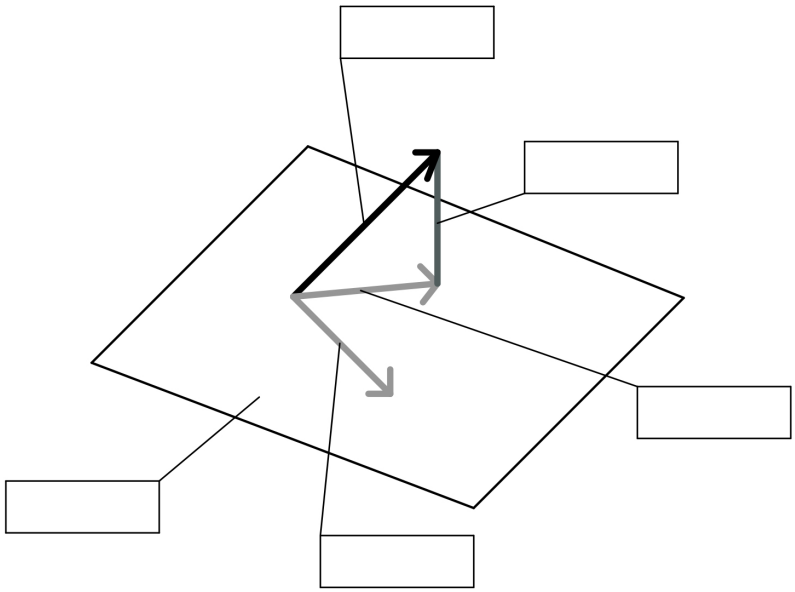
Data 100, Spring 2023

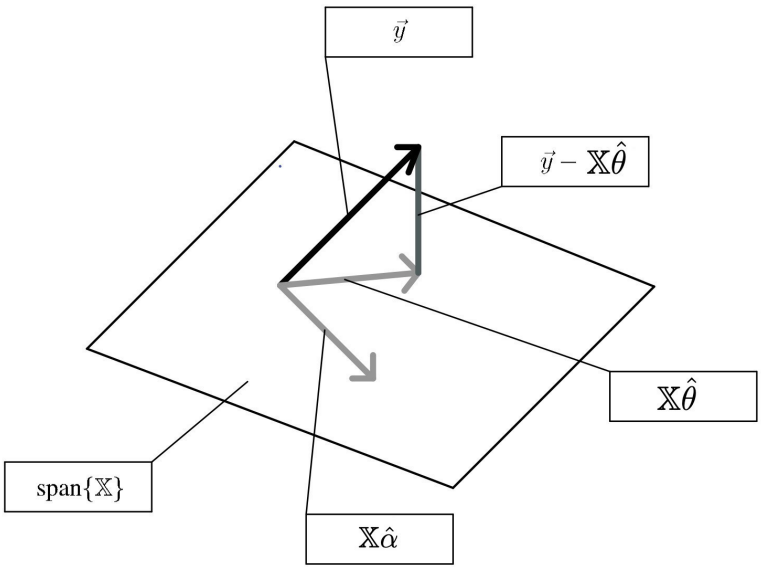
Discussion #6 Solutions

Geometry of Least Squares

1. Suppose we have a dataset represented with the design matrix span($\mathbb{X}$) and response vector $\mathbb{Y}$. We use linear regression to solve for this and obtain optimal weights as $\hat{\theta}$. Label the following terms on the geometric interpretation of ordinary least squares:

    - $\mathbb{X}$ (i.e., span($\mathbb{X}$))
    - The response vector $\mathbb{Y}$
    - The residual vector $\mathbb{Y} - \mathbb{X}\hat{\theta}$
    - The prediction vector $\mathbb{X}\hat{\theta}$ (using optimal parameters)
    - A prediction vector $\mathbb{X}\alpha$ (using an arbitrary vector α).





Solution:

2. Using the geometry of least squares, let's answer a few questions about Ordinary Least Squares (OLS)!

   (a) Which of the following are true about the optimal solution $\hat{\theta}$ to OLS? Recall that the least squares estimate $\hat{\theta}$ solves the normal equation $(\mathbb{X}^T\mathbb{X})\theta = \mathbb{X}^T\mathbb{Y}$.

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

   ☐ A. Using the normal equation, we can derive an optimal solution for simple linear regression with an L_2 loss.

   ☐ B. Using the normal equation, we can derive an optimal solution for simple linear regression with an L_1 loss.

