DS 100/200: Principles and Techniques of Data Science **Date: October 23, 2019**

Discussion #9

Name:

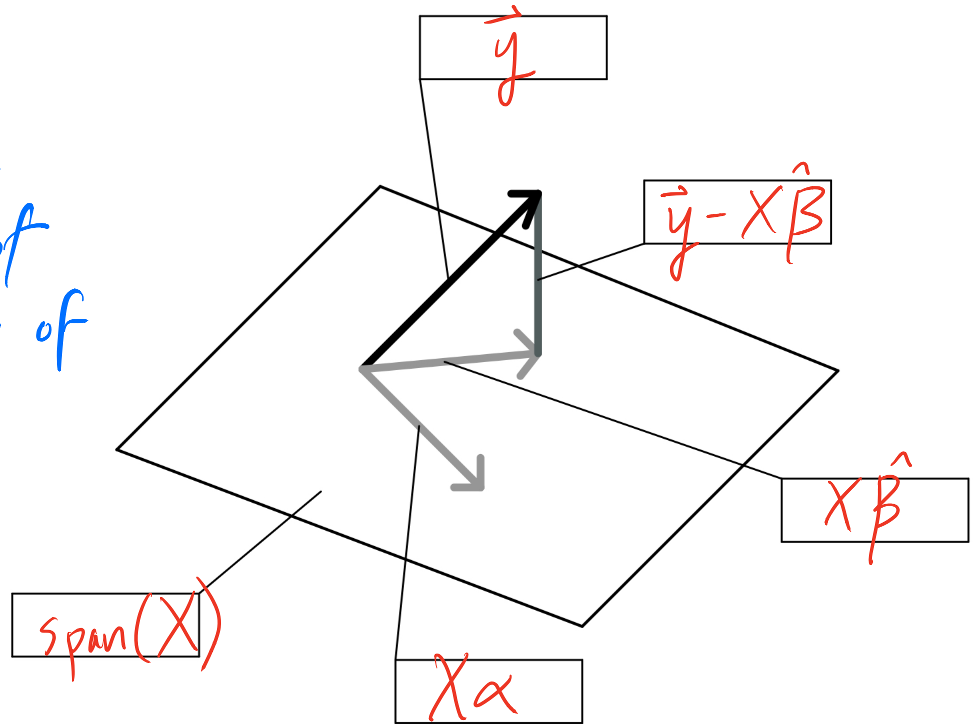
Geometry of Least Squares

1. Consider the following diagram for the geometry of least squares. Fill in the blanks on the diagram with one of the following: (Note that $\hat{\beta}$ is the optimal β, and α is an arbitrary vector.)

- span$\{\mathbb{X}\}$ → set of all linear combinations of columns of X
- $\vec{y}$
- $\mathbb{X}\vec{\alpha}$
- $\mathbb{X}\hat{\beta}$
- $\vec{y} - \mathbb{X}\hat{\beta}$

residual/error

note: Xα is just a linear combination of the columns of X



2. Use the figure above, to explain why, for all $\alpha \in \mathbb{R}^p$,

$$\|\vec{y} - \mathbb{X}\alpha\|_2^2 \geq \|\vec{y} - \mathbb{X}\hat{\beta}\|_2^2$$

$\hat{\beta}$ is optimal : $\|y - X\hat{\beta}\|_2^2$ is as small as possible

3. From the figure above, what can we say about the residuals and the column space of X? Explain your statement using linear algebra ideas.

$$X = \begin{bmatrix} | & | & & | \\ \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_p \\ | & | & & | \end{bmatrix}$$

$\vec{x}_1^T(y - X\hat{\beta}) = 0$

$\vec{x}_2^T(y - X\hat{\beta}) = 0$

$\vdots$

$\boxed{X^T(y - X\hat{\beta}) = 0}$

error orthogonal to span{X}

Recall : $\vec{a}$, $\vec{b}$ orthogonal iff

$a \cdot b = 0$

$a^T b = 0$

4. Derive the normal equations from the fact above. That is, starting from the orthogonality of the residuals and column space of $\mathbb{X}$, derive $\mathbb{X}^t\vec{y} = \mathbb{X}^t\mathbb{X}\vec{\hat{\beta}}$.

