F-tests / ANOVA in MLR Recap

We can use `sm.stats.anova_lm(model, typ = _)` (typ = 1, 2, or 3) to generate ANOVA tables in MLR.

The typ argument is super important — it determines if we get sequential or partial F-test results.

Typ = 1 ⇒ uses **sequential sum of squares**

reduction in SSE from H_0 to H_1

source	SS	df	F	p
X_1	$SS(X_1)$	1	$\frac{SSX_1/df}{SSE_F/n-p}$	
X_2	$SS(X_2 \mid X_1)$	1	$\frac{SS/1}{SSE_F/n-p}$	.02
X_3	$SS(X_3 \mid X_1, X_2)$	1		
resid	SSE_F	$n-p$		

H_0	H_1
$n-1$	$n-2$
$Y \sim 1$	$Y \sim X_1$
$n-2$	$n-3$
$Y \sim X_1$	$Y \sim X_1 + X_2$
$Y \sim X_1 + X_2$	$Y \sim X_1 + X_2 + X_3$

(*) order matters for typ = 1

For ex, if X_2's row has $p \approx .02 \Rightarrow$

X_2 is a sig. predictor GIVEN that

X_i is included in the model.

$$\textcircled{*} \quad \frac{(SSE_{H_0} - SSE_{H_1}) / (df_{H_0} - df_{H_1})}{SSE_F / df_F}$$

For typ=2 ⇒ uses partial sum of squares

reduction in SSE from H_0 to H_1

source	SS	df	F	p	H_0	H_1
X_1	$SS(X_1 \mid X_2, X_3)$	1	$\frac{SS/df}{SSE_F/n-p}$	1.4	$Y \sim X_2 + X_3$	$Y \sim X_1 + X_2 + X_3$
X_2	$SS(X_2 \mid X_1, X_3)$	1		1.5	$Y \sim X_1 + X_3$	$Y \sim X_1 + X_2 + X_3$
X_3	$SS(X_3 \mid X_1, X_2)$	1		1.0001	$Y \sim X_1 + X_2$	$Y \sim X_1 + X_2 + X_3$
resid	SSE_F	$n-p$				

