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

Discussion #9

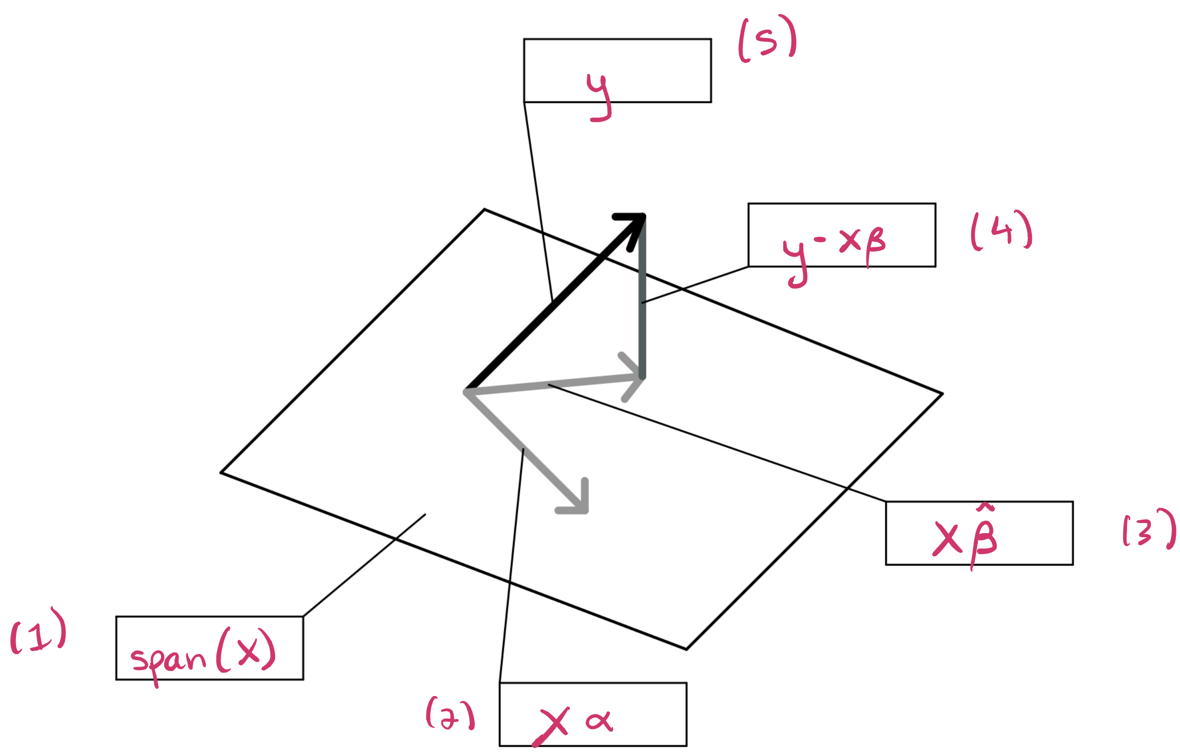
Name: Raguvir Kunani

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$\{\mathbb{X}\}$
- ✓ $\vec{y}$
- ✓ $\mathbb{X}\vec{\alpha}$
- ✓ $\mathbb{X}\hat{\beta}$
- ✓ $\vec{y} - \mathbb{X}\hat{\beta}$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2$$



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

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

By definition, $\hat{\beta}$ is such that $X\beta$ has closest distance to y. this means $\|y - X\beta\|^2$ is minimized when $\beta = \hat{\beta}$. Any other $\beta = \alpha$ when $\alpha \neq \hat{\beta}$ will have a greater or equal value for $\|y - X\beta\|^2$.

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.

residuals orthogonal to column space of X

VERY IMPORTANT!!

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 $\mathbb{X}^T\vec{y} = \mathbb{X}^T\mathbb{X}\vec{\hat{\beta}}$.

See end of worksheet for solution

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?

Normal equation: $X^T X \hat{\beta} = X^T y$

If $(X^T X)^{-1}$ exists, $\hat{\beta} = (X^T X)^{-1} X^T y$

$(X^T X)^{-1}$ exists only if X is full column rank, which also means the features X contains must be linearly independent

See end of worksheet for why $(X^T X)^{-1}$ exists only when X is full column rank

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.

$$\begin{array}{c} \begin{array}{ccc} x_A & x_B & x_C \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

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).

