MATH 342W / 650.4 Spring 2024 Homework #2

Loyd Flores.

Monday 26th February, 2024

Problem 1

These are questions about Silver's book, chapter 2, 3. Answer the questions using notation from class (i.e. $t, f, g, h^*, \delta, \epsilon, e, t, z_1, \ldots, z_t, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1\cdot}, \ldots, x_{n\cdot}$, etc).

(a) [harder] If one's goal is to fit a model for a phenomenon y, what is the difference between the approaches of the hedgehog and the fox? Connecting this to the modeling framework should really make you think about what Tetlock's observation means for political and historical phenomena.

Answer:

Philip Tetlock observed that experts' predictions varied, with some performing better than others. He classified these experts into two groups: *Hedgehogs and Foxes*. Hedgehogs tend to approach modeling phenomena with rigidity, similar to hunting one large prey, relying on a single overarching theory to explain phenomena. They often struggle to separate their own biases from their analysis, leading to a fusion of facts and values. In terms of fitting a model, hedgehogs tend to over-fit. In contrast, Foxes possess a multidisciplinary approach, considering many diverse ideas and embracing uncertainty. They recognize the limitations of a single theory and adapt their views based on new information. Tetlock's findings emphasize the importance of open-mindedness and consideration of different perspectives, particularly in fields such as politics, history, and model development. Fox-like thinkers adjust their views based on new information and are better equipped to understand complex issues, while Hedgehog thinkers remain rigid and oversimplify problems, allowing biases to influence their conclusions. Recognizing these different thinking styles helps us strive for a clearer and more truthful understanding of the world as it is, rather than how we wish to perceive it.

(b) [easy] Why did Harry Truman like hedgehogs? Are there a lot of people that think this way?

Answer:

Foxes face challenges in various domains such as television, business, and politics. Their modest predictions regarding complex issues are often misinterpreted as a lack of confidence and conviction. However, this modesty stems from Foxes' acknowledgment of the uncertainties and complexities inherent in such problems. Foxes frustrate

Truman due to their inability to provide straightforward answers. Truman's preference for Hedgehog-like responses, evident in his impatience with his Fox-minded administration members, reflects a desire for simplified solutions. It appears that Truman prioritizes speed over quality, seeking quick, definitive answers, even if they may be incorrect. Foxes, on the other hand, are perceived as too slow and hesitant, as they prioritize thorough analysis over hasty conclusions.

(c) [difficult] Why is it that the more education one acquires, the less accurate one's predictions become?
Answer:

This holds particularly true for individuals with a Hedgehog-like mindset. When presented with more facts, there arises a greater potential for these facts to be twisted and manipulated to align with existing biases. This merging of facts and values becomes a significant concern. However, if one can effectively separate biases from the analysis, additional information can indeed prove beneficial in comprehending the intricacies of data relationships.

(d) [easy] Why are probabilistic classifiers (i.e. algorithms that output functions that return probabilities) better than vanilla classifiers (i.e. algorithms that only return the class label)? We will move in this direction in class soon.
Answer:

In my view, *probabilistic classifiers* outperform *vanilla classifiers* because they don't just predict the best possible class, which can only result in a binary outcome of right or wrong. Instead, probabilistic classifiers predict the probabilities associated with each potential class. This is particularly advantageous when uncertainty is significant because the model not only makes predictions but also provides information about its confidence level in those predictions. This feature empowers engineers with actionable insights, enabling them to make necessary adjustments to the model or take further steps in data exploration.

(e) [easy] What algorithm that we studied in class is PECOTA most similar to?
Answer:

PECOTA appears quite similar to the NN or KNN (Nearest Neighbor or K Nearest Neighbor) algorithm. PECOTA employs a concept known as *similarity scores*, designed to evaluate the statistical similarity between the career statistics of any two major-league baseball players. This concept mirrors how NN or KNN functions in theory. Instead of similarity scores, NN/KNN uses a distance function to determine the closest data point or class, which is then used to make a prediction. Similarly, PECOTA groups together players with the closest similarity scores and predicts a player's career trajectory based on the players they are associated with.

(f) [easy] Is baseball performance as a function of age a linear model? Discuss.

Answer:

In a way, you could argue yes, but ultimately, it is not. If you frame the problem as predicting whether someone is good with a yes/no outcome and use age as the main feature, you would need to determine that age and set it as a threshold for the model to predict player performance solely based on age. This approach might make the data appear linearly separable, but it would be an oversimplification. There are numerous factors influencing a baseball player's performance, such as the current environment, injuries, and mental maturity. With such complexity in the data, a linear model would struggle to capture the relationships and would oversimplify the issue at hand.

