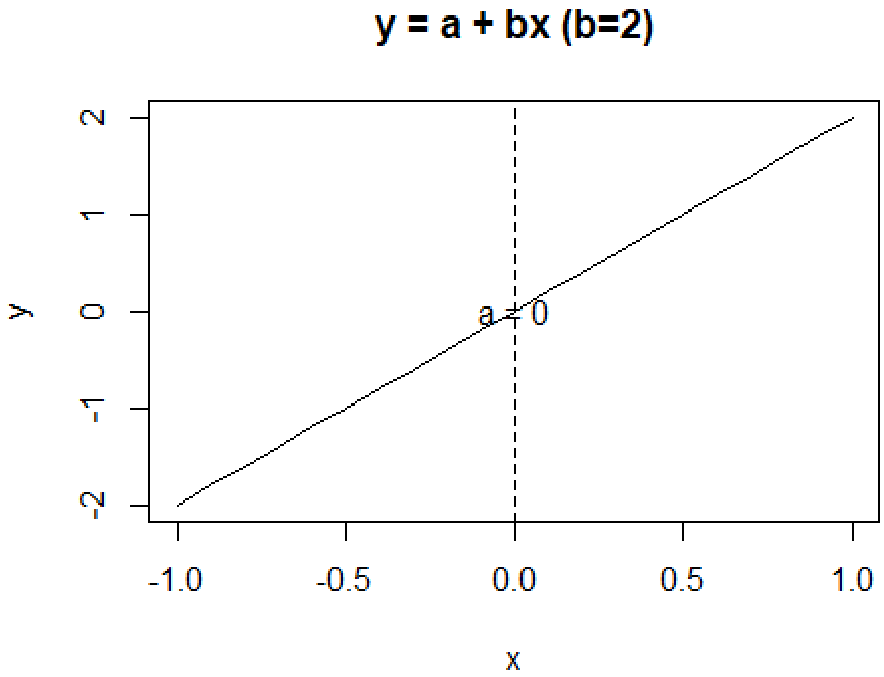
Adam Tallman

2022-11-28

Simple linear model

$$y = a + bx$$

In this formula a is the intercept, b is the coefficient, y is the dependent variable and x is the independent variable.



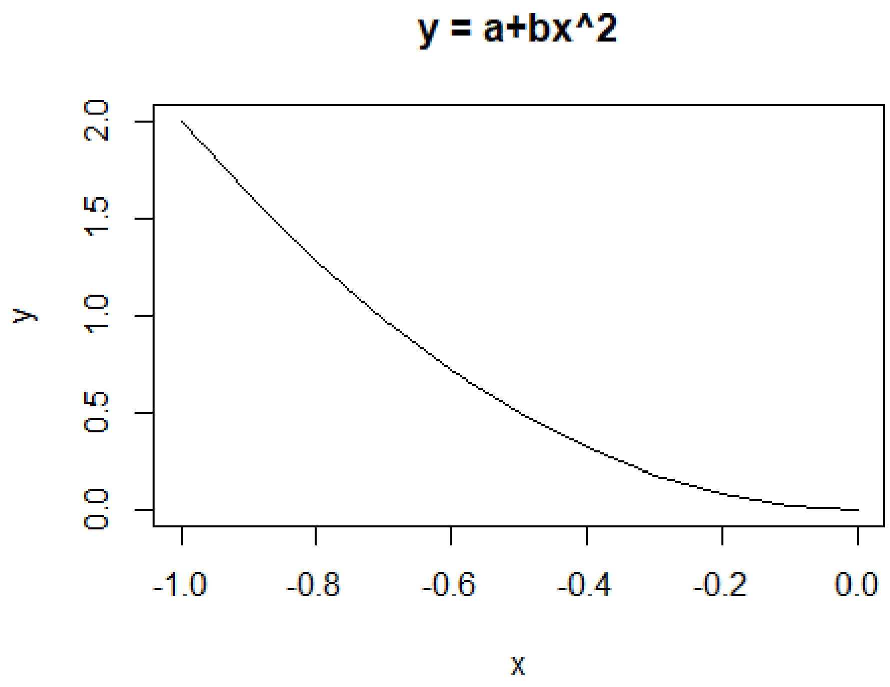
This is the formula for a straight line - if we change the formula slightly will it still be a straight line?

$$y = a + bx^2$$

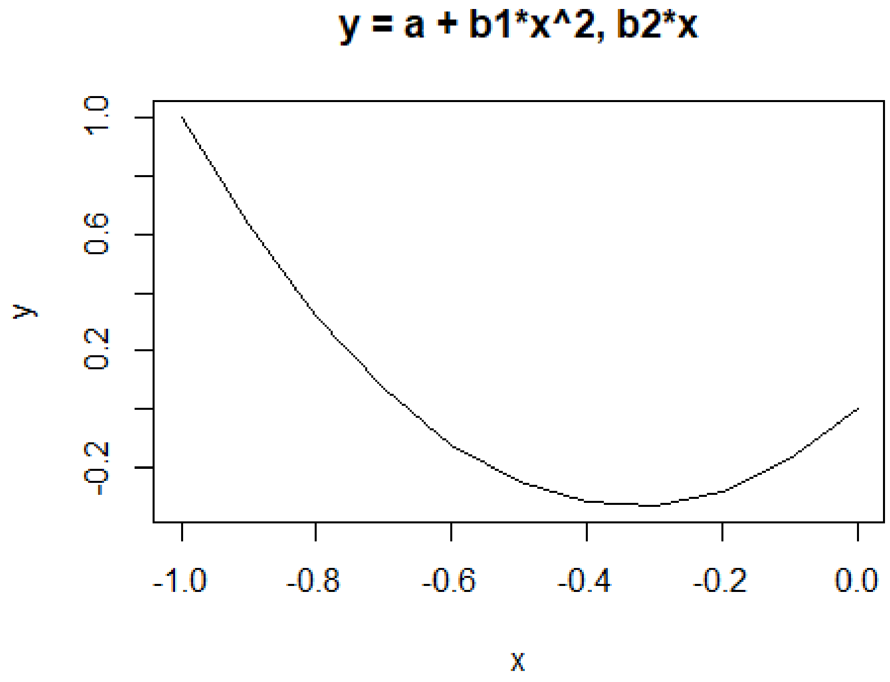
And this?

$$y = a + b_1x^2 + b_2x$$

Let's model an exponential curve.



