STATS305A - Lecture 8

John Duchi
Scribed by Michael Howes

10/14/21

These were updated on 10/15/21 to reflect the new notes on Scheffe's method that John uploaded after class.

Contents

1 Announcements

Etude due tonight. John's null hypothesis for today H_0 : It is intelligent to wear headphones while biking! John is confident that this null hypothesis can be rejected.

2 Multiple Hypothesis Testing

2.1 Motivation

In the linear model $Y = X\beta + \varepsilon$ we are often interested in many parameters. For example if the coordinate j associated with the response Y for $j = 1, \ldots, d$.

Another way to say this is that we wish to test many nulls. For example

$$\begin{aligned} &H_{0,1} : \beta_1 = 0, \beta_0, \beta_2, \ldots, \beta_d \text{ arbitrary.} \\ &H_{0,2} : \ldots \\ &\quad \vdots \\ &H_{0,j} : \beta_j = 0, \beta_{\setminus j} \text{ arbitrary} \\ &\quad \vdots \\ &H_{0,d} : \ldots \end{aligned}$$

The notation $\beta_{\setminus j}$ means all β_i apart from β_j.

Issue/problem If we perform d distinct hypothesis tests, each of which we reject at level α, then we expect as many as $d\alpha$ false rejections.

2.2 Why most published research is false

See: Ioanndis 2005 on course webpage. Setting: $H_{0,1}, H_{0,2}, \ldots, H_{0,d}$ null hypotheses. Rejecting a null means making a discovery. We would like to understand

$$\mathbb{P}(H \text{ is false} | H \text{ is rejected}).$$

Saw we use α-level test with β power. That is

$$\beta = \mathbb{P}(\text{reject } | H \text{ is false}).$$
$$\alpha = \mathbb{P}(\text{reject } | H \text{ is true}).$$

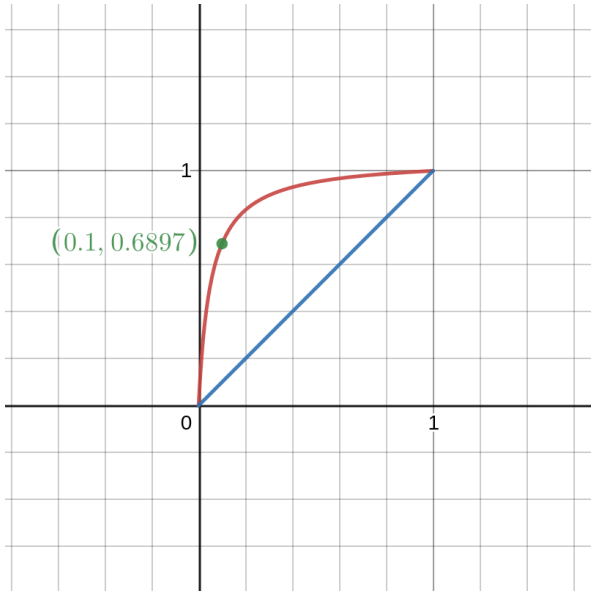
Suppose also there is some value ϕ = frequency of *actually* false hypothesis = frequency of true scientific discoveries. By Bayes rule

$$\begin{aligned}\mathbb{P}(H \text{ is false } | H \text{ rejected}) &= \frac{\mathbb{P}(H \text{ rejected } | H \text{ is false})\mathbb{P}(H \text{ is false})}{\mathbb{P}(H \text{ rejected})} \\ &= \frac{\mathbb{P}(H \text{ rejected } | H \text{ is false})\mathbb{P}(H \text{ is false})}{\mathbb{P}(H \text{ rejected } | H \text{ is false})\mathbb{P}(H \text{ is false}) + \mathbb{P}(H \text{ rejected } | H \text{ is true})\mathbb{P}(H \text{ is true})} \\ &= \frac{\beta\phi}{\beta\phi + \alpha(1-\phi)}\end{aligned}$$

Define $p_\beta(\phi) = \frac{\beta\phi}{\beta\phi+\alpha(1-\phi)}$. We wish to understand this curve as a function of ϕ. Suppose that $\beta = 1$ and so if the null is false we will always correctly reject (if there is a true discovery to be discovered, we will discover it.) Then

$$p(\phi) = \frac{\phi}{\phi + \alpha(1-\phi)}.$$

If $\alpha = 0.05$, the plot of this function looks like this:



If $\phi = 0.1$ so 1/10 of the hypotheses we generate correspond to true discoveries. Then

$$p(0.1) = \frac{0.1}{0.1 + 0.05 \times 0.9} \approx \frac{2}{3}.$$

Thus given that we reject a hypothesis, there is a $\frac{2}{3}$ chance that we have made a true discovery (ie a $\frac{1}{3}$ chance we haven't discovered anything).

