10/05/22

Lecture 8

Recap - ANOVA & Multiple Testing

ANOVA:
- Cell means model:

$$y_{ij} \sim N(\mu_i, \sigma^2) \quad \begin{matrix} i = 1, \dots, k \\ j = 1, \dots, n_i \end{matrix}$$

- we used an F-test

$$F = \frac{MSE_{between}}{MSE_{within}}$$

to test: $H_0: \mu_1 = \dots = \mu_k$ [global testing]

comparisons:

- we can check differences between individual groups

$$t = \frac{\bar{y}_{i\cdot} - \bar{y}_{\ell\cdot}}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_\ell}}} \sim t_{n-k},$$

$$s^2 = \frac{SS_{fit}}{n-k} = \frac{SS_{within}}{n-k} \quad (MSE_{within})$$

- we can also check contrasts:

$$t = \frac{\sum_{i=1}^{k} \lambda_i \bar{y}_{i\cdot}}{s\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2}{n_i}}} \qquad \sum \lambda_i = 0 \quad \sum \lambda_i^2 \neq 0$$

$$t \sim t_{n-k}$$

Muliple Comparisons

- Goal: find which groups are

motivating rejection of global null

- This requires us to test "jointly" a family of null hypotheses $\{H_{0,i}\}_{i=1}^{m}$ (e.g. i is the i-th contrast)

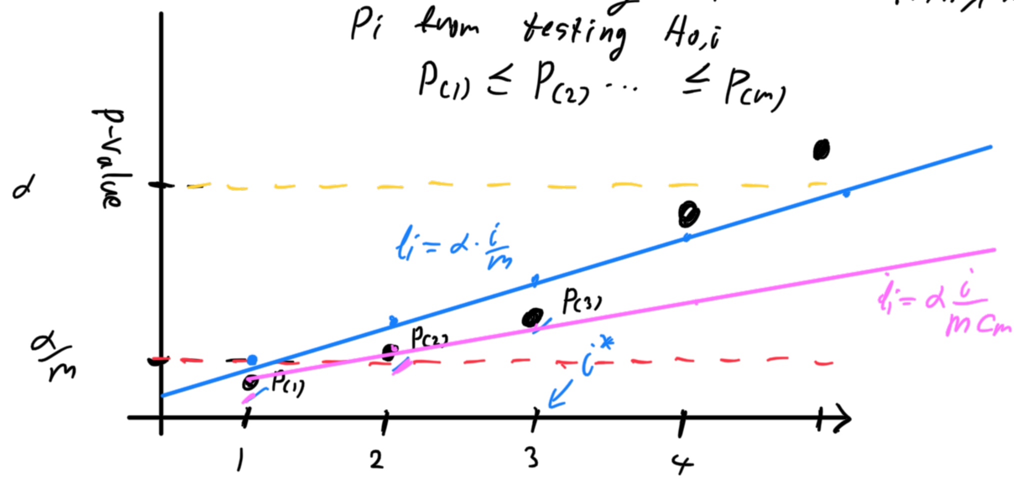
- Bonferroni's union bound: set significance level in each test to be α/m

- Bonferroni's method guarantees

$$\Pr(\text{reject something} \mid \text{all } H_{0,i} \text{ are true}) \leq \alpha$$

- However, this can be "too conservative", meaning you may not reject some non nulls.

We have many p-values $P_1, \dots, P_m$
P_i from testing $H_{0,i}$
$P_{(1)} \leq P_{(2)} \cdots \leq P_{(m)}$



False-Discovery Rate (FDR)

- Suppose we make m hypotheses tests $\{H_{0,i}\}_{i=1}^{m}$. Each has either rejected or not.

We summarize the situation in a table:

	# not rej.	# rej.	Total
$H_{0,i}$ true	U	V	m_0

$H_{0,i}$ false	T	S	m_1
Total	$m-R$	R	m

- Def. false discovery proportion is

$$FDP := \begin{cases} V/R & R>0 \\ 0 & R=0 \end{cases}$$

- Def. false discovery rate is

$$E[FDP]$$

FDR controlling using Benjamini & Hochberg (BH)

- Perform each test; sort the p-values from lowest to highest
  $$p_{(1)} \le p_{(2)} \le \dots p_{(m)}$$
- Define $l_i = \alpha \cdot \frac{i}{m}$ (line with slope $\frac{\alpha}{m}$)
- Define $i^* = \max\{i : p_{(i)} \le l_i\}$
- Reject all $H_{0,i}$ $p_{(i)} \le p_{(i^*)}$

Theorem (BH '95)

If all p-values are independent,

then $FDR \le \frac{m_0}{m}\alpha \le \alpha$

FDR with w/o independence assumption

- If the tests statistics satisfy "positive regression dependency" property (PRD) then BH procedure still controls FDR at level α.