Let's model a parabolic curve.

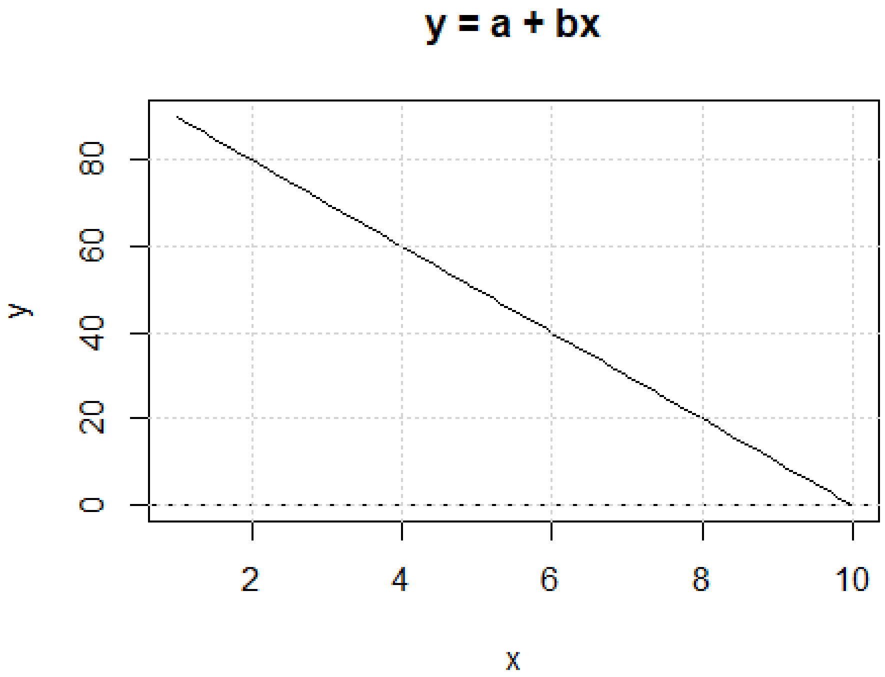


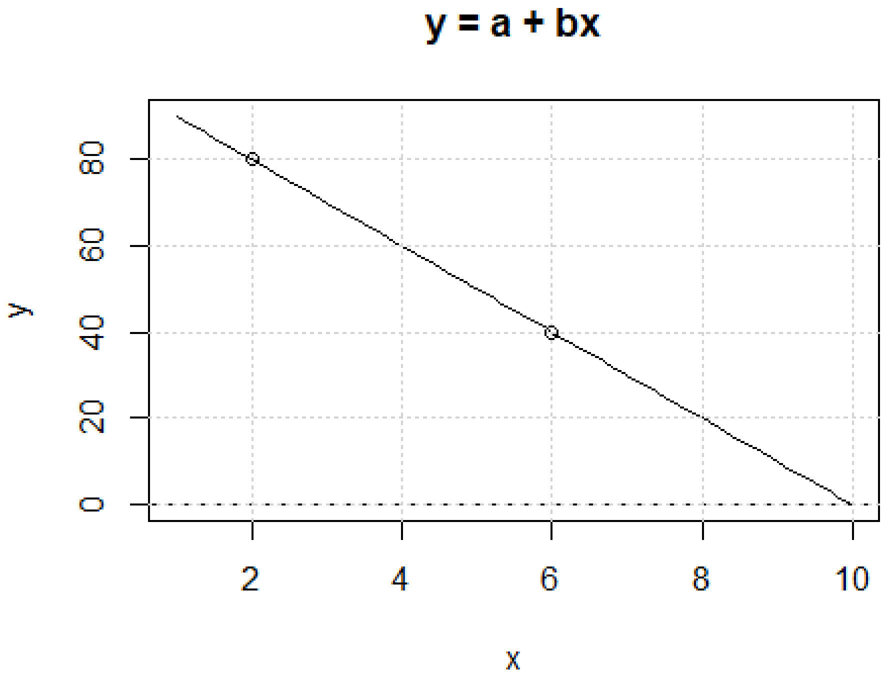
Part of "model choice" is what type of relationship you think x and y display with one another.