Example 1. Consider a genetics experiment such as a GWAS. We may have $d = 10000$ genes and we wish to find the genes that are implicated in a disease. There may only be one such gene. Thus $\phi = 1/10000$. Thus

$$\begin{aligned}
\mathbb{P}(\text{false discovery}) &= \mathbb{P}(H \text{ is false } | H \text{ rejected}) \\
&= 1 - \frac{1}{1 + \alpha \frac{1-\phi}{\phi}} \\
&\approx 1 - \frac{1}{1 + \alpha \cdot 10000} \\
&\approx 1 \quad \text{unless } \alpha \text{ is very small.}
\end{aligned}$$

Even if we do make α very very small, there are problem since multiple people are doing multiple studies and so again the false discovery rate goes up.

Some comments from students: In actual GWAS studies, the first experiement is done to find a gene that might be implicated in a disease. In a second independent experiment that one gene is tested for significance. This brings the number of hypotheses down from 10000 to 1 and helps control false discoveries.

Another comment: Sometimes the genes are grouped into clusters and we then examine which clusters are associated with the disease. This again reduces the number of tests. We will see this idea of grouping inputs later in class.

We will still have the problem of multiple people performing multiple experiments which increases the false discovery rate.

3 Corrections

How can we deal with these problems?

3.1 Bonferroni correction (union bound)

Say we have nulls $H_{0,j}, j = 1, \ldots, k$. Then

$$\begin{aligned}
\mathbb{P}(\text{any } H_{0,j} \text{ falsely rejected}) &\leq \sum_{j=1}^{k} \mathbb{P}(H_{0,j} \text{ falsely rejected}) \\
&\leq \sum_{j=1}^{k} a_j \quad \text{if test } j \text{ is level } \alpha_j.
\end{aligned}$$

Thus one solution is to choose level α_j tests so that $\sum_{j=1}^{k} \alpha_j = \alpha$. Usually when people do Bonferroni corrections they take $\alpha_j = \alpha/k$ for all j. This correction controls the FWER (family wise error rate). With this correction we know

$$\mathbb{P}(\textit{any} \text{ false discovery}) \leq \alpha.$$

3.2 Scheffe's method

If we go back to the linear model we can make more specialised tests/corrections.

Suppose we want to simultaneously test a bunch of constrasts (ie linear functions). For $\Lambda \in \mathbb{R}^{d\times k}$ we want to test

$$\lambda^T\beta = 0,$$

for all $\lambda \in \text{span}(\Lambda)$. This is *a lot* of hypotheses.

Suppose that $\text{rank}(\Lambda) = r$ and that $\text{span}(\Lambda) \subseteq \text{span}(X^T)$ and so Λ is estimable. We have our usual t-test to test $\lambda^T\beta = 0$. Note that

$$\lambda^T\beta = 0 \text{ for all } \lambda \in \text{span}(\Lambda) \iff \Lambda^T\beta = 0.$$

We can test $\Lambda^T\beta = 0$ by the F-test. Find $C \in \mathbb{R}^{n\times k}$ so that $\Lambda = X^TC$ and define $\widehat{\theta} = C^TY$. We know $\mathbb{E}[\widehat{\theta}] = C^TX\beta = \Lambda^T\beta$. The F-test tells us we need to look at the ratio between $\left\|\widehat{\theta}\right\|_2^2$ and $\left\|Y - \widehat{Y}\right\|_2^2$.

How do we actually implement this? We will first look at the distribution of $\Lambda^T\widehat{\beta}$ under the null $H_0 : \Lambda^T\beta = 0$. We will use the fact that $(X^TX)^\dagger = X^\dagger(X^\dagger)^T$. This is true because if $X = U\Sigma V^T$, then

$$X^\dagger(X^\dagger)^T = V\Sigma^\dagger U^TU\Sigma^\dagger V^T = V(\Sigma^2)^\dagger V^T = (V\Sigma^2V^T)^\dagger = (X^TX)^\dagger.$$

We claim that $\Lambda^T\widehat{\beta} \sim \mathsf{N}(0, \sigma^T\Lambda^T(X^TX)^\dagger\Lambda)$. Note that $\beta = X^\dagger Y$ and that $X^TX(X^TX)^\dagger\Lambda = \Lambda$ since Λ is in the range of X^T. and $(X^TX)(X^TX)^\dagger$ is the projection onto range(X^T). It follows that

$$\Lambda^T\widehat{\beta} = \Lambda^TX^\dagger Y = \Lambda^TX^\dagger X\beta + \Lambda^TX^\dagger\varepsilon = \Lambda^T\beta + \Lambda^TX^\dagger\varepsilon = \Lambda^TX^\dagger\varepsilon,$$

under H_0. It follows that $\Lambda^T\widehat{\beta}$ is normal with mean 0 and variance $\sigma^r\Lambda^TX^\dagger(X^\dagger)^T\Lambda = \sigma^2\Lambda^T(X^TX)^\dagger\Lambda$.

