STATS305A - Lecture 6

John Duchi
Scribed by Michael Howes

10/07/21

Contents

1 Anouncements **1**

2 Setting for today **1**
2.1 ANOVA Example 1

3 Estimable parameters **2**
3.1 ANOVA example 3
3.2 Construction of estimators 3

4 Psuedo-inverses **3**
4.1 Consequences 4

5 Best linear estimators **5**

1 Anouncements

- HW1 due tomorrow.
- Etude 1 out tomorrow.

2 Setting for today

Sometimes it is interesting ot estimate *functions* of β in the linear model $Y = X^T\beta + \varepsilon$ i.e estimating the parameter $\theta = c^T\beta$ for some $c \in \mathbb{R}^n$. (Sometimes we will have $\theta \in \mathbb{R}^k$ and $\theta = C^T\beta$, $C \in \mathbb{R}^{d\times k}$).

2.1 ANOVA Example

The following example is called one way ANOVA (ANalysis Of VAriance). Assume $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where $\varepsilon_{ij} \overset{\text{iid}}{\sim} (0, \sigma^2)$. Here $i = 1, \ldots, k$ are the treatment classes/treatments/groups, N_i is the number of members in the treatment class i and Y_{ij} are the measurements for $j = 1, \ldots N_i$ for the N_i different members of class i. α_i are the treatment effects/group effects/fixed effects and μ is the full population mean. In the historical setting we would have fields with different treaments and Y_{ij} would be the yield of the j^{th} field in the i^{th} treatment class.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

μ is the mean over all fields, α_i is the effect of treatment i and ε_{ij} is the noise for field i, j. As a linear model we can represent this as

$$\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix} \in \mathbb{R}^{k+1},$$

and

$$X = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix} = [\mathbf{1}, x^{(1)}, \dots, x^{(k)}],$$

where the blocks of 1's in the different columns are disjoint and column $x^{(i)}$ has N_i 1's. We have a problem in that this matrix is low rank since

$$\mathbf{1} = \sum_{i=1}^{k} x^{(i)},$$

This causes problems with trying to estimate β since $X^T X$ is not invertible. What we would like to know is $\alpha_i - \alpha_j = (e_i - e_j)^T \beta$ which is a function of β and represents the difference between treatments.

3 Estimable parameters

Definition 1. A parameter $\theta = c^T \beta$ is (linearly) *estimable* if there exists $a \in \mathbb{R}^n$ such that

$$\mathbb{E}_\beta[a^T Y] = \theta = c^T \beta,$$

for all possible β. $\mathbb{E}_\beta$ means expectation with respect to β.

Proposition 1. *$c \in \mathbb{R}^d$ is estimable if and only if $c \in \text{range}(X^T)$ = row space of X.*

Proof. Suppose c is estimable so for some a, $\mathbb{E}_\beta[a^T X] = c^T \beta$ for β. Thus

$$c^T \beta = \mathbb{E}_\beta[a^T (X\beta + \varepsilon)] = a^T X \beta = (X^T a)^T \beta,$$

For all $\beta \in \mathbb{R}^d$. Thus $c = X^T a$ so $c \in \text{range}(X^T)$.

Conversely if $c = X^T a$, then

$$\mathbb{E}_\beta[a^T Y] = \mathbb{E}_\beta[a^T (X\beta + \varepsilon)] = (X^T a)^T \beta = c^T \beta,$$

thus c is estimable. □

3.1 ANOVA example

Returning to our ANOVA example suppose that $k = 2$ and so $\beta = \begin{bmatrix} \mu \\ \alpha_0 \\ \alpha_1 \end{bmatrix}$ and we are interested in $\theta = c^T\beta = \alpha_1 - \alpha_0$ and so $c = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$. Note that we have

$$X^T = \begin{bmatrix} \mathbf{1}_{N_0}^T & \mathbf{1}_{N_1}^T \\ \mathbf{1}_{N_0}^T & 0 \\ 0 & \mathbf{1}_{N_1}^T \end{bmatrix},$$

where $\mathbf{1}_N$ is the all 1's vector of length N. There are infinitely many ways we can write $c = X^T a$. Two possibilities are

$$a = \begin{bmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad a' = \begin{bmatrix} -\frac{1}{N_0} \\ -\frac{1}{N_0} \\ \vdots \\ -\frac{1}{N_0} \\ \frac{1}{N_1} \\ \frac{1}{N_1} \\ \vdots \\ \frac{1}{N_1} \end{bmatrix},$$

that is a' is $-\frac{1}{N_0}$ for the first N_0 entries and then $\frac{1}{N_1}$ for the remaining N_1 entries.

