

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

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

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

comparisons:

- we can check differences between individual groups

$$t = \frac{\bar{y}_{i.} - \bar{y}_{l.}}{S \sqrt{\frac{1}{n_i} + \frac{1}{n_l}}} \sim t_{n-k}, \quad \begin{matrix} MSE_{\text{within}} \\ S^2 = \frac{SS_{\text{fit}}}{n-k} = \frac{SS_{\text{within}}}{n-k} \end{matrix}$$

- we can also check contrasts:

$$t = \frac{\sum_{i=1}^k \lambda_i \bar{y}_{i.}}{S \sqrt{\sum_{i=1}^k \frac{\lambda_i^2}{n_i}}} \quad \begin{matrix} \sum \lambda_i = 0 & \sum \lambda_i^2 \neq 0 \\ t \sim t_{n-k} \end{matrix}$$

Multiple 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)

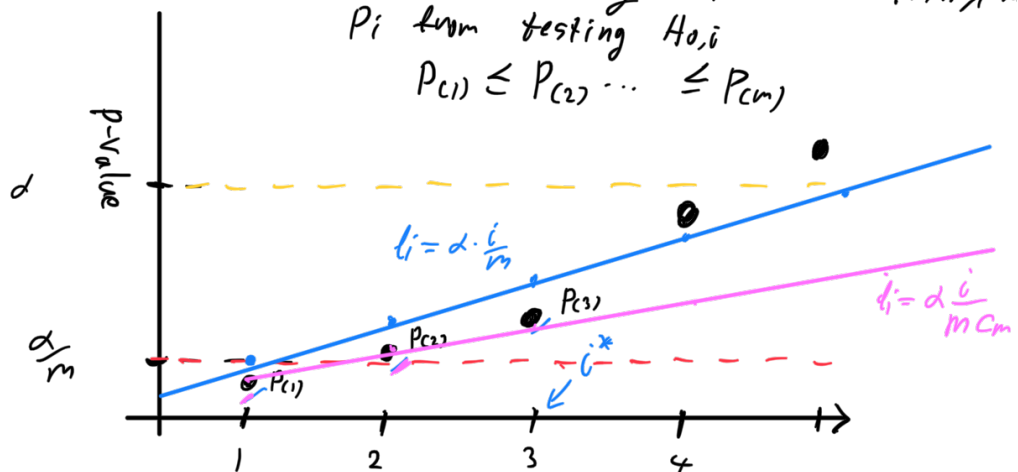
- Bonferroni's union bound:
set significance level in each test to be α/m

- Bonferroni's method guarantees

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

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

We have many p-values p_1, \dots, p_n
 p_i from testing $H_{0,i}$
 $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$



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	$1)$	V	m_0

$H_{0,i} \text{ false}$	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

$$E[FDP]$$

(BH)

FDR controlling using Benjamini & Hochberg

- Perform each test; sort the p-values from lowest to highest
 $P_{(1)} \leq P_{(2)} \leq \dots P_{(m)}$

- Define $l_i = \alpha \cdot \frac{i}{m}$ (line with slope $\frac{\alpha}{m}$)

- Define $i^* = \max \{i : P_{(i)} \leq l_i\}$

- Reject all $H_{0,i}$ $P_{(i)} \leq P_{(i^*)}$

Theorem (BH '95)

If all p-values are independent,

$$\text{then } FDR \leq \frac{m_0}{m} \alpha \leq \alpha$$

FDR with w/o independence assumption

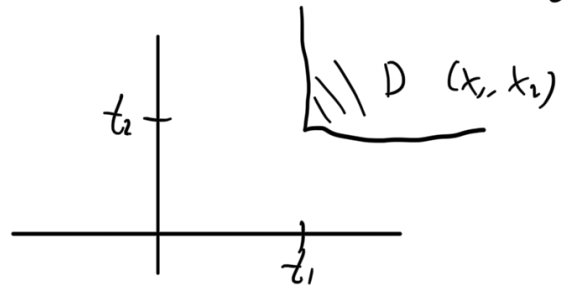
- If the tests statistics satisfy
 "positive regression dependency"
or PRD

properly (FDR) even if procedure still controls FDR at level α .
 (Benjamini & Yekutieli '01)
 PRD: For any increasing set D