Furthermore if we define $H = XX^\dagger =$ projection onto span(X), then

$$(I - H)Y = (I - H)\varepsilon \sim N(0, \sigma^2(I - H))$$

and

$$\text{Cov}((I - H)Y, \Lambda^T\widehat{\beta}) = \mathbb{E}[(I - H)\varepsilon\varepsilon^T(X^\dagger)^T\Lambda] = \sigma^2(I - H)(X^\dagger)^T\Lambda = 0.$$

Thus $(I - H)Y$ and $\Lambda^T\widehat{\beta}$ are independent under the null $H_0 : \Lambda^T\beta = 0$. Thus we have the following:

Proposition 1. *Under the null $H_0 : \Lambda^T\beta = 0$, the statistic*

$$T := \frac{\frac{1}{r}(\Lambda^T\widehat{\beta})^T(\Lambda^T(X^TX)^\dagger\Lambda)^\dagger\Lambda^T\widehat{\beta}}{\frac{1}{n-r(X)}\left\|(I - H)Y\right\|_2^2}, \tag{1}$$

has an $F_{r,n-r(X)}$ distribution where r is the rank of Λ and $r(X)$ is the rank of X.

Note that this result holds even if X and Λ are low rank. The only requirement is that Λ is estimable. That is $\text{span}(\Lambda) \subseteq \text{span}(X^T)$.

Example 2. Suppose we are doing ANOVA with k groups. And so

$$X = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}$$

$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$, $i = 1, \dots, k$, $j = 1, \dots, n_i$ and $X \in \{0,1\}^{N \times k+1}$ where $N = \sum_{i=1}^{k} n_i$. Define

$$\Lambda = [e_2 - e_3, e_2 - e_4, \dots, 3_2 - e_{k+1}, e_3 - e_4, \dots, \dots, e_k - e_{k+1}] \in R^{(k+1) \times \frac{k(k-1)}{2}}.$$

That is Λ is a matrix of differences that tests the hypotheses $\alpha_i - \alpha_{i'} = 0$. Note $\text{rank}(\Lambda) << \frac{k(k-1)}{2}$ and $\text{rank}(X) = k < k+1$, so both are low rank. We also have

$$X^T X = \begin{bmatrix} N & n_1 & n_2 & \dots & n_k \\ n_1 & n_1 & 0 & \dots & 0 \\ n_1 & 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_k & 0 & 0 & \dots & n_k \end{bmatrix} \in \mathbb{R}^{(k+1) \times (k+1)},$$

and

$$\Lambda^T \widehat{\beta} = \begin{bmatrix} \widehat{\alpha}_1 - \widehat{\alpha}_2 \\ \widehat{\alpha}_1 - \widehat{\alpha}_3 \\ \vdots \\ \widehat{\alpha}_{k-1} - \widehat{\alpha}_k \end{bmatrix} \in \mathbb{R}^{k(k-1)/2},$$

and

$$\begin{aligned} \Lambda^T (X^T X)^\dagger \Lambda &= \Lambda^T \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & n_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_k \end{bmatrix}^\dagger \Lambda \\ &= \Lambda^T \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 1/n_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/n_k \end{bmatrix} \Lambda \end{aligned}$$

We are now ready to state Scheffe's method:

Declare $\lambda \in \text{span}(\Lambda)$ significant if

$$\frac{\frac{(\lambda^T \widehat{\beta})^2}{r \lambda^T (X^T X)^\dagger \lambda}}{\frac{1}{r - \text{rank}(X)} \|(I - H)Y\|_2^2} > F \text{ threshold with } \alpha \text{ level, } r, n - \text{rank}(X) \text{ degrees of freedom.} \tag{2}$$

Scheffe's test rejects at level α iff the F test in equation (1) rejects at the level α.