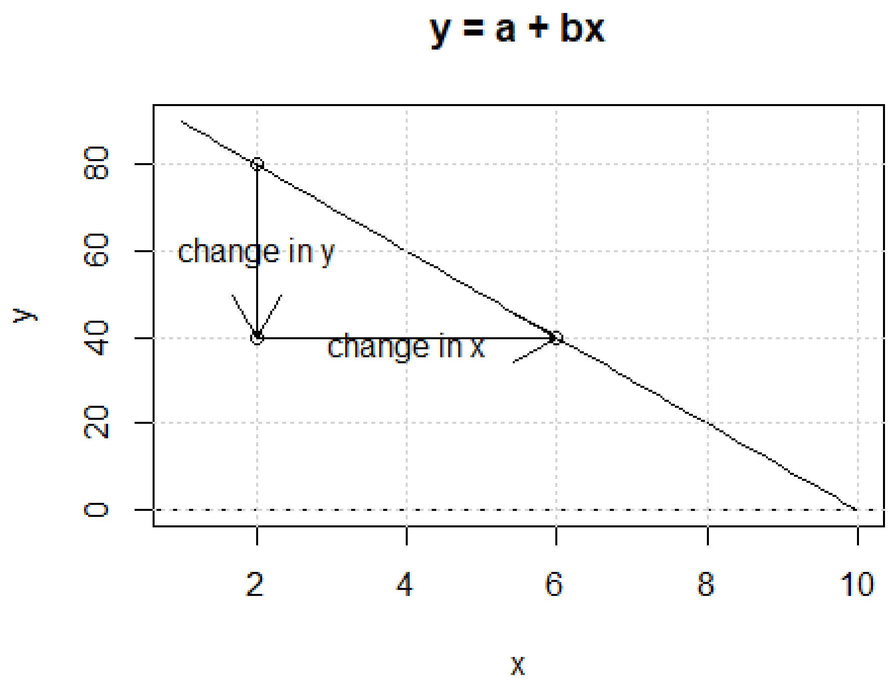
$$y = a + bx$$

- So how do we get the slope coefficient `b’?

$$b = \frac{\text{change in y}}{\text{change in x}}$$


y = a + bx
y
0
20
40
60
80
x
2
4
6
8
10


y = a + bx
y
0
20
40
60
80
x
2
4
6
8
10



So this is how we would calculate this.

$$b = \frac{-40}{4}$$

Word length and reaction times

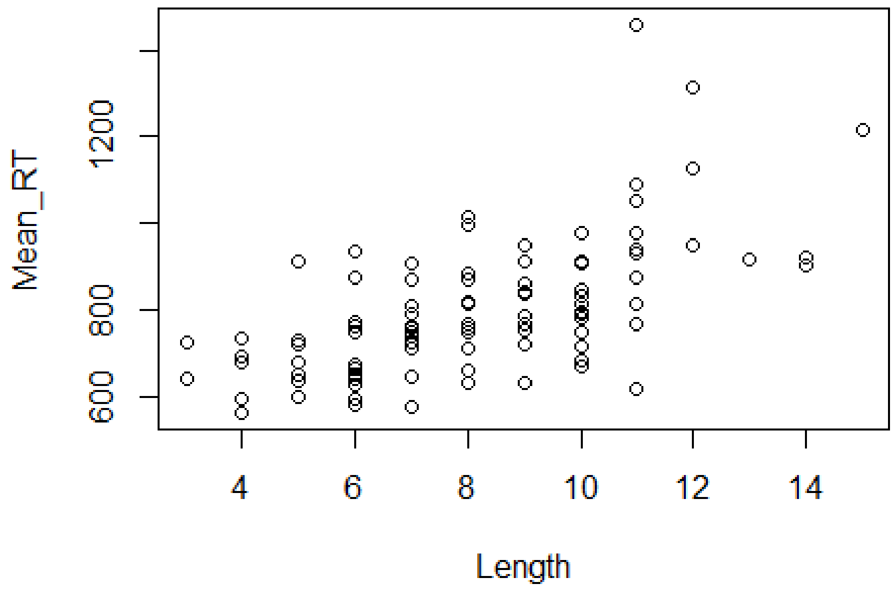
First download the file Rling_1.0.tar.gz from Levshina's github page. Then you have to install it in R.

```
## Warning: package 'Rling' is in use and will not be installed
```

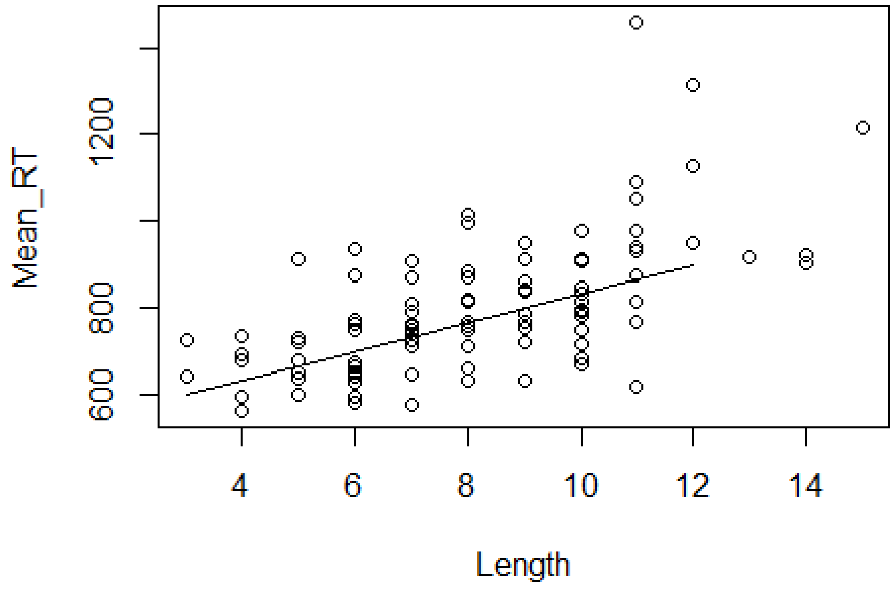
We load the Rling package and then load the data of the Rling package called ldt.

```
##      Length           Freq             Mean_RT      
##  Min.   : 3.00   Min.   :    0.0   Min.   : 564.2  
##  1st Qu.: 6.00   1st Qu.:   53.5   1st Qu.: 713.1  
##  Median : 8.00   Median :  310.5   Median : 784.9  
##  Mean   : 8.23   Mean   : 3350.3   Mean   : 808.3  
##  3rd Qu.:10.00   3rd Qu.: 2103.2   3rd Qu.: 905.2  
##  Max.   :15.00   Max.   :75075.0   Max.   :1458.8
```

So there's a data with response times in relation to word length.