This leaves us with two questions:

(a) How do we construct unbiased linear estimators?

(b) Can we say that an unbiased linear estimator is "optimal"? How we choose a "best" estimator?

3.2 Construction of estimators

In the simplest case $X \in \mathbb{R}^{n\times d}$ is rank d (full rank). We can then invert $(X^TX)^{-1}$ and as we have seen

$$\widehat{\beta} = (X^TX)^{-1}X^TY,$$

is unbiased for β and $\widehat{\theta} = c^T\widehat{\beta}$ satisfies $\mathbb{E}_\beta[\widehat{\theta}] = c^T\beta$.

More generally, say $c \in \text{range}(X^T)$ then for any choice of $\lambda \in \mathbb{R}^n$ such that $c = X^T\lambda$, then we will have

$$\mathbb{E}[\lambda^TY] = \mathbb{E}[\lambda^T(X\beta + \varepsilon)] = \lambda^TX\beta = (X^T\lambda)^T\beta = c^T\beta.$$

What do we do if there are multiple choices of λ that satisfy $c = X^T\lambda$?

4 Psuedo-inverses

How can we compute things that are close to inverses for non-invertible matrices?

Definition 2. Suppose $D = \text{diag}(d_1, \ldots, d_n)$ is diagonal. Define a diagonal matrix $D^\dagger$ by

$$(D^\dagger)_{ii} = \begin{cases} \frac{1}{d_i} & \text{if } d_i \neq 0, \\ 0 & \text{else.} \end{cases}$$

The matrix $D^\dagger$ is the *psuedo inverse of* D.

Note that $D^\dagger D = DD^\dagger = \text{diag}\left(\begin{cases} 1 & \text{if } d_i \neq 0, \\ 0 & \text{if } d_1 = 0, \end{cases}\right)$. In general,

Definition 3. Let $A = U\Sigma V^T$ be the SVD of A where $\Sigma = \text{diag}(s_1, \ldots, s_r, 0, \ldots, 0)$ $s_r > 0$. Then

$$A^\dagger = V\Sigma^\dagger U^T,$$

is the pseudo inverse of A.

4.1 Consequences

Remark 1. A quick comment on the SVD.

- If $A \in \mathbb{R}^{n\times d}$, then Σ is always square. Σ is $n \times n$ if $n \leq d$ and Σ is $d \times d$ if $d \leq n$.
- If A is square ($n = d$), then Σ, U and V are all also square.
- If A is tall ($n > d$), then U is tall ($n \times d$) and V is square ($d \times d$) and Σ is $d \times d$.
- If A is wide ($n < d$), then U is square ($n \times n$), Σ is square ($n \times n$) and V is tall ($d \times n$) and so V^T is wide ($d \times n$).

With this in mind we can investigate the psuedo inverse of A.

(a) If A is square and full rank, then $A^\dagger = A^{-1}$ since $\Sigma^\dagger = \Sigma^{-1}$ and hence

$$A^\dagger A = V\Sigma^{-1}U^T U\Sigma V^T = VV^T = I,$$

and similarly $AA^\dagger = I$.

(b) If $A \in \mathbb{R}^{n\times d}$ with $n \geq d$ (A is tall) is full rank ($\text{rank}(A) = d$), then

$$A^\dagger = V\Sigma^{-1}U^T \in \mathbb{R}^{d\times n},$$

and

$$A^\dagger A = V\Sigma^{-1}U^T U\Sigma V^T = VV^T = I$$

and

$$AA^\dagger = U\Sigma V^T V\Sigma^{-1}U^T = UU^T.$$

Thus $A^\dagger$ is a left inverse of A and $AA^\dagger$ is the orthogonal projection onto the range of A. Thus one can check that $A^\dagger = (A^TA)^{-1}A^T$ (exercise).

