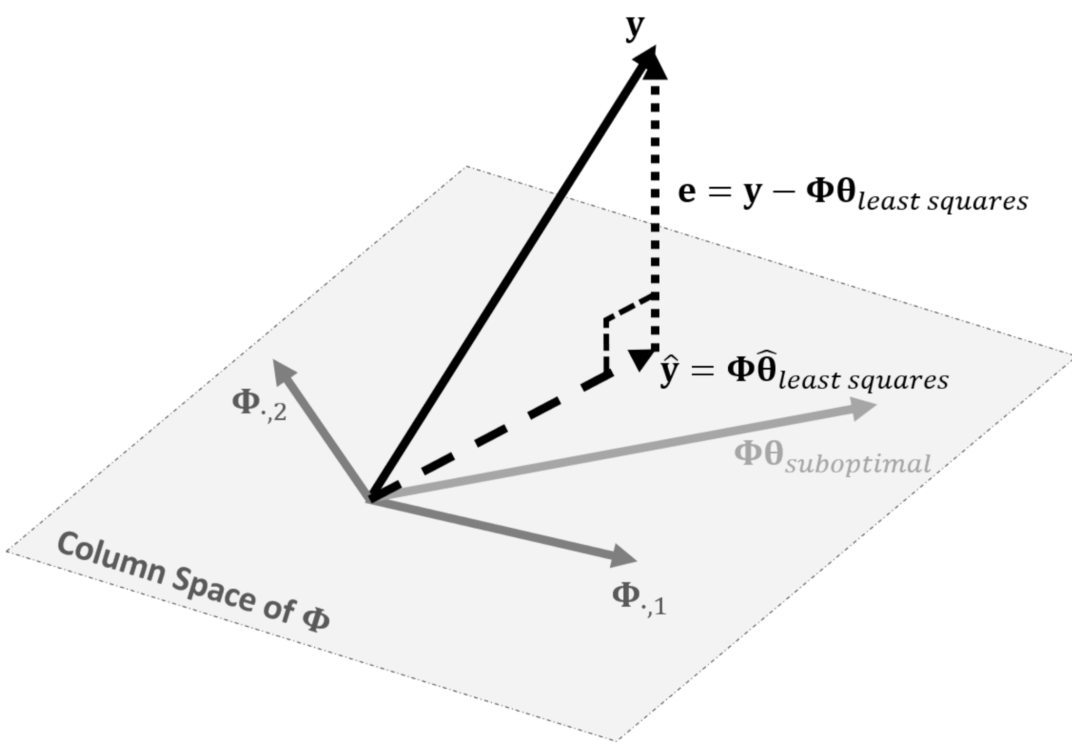
DS 100/200: Principles and Techniques of Data Science **Date: March 15, 2019**

Discussion #7

Name:

Geometry of Least Squares

1. This diagram shows the geometry of 3 observations with 2 features. Φ_1 is the column vector of the three values for feature 1, and Φ_2 is the column vector of values for feature 2. We're fitting a model with parameters θ, a two-element vector, that determines a linear combination of the 2 features. A choice of θ gives fitted values for the 3 observations, and these fitted values are always in the column space of Φ. The observed y, a vector of the response values for the 3 observations, is not in the column space of Φ. The least-squares choice for θ is the one for which $\Phi\theta$ is closest to y. This diagram is analogous to a setting with more observations and more features.



(a) From the image above, what can we say about the residuals and the column space of $\mathbf{\Phi}$? Write this mathematically and prove this statement using a calculus-based argument about minimizing the linear regression loss function.

(b) Show that $\theta = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T Y$. from the fact above for the least squares solution $\mathbf{\Phi}$.

(c) Let $\mathbf{\Phi}$ be a $n \times p$ design matrix with full column rank (the rank is equal to the number of columns). In this question, we will look at properties of matrix $H = \mathbf{\Phi}(\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T$ that appears in linear regression.

i. Recall for a vector space V that a projection $\mathbf{P} : V \to V$ is a linear transformation such that $\mathbf{P}^2 = \mathbf{P}$. Show that $\mathbf{H}$ is a projection matrix.

ii. This is often called the "hat matrix" because it puts a hat on $\mathbf{y}$, the observed responses used to train the linear model. Show that $\mathbf{Hy} = \hat{\mathbf{y}}$

iii. Show that $\mathbf{M} = \mathbf{I} - \mathbf{H}$ is a projection matrix.

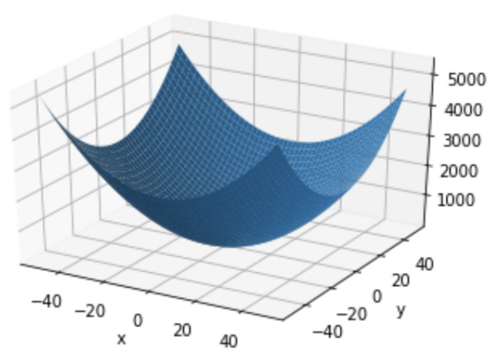
iv. Show that $\mathbf{My}$ results in the residuals of the linear model.

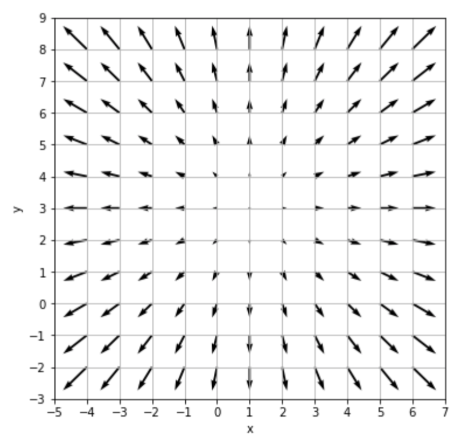
v. Notice that the hat matrix is a function of our observations $\mathbf{\Phi}$ rather than our response variable $\mathbf{y}$. Intuitively, what do the values in our hat matrix represent? It might be helpful to write $\hat{y}_i$ as a summation.

(d) We can show that $rank(\mathbf{\Phi}) = rank(\mathbf{\Phi}^T\mathbf{\Phi})$ by showing that these two matrices have the same null space. List some reasons why $\mathbf{\Phi}$ might not have full column rank, which would make $\mathbf{\Phi}^T\mathbf{\Phi}$ not invertible.

Gradients

2. On the left is a 3D plot of $f(x, y) = (x-1)^2 + (y-3)^2$. On the right is a plot of its gradient field. Note that the arrows show the relative magnitudes of the gradient vector.





(a) Is this function convex? Make a visual argument—it doesn't have to be formal.

(b) Superimpose a contour plot of this function for $f(x, y) = 0, 1, 2, 3, 4, 5$ onto the gradient field.

(c) What do you notice about the relationship between the level curves and the gradient vectors?

(d) From the visualization, what do you think is the minimal value of this function and where does it occur?

(e) Calculate the gradient $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T$.

(f) When $\nabla f = \mathbf{0}$, what are the values of x and y?

3. In this question, we will explore some basic properties of the gradient.

   Note: In this class, we use the following conventions:

   - x represents a scalar
   - X represents a random variable
   - **x** represents a vector
   - **X** represents a matrix or a random vector (context will tell)

(a) Determine the derivative of $f(x) = a_0 + a_1 x$ and gradient of $g(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$.

(b) Suppose $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & ... & x_n \end{bmatrix}^T$, and $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$, where $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$. Determine ∇h.

(c) Determine the gradient of $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$. *(Hint: f is a scalar-valued function. How can you write $\mathbf{x}^T\mathbf{x}$ as a sum of scalars?)*