$$b = \frac{dy}{dx} = \frac{900 - 600}{12 - 3} = 33.3$$



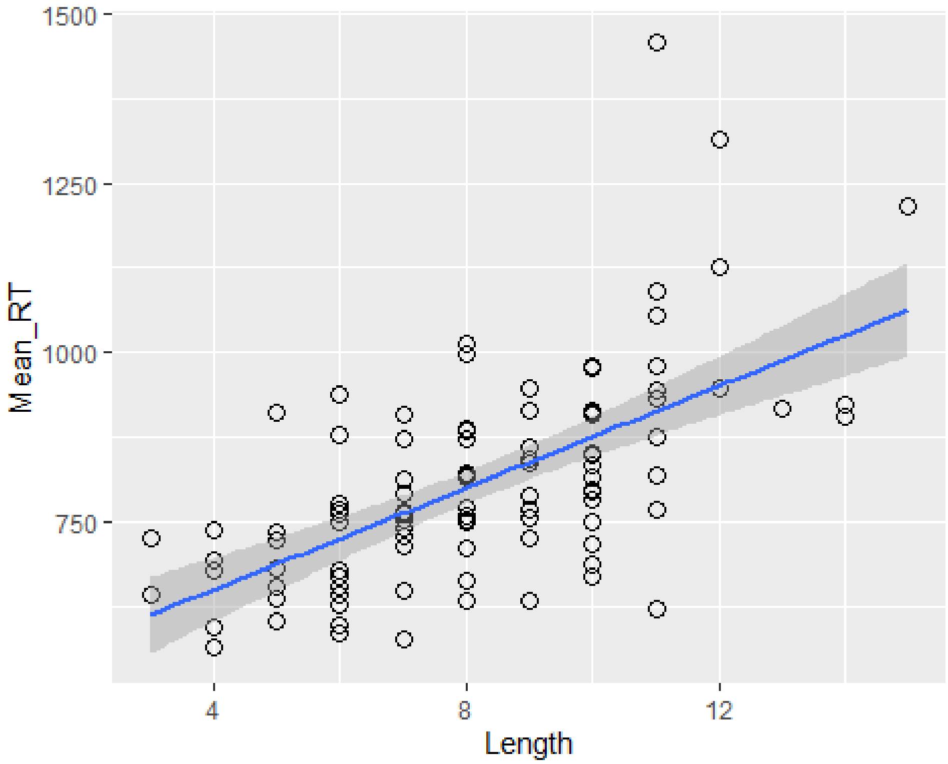
We can use the function ‘lm’ to provide a simple linear model.

```
##
## Call:
## lm(formula = Mean_RT ~ Length)
##
## Coefficients:
## (Intercept)       Length
##      498.44        37.64
```

In ggplot the code is as follows.

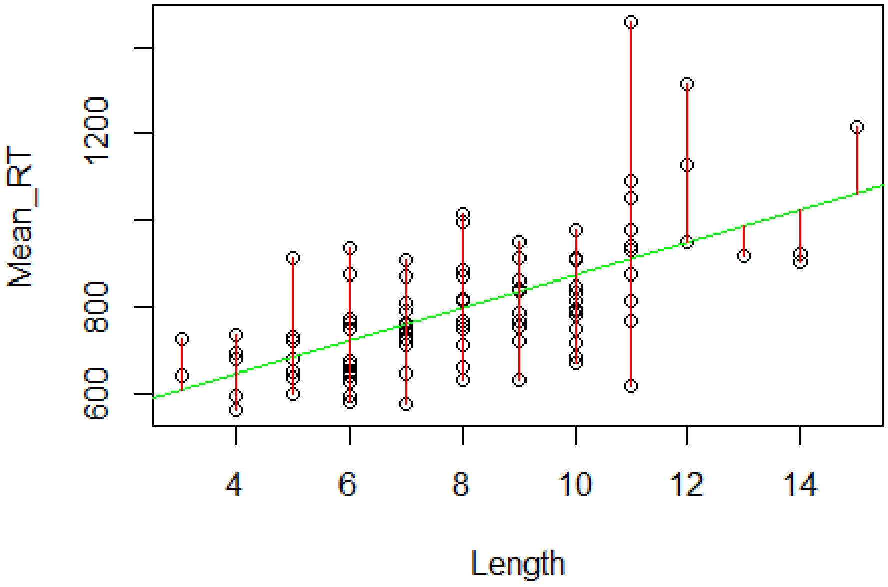
```
## The following objects are masked from ldt (pos = 3):
##
##     Freq, Length, Mean_RT

## `geom_smooth()` using formula 'y ~ x'
```



Residuals

A residual is the difference between each data point and the value predicted by the model at the same value of x. Some residuals are positive and others are negative.



$\hat{y}$ refes to the predicted value. d refers to the residual of a specific data point.

$$d = y - \hat{y}$$

$$d = y - (a + bx)$$

$$d = y - a + bx$$

A regression calculates a line with intercept and slope that make the sum of the residuals eqaul to 0 (or as close to 0 as possible).