Proof. Consider maximizing

$$\frac{(\lambda^T \widehat{\beta})^2}{\lambda^T X ({}^T X)^\dagger \lambda}$$

over $\lambda \in \text{span}(\Lambda)$. We have assumed $\text{span}(\Lambda) \subseteq \text{span}(X^T)$. Let $X = U \Sigma V^T$ and take $\lambda = \Lambda w$ for some vector w. Define

$$v := (\Lambda^T (X^T X)^\dagger \Lambda)^{1/2} w,$$

and so

$$w = ((\Lambda^T (X^T X)^\dagger \Lambda)^{1/2})^\dagger v = (\Lambda^T (X^T X)^\dagger \Lambda)^{\dagger/2} v.$$

Thus

$$\frac{(w^T \Lambda^T \widehat{\beta})^2}{w^T \Lambda^T (X^T X)^\dagger \Lambda w} = \frac{(v^T (\Lambda^T (X^T X)^\dagger \Lambda)^{\dagger/2} \Lambda^T \widehat{\beta})^2}{\|v\|_2^2}.$$

By Cauchy-Schwarz, we have

$$\begin{aligned}((v^T(\Lambda^T(X^TX)^\dagger\Lambda)^{\dagger/2})\Lambda^T\widehat{\beta})^2 &\le \|v\|_2^2 \left\|(\Lambda^T(X^TX)^\dagger\Lambda)^{\dagger/2})\Lambda^T\widehat{\beta}\right\|_2^2 \\ &= \|v\|_2^2 \left((\Lambda^T(X^TX)^\dagger\Lambda)^{\dagger/2}\Lambda^T\widehat{\beta}\right)^T(\Lambda^T(X^TX)^\dagger\Lambda)^{\dagger/2}\Lambda^T\widehat{\beta} \\ &= \|v\|_2^2 \,(\Lambda^T\widehat{\beta})^T(\Lambda^T(X^TX)^\dagger\Lambda)^\dagger\Lambda^T\widehat{\beta}.\end{aligned}$$

Thus combining all we have done we have

$$\frac{(\lambda^T\widehat{\beta})^2}{\lambda^T X({}^TX)^\dagger\lambda} \le \frac{\|v\|_2^2}{\|v\|_2^2}(\Lambda^T\widehat{\beta})^T(\Lambda^T(X^TX)^\dagger\Lambda)^\dagger\Lambda^T\beta = (\Lambda^T\widehat{\beta})^T(\Lambda^T(X^TX)^\dagger\Lambda)^\dagger\Lambda^T\beta.$$

And we can conclude that the ratio in (2) is always less than or equal to the ratio in (1). Furthermore we have equality when

$$v = (\Lambda^T(X^TX)^\dagger\Lambda)^{\dagger/2}\Lambda^T\widehat{\beta},$$

and so we are done. □

Note: Often Scheffe's test is too conservative and it fails to reject some interesting stuff.

Example 3. Continuing our ANOVA example. In our example span(Λ) is equal to $\lambda = [\lambda_0, \lambda_1, \ldots, \lambda_k]^T$ such that $\lambda_0 = 0$ and $\mathbf{1}^T\lambda = 0$. Thus

$$\lambda^T\widehat{\beta} = \sum_{i=1}^k \lambda_i\widehat{\alpha}_i = \sum_{i=1}^k \lambda_i\bar{Y}_i.$$

Also recall that

$$X^TX = \begin{bmatrix} N & n_1 & n_2 & \ldots & n_k \\ n_1 & n_1 & 0 & \ldots & 0 \\ n_1 & 0 & n_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_k & 0 & 0 & \ldots & n_k \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)},$$

and

$$\lambda^T(X^TX)^\dagger\lambda = \lambda^T \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & 1/n_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/n_k \end{bmatrix} \lambda = \sum_{i=1}^k \frac{\lambda_i^2}{n_i}.$$

Thus our statistic from Scheffe's method is

$$\frac{\frac{(\lambda^T\widehat{\beta})^2}{r\lambda^T(X^TX)^{-1}\lambda}}{\frac{1}{r-\text{rank}(X)}\|(I-H)Y\|_2^2} = \frac{\frac{(\sum_{i=1}^k \lambda_i\bar{Y}_i)^2}{r\sum_{i=1}^k \frac{\lambda_i^2}{n_i}}}{\frac{1}{r-k}\|(I-H)Y\|_2^2}.$$

Exercise find the λ that maximizes the ratio:

$$\frac{(\sum_{i=1}^k \lambda_i\bar{Y}_i)^2}{r\sum_{i=1}^k \frac{\lambda_i^2}{n_i}},$$

subject to $\lambda^T\mathbf{1} = 0$.