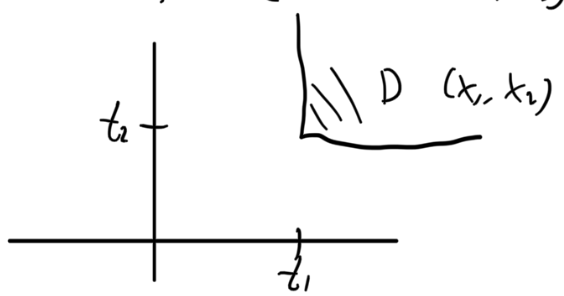
(Benjamini & Yekutieli '01)

PRD: For any increasing set D

$$\Pr(X \in D \mid X_1 = x_1, \ldots, X_n = x_n)$$

is non-decreasing in $x_1, \ldots, x_n$

(X_1, X_2) $\quad\quad D = ((t_1, \infty) \times (t_2, \infty)]$



Under general depency structure:

- we should be more conservative
- we use $l_i = \alpha \frac{i}{m} \cdot \frac{1}{C_m}$

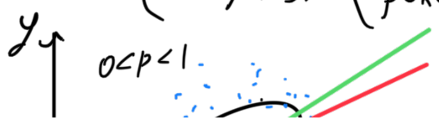
$$C_m = \sum_{i=1}^{m} \frac{1}{i} \approx \ln(m) + \gamma - \frac{1}{2m} \approx \ln m$$

γ: Euler constant ≈ 0.57

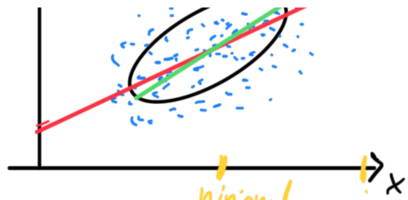
Simple Regression

- $(x_i, y_i) \in \mathbb{R}^2$

- $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right]$



$-1 \le \rho \le 1$



- Regression line:

$$\boxed{E[Y \mid X = x]} = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

- The "45°-degree" line

$$y(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

- Regression to the mean effect

The linear Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$Z = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \hat{\beta} = (Z^T Z)^{-1} Z^T y$$

We have:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\overbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})y_i}^{S_{XY}}}{\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}_{S_{XX}}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof: HW5

Variance of $\hat{\beta}$

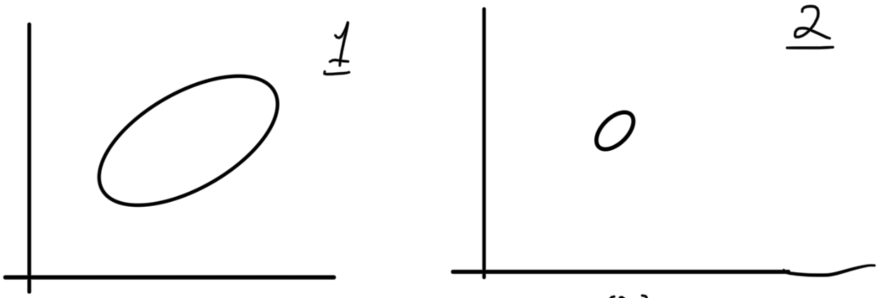
Variance of $\hat{\beta}_1$

- When $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \iff Y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 X_i,$

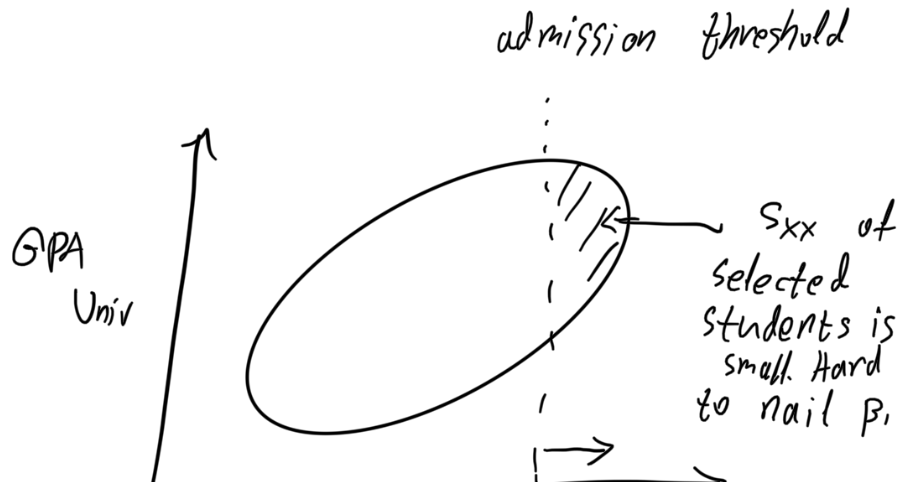
$$\mathrm{var}(\hat{\beta}_1) = \mathrm{Var}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\quad}\right] = \sum^{n} \mathrm{Var}((x_i - \bar{x})$$

$$= \left(\sum_{i=1}^{n} \underbrace{(x_i-\bar{x})^2}_{S_{xx}^2} \right) Var(Y_i) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{n\Big(\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i-}_{S_{xx}}}$$

- What affects $Var(\hat{\beta}_1)$:
  - σ^2 (increases with)
  - n (decreases with)
  - $S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ measures "variability" in x in the data.



