Forward Selection: we start with Model 0 and build up.

Model 1 vs Model 0 $\quad F = 412.568$

Model 2 vs Model 0 $\quad F = 940.846$

It seems that Model 2 is the better improvement, so we try the selection path

$$\text{Model } 0 \longrightarrow \text{Model } 2 \longrightarrow \text{Model } 3 \longrightarrow \text{Model } 4$$

Model	SSE	$SSE_R - SSE_C$
0	28.504	-
2	3.738	24.766
3	1.472	2.266
4	1.318	0.154

ie we work down the table, $28.504 - 3.738 = 24.766$ etc.

Comparison	k	g	SSE_C	$SSE_R - SSE_C$	F
2 vs 0	1	0	3.738	24.766	940.82
3 vs 2	3	1	1.472	2.266	107.76
4 vs 3	5	3	1.318	0.154	8.06

Recall that $n = 144$, and

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

Under each H_0,

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

- $F_{0.05}(1, 142) \doteqdot 3.92 < 940.82$
  Therefore Model 0 is **NOT** an adequate simplification of Model 2
- $F_{0.05}(2, 140) \doteqdot 3.07 < 107.76$
  Therefore Model 2 is **NOT** an adequate simplification of Model 3
- $F_{0.05}(2, 138) \doteqdot 3.07 < 8.06$
  Therefore Model 3 is **NOT** an adequate simplification of Model 4

All of the null hypotheses are **rejected**.

Therefore by both forward and backward selection, we select Model 4

$$X_1 + X_2 + X_1.X_2$$

as the most appropriate model.

Note: In this sequence of hypothesis tests, the convention is **not** to correct for multiple testing (we don't know how many tests we are going to do), although a correction could be used.

Example: Potato Damage Data.

The damage to potato plants caused by cold temperatures is to be studied.

In this experimental study, three binary factor predictors were used: we label them A, B and C rather than X_1, X_2, X_3 to recall the link with Factorial Designs in ANOVA. Each factor takes two levels:

	Factor	Levels
A	Potato Variety	0- Variety 1, 1- Variety 2
B	Acclimatization Routine	0- Room Temp, 1- Cold Room
C	Preparation Treatment	0- -4C, 1- -8C

Thus we have a $2 \times 2 \times 2$ three-way factorial design.

However, the design is **unbalanced**; we have different numbers of replicates in each of the 8 factor-level combinations.

This means we cannot use conventional 3-way ANOVA; the lack of balance means that the stated *p*-values will be **wrong** if we perform a standard ANOVA.

Thus we are forced to use the General Linear Model F-test approach.

We begin with the most complex model and do backward selection.

Here the most complex model can be written

$$A + B + C + A.B + A.C + B.C + A.B.C$$

that is,

- all main effects (terms 1,2 and 3)
- all two-way interactions (terms 4,5 and 6)
- all three-way interactions (term 7)

We may write this model

$$A * B * C$$

Counting the numbers of parameters

Term	Parameters	
A	$(a-1)$	1
B	$(b-1)$	1
C	$(c-1)$	1
$A.B$	$(a-1)(b-1)$	1
$A.C$	$(a-1)(c-1)$	1
$B.C$	$(b-1)(c-1)$	1
$A.B.C$	$(a-1)(b-1)(c-1)$	1
Total		7

where $a = b = c = 2$.

We have 7 parameters in total when all terms are considered, so

$$k = 7$$