   ☐ C. Using the normal equation, we can derive an optimal solution for a constant model with an L_2 loss.

   ☐ D. Using the normal equation, we can derive an optimal solution for a constant model with an L_1 loss.

   ☐ E. Using the normal equation, we can derive an optimal solution for the model $\hat{y} = \theta_1 x + \theta_2 \sin(x^2)$.

   Solution: We can derive solutions to both simple linear regression and constant model with an L_2 loss since they can be represented in the form $y = x^T\theta$ in some way. Specifically, one of the two entries of x would be 1 for SLR (and the other would be the explanatory variable). The only entry for the constant model would be 1.

We cannot derive solutions for anything with the L_1 loss since the normal equation optimizes for MSE.
Since option B in question 1 is a linear model, we can use the normal equation.

(b) Which of the following conditions are required for the least squares estimate in the previous subpart?

☐ A. $\mathbb{X}$ must be full column rank.
☐ B. $\mathbb{Y}$ must be full column rank.
☐ C. $\mathbb{X}$ must be invertible.
☐ D. $\mathbb{X}^{\mathbb{T}}$ must be invertible.

Solution: $\mathbb{X}$ must be full column rank in order for the normal equation to have a unique solution. Otherwise, if $\mathbb{X}$ is not full column rank, then $\mathbb{X}^T\mathbb{X}$ will not be invertible, and there will be infinite least squares estimates. Note that invertibility is required of $\mathbb{X}^T\mathbb{X}$: neither $\mathbb{X}$ nor $\mathbb{X}^T$ need to be invertible (i.e. a counterexample is $\mathbb{X}^T = [1, 0]$). $\mathbb{Y}$ does not need to be full column rank for a unique least squares estimate.

(c) What is always true about the residuals in the least squares regression? Select all that apply.

☐ A. They are orthogonal to the column space of the design matrix.
☐ B. They represent the errors of the predictions.
☐ C. Their sum is equal to the mean squared error.
☐ D. Their sum is equal to zero.
☐ E. None of the above.

Solution: (A), (B)

(C): (C) is wrong because the mean squared error is the *mean* of the sum of the *squares* of the residuals.

(D): A counter-example is: $\mathbb{X} = \begin{bmatrix} 2 & 3 \\ 1 & 5 \\ 2 & 4 \end{bmatrix}, \mathbb{Y} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$. After solving the least squares problem, the sum of the residuals is -0.0247, which is not equal to zero. However, note that this statement is, in general, true if every feature contains the same constant intercept term.

(E): is wrong since (A) and (B) are correct.

(d) Which are true about the predictions made by OLS? Select all that apply.

☐ A. They are projections of the observations onto the column space of the design matrix.

☐ B. They are linear combinations of the features.

☐ C. They are orthogonal to the residuals.

☐ D. They are orthogonal to the column space of the features.

☐ E. None of the above.

Solution: (A), (B), (C)

(A) is correct because they are linear projections onto the column space. This fact also makes (C) correct, (E) incorrect, and (D) incorrect.

(B) is correct based on the definition of OLS.

(e) We fit a simple linear regression to our data (x_i, y_i) for $i \in \{1, 2, \ldots, n\}$, where n is the number of samples, x_i is the independent variable, and y_i is the dependent variable. Our regression line is of the form $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$. Suppose we plot the relationship between the residuals of the model and the $\hat{y}_i$'s and find that there is a curve. What does this tell us about our model?

☐ A. The relationship between our dependent and independent variables is well represented by a line.

☐ B. The accuracy of the regression line varies with the size of the dependent variable.

☐ C. The variables need to be transformed, or additional independent variables are needed.

Solution:
If we see a curve in our residual plot, then the relationship is not well represented by a line. Either more independent variables are needed, or transformations of the current variables are necessary.

(f) Which are the following is true of the mystery quantity $\vec{v} = (I - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T)\mathbb{Y}$?

☐ A. The vector $\vec{v}$ represents the residuals for any linear model.

☐ B. If the $\mathbb{X}$ matrix contains the $\vec{1}$ vector, then the sum of the elements in vector $\vec{v}$ is 0 (i.e. $\sum_i v_i = 0$).

☐ C. All the column vectors x_i of $\mathbb{X}$ are orthogonal to $\vec{v}$.

☐ D. If $\mathbb{X}$ is of shape n by p, there are p elements in vector $\vec{v}$.

☐ E. For any α, $\mathbb{X}\alpha$ is orthogonal to $\vec{v}$.