(c) If $A \in \mathbb{R}^{n\times d}$ with $d \geq n$ (A is wide) and $\text{rank}(A) = n$ (full rank), then U is square and V is tall and $A^\dagger = V\Sigma^{-1}U^T = A^T(AA^T)^{-1}$ (another exercise). Further more

$$AA^\dagger = U\Sigma V^T V\Sigma^{-1}U^T = UU^T = I,$$

and,

$$A^\dagger A = V\Sigma^{-1}U^T U\Sigma V^T = VV^T.$$

Thus $A^\dagger$ is a right inverse of A and $A^\dagger A$ is the projection onto the range of A^T.

We have the following intuition, if $A \in \mathbb{R}^{n\times d}$ is tall, then $A^\dagger$ is a left inverse and $AA^\dagger$ gets as close as possible to the identity in the space range(A).

5 Best linear estimators

Let's use psuedo inverses to construct estimators. $X \in \mathbb{R}^{n\times d}$, $X = U\Sigma V^T$, we want $c = X^T\lambda = V\Sigma U^T\lambda$. A natural choice is $\lambda = (X^T)^\dagger c = (X^\dagger)^T = U\Sigma^\dagger V^T c$. If c is estimable, then c is the range of X^T and $(X^T)(X^T)^\dagger$ is the projection onto range(X^T). Thus

$$X^T\lambda = X^T(X^T)^\dagger c = c,$$

since c is already in range(X^T).

Definition 4. A linear estimator $\widehat{\theta} = A^TY$ is *best linear unbiased estimator (BLUE)* for θ if

- $\mathbb{E}[\widehat{\theta}] = \mathbb{E}[A^TY] = \theta$ and
- If B is such that $\mathbb{E}[B^TY] = \theta$, then

$$MSE(A) = \mathbb{E}\left[\left\|A^TY - \theta\right\|_2^2\right] \le \mathbb{E}\left[\left\|B^TY - \theta\right\|_2^2\right] = MSE(B).$$

A comment: we are making two restrictions. We are requiring that our estimator is linear and unbiased. There may be better unbiased estimator that are not linear or there may be better linear estimators that are not unbiased or the "best" estimator may be neither linear nor unbaised. We will talk about biased linear estimators later in the course.

Theorem 1. *Assume that $Y = X\beta + \varepsilon$ where $\varepsilon \sim (0, \sigma^2 I)$ (this is our weakest assumption on ε). If $\theta = C^T\beta \in \mathbb{R}^k$ is estimable, then $A = (X^\dagger)^T C$ is the BLUE. Furthermore for any given B such that $\mathbb{E}[B^TY] = C^T\beta$, then $B_* = \Pi_X B$ is BLUE where $\Pi_X = XX^\dagger$ =projection onto* range(X).

Proof. Let start with a B such that $\mathbb{E}[B^TY] = C^T\beta$. Note that

$$\begin{aligned}
\mathbb{E}[B_*^TY] &= \mathbb{E}[B^T\Pi_X Y] \\
&= \mathbb{E}[B^T(\Pi_X(X\beta + \varepsilon))] \\
&= \mathbb{E}[B^T(X\beta) + B^T\Pi_X\varepsilon] \\
&= B^T(X\beta) \\
&= C^T\beta.
\end{aligned}$$

Thus B_* is unbiased. To see that B_* is BLUE, let A be any other unbiased linear estimator. Thus

$$\mathbb{E}[A^TY] = C^T\beta.$$

Now note

$$\begin{aligned}
MSE(A) &= \mathbb{E}\left[\left\|A^TY - \theta\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|(A^T - B_*)^TY + B_*^TY - \theta\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|(A^T - B_*)^TY\right\|_2^2\right] + \mathbb{E}\left[\left\|B_*^TY - \theta\right\|_2^2\right] + 2\mathbb{E}\left[\left((A^T - B_*)^TY\right)^T\left(B_*^TY - \theta\right)\right] \\
&= \mathbb{E}\left[\left\|(A^T - B_*)^TY\right\|_2^2\right] + MSE(B) + 2\mathbb{E}\left[\left((A^T - B_*)^TY\right)^T\left(B_*^TY - \theta\right)\right] \\
&\ge 0 + MSE(B_*) + 2\mathbb{E}\left[\left((A^T - B_*)^TY\right)^T\left(B_*^TY - \theta\right)\right] \\
&= MSE(B_*) + 2\mathbb{E}\left[\left((A^T - B_*)^TY\right)^T\left(B_*^TY - \theta\right)\right]
\end{aligned}$$

