Method

1. Fit the **COMPLETE MODEL** and obtain the sum of squared errors, SSE_C, available from the ANOVA table.
2. Fit the **REDUCED MODEL** and obtain the sum of squared errors, SSE_R, available from the ANOVA table.
3. Form the test statistic

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

If H_0 is **true**, then $F \sim \text{Fisher-F}(k-g, n-k-1)$

Note: $k-g$ is the number of parameters we set equal to zero when moving from complete to reduced model.

Using this F statistic, we can assess whether there is evidence to support the reduced model over the complete model.

Complete Model ANOVA table:

SOURCE	DF	SS	MS	F
COMPLETE MODEL	k	SSR_C	MSR_C	F_C
ERROR$_C$	$n-k-1$	SSE_C	MSE_C	
TOTAL	$n-1$	SS		

Reduced Model ANOVA table:

SOURCE	DF	SS	MS	F
REDUCED MODEL	g	SSR_R	MSR_R	F_R
ERROR$_R$	$n-g-1$	SSE_R	MSE_R	
TOTAL	$n-1$	SS		

The result holds for comparing any two nested models where the standard model assumptions hold:

- ϵ uncorrelated
- ϵ independent of $x_1, \ldots, x_k$
- ϵ has constant variance
- $\epsilon \sim N(0, \sigma^2)$

Note: It does not allow us to compare non-nested models; for example

$$\begin{aligned} \text{MODEL 1} &: \quad y = \beta_0 + \beta_1 x_1 + \epsilon \\ \text{MODEL 2} &: \quad y = \beta_0 + \beta_2 x_2 + \epsilon \end{aligned}$$

- **NOT NESTED !**

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{①/②}{③/④}$$

① - $SSE_R - SSE_C$: this is the improvement in fit when the reduced model is extended to the complete model

② - $k - g$: this is the number of extra parameters needed to extend the reduced model to the complete model

③ - SSE_C
④ - $n - k - 1$

③/④ - this is the best guess we have at the true value of σ^2, that is, the estimate of σ^2 constructed using as much information as possible, once the effects of

$$x_1, \ldots, x_k$$

have been accounted for.

Example: Hooker's Data.

We consider the two models:

$$
\begin{aligned}
\text{MODEL 1} &: \quad y = \beta_0 + \beta_1 x + \epsilon \\
\text{MODEL 2} &: \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon
\end{aligned}
$$

Here

- MODEL 1: Reduced Model
- MODEL 2: Complete Model

$k = 2$, $g = 1$.

IS THE QUADRATIC TERM NEEDED ?

Example: Hooker's Data.

COMPLETE MODEL	SSR_C	2286.933
	SSE_C	4.382
REDUCED MODEL	SSR_R	2272.474
	SSE_R	18.840

$n = 31, k = 2, g = 1$

$$\Longrightarrow k - g = 1, n - k - 1 = 28$$

So

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{(18.840 - 4382)/1}{4.382/28} = 92.383$$

Example: Hooker's Data.

We compare F with the

$$\text{Fisher-F}(k - g, n - k - 1) \equiv \text{Fisher-F}(1, 28)$$

distribution.

$$F_{0.05}(1, 28) = 4.20$$

Thus

$$92.383 = F > F_{0.05}(1, 28) = 4.20$$

and $H_0 : E[Y] = \beta_0 + \beta_1 x$ is **REJECTED** in favour of $H_a : E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2$.

i.e. the **quadratic model** fits better than the straight-line model.

NOTE: From the original ANOVA tables, we already know that Model 1 and Model 2 both fit better than the null model

$$\begin{aligned} \text{MODEL 0} E[Y] &= \beta_0 \\ y &= \beta_0 + \epsilon \end{aligned}$$

where there is no dependence on x.

We have now confirmed that Model 2 fits better than Model 1.

Example: Diabetes Data.

Factor Predictor: **group** (X_2)
Continuous Covariate: **loggluf** (X_1)
Response: **logglut** (Y)

We have five models to confirm:

$$\begin{aligned} \text{MODEL 0} &: \quad 1 \\ \text{MODEL 1} &: \quad X_2 \\ \text{MODEL 2} &: \quad X_1 \\ \text{MODEL 3} &: \quad X_1 + X_2 \\ \text{MODEL 4} &: \quad X_1 + X_2 + X_1.X_2 \end{aligned}$$

Example: Diabetes Data.

MODEL 4 us the most complex model with 6 parameters

$$\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \beta_{30}, \beta_{31}$$

MODEL 4:

$$E[Y] = \begin{cases} \beta_{10} + \beta_{11} x_1 & \text{GROUP 1} \\ \beta_{20} + \beta_{21} x_1 & \text{GROUP 2} \\ \beta_{30} + \beta_{31} x_1 & \text{GROUP 3} \end{cases}$$

All of the other models are nested inside Model 4; we can obtain them all by setting parameters equal to zero.