Solution: (B), (C), and (E) are correct.

(A) is incorrect because any linear model does not create the residual vector v; only the optimal linear model does.

(D) is incorrect because the vector v is of size n since there are n data points.

The rest are correct by properties of orthogonality as given by the geometry of least squares.

(g) Derive the least squares estimate $\hat{\theta}$ by leveraging the geometry of least squares.

Note: While this isn't a "proof" or "derivation" class (and you certainly will not be asked to derive anything of this sort on an exam), we believe that understanding the geometry of least squares enough to derive the least squares estimate shows great understanding of all the linear regression concepts we want you to know! Additionally, it provides great practice with tricky linear algebra concepts such as rank, span, orthogonality, etc.

Solution: We know that the best estimate of $\mathbb{Y}$ is such that everything that we *cannot* represent or get to (i.e. the residuals) is orthogonal to what we can represent or get to (i.e. the span). Mathematically, we then know that every vector in $\mathbb{X}$ and $\mathbb{Y} - \hat{\mathbb{Y}}$ are orthogonal.

We know that when two vectors u, v are orthogonal, $u^T v = 0$. Using this fact, we can then know that for any (column) vector x_i in $\mathbb{X}$, that $x_i^T(\mathbb{Y} - \hat{\mathbb{Y}}) = 0$.

When we write this out for all our column vectors, we know that this entire quantity is just the zero vector!

$$\begin{bmatrix} x_1^T(\mathbb{Y} - \hat{\mathbb{Y}}) \\ x_2^T(\mathbb{Y} - \hat{\mathbb{Y}}) \\ \cdots \\ x_p^T(\mathbb{Y} - \hat{\mathbb{Y}}) \end{bmatrix} = \mathbb{X}^T(\mathbb{Y} - \hat{\mathbb{Y}}) = 0$$

From here, all that is left is algebra! Recall from our linear model that $\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$.

$$\mathbb{X}^T\mathbb{Y} = \mathbb{X}^T\hat{\mathbb{Y}} = \mathbb{X}^T\mathbb{X}\hat{\theta}$$

We know for a fact that $\mathbb{X}^T\mathbb{X}$ has to be square, but it may or may not be invertible depending on one particular condition (take a look at Question 2 part b). In this case, it is:

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

Driving with a Constant Model

3. Adam is trying to use modeling to drive his car autonomously. To do this, he collects a lot of data where he drives around his neighborhood, and he wants your help to design a model that can drive on his behalf in the future using the outputs of the models you design. We will tackle two aspects of this autonomous car modeling framework: going forward and turning.

   We show some statistics from the collected dataset below using `pd.describe`, which returns the mean, standard deviation, quartiles, minimum, and maximum for the two columns in the dataset: `target_speed` and `degree_turn`.

	target_speed	degree_turn
count	500.000000	500.000000
mean	32.923408	143.721153
std	46.678744	153.641504
min	0.231601	0.000000
25%	12.350025	6.916210
50%	25.820689	45.490086
75%	39.788716	323.197168
max	379.919965	359.430309

   (a) Suppose the model predicts the target speed of the car. Using constant models trained on the speeds of the collected data shown above with L_1 and L_2 loss functions, which of the following is true?

   ◯ A. The model trained with the L_1 loss will always drive slower than the model trained with L_2 loss.

   ◯ B. The model trained with the L_2 loss will always drive slower than the model trained with L_1 loss.

   ◯ C. The model trained with the L_1 loss will sometimes drive slower than the model trained with L_2 loss.

   ◯ D. The model trained with the L_2 loss will sometimes drive slower than the model trained with L_1 loss.

   (b) Finding that the model trained with the L_2 loss drives too slowly, Adam changes the loss function for the constant model where the loss is penalized **more** if the speed is higher.

That way, the model wants to optimize more for the case where we wish to drive faster since the loss is higher, accomplishing his goal. Adam writes this as $L(y, \hat{y}) = y(y - \hat{y})^2$. Find the optimal $\hat{\theta}$ for the constant model using the new empirical risk function $R(\theta)$ below:

$$R(\theta) = \frac{1}{n} \sum_i y_i (y_i - \theta)^2$$

Solution: Take the derivative:

$$\frac{dR}{d\theta} = \frac{1}{n} \sum_i \frac{d}{d\theta} y_i (y_i - \theta)^2 \tag{1}$$

$$= \frac{1}{n} \sum_i -2y_i(y_i - \theta) = -\frac{2}{n} \sum_i y_i^2 - y_i\theta \tag{2}$$

Set the derivative to 0:

$$-\frac{2}{n} \sum_i y_i^2 - y_i\theta = 0 \tag{3}$$

$$\theta \sum_i y_i = \sum_i y_i^2 \tag{4}$$

$$\theta = \frac{\sum_i y_i^2}{\sum_i y_i} \tag{5}$$

(c) Suppose he is working on a model that predicts the degree of turning at a particular time between 0 and 359 degrees using the data in the `degree_turn` column. Explain why a constant model is likely inappropriate in this use case.