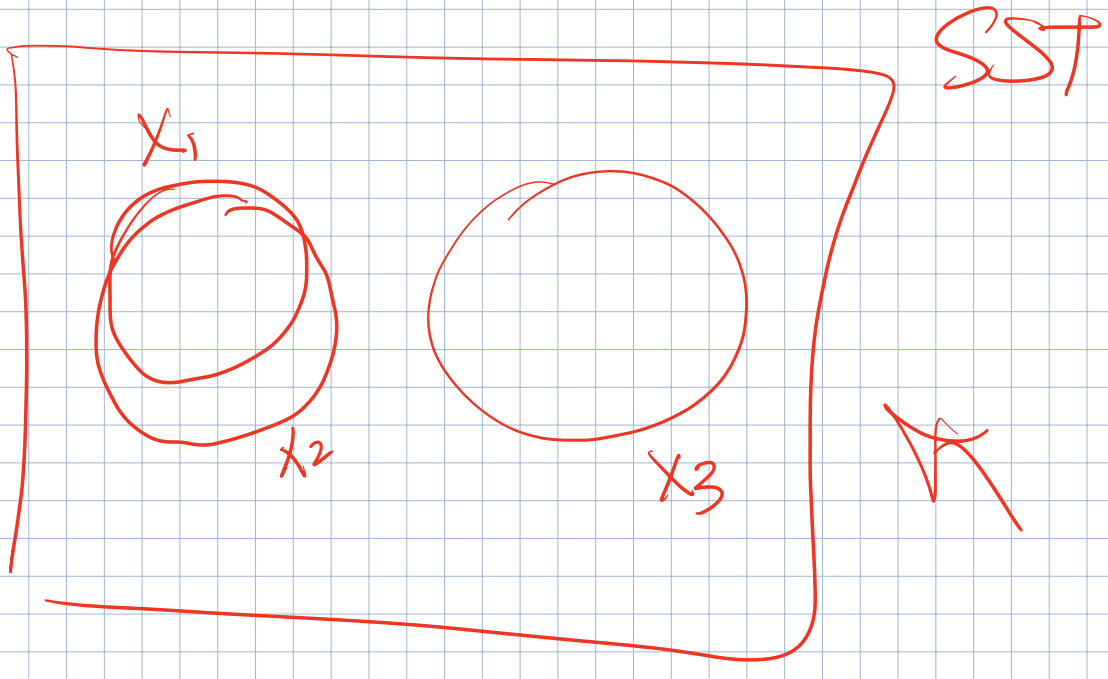
⊛ order doesn't matter for typ=2

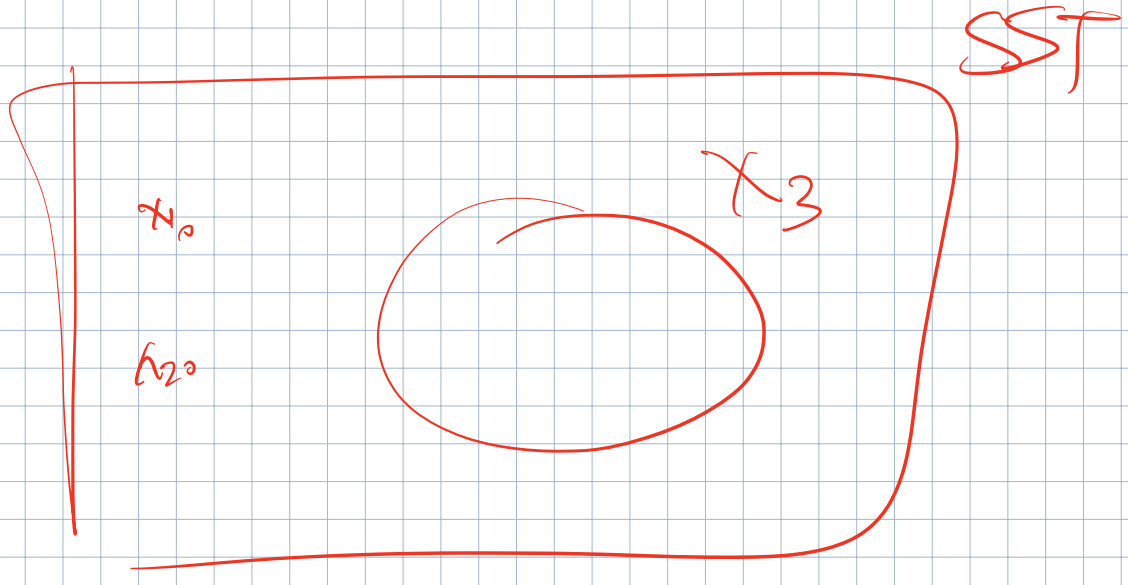
For ex: if $p < .0001$ for X_3's row:

We have evidence that X_3 is a sig. predictor GIVEN all other predictors are in the model.

Note: ① Typ=2 ⇒ order doesn't matter

② typ=2 ⇒ $F = t^2$





Dealing w/ Categorical Predictors

What happens if I have categorical preds:

- GENDER: M, F, NB,
- RACE: ...
- EYE COLOR: ...

The problem is in their current form, this data isn't compatible w/ the OLS solution $\hat{\beta} = (X^T X)^{-1} X^T y$

X need numbers!

Indicators/Dummy Vars

The most common way to "recode" this data is to define the indicators:

Ex: Eye color has 4 levels:
- blue
- green
- hazel
- brown

$$X_i = \begin{cases} X_{i\_blue} \\ X_{i\_green} \\ X_{i\_hazel} \end{cases}$$

$$X_{i\_blue} = \begin{cases} 1 & \text{if } X_i = \text{"blue"} \\ 0 & X_i \neq \text{"blue"} \end{cases}$$

$$X_{i\_green} = \begin{cases} 1 & \text{if } X_i = \text{"green"} \\ 0 & X_i \neq \text{"green"} \end{cases}$$

$$x_{i\_hazel} = \begin{cases} 1 & \text{if } x_i = \text{"hazel"} \\ 0 & x_i \neq \text{"hazel"} \end{cases}$$

	x_i	x_{i_blue}	x_{i_green}	x_{i_hazel}
1	green	0	1	0
2	brown	0	0	0
3	brown	0	0	0
4	brown	0	0	0
5	purple	0	0	0
6.	blue	1	0	0

(11 / 0)

we call this indicator variable:

$$x_{i\_green} = \mathbb{1}(x_i = \text{green}) = \begin{cases} 1 & \text{if } x_i = \text{green} \\ 0 & \text{if } x_i \neq \text{green} \end{cases}$$

EX: Predict Y = dexterity

X_1 = R/L-handed

X_2 = brown / not

$$Y = \beta_0 + \beta_1 \mathbb{1}_{(LH)} + \beta_2 \mathbb{1}_{(\text{not brown})} + \varepsilon_i$$

↳ fit model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \mathbb{1}_{(LH)} + \hat{\beta}_2 \mathbb{1}_{(\text{not brown})}$$

int	6
LH	1
Not Brown	0

Interpreting the coefs:

Baseline:

We predict an avg dexterity score of

$\hat{\beta}_0$ = 6 for right handed brown-eyed people.

We predict an avg. dexterity score of

$\hat{\beta}_0 + \hat{\beta}_1 = 6 + 1 = 7$ for left handed brown eyed individuals.

Example Design Matrix

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \vdots & & & \\ & & & \\ & & & \\ & & & \\ 1 & 0 & 0 & 1 \end{bmatrix}$$