(http://rinterested.github.io/statistics/index.html)

MANUAL CALCULATION OF ANOVA BETWEEN REGRESSION MODELS:

I will answer your question with an example that (I hope) you can follow in [R]. If you don't use [R] you can still follow the results on this post.

I'll use the data set `mtcars`. You can find documentation of what it is about here (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html). But just remember that there are 32 models, and for each one the miles-per-gallon, horse-power, and other variables are recorded. This is the beginning of it:

```
                 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
```

MODELS:

We'll run two almost random OLS regressions as follows:

```
fit1 <- lm(mpg ~ wt, mtcars)          #mpg regressed on weight of the car
fit2 <- lm(mpg ~ wt + qsec, mtcars)   #mpg regressed on weight and qsec
```

Notice that `fit1` is a ***constrained*** model in the way that we force the regression coefficient for `qsec` in `fit2` to be zero. `fit2`, conversely, is ***unconstrained***.

ANOVA:

```
anova(fit1, fit2)

Analysis of Variance Table

Model 1: mpg ~ wt
Model 2: mpg ~ wt + qsec

     Res.Df    RSS      Df    Sum of Sq      F            Pr(>F)
1     30      278.32
2     29      195.46    1     82.858       12.293         0.0015 **
```

I won't enter into a lengthy explanation of what these values signify, but seeing where they come from will probably help you.

DEGREES OF FREEDOM:

1. Error or Residual Degrees of Freedom: We see them in the output of the `anova` call as `Res. Df 30` and `Res. Df 29`. They are calculated as:

$\text{no. observations} - \text{no. indepen't variables} - 1 = 32 - 1 - 1 = 30$ for `fit1`, and $32 - 2 - 1 = 29$ for `fit2`. Remember that we have 32 car models.

2. Model Degrees of Freedom: It is equal to $\text{no. inepen't variables.}$

3. Total Degrees of Freedom: $\text{no. observations} - 1.$

4. Constraints: The unconstrained model (`fit2`) has two independent variable, and hence, it is a model with 2 degrees of freedom. In contrast, the constrained model (`fit1`) has only 1 degree of freedom. The difference between $\text{model unconstrained df} - \text{model constrained df} = 1$ is the number of constraints, shown on the output of the anova table as `Df 1`.

RESIDUAL SUM OF SQUARES & R SQUARED:

Let's calculate the **RSS** (residual sum of squares (https://en.wikipedia.org/wiki/Residual_sum_of_squares)), also known as sum of squared errors (SSE), and the **F value**. To do so these are the pertinent manual calculations:

Mean dependent variable: $\bar{y}$

```
mu_mpg <- mean(mtcars$mpg)                       # Mean mpg in dataset
```

Total Sum of Squares (TSS): $\sum_1^n (y_i - \bar{y})^2$

```
TSS <- sum((mtcars$mpg - mu_mpg)^2)              # Total sum of squares
```

Model Sum of Squares (MSS): $\sum_1^n (\hat{y}_i - \bar{y})^2$

```
MSS_fit1 <- sum((fitted(fit1) - mu_mpg)^2)      # Variation accounted for by model
MSS_fit2 <- sum((fitted(fit2) - mu_mpg)^2)      # Variation accounted for by model
```

Residual Sum of Squares (RSS, also SSE): $\sum_1^n (y_i - \hat{y})^2$

```
RSS_fit1 <- sum((mtcars$mpg - fitted(fit1))^2)  # Error sum of squares fit1
```

`RSS_fit1` 278.3219

```
RSS_fit2 <- sum((mtcars$mpg - fitted(fit2))^2)  # Error sum of squares fit2
```

`RSS_fit2` 195.4636

Notice that the `RSS` column in the *ANOVA* table correspond to `RSS_fit1 = 278.3219` and `RSS_fit2 = 195.4636` of the manual calculations above.

In the *ANOVA* table we also have the difference in RSS: `sum(residuals(fit1)^2)-sum(residuals(fit2)^2) = 82.85831`, or calculated as indicated above:

$\text{RSS\_fit1 - RSS\_fit2} = 82.85831$, indicated in the anova table as `Sum of Sq`.

Fraction RSS/TSS:

```
Frac_RSS_fit1 <- RSS_fit1 / TSS                  # % Variation secndry to residuals fit1
Frac_RSS_fit2 <- RSS_fit2 / TSS                  # % Variation secndry to residuals fit2
```

R-squared of the model: $1 - RSS/TSS$

```
R.sq_fit1 <- 1 - Frac_RSS_fit1                  # % Variation secndry to Model fit1
```

`R.sq_fit1` 0.7528328 Compare to summary(fit1)$r.square 0.7528328

```
R.sq_fit2 <- 1 - Frac_RSS_fit2                  # % Variation secndry to Model fit2
```