Dot Product of x and y : $x^T y = \sum_i x_i y_i$

For any two vectors we choose from x_A, x_B, x_C, the i^{th} entry of those 2 vectors is never both 1 (by the way we constructed x_A, x_B, and x_C. thus, the product of the i^{th} entries is always 0, so the dot product is always 0.

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.

$$X = \begin{bmatrix} | & | & | \\ x_A & x_B & x_C \\ | & | & | \end{bmatrix} \Rightarrow X^T = \begin{bmatrix} — x_A — \\ — x_B — \\ — x_C — \end{bmatrix} \Rightarrow X^T X = \begin{bmatrix} — x_A — \\ — x_B — \\ — x_C — \end{bmatrix} \begin{bmatrix} | & | & | \\ x_A & x_B & x_C \\ | & | & | \end{bmatrix} = \begin{bmatrix} x_A \cdot x_A & x_A \cdot x_B & x_A \cdot x_C \\ x_B \cdot x_A & x_B \cdot x_B & x_B \cdot x_C \\ x_C \cdot x_A & x_C \cdot x_B & x_C \cdot x_C \end{bmatrix}$$

$$= \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

8. Show that

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

$$X^T = \begin{bmatrix} — x_A — \\ — x_B — \\ — x_C — \end{bmatrix} \begin{bmatrix} | \\ y \\ | \end{bmatrix} = \begin{bmatrix} x_A \cdot y \\ x_B \cdot y \\ x_C \cdot y \end{bmatrix} = \begin{bmatrix} \sum_{i\in A} y_i \\ \sum_{i\in B} y_i \\ \sum_{i\in C} y_i \end{bmatrix}$$

This is because x_A only has 1's where the corresponding data point is an A

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

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

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

(you can look this up to verify)

$$(X^T X)^{-1} X^T y = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix} \begin{bmatrix} \sum_{i\in A} y_i \\ \sum_{i\in B} y_i \\ \sum_{i\in C} y_i \end{bmatrix} = \begin{bmatrix} \frac{1}{n_A} \sum_{i\in A} y_i \\ \frac{1}{n_B} \sum_{i\in B} y_i \\ \frac{1}{n_C} \sum_{i\in C} y_i \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$

(4) Derive $X^TX\hat{\beta} = X^Ty$ using the fact that the residuals are orthogonal to the columns of X.

Note that a matrix vector product Ax can be viewed as the dot product of the rows of A with x:

$$A = \begin{bmatrix} - a_1 - \\ - a_2 - \\ \vdots \\ - a_n - \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 \cdot x \\ a_2 \cdot x \\ \vdots \\ a_n \cdot x \end{bmatrix}$$

The residual is $y - X\hat{\beta}$, and we know this is orthogonal to the columns of X. If $x_1, \ldots, x_p$ are the columns of X, then $x_1 \cdot (y - X\hat{\beta}) = x_2 \cdot (y - X\hat{\beta}) = \ldots = x_p \cdot (y - X\hat{\beta}) = 0$. In matrix notation, this translates to $X^T(y - X\hat{\beta}) = 0$. The X^T results from the fact that matrix vector products contain the dot products of the rows of the matrix with the vector, and in this case we care about the columns.

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

$$X^Ty - X^TX\hat{\beta} = 0$$

$$\boxed{X^Ty = X^TX\hat{\beta}}$$

☆ $(X^TX)^{-1}$ exists only when X is full column rank

Remember that $(X^TX)^{-1}$ exists only if X^TX is full column rank. Thus, all we need to show is that X^TX is full column rank.

First, remember that X is full column rank means X has a trivial (empty) nullspace. Thus, one way of showing that X^TX is full column rank is by showing X^TX has a trivial nullspace.

nullspace(X) is defined by $Xu = 0$.
nullspace(X^TX) is defined by $(X^TX)u = 0$.

But $(X^TX)u = X^T(Xu) = X^T0 = 0$. Thus, if u is in the nullspace of X, it must also be in the nullspace of X^TX. We can also say that $(X^TX)u = 0 \Rightarrow X^T(Xu) = 0 \Rightarrow Xu = 0$. This shows that if u is in the nullspace of X^TX, it must also be in nullspace of X. Thus, X and X^TX have the same nullspace. We already know X has a trivial nullspace, so X^TX has a trivial nullspace. This also means X^TX is full column rank.