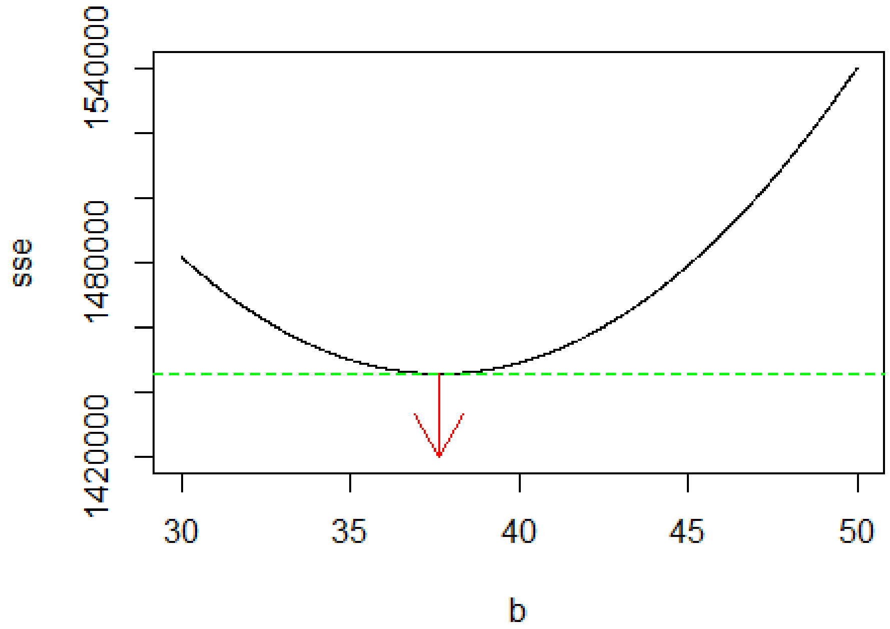
$$\sum d = \sum(y - a - bx) = 0$$

So how do we (or how does R) arrive at this number?

$$SSE = \sum d^2$$

Imagine guessing what the slope. After this we change the value of the slope. Then we work out the new intercept $ a = y -bx - d $. Then we predict the fitted values of y (reaction time) for the new b. After this we work out the residuals $ y - a bx $ Then we calculate the SSE. Then we repeat for values of b until we arrive at the smallest number.

We can actually simulate what this looks like.



When we use lm() it gets the same result.

```
##
## Call:
## lm(formula = Mean_RT ~ Length)
##
## Coefficients:
## (Intercept)       Length
##      498.44        37.64
```

This is the brute force way to calculate coefficients. There is another way to calculate the coefficients using the magic of calculus. So we want to find the minimum of

$$\sum(y - a - bx)^2$$

that translates to wanting to find the derivative of ...

$$\frac{dSSE}{db} = -2\sum(y - a - bx)$$

If you solve for b, you get

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

In this equation the value in the numerator is called the `corrected sum of products'' (SSXY). The value in the denominator is called the`corrected sum of squares of x" (SSX).

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

The corrected sum of squares of the slope is represented as

$$b = \frac{SSXY}{SSX}$$

We can also solve for the intercept with the following equation.

$$a = \overline{y} - b\overline{x} = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

From all of this we can get the best line for fitting the data, but we do not yet have a measure of the unreliability or uncertainty of our model.

We are now going to want the following:

The Error sum of squares: SSE
Regression sum of squares: SSR
F-statistic
p-value
R squared
Confidence intervals

s^2 is the error variance

$$SSE = SSY - SSR$$

$$s^2 = \frac{SSE}{n-2}$$

$$SSR = b \cdot SSXY$$

The F statistic is the regression variance divided by the error variance.

$$F = \frac{\text{Regression error}}{\text{Error variance}}$$

$$F = \frac{SSR}{s^2}$$

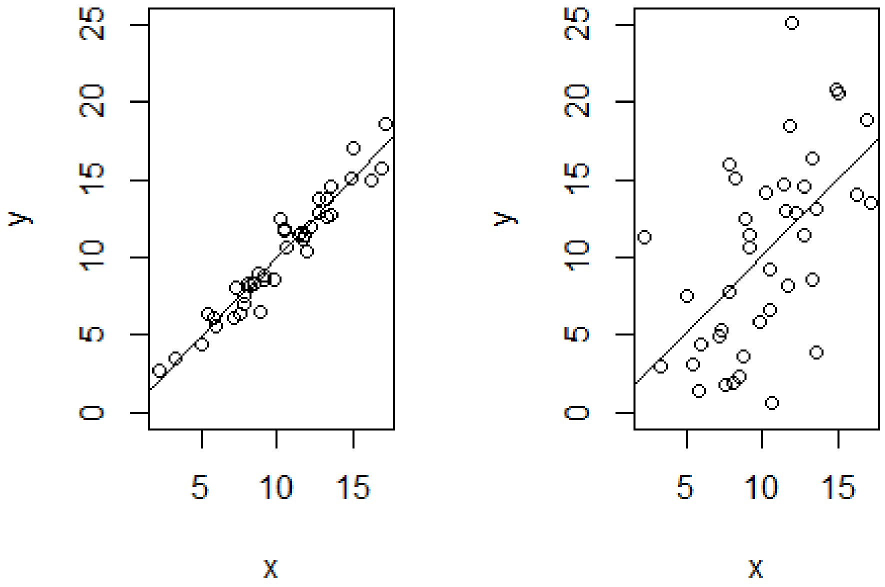
This is how we can calculate everything ``by hand".

```
##
## Call:
## lm(formula = Mean_RT ~ Length)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -291.74  -77.81   -3.69   47.92  546.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  498.443     41.949  11.882  < 2e-16 ***
## Length        37.644      4.879   7.716 1.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121.5 on 98 degrees of freedom
## Multiple R-squared:  0.3779, Adjusted R-squared:  0.3716
## F-statistic: 59.53 on 1 and 98 DF,  p-value: 1.019e-11
```

From model fitting to statistical inference

Two regression lines can have the same slope and intercept, but be different with respect to their residuals.