$X^T y - X^T X \hat{\beta} = 0$

$X^T y = X^T X \hat{\beta}$

5. What must be be true about $\mathbb{X}$ for the normal equation to be solvable, i.e., to get a solution for $\vec{\hat{\beta}}$? What does this imply about the rank of $\mathbb{X}$ and the features that it represents?

$\Rightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T y$

- $X^T X$ needs to be invertible

$\rightarrow$ $X^T X$ needs to be full rank

$\rightarrow$ $\boxed{X \text{ needs to be full rank}}$

all columns lin. ind.

need $n > p$

rank(X) = rank$(X^T X)$

(can show nullspaces are the same)

Dummy Variables/One-hot Encoding

In order to include a qualitative variable in a model, we convert it into a collection of dummy variables. These dummy variables take on only the values 0 and 1. For example, suppose we have a qualitative variable with 3 levels, call them A, B, and C, respectively. For concreteness, we use a specific example with 10 observations:

$$[A, A, A, A, B, B, B, C, C, C]$$

In linear modeling, we represent this variable with 3 dummy variables, $\vec{x}_A$, $\vec{x}_B$, and $\vec{x}_C$ arranged left to right in the following design matrix. This representation is also called one-hot encoding.

y_1, y_2, y_3, y_4 : class A

y_5, y_6, y_7 : class B

y_8, y_9, y_{10} : class C

$$\begin{array}{c} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \\ \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \end{array} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{10} \end{bmatrix}$$

We will show that the fitted coefficients for $\vec{x}_A$, $\vec{x}_B$, and $\vec{x}_C$ are $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$, the average of the y_i values for each of the groups, respectively.

6. Show that the columns of $\mathbb{X}$ are orthogonal, (i.e., the dot product between any pair of column vectors is 0).

In each row: exactly one 1, rest zero

e.g.

$$X_1 \cdot X_2 = 1\cdot 0 + 1\cdot 0 + 1\cdot 0 + 1\cdot 0 + 0\cdot 1 + 0\cdot 1 + 0\cdot 1 + 0\cdot 0 + \underbrace{0\cdot 0}_{\text{each term includes a } 0} + 0\cdot 0 = 0$$

note: $x_1 \cdot x_1$
$= 1^2+1^2+1^2+1^2+0^2+\dots+0^2$
$= 1 \cdot 4$
$= 4$ = # of points belonging to class $A = n_A$

7. Show that

$$\mathbb{X}^t\mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

Here, n_A, n_B, n_C are the number of observations in each of the three groups defined by the levels of the qualitative variable.

X^TX, in general, is a matrix of dot products

$$X^TX = \begin{bmatrix} x_1 \cdot x_1 & x_1 \cdot x_2 & x_1 \cdot x_3 \\ x_2 \cdot x_1 & x_2 \cdot x_2 & x_2 \cdot x_3 \\ x_3 \cdot x_1 & x_3 \cdot x_2 & x_3 \cdot x_3 \end{bmatrix} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

8. Show that

$$\mathbb{X}^t\vec{y} = \begin{bmatrix} \sum_{i\in A} y_i \\ \sum_{i\in B} y_i \\ \sum_{i\in C} y_i \end{bmatrix}$$

$$X^Ty = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_9 \\ y_{10} \end{bmatrix}$$

$$= \begin{bmatrix} y_1 + y_2 + y_3 + y_4 \\ y_5 + y_6 + y_7 \\ y_8 + y_9 + y_{10} \end{bmatrix} \longrightarrow \begin{bmatrix} \sum_{i\in A} y_i \\ \sum_{i\in B} y_i \\ \sum_{i\in C} y_i \end{bmatrix}$$

9. Use the results from the previous questions to solve the normal equations for $\hat{\beta}$, i.e.,

$$\begin{aligned} \hat{\beta} &= [\mathbb{X}^t\mathbb{X}]^{-1}\mathbb{X}^t\vec{y} \\ &= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} \end{aligned}$$

Inverse of $X^TX = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$ is $(X^TX)^{-1} = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix}$

(multiplying the two yields $I_{3\times 3}$)

Then: $\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix} \begin{bmatrix} \sum_{i\in A} y_i \\ \sum_{i\in B} y_i \\ \sum_{i\in C} y_i \end{bmatrix}$

$= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$

note: $\frac{1}{n_A} \cdot \sum_{i\in A} y_i$ is the mean of all y_i corresponding to class A, i.e. $\bar{y}_A$