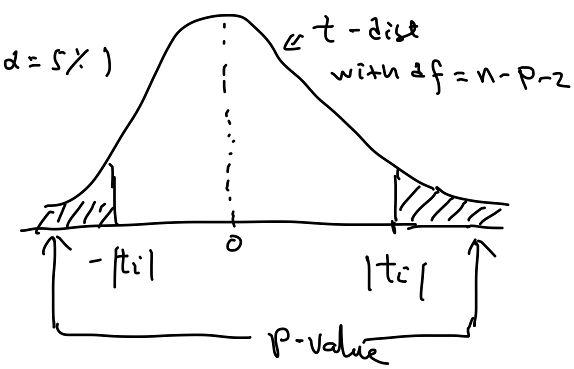
§1. Outliers. (cont'd from lecture 11)

☆ Suppose we have some prior information and aim to test the ith case (pre-specified).

H_0 : ith case is not outlier ; H_A : ith case is outlier.

We can use the p-value calculated from the t-test in lecture 11.

Declare ith case is an
$$\begin{cases} \text{outlier} & \text{if } \ \text{p-value} < \alpha \ (\text{e.g. } \alpha = 5\%) \\ \text{not outlier} & \text{otherwise.} \end{cases}$$



☆ If we don't have prior information and want to find/test an outlier
then we usually choose the critical value (for the p-value)

to be $\frac{\alpha}{n} \times 100\%$ (usually $\alpha = 5\%$)

- This is called Bonferroni Correction (BC), popularly used in Multiple Testing Problems to control the Type I error.

- In this case,

H_0: No outlier in the data, i.e., case 1, ... case n all not outliers

H_A: at least one outlier.

- Our BC test procedure :
(based on the t-test)

For $i = 1, \cdots, n$

Step 1 we compute t-statistics values $t_1 \cdots t_n$ (from lecture 11)

Step 2 from $t_i \overset{H_0}{\sim}$ t-dist with $df = n-p-2$, we obtain their p-values
$i=1,\cdots,n$

$p_1, \cdots, p_n$

Step 3 If $\underbrace{\text{all p-values} \geq \frac{\alpha}{n}}$ $\Longrightarrow$ no outliers

$\Leftrightarrow \min_{i=1,\cdots,n} p_i \geq \frac{\alpha}{n}$

otherwise ($\min_i p_i < \frac{\alpha}{n}$) $\Longrightarrow$ exist outlier.

Moreover, for $i = 1, \cdots, n$, if $p_i < \frac{\alpha}{n}$ $\Rightarrow$ ith ~~case~~ case is an outlier.

- We next show that why the critical value is taken as $\frac{\alpha}{n}$.

We want to control the Type I error, s.t.

$$P(\text{Reject } H_0 \mid H_0 \text{ is true}) \leq \alpha$$

$\uparrow$ $\min_{i=1,\cdots,n} p_i < \alpha^*$ $\leftarrow$ critical value

$\uparrow$ no outlier

and we want to find α^* s.t. the above ~~eq~~ inequality holds.

- BC takes $\alpha^* = \frac{\alpha}{n}$, and this ensures the above inequality.

Why?

$$P(\min_i p_i < \frac{\alpha}{n} \mid H_0)$$

$$= 1 - P(\min_i p_i \geq \frac{\alpha}{n} \mid H_0)$$

$$= 1 - P(p_i \geq \frac{\alpha}{n} \text{ for all } i = 1, \cdots, n \mid H_0)$$

If all p-values are independent

$$= 1 - \prod_{i=1}^{n} \underbrace{P(p_i \geq \frac{\alpha}{n})}_{\shortparallel}$$

Fact: under H_0, the p-value follows a uniform distribution on $[0,1]$ (practice)

$$= 1 - \prod_{i=1}^{n} (1 - \frac{\alpha}{n})$$

If $p_i \sim \text{Unif}[0,1]$

$p(p_i < x) = x$

$x \in [0,1]$

$$= 1 - (1-\frac{\alpha}{n})^n$$

$$\leq 1 - (1 - \frac{\alpha}{n} \cdot n)$$

$$= \alpha .$$

<u>Example</u>: If $n = 65$, $p = 3$. then $t_i \overset{H_0}{\sim}$ t-dist $df = 65-3-2 = 60$

$$p(\ p_i < \alpha = 5\% \ | H_0) \approx 5\%$$

$\underset{\text{approx.}}{\Leftrightarrow}$ $|t_i| > 2$

However, $p(\min_{i=1,\cdots,n} p_i < 5\% \ | H_0) = 96.4\%$

If we use BC. $p(\min_i p_i < \frac{5\%}{65} \ | H_0) \leq 5\%$ ($\frac{\alpha}{n}$)

☆ A more efficient way to compute t_i's.

$$t_i \overset{\text{from lecture 11}}{=\!=} \frac{\hat{e}_i}{\hat{\sigma}_{(i)} \sqrt{1-h_{ii}}} = \underbrace{r_i \left(\frac{n-p-2}{n-p-1-r_i^2} \right)^{\frac{1}{2}}}_{(**)}$$

$\hat{e}_i$ ← from reg with all n obs.

$\hat{\sigma}_{(i)}$ ↑ from reg with all but the ith obs.

where r_i is the <u>standardized residual</u> for the ith case

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

where $\hat{\sigma}$, $\hat{e}_i$ are from regression with <u>all</u> data.

If we use the <u>first equation</u> (above), we need to do $(n+1)$ regressions to get $t_1 \cdots t_n$.
(from lecture 11)

On the other hand, using the <u>second equation</u> $(**)$, we only need to do one regression

§ 2. Influentical cases. (§9. Applied linear Reg).

- A single case or small groups can be strongly influence the fit of a regression model.

  Cases whose removal will cause major changes in the regression analysis are called influentical cases.

- How to detect?

  We will use $\underbrace{\text{Cook's Distance}}_{D}$

- For ith case, its cook's distance is

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^T X)(\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)\,\hat{\sigma}^2}$$

[Recall that $Var(\hat{\beta} \mid X) = \sigma^2 (X^T X)^{-1}$]

where $\hat{\sigma}$, $\underbrace{\hat{\beta}}_{(p+1)\times 1}$ are OLS estimates from the regression with all observations

$\underbrace{\hat{\beta}_{(i)}}_{(p+1)\times 1}$ is the OLS estimates with all observation but the ith obs.

$y_j \sim x_j$ for $j = 1, \dots, i-1, i+1, \dots n$

- In practice, usually ~~1~~ or "approximately 1" is used as a critical value for potential influentical cases.

- A more efficient way to compute D_i is

$$D_i = \frac{1}{p+1} r_i^2 \frac{h_{ii}}{1-h_{ii}}$$

where r_i is the standardised residual and h_{ii} is the leverage