Measuring the Degree of Fit

We use the r^2 for measuring the degree of fit. The old school name for this is the ``coefficient of determination". If r^2 = 1, the regression line explains all of the variation. If r^2 = 0, the regression line explains none of the variation. The equation for the r^2 is as follows.

$$r^2 = \frac{SSR}{SSY}$$

The top of the equation is the regression sum of squares which as we saw previously is calculated as

$$SSR = b \cdot SSXY$$

The SSXY is the corrected sum of products:

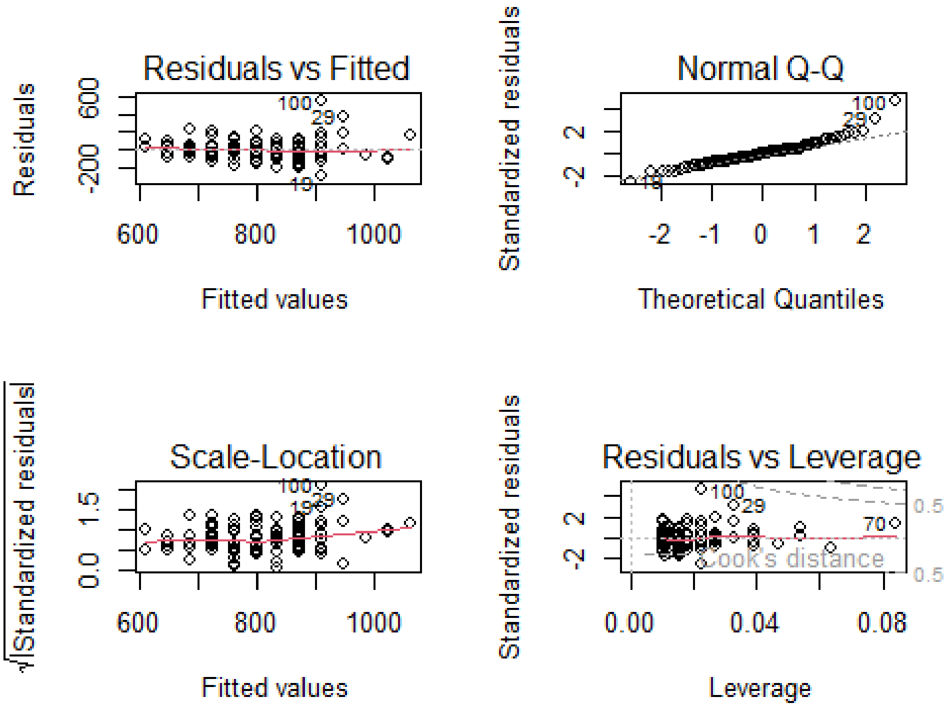
$$\sum xy - \frac{\sum x \sum y}{n}$$

The bottom of the equation is the corrected sums of squares of y and x, which we say above.

$$SSY = \sum(y^2) - \frac{(\sum y)^2}{n}$$

Model checking

In principle when you run a regression it is good practice to look at the residuals. We can do this by plotting the model with plot().



The most important things to look at for now are the Residuals vs. Fitted and the Normal Q-Q plot. The residuals against fitted should not show a pattern: there should be an increase in spread at different values of y, it should be a flat line. The quantile-quantile plot should be a straight line.

The degree to which the residuals vs fitted and the normal Q-Q plot do not look like they are supposed to is the degree to which the model might be misspecified.

The plots above actually suggest that the model might be misspecified. Actually this is not so surprising, because at a certain threshhol of word length, the reaction time increases more dramatically, but we were pretending the relationship is one of a straightline.

You might wonder why this matters. Well, for this, we have to take a step back and ask what a statistical model is doing vis-a-vis our scientific hypothesis. Scientific hypotheses are supposed to make novel predictions. But when a model is “misspecified” (the residuals aren’t even across different values), it means there are limits regarding what predictions we can make about certain values in the future.

Analysis of variance (again)