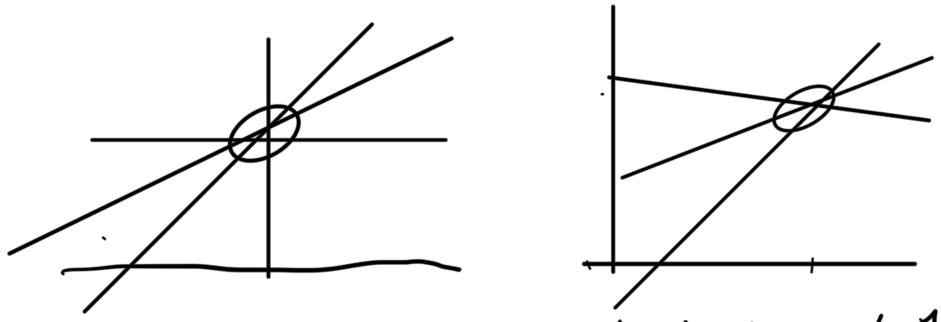
$$S_{xx}^{(1)} > S_{xx}^{(2)}$$



<u>Variance in $\hat{\beta}_0$</u>

$$Var(\hat{\beta}_0) = \ldots = \frac{\sigma^2}{n} \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{S_{xx}}$$

- tends to be large when we are dealing with large values of X that have small S_{XX}



Hard to nail $\hat{\beta}_0$

Variance of $\hat{\beta}_0 + \hat{\beta}_1 x$

- We have $\mathbb{E}[Y \mid X = x] = \hat{\beta}_0 + \hat{\beta}_1 x$
- We have:

$$\operatorname{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) = \dots = \frac{\sigma^2}{n}\left(1 + \left[\sqrt{n}\frac{x - \bar{x}}{\sqrt{S_{XX}}}\right]^2\right)$$

- $\frac{x - \bar{x}}{\sqrt{S_{XX}}}$ measures how many standard deviations x is away of $\bar{x}$.
- The variance is minimal at $x = \bar{x}$
- We have

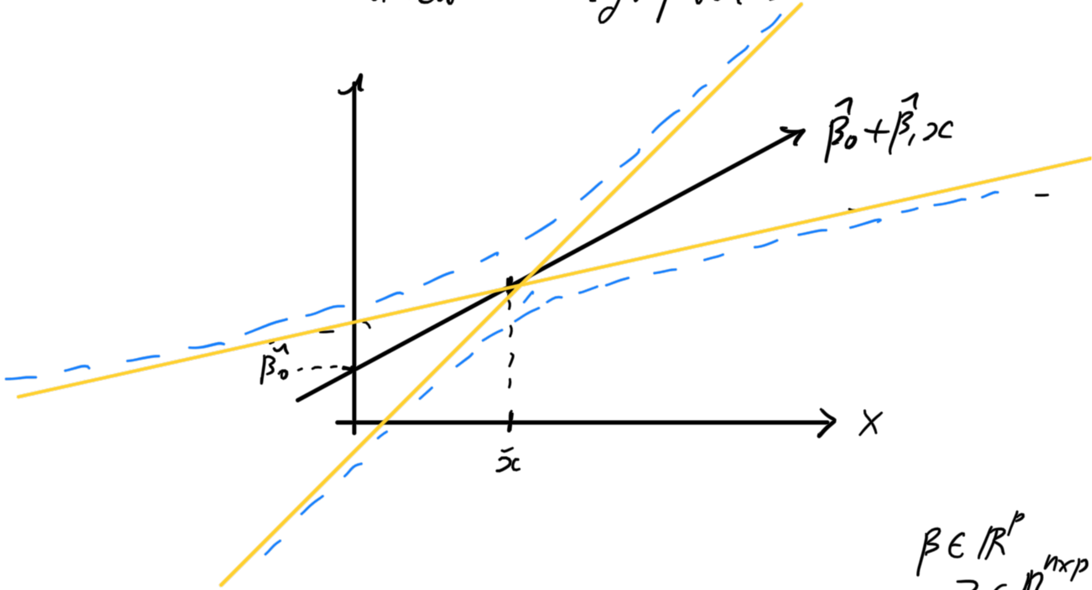
$$\frac{\overbrace{\hat{\beta}_0 + \hat{\beta}_1 x}^{\hat{y}(x)} - (\beta_0 + \beta_1 x)}{s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}}} \sim t_{n-2}$$

- confidence interval for $\beta_0 + \beta_1 \cdot x$:

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2}^{1-\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}}$$

- If we calculate this across all x,

we get a <u>confidence band</u>
(it is a hyperbola with linear asymptotes)



$\beta \in \mathbb{R}^p$
$Z \in \mathbb{R}^{n \times p}$

More generally: for $\underline{Y = Z\beta + \varepsilon}$
and $z_0 \in \mathbb{R}^p$

$$z_0^T \hat{\beta} \pm t_{n-p}^{1-\frac{\alpha}{2}} \cdot s \sqrt{z_0^T (Z^T Z)^{-1} z_0}$$

is a confidence interval for $z_0^T \beta$
(hyperboloid in p dimentions)