Thus to show that $MSE(A) \geq MSE(B_*)$, it suffices to show that $\mathbb{E}\left[\left((A^T - B_*)^T Y\right)^T \left(B_*^T Y - \theta\right)\right] = 0$. Note firstly that

$$\begin{aligned} B_*^T Y - \theta &= B^T \Pi_X (X\beta + \varepsilon) - \theta \\ &= \theta + B^T \Pi_X \varepsilon - \theta \\ &= B^T \Pi_X \varepsilon, \end{aligned}$$

because B is unbiased. Also

$$(A - B_*)^T Y = (A - B_*)^T (X\beta + \varepsilon) = (A - B_*)^T \varepsilon,$$

since $A^T X\beta = B_*^T X\beta = \theta$ by unbiasedness. Thus we have

$$\mathbb{E}\left[\left((A^T - B_*)^T Y\right)^T \left(B_*^T Y - \theta\right)\right] = \mathbb{E}\left[((A - B_*)\varepsilon)^T B_*^T \varepsilon\right].$$

We will use the indentity $u^T v = \text{tr}(uv^T)$ where $\text{tr}(D)$ is the trace of a matrix D. Thus

$$\begin{aligned} \mathbb{E}\left[((A - B_*)\varepsilon)^T B_*^T \varepsilon\right] &= \text{tr}\left(\mathbb{E}[(A - B_*)\varepsilon\varepsilon^T B_*]\right) \\ &= \sigma^2 \text{tr}\left((A - B_*)^T B_*\right), \end{aligned}$$

since $\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 I$. Also note that $B_* = \Pi_X B_*$ since Π_X is a projection so

$$\Pi_X B_* = \Pi_X^2 B = \Pi_X B = B_*.$$

Also recall that $\Pi_X = XX^\dagger$. Thus

$$\begin{aligned} \sigma^2 \text{tr}\left((A - B_*)^T B_*\right) &= \sigma^2 \text{tr}\left((A - B_*)^T \Pi_X B_*\right) \\ &= \sigma^2 \text{tr}\left((A - B_*)^T X X^\dagger B_*\right) \\ &= \sigma^2 \text{tr}\left((X^T A - X^T B_*)^T X^\dagger B_*\right) \\ &= \sigma^2 \text{tr}\left((C - C)^T X^\dagger B_*\right) \\ &= 0, \end{aligned}$$

where we have once again used that A and B_* are both unbiased and so $X^T A = X^T B_* = C$. Thus we can conclude that

$$MSE(A) \geq MSE(B_*).$$

Thus B_* is BLUE. We now have to show that $A = (X^\dagger)^T C$ is BLUE. Note that A is unbiased since $X^T A = X^T (X^\dagger)^T C = X^T (X^T)^\dagger C$ since $C \in \text{range}(X^T)$. Also

$$\begin{aligned} \Pi_X A &= UU^T (X^\dagger)^T C \\ &= UU^T (V\Sigma^\dagger U^T)^T C \\ &= UU^T U \Sigma^\dagger V^T C \\ &= U\Sigma^\dagger V^T C \\ &= (V\Sigma^\dagger U^T)^T C \\ &= (X^\dagger)^T C. \end{aligned}$$

Thus $A_* = A$ and A is BLUE. □

If X is full rank then we have the following simplification.

Proposition 2. *If X is full rank, then $\widehat{\beta} = (X^T X)^{-1} X^T Y$ is BLUE and $C^T \widehat{\beta}$ is also BLUE.*

Proof. Note that $\mathbb{E}[C^T \widehat{\beta}] = C^T \beta$ so $C^T \widehat{\beta}$ is unbiased. To see that $C^T \widehat{\beta}$ is best unbiased. Note that we know $X^\dagger = (X^T X)^{-1} X^T$ by the exercise. Thus $(X^\dagger)^T = X(X^T X)^{-1}$. If we let $A = (X^\dagger)^T C = X(X^T X)^{-1} C$, then

$$A^T Y = C^T (X^T X)^{-1} X^T Y = C^T \widehat{\beta}.$$

To see that $\widehat{\beta}$ is BLUE, take $C = I$. □