An analysis of variance table provides information about variance and error in a regression analysis

```
##               Df  Sum Sq Mean Sq F value   Pr(>F)    
## Length         1  878171  878171   59.53 1.02e-11 ***
## Residuals     98 1445574   14751                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F statistic is calculated by the corrected sums over the variance.

$$SSR = b \cdot SSXY$$

$$F = \frac{SSR}{s^2}$$

## ANOVA and regression

What is typically referred to as ANOVA (Analysis of variance) often refers to a type of statistical analysis when all the predictor variables categorical. There is some controversy with respect to whether ANOVA can just be understood as a special case of regression analysis. Compare

- Cottingham et al. 2005 ``Regression versus ANOVA” in Frontiers in Ecology and the Environment

- Gelman 2006 ``Analysis of variance - why it is more important than ever” in The Annals of Statistics

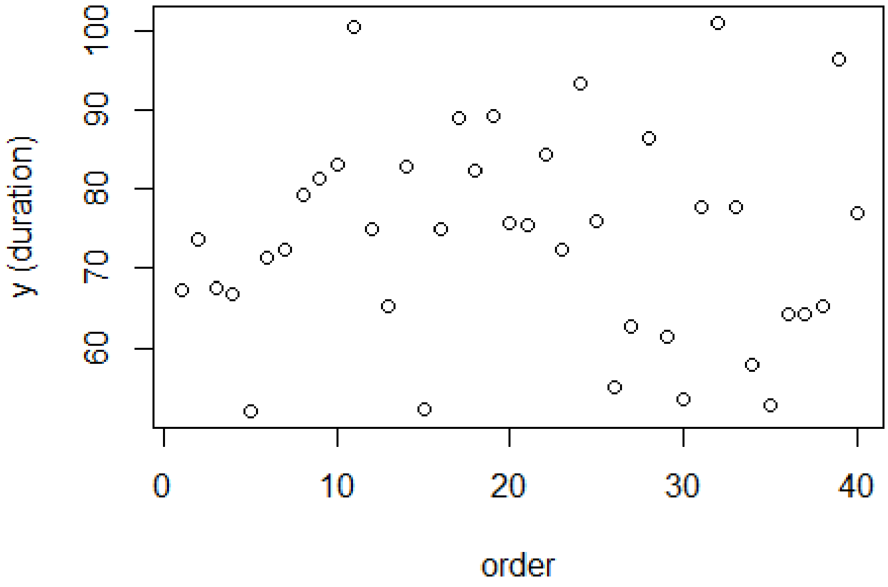
ANOVA has the same mathematical structure as regression,but some regard it as conceptually different, ANOVA can also mean the analysis of data into batches and groups (hierarchical modeling), rather than narrowly as the classical ANOVA test. For historical reasons the ANOVA is more often used in experimental settings and regression analysis for observational data.

Sometimes classical ANOVA is used, when regression really should be: the research bins the data of the predictor variable into groups to make it categorical (see Stoll & Gries 2009) in Journal of Language Acquisition.

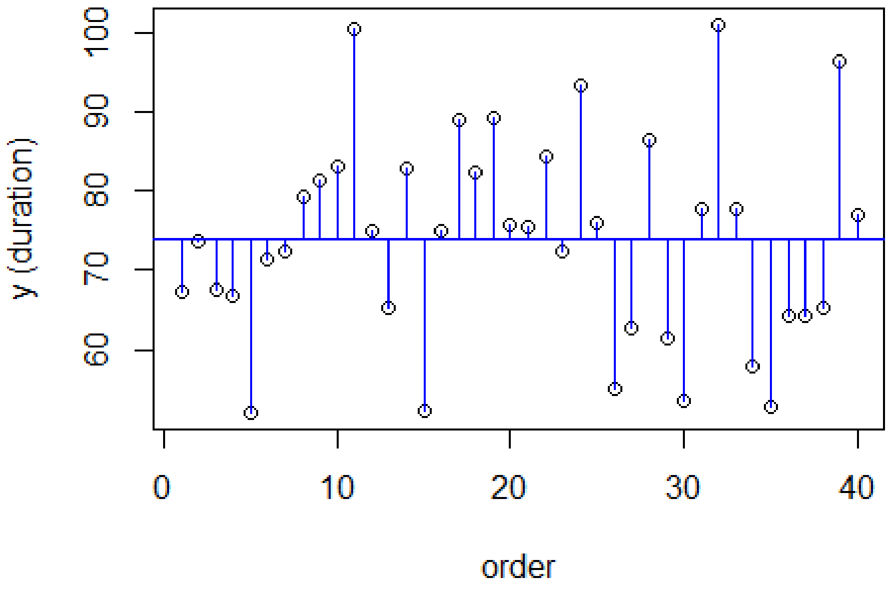
One-way ANOVA

For the stressed versus unstressed distinction, we would typically use an ANOVA if we had production data. In such a model we try to predict the duration from the stressed/unstressed distinction.

We know from this simulation that all the data points in d actually come from two groups. But let's make a plot that assumes that we do not know this.



A one way ANOVA starts by getting the residuals assuming that everything is in one group. To do this we just estimate the distance that each point is from the mean.



We start with an overall measure of variance, total sum of squares or SSY.

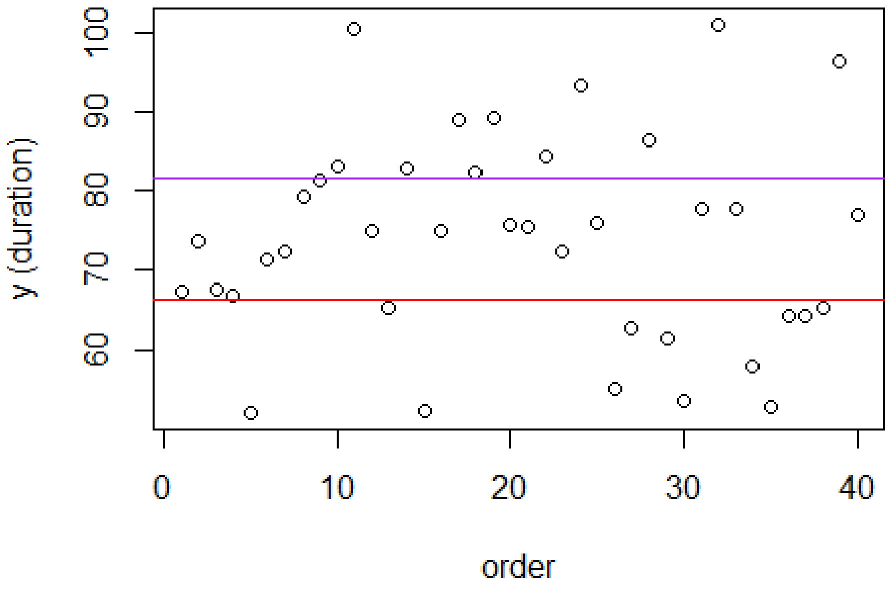
$$SSY = \sum(y - \overline{y})^2$$

Note that earlier we represented this as - its the same thing.

$$SSY = \sum y^2 - \frac{\sum y^2}{n}$$

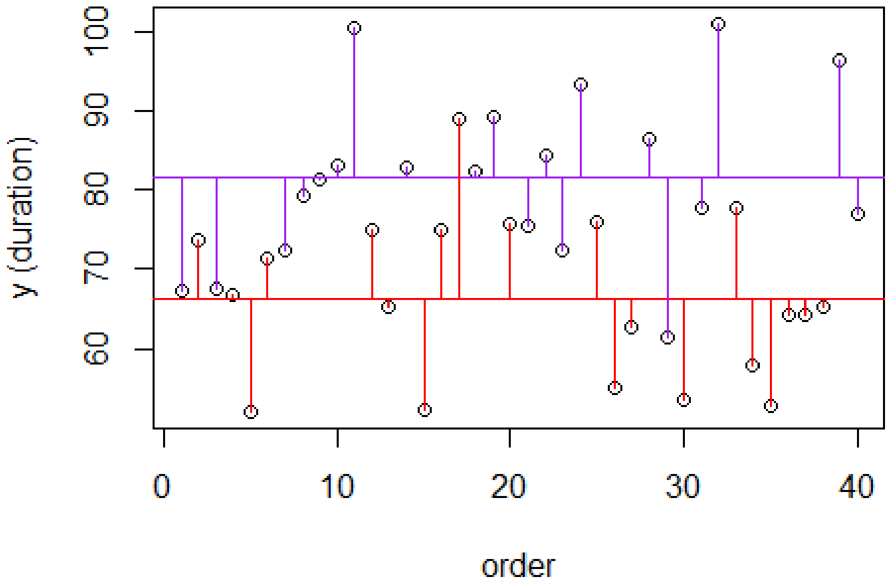
Plotting the means of the variation within the two groups you get.