Extra: If you've studied some physics, you may recognize the behaviour of our constant model!

Solution: Any constant model will essentially be always turning at an angle and will be unable to turn either direction or go straight (i.e. it'll essentially go in a circle forever).

(d) Suppose we finally expand our modeling framework to use simple linear regression (i.e. $f_\theta(x) = \theta_{w,0} + \theta_{w,1}x$). For our first simple linear regression model, we predict the turn angle (y) using target speed (x). Our optimal parameters are: $\hat{\theta}_{w,1} = 0.019$ and $\hat{\theta}_{w,0} = 143.1$.

However, we realize that we actually want a model that predicts target speed (our new y) using turn angle, our new x (instead of the other way around)! What are our new optimal parameters for this new model? (*Hint: use the information in the table.*)

Solution: Recall that $\hat{\theta}_1 = \frac{r\sigma_{\text{speed}}}{\sigma_{\text{turn}}}$. Currently, we have a quantity $\hat{\theta}_{w,1} = \frac{r\sigma_{\text{turn}}}{\sigma_{\text{speed}}}$. Then, $\hat{\theta}_1 = \hat{\theta}_{w,1}\frac{\sigma^2_{\text{speed}}}{\sigma^2_{\text{turn}}} = 0.019 * \frac{46.678744^2}{153.641504^2} = 0.00175$.

Then, $\hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x} = 32.92 - 0.00175 * 143.72 = 32.67$.

Modeling using Multiple Regression

4. We wish to model exam grades for *Data 100* students. We collect various information about student habits, such as how many hours they studied, how many hours they slept before the exam, and how many lectures they attended, and observe how well they did on the exam. Suppose you collect such information on n students and wish to use a multiple-regression model to predict exam grades.

$$\begin{bmatrix} 1 & study_1 & sleep_1 & lectures_1 \\ 1 & ... & ... & ... \\ 1 & ... & ... & ... \end{bmatrix}$$

   (a) Suppose on our n individuals, we construct our design matrix $\mathbb{X}$, adding an **intercept term**, and use the OLS formula to obtain the following $\hat{\theta}$:

$$\hat{\theta} = \begin{bmatrix} 0.5 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

   Suppose our design matrix $\mathbb{X}$ was constructed such that the first is the bias, the second column represents how many hours each of the n students studied, the third contains how many hours each student slept before the exam, and the fourth represents how many lectures each student attended. With this knowledge, give an interpretation of what each entry of $\hat{\theta}$ means in context.

   Solution:
   Each regression coefficient (a component of the θ vector), except for θ_0, represents the amount we expect an individual's exam score to go up when increasing the corresponding variable by one unit and holding all other variables fixed. For instance, for the first component, when holding 'hours of sleep' and 'lectures attended' constant, an individual's score is expected to go up by 3 per extra hour spent studying. The other components can be interpreted similarly.

   (b) After fitting this model, suppose we have two individuals for which we would like to predict their exam grades using these variables. Suppose Individual 1, slept 10 hours, studied 15 hours, and attended 4 lectures. Suppose also Individual 2, slept 5 hours, studied 20 hours, and attended 10 lectures. Construct a matrix $\mathbb{X}'$ such that, if you computed $\mathbb{X}'\hat{\theta}$, you would obtain a vector of each individual's predicted exam scores.

Solution:

$$\mathbb{X}' = \begin{bmatrix} 1 & 15 & 10 & 4 \\ 1 & 20 & 5 & 10 \end{bmatrix}$$

(c) Denote $\mathbb{Y}'$ as a 2×1 vector that represents the actual exam scores of the individuals we are predicting on. Write out an expression that evaluates to the MSE of our predictions, written as a function of the squared L2-norm of a vector.

Solution:

$$\text{MSE} = \frac{1}{2}\|\mathbb{Y}' - \mathbb{X}'\hat{\theta}\|^2$$