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

is non-decreasing in x_1, \dots, x_n

$$(x_1, x_2) \quad D = ([t_1, \infty) \times (t_2, \infty)]$$



Under general dependency structure:

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

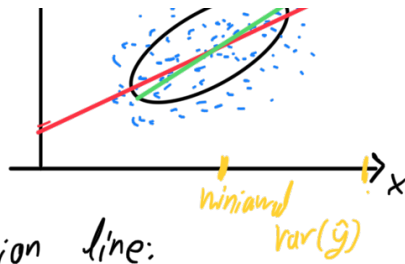
$$c_m = \sum_{i=1}^m \frac{1}{i} \approx \ln(m) + \underbrace{\gamma}_{\text{Euler constant} \approx 0.57} - \frac{1}{2m} \approx \ln m$$

Simple Regression

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

- $\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix} \right)$

$y \uparrow$ $0 < \rho < 1$ $-1 \leq \rho \leq 1$



- Regression line:

$$E(Y | X=x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

- The "45°-degree" line

$$y(x) = \mu_y + \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

- Regression to the mean effect

The linear model:

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

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

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: AHS

Variance of $\hat{\beta}$

Variance of $\hat{\beta}_1$

- When $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2) \Leftrightarrow Y_i \stackrel{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$

$$\text{Var}(\hat{\beta}_1) = \text{Var} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n \text{Var}(x_i - \bar{x}) Y_i$$

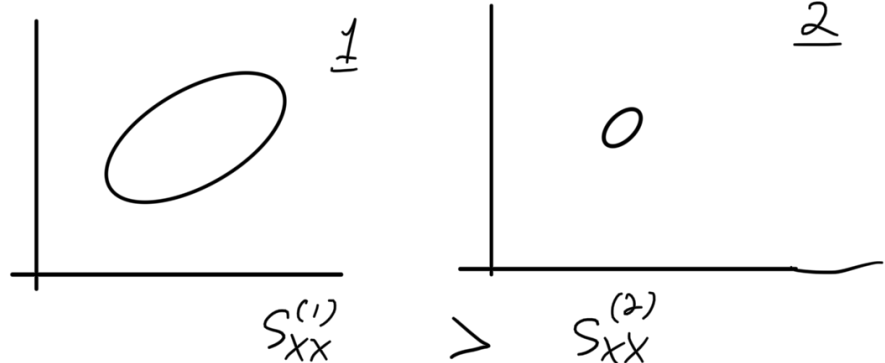
$$s_{xx} = \sum_{i=1}^n \underbrace{(x_i - \bar{x})^2}_{s_{xx}^2} \text{Var}(Y_i) = \frac{\sigma^2}{s_{xx}} = \frac{\sigma^2}{n \underbrace{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)}_{s_{xx}}}$$

- What affects $\text{Var}(\hat{\beta}_1)$:

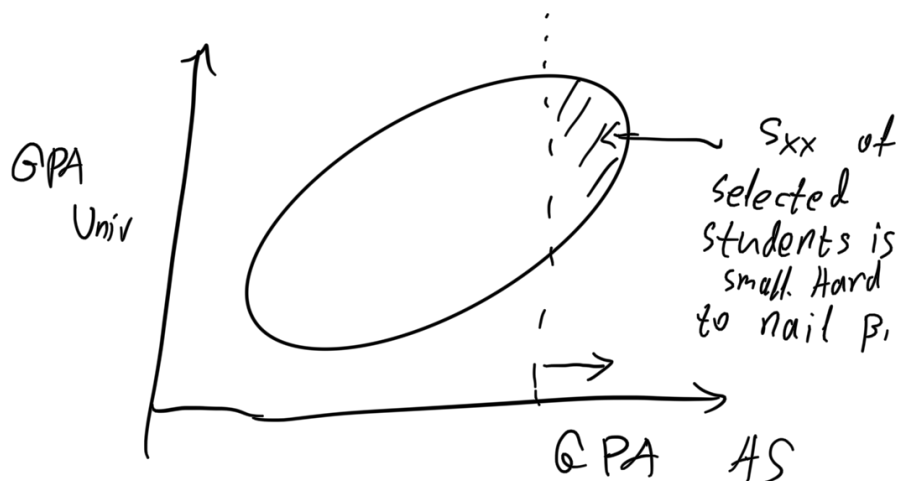
- σ^2 (increases with)

- n (decreases with)

- $s_{xx} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ measures "variability" in x in the data.