```
## Warning in plot.xy(xy, type, ...): NAs introduced by coercion
```



And then adding lines for the distances between these two lines.

```
## Warning in plot.xy(xy, type, ...): NAs introduced by coercion
```



We ask two questions about the splitting of the variance in to two groups.

- If there is no significant difference between stressed and unstressed vowels with respect to duration, what would the overall difference in length of the red/purple lines versus the green lines?

- If there was a significant difference between stressed and unstressed vowels, would the residuals of represented by the green lines be larger or smaller than those of the purple and red lines?

$$SSY = \sum(y - \overline{y})^2$$

$$SSE_{strong} = \sum\left(y_{strong} - \overline{y_{strong}}\right)^2$$

$$SSE_{weak} = \sum(y_{weak} - \overline{y_{weak}})^2$$

$$SSE_{total} = SSE_{strong} + SSE_{weak}$$

We are interested to see if the dispersion with one line (mean) is more or less than the dispersion with two lines (mean). We want to measure how much variance is accounted for by having two factors as opposed to one. SSY is just variation in the population. SSE is the sum of variation i the two popoulations assuming they are different. If they are not different than SSS (treatment sum of squares) should be zero (or not statistically different from 0)

$$SSE_{treatment} = SSY_{pooled} - SSE_{SSE_{stressed}+SSE_{unstressed}}$$

In order to see whether this number if large we have to calculate the F statistic.

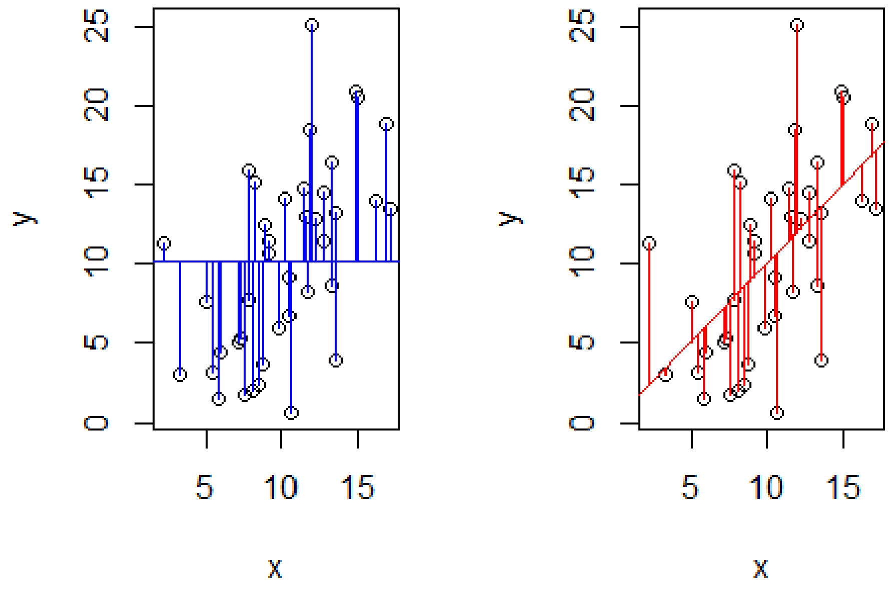
$$s^2 = \frac{SSE}{n-2}$$

$$F = \frac{SSE_{treatment}}{s^2}$$

So what is the difference between regression and ANOVA. One-way ANOVA is just `how well do two horizontal (intercept only) lines, one for each group, account for the total variation." Regression is just`how well does a line with an intercept and a slope account for the total variation”. We’ll see later that we can also add more lines with for the regression …

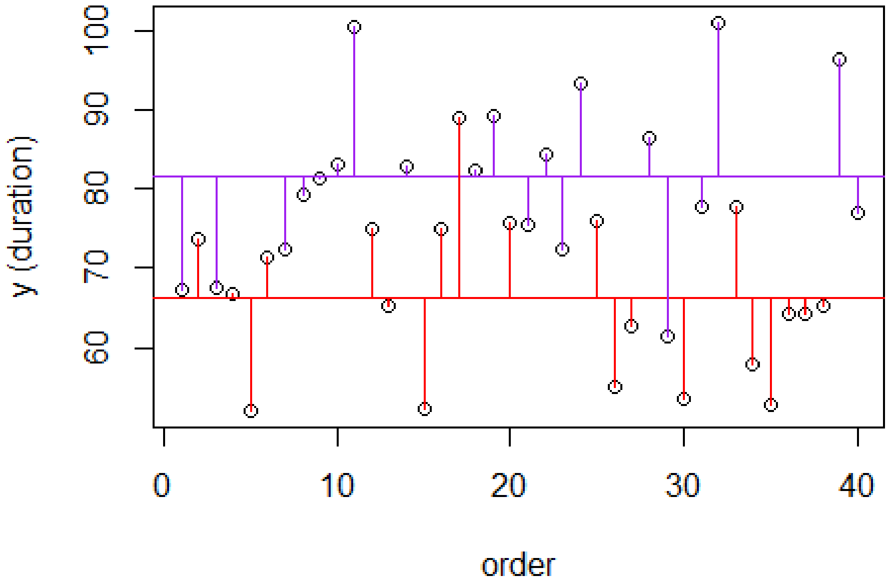
For a regression coefficient, the model asks, what is better.

$$y = a + bx \text{ or } y = a$$



For an ANOVA the model asks what is better, its just that b*x is just a binary partition into two lines.

```
## Warning in plot.xy(xy, type, ...): NAs introduced by coercion
```


y (duration)
order