ANOVA table and F tests

1. For the model: $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$ we can write the ANOVA table as:

Source	df	SS	MS
Regression	k	$\hat{\beta}' X' Y - n\bar{y}^2$	SSReg$/k$
Error	$n-k-1$	$\sum \hat{\epsilon}_i^2 = \hat{\epsilon}'\hat{\epsilon}$	SSErr$/(n-k-1) = s_\epsilon^2$
Total	$n-1$	$\sum (y_i - \bar{y})^2 = Y'Y - n\bar{y}^2$	

   Note that $F =$ MSReg/MSError is a test of H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0 \mid \beta_0$.

2. To focus on the sequential fitting order of parameters, we can write the ANOVA table as, for example:

Source	df
$\beta_1 \vert \beta_0$	1
$\beta_2 \vert \beta_0, \beta_1$	1
$\vdots$	$\vdots$
$\beta_k \vert \beta_0, \ldots, \beta_{k-1}$	1
Error	$n-k-1$
Total	$n-1$

3. To test H_0: $C\beta = t$ we use an "additional sum of squares test." The test has three steps.

   (a) Fit the "full model", which corresponds to the *alternative hypothesis*: get its SSError and dfError.

   (b) Fit the "reduced model", which agrees with the *null hypothesis*: get its SSError and dfError.

   (c) Form the F-statistic and perform an F-test. Let

$$F = \frac{(\text{SSError}_{\text{reduced}} - \text{SSError}_{\text{full}}) \,/\, (\text{dfError}_{\text{reduced}} - \text{dfError}_{\text{full}})}{\text{SSError}_{\text{full}}/\text{dfError}_{\text{full}}}$$

   The degrees of freedom for the F-test are $(\text{dfError}_{\text{reduced}} - \text{dfError}_{\text{full}})$ in the numerator and $\text{dfError}_{\text{full}}$ in the denominator.

   The quantity

$$(\text{SSError}_{\text{reduced}} - \text{SSError}_{\text{full}})$$

   in the numerator of the F-test is called the "additional sum of squares." It refers to the change in the error sum of squares when a hypothesis is assumed to be true. The denominator of the F-test could also be called $\text{MSError}_{\text{full}}$.

Residuals

There are several different definitions of residual. Here is a summary table of some of these definitions, including their names in R and SAS.

Definition	Type	R	SAS
$\hat{\epsilon}_i = y_i - \hat{y}_i$	raw	`residuals`	`residual`
$r_i = \frac{\hat{\epsilon}_i}{s_\epsilon \sqrt{1-h_{ii}}}$	internally studentized	`rstandard`	`student`
$t_i = \frac{\hat{\epsilon}_i}{s_{(i)} \sqrt{1-h_{ii}}}$	externally studentized	`rstudent`	`rstudent`

1. $\text{var}(\hat{\epsilon}_i) = \sigma_\epsilon^2(1 - h_{ii})$ where h_{ii} is from the "hat" matrix $H = X(X'X)^{-1}X'$.
2. In simple linear regression,
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$
3. In the formula for t_i, $s_{(i)}$ is the "leave one out" estimate of σ_ϵ based on the regression that is conducted on all the data except the ith case.
4. t_i can also be calculated as
$$t_i = \frac{\hat{\epsilon}_{(i)}}{\sqrt{\widehat{\text{var}}(\hat{\epsilon}_{(i)})}}$$
where $\hat{\epsilon}_{(i)}$ is the residual corresponding to the ith observation, but based on the regression calculated using all data except observation i.

   In either case, t_i has a T distribution with $n - k - 2$ df. This suggests a formal test for an outlier:

   - We identify the ith observation as an outlier if $|t_i| > T_{n-k-2,\alpha/2}$ where $T_{n-k-2,\alpha/2}$ satisfies $P(T_{n-k-2} > T_{n-k-2,\alpha/2}) = \alpha/2$.
   - It is common in this situation to apply a Bonferroni correction and therefore use $\alpha = 0.05/n$.