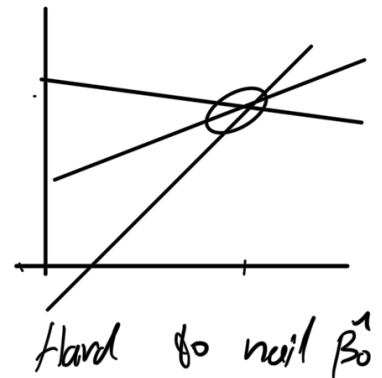
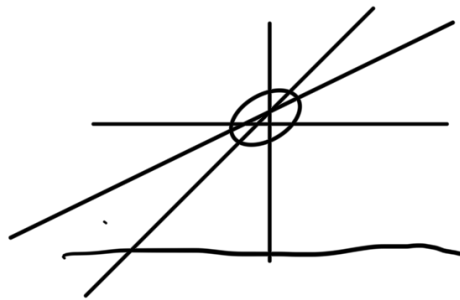
admission threshold



Variance in $\hat{\beta}_0$

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

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



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

- we have $E[Y | X=x] = \hat{\beta}_0 + \hat{\beta}_1 x$

- we have:

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

- $\frac{x - \bar{x}}{\sqrt{S_{xx}}}$ measures how many standard deviations x is away of \bar{x} .

- The variance is minimal at $x = \bar{x}$

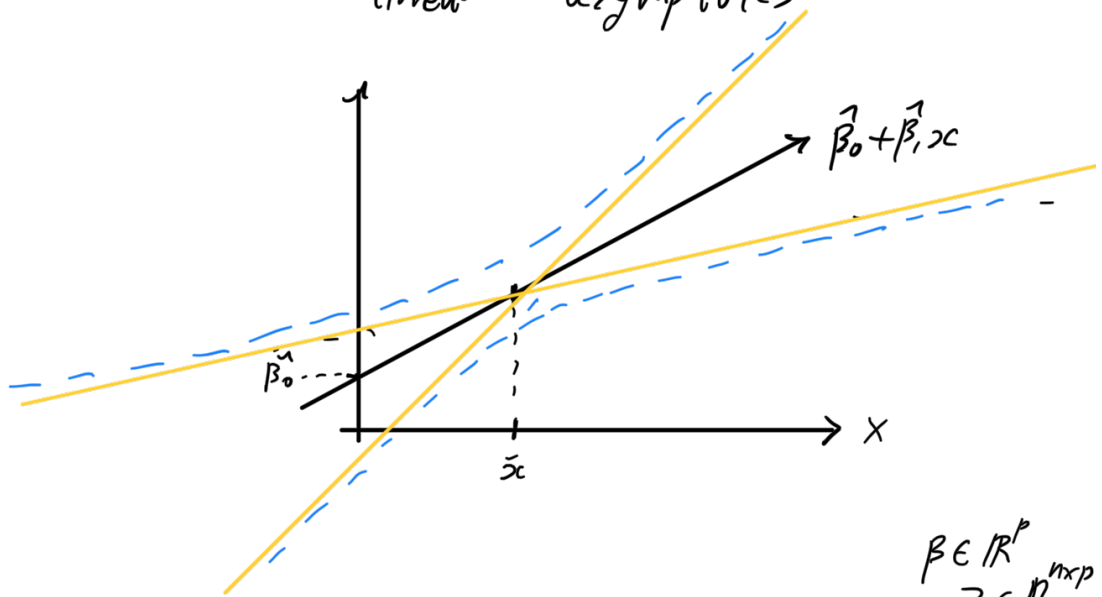
- we have
$$\frac{\hat{y}(x) - (\hat{\beta}_0 + \hat{\beta}_1 x)}{S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

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

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

- If we calculate this across all x ,

we get a confidence band
 (it is a hyperbola with
 linear asymptotes)



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

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

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