`R.sq_fit2` 0.8264161 Compare to summary(fit2)$r.square 0.8264161

F VALUE:

```
n <- nrow(mtcars)                               # Number of subjects or observations

Constraints <- 1                 # Constraints imposed or difference in iv's fit2 vs. fit1
UnConstrained <- 2               # Independent variables uncontrained model (fit2)
```

$$F = \frac{(R^2_{\text{mod.}2} - R^2_{\text{mod.}1}) \times (N - \text{no. unconstrained}_{\text{mod.}2} - 1)}{((1 - R^2_{\text{mod.}2}) \times \text{no. constraints})}$$

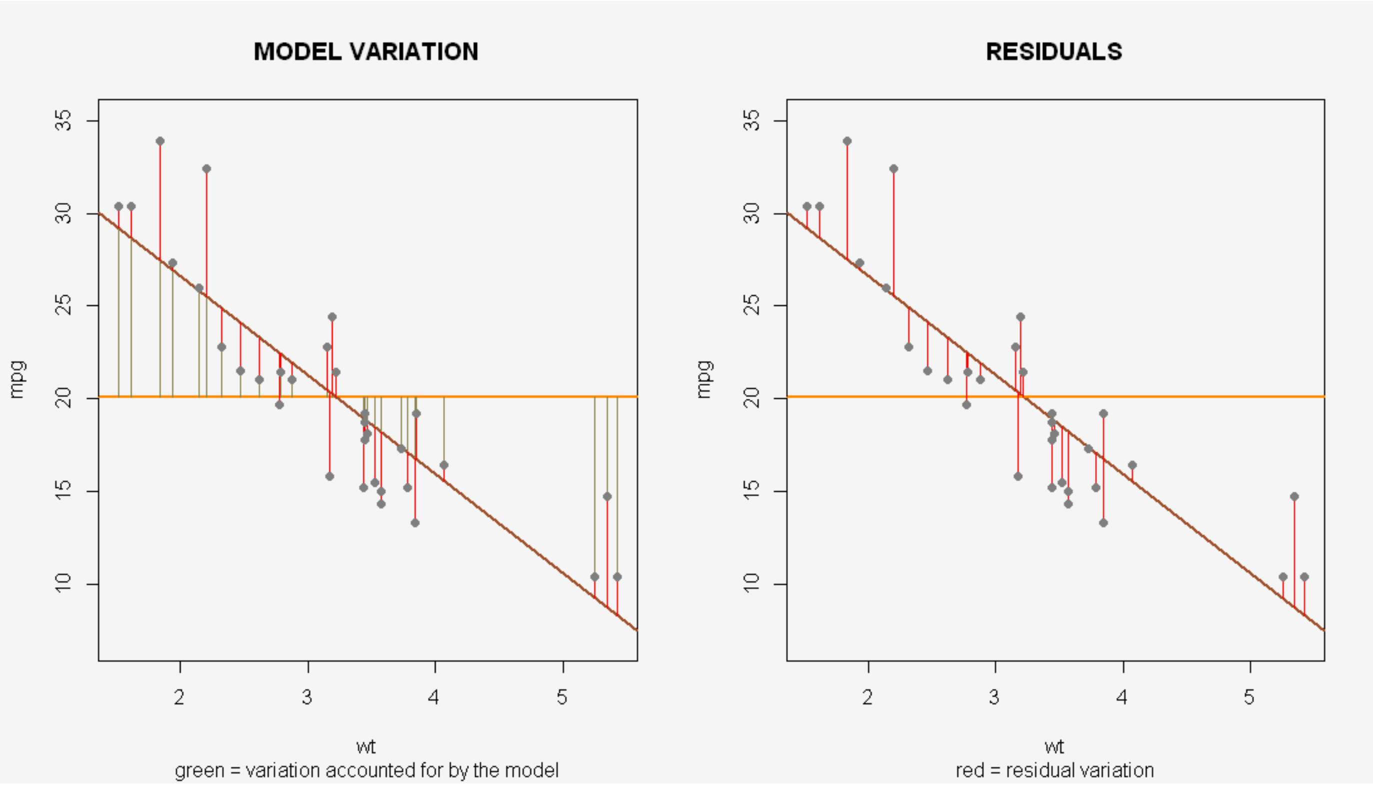
with N corresponding to the number of observations; no. unconstrained, the number of independent variable in the full model; and no. constraints, the difference in independent variables between the full and the reduced model.

```
F_value=(R.sq_fit2 - R.sq_fit1) * (n - UnConstrained - 1) / ((1 - R.sq_fit2) * Constraints)
```

`F_value #` 12.29329

And the `p-value`, which in this case is `0.0015`, which is significant. [R] has a system of stars to point out the level of significance, in this case `p < 0.01`.

In terms of a more graphical interpretation of the *ANOVA* of an OLS regression, we can visualize the model squared variation (`MSS`) for `fit1` as the green lines in the plot below (equivalent to the “between groups” variance or signal). The `RSS` is exactly the sum of the length of the red segments separating the individual points from the fitted regression line (and corresponds to the “within group” variance or noise):



(http://i.stack.imgur.com/WEvJl.png)

(Code here (https://github.com/RInterested/PLOTS/blob/master/LINES%20FROM%20POINTS%20TO%20LINES))

Home Page (http://rinterested.github.io/statistics/index.html)