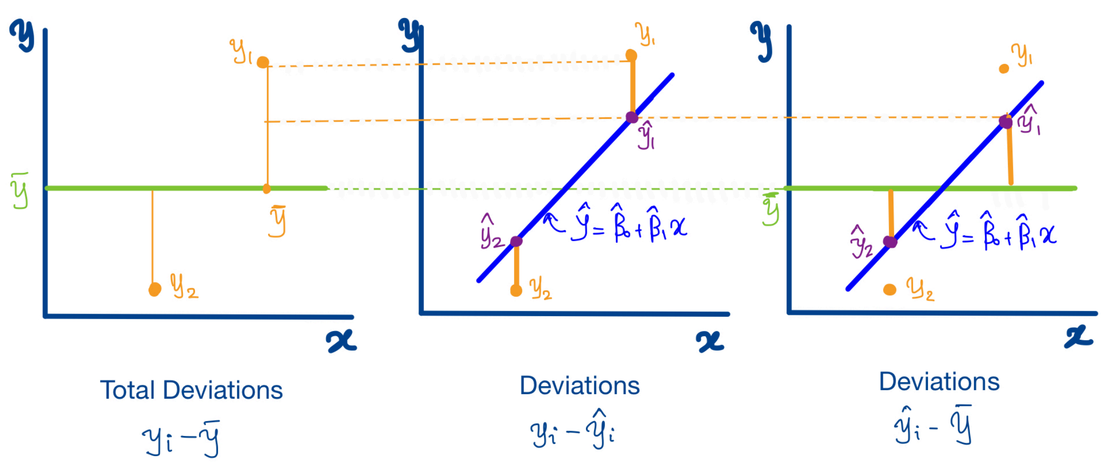
1.6 ANOVA Table & the F-Test

Recall the decomposition of the Total Sum of Squares (*TSS*)

$$\begin{array}{ccccc} TSS & = & FSS & + & RSS \\ \uparrow & & \uparrow & & \uparrow \\ \text{Total variation} & & \text{Variation accounted} & & \text{Variation left} \\ \text{in } y & & \text{through the model} & & \end{array}$$

- $TSS = \sum_i (y_i - \bar{y})^2$ *measures the total variation in ys*: the greater TSS is, the more variation there is in the y values.
- $RSS = \sum_i (y_i - \hat{y}_i)^2$ *measures the variation in the data using the stated regression model*: the larger RSS is, the more y_is vary around the estimated regression line.
- $FSS = \sum_i (\hat{y}_i - \bar{y})^2$ measures *how far* the predicted center of each probability distribution is from the overall center of all y's together.

The partition of the total variation is also illustrated below:



Breakdown of Degrees of Freedom

Each of the terms in the decomposition of the *TSS* is associated with its own degrees of freedom as follows:

- $df_{TSS} = n - 1$: one df is lost, because the sample mean is used to estimate the population mean.
- $df_{RSS} = n - 2$: two df are lost, because the two parameters are estimated in obtaining the fitted values $\hat{y}$.
- $df_{FSS} = 1$: there are n deviations $\hat{y}_i - \bar{y}$, but all the fitted values are associated with the same regression line.

The degrees of freedom are **additive**, so

$$df_{TSS} = df_{RSS} + df_{FSS}$$

Sum of Squares	Expression	*df*
FSS	$\sum_i (\hat{y}_i - \bar{y})^2$	1
RSS	$\sum_i (y_i - \hat{y}_i)^2$	$n-2$
TSS	$\sum_i (y_i - \bar{y})^2$	$n-1$

The Sume of Squares, their corresponding degrees of freedom and the corresponding Mean Squares are all summarized in the famous **ANOVA Table**

Source	SS	df	MS	*F*
Regression (model)	*FSS*	1	$MSReg = \frac{FSS}{1}$	$F = \frac{MSReg}{MSE}$
Error	*RSS*	$n-2$	$MSE = \frac{RSS}{n-2}$	
Total	*TSS*	$n-1$		

As you can see the *Mean Squares* are defined as the ratio of the *SS* divided by its own degrees of freedom:

$$MS = \frac{SS}{df}$$

Be careful, because the Mean Squares **are not additive**.

On the ANOVA table above, we also see an F test, but *to which hypothesis this test corresponds to?*

In **Simple** Linear Regression the answer is easy: it is an alternative test statistic that can be used to test whether the slope is 0 or not, i.e.

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_\alpha : \beta_1 \neq 0 \end{cases}$$

Under the H_0, the F-test statistic is

$$F = \frac{MSReg}{MSE} = \frac{FSS}{RSS/(n-2)} \sim F_{1,n-2}$$

So, if you ask which test to use when you want to test for the statistical significance of the slope, the answer is: either! In fact, if can be shown that the F-test statistic is **equal** to the square of the t-test statistic and their p-values are the same.

$$F = \frac{FSS}{RSS/(n-2)} = \left(\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \right)^2 = t^2$$

So, *this test is equivalent to the t-test before*.

In `R`, the ANOVA table is obtained using the `anova` function.

University Admissions Example (Revisited)

The results of this F test are shown at the bottom of the `lm` output. We can also obtain the full ANOVA table, as discussed in the lecture as follows:

```
admissions.anova = anova(admissions.lm)
admissions.anova
```

```
## Analysis of Variance Table
##
## Response: gpa
##                Df Sum Sq Mean Sq F value    Pr(>F)
## entrance_score  1  5.963  5.9630  33.998 1.597e-05 ***
## Residuals      18  3.157  0.1754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the p-value of the F test is exactly the same as the one obtained in the regression table from the t test for the slope. We can also verify that the square of the t test statistic is equal to the F test statistic. Indeed,

```
admissions.anova[1,4]   ## F- value from ANOVA Table
```

```
## [1] 33.9982
```

```
admissions.coef[2,3]^2  ## Square of t test
```

```
## [1] 33.9982
```