F-tests / ANOVA in MLR Recap

We can use sm.stats.anova_lm(model, 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	H_0	H_1
X_1	$SS(X_1)$	1	$\frac{SS(X_1)/df}{SSE/n-p}$		$Y \sim 1$	$Y \sim X_1$
X_2	$SS(X_2 \mid X_1)$	1		.001	$Y \sim X_1$	$Y \sim X_1 + X_2$
X_3	$SS(X_3 \mid X_1, X_2)$	1			$Y \sim X_1 + X_2$	$Y \sim X_1 + X_2 + X_3$
resid		$n-p$				

★ order matters for typ = 1

Ex. If I look @ X_2's row:

If $p < \alpha$, then I conclude that X_2 is a sig. predictor of Y, GIVEN that X_1 is already in the model.

Notes:

1. SSE in the table is the resid. sum of squares for the full model

2. summation of SS terms $\neq$ SST

$$SS(X_1) + SS(X_2) + SS(X_3) + SSE \neq SST$$

In general, the form of the F stat is

$$F = \frac{(SSE_{H_0} - SSE_{H_1}) / (df_{H_0} - df_{H_1})}{SSE_{H_1} / df_{H_1}}$$

Type = 2 uses partial SS

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	$\frac{SS/df}{SSE/n-p}$	1×10^{-64}	$Y \sim X_2 + X_3$	$Y \sim X_1 + X_2 + X_3$
X_2	$SS(X_2 \mid X_1, X_3)$	1		.45	$Y \sim X_1 + X_3$	$Y \sim X_1 + X_2 + X_3$
X_3	$SS(X_3 \mid X_1, X_2)$	1		.0005	$Y \sim X_1 + X_2$	$Y \sim X_1 + X_2 + X_3$
resid		$n-p$				

For example if u look @ X3's col & see $p < \alpha$, then u conclude that X_3 is a sig. predictor given all other predictors are already in the model.

Warning!

Looking @ the p-values only in the t table is not a valid way to decide 1-by-1 which indiv. predictors make it into the final model.

This idea is the one of "model selection" & we'll get there soon.

The t-table is just useful if you have very specific hyp. that are assoc. w/ each row.

Notes

① Typ=2 ⇒ order doesn't matter.

② $F_{stat} = \dfrac{(SSE_R - SSE_F)/1}{MSE_F} = t^2$ ← t-stat for testing indiv β_j's.

Dealing with Categorical Predictors

So far we have only seen quantitative predictors in $X \Rightarrow$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \text{ is straightforward.}$$

What if we have categorical predictors?

EX: GENDER = M, F, NB, ...

RACE = ...

EYE COLOR = ...

How can I represent these with some numeric form to reasonably convey their meaning.

Indicators / Dummy Variables

The most common way to "recode" cat. predictors is to define the dummy vars:

eye color

$$X_i = (\text{blue}, \text{green}, \text{brown})$$

$\downarrow$

$$X_{i,blue} = \mathbb{1}(X_i = \text{blue}) = \begin{cases} 1 & \text{if } X_i = \text{blue} \\ 0 & \text{if } X_i \neq \text{blue} \end{cases}$$

$$X_{i,green} = \mathbb{1}(X_i = \text{green}) = \begin{cases} 1 & \text{if } X_i = \text{green} \\ 0 & X_i \neq \text{green} \end{cases}$$

Technically we don't need the last dummy:

$X_{i,brown}$ ... b/c a 0 for $X_{i,blue}$ & $X_{i,green}$ implies $X_i = \text{brown}$

In general if you have k cat.
then you will encode (k-1)
indicators / dummy vars.

Ex of what this looks like in practice

Eye color of 5 subjects:

Obs	Color
1	Brown
2	Blue
3	Blue
4	G
5	Brown

$\longrightarrow$

Obs	Color_blue	Color_green
1	0	0
2	1	0
3	1	0
4	0	1
5	0	0

btw: This is technically the model:

$$y_i = \beta_0 + \beta_1 \mathbb{1}_{(x_i = blue)} + \beta_2 \mathbb{1}_{(x_i = green)} + \varepsilon_i$$

After fitting,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathbb{1}_{(x_i = \text{blue})} + \hat{\beta}_2 \mathbb{1}_{(x_i = \text{green})}$$

When x_i = brown, we predict

$\hat{y}_i = \hat{\beta}_0$ for brown-eyed indiv.

When x_i = blue, we predict

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1$ for blue-eyed indiv.

$\therefore$ $\hat{\beta}_1$ measures the change in y when the observation is in the blue eye category, compared to the brown eye baseline.