(g) [harder] How can baseball scouts do better than a prediction system like PECOTA?

Answer:

Both scouts and PECOTA have the ability to analyze statistical information, such as batting average. A model like PECOTA may be more efficient and accurate than a scout in terms of statistical computation and the elimination of existing biases. However, scouts have an advantage in accessing additional information that models do not have access to. Scouts can gather more information over time, such as the actual speed of pitches and hits in real-time, rather than solely relying on available statistical data. This multi-source information enables scouts to make more informed decisions compared to models, which can only predict based on available data.

(h) [harder] Why hasn't anyone (at the time of the writing of Silver's book) taken advantage of Pitch f/x data to predict future success?

Answer:

To my understanding, at the time of the publication of Silver's book, Pitch f/x was relatively new to the game, introducing many new measurable metrics such as the horizontal and vertical movement of a pitched ball. These metrics could be utilized as features to enhance our models' ability to predict a target 'y'. However, at that time, these metrics were more applicable in the scouting realm, as people were still understanding the relationships they would create and their correlation to the target 'y'. Once people figure out how to effectively utilize this data, it could be applied in the realm of forecasting. However, simply inserting this newfound data into our models might produce garbage or overfit our model, as it would introduce unnecessary noise. To conclude like stated in Silver's introduction it takes time for people to utilize newfound data, like the invention of the printing press it took another 300 years for its effects to show. In time people will learn how to utilize the new data and produce more accurate and descriptive models.

Problem 2

These are questions about the SVM.

(a) [easy] State the hypothesis set $\mathcal{H}$ inputted into the support vector machine algorithm. Is it different than the $\mathcal{H}$ used for $\mathcal{A}$ = perceptron learning algorithm?

The objectives of the Perceptron Learning Algorithm and Support Vector Machines are quite similar; both aim to draw a line that linearly separates our data. However, their approaches differ. Support vectors act as balloons that can be enlarged, guiding and enhancing flexibility in finding the optimal hyperplane by maximizing the margin, or the space between the hyperplane and the data points. In contrast, the Perceptron focuses solely on adjusting weights to achieve the best line that separates the classes without margin optimization. Despite this distinction, both algorithms share an H consisting of all linear functions. However, SVM's H comprises a set of linear functions representing hyperplanes supported by support vectors, while the Perceptron's H consists of linear functions that determine the best line by adjusting weights.

(b) [E.C.] Prove the max-margin linearly separable SVM converges. State all assumptions. Write it on a separate page.

Assumptions:
1. D is linearly separable, therefore there exists a hyperplane that separates the data points
2. There exists an optimization software / formula
3. Iteratively our optimization formula will continuously adjust the parameters (weights and intercept) to find the best hyperplane
4. Since data is linearly separable SVM will always converge to find an h.

(c) [difficult] Let $\mathcal{Y} = \{-1, 1\}$. Rederive the cost function whose minimization yields the SVM line in the linearly separable case.

(d) [easy] Given your answer to (c) rederive the cost function using the "soft margin" i.e. the hinge loss plus the term with the hyperparameter λ. This is marked easy since there is just one change from the expression given in class.

Problem 3

These are questions are about the k nearest neighbors (KNN) algorithm.

(a) [easy] Describe how the algorithm works. Is k a "hyperparameter"?
Answer:
The K nearest neighbor algorithm works by identifying the closest data points and classifying them together. When a model employs KNN, the prediction is determined by the class associated with the nearest data point. Different distance functions, such

as Manhattan and Euclidean distance, are used to calculate the distance. K serves as a hyperparameter of KNN, representing the number of neighbors the model will consider when grouping points together. The choice of K significantly impacts the model's performance, and it is the engineer's responsibility to determine an appropriate value. Maximizing K may result in overfitting, as everyone becomes a neighbor, while selecting a K that is too low may lead to inaccurate assumptions.

(b) [difficult] [MA] Assuming $\mathcal{A}$ = KNN, describe the input $\mathcal{H}$ as best as you can.

Answer:

For example we have a simple 2-Dimensional Dataset and each data point will either only red or blue. H in this case will be composed of functions that all possible classifications based on the nearest neighbor provided by the Distance Function selected. In summary H encompases all possible classifications determined by class of its neighbor (red or blue), that changes depending on the hyper-parameter K selected.

(c) [easy] When predicting on $\mathbb{D}$ with $k = 1$, why should there be zero error? Is this a good estimate of future error when new data comes in? (Error in the future is called *generalization error* and we will be discussing this later in the semester).

Answer:

When the hyperparameter k is set to 1, it means we're considering only one neighbor. The closest neighbor will then be the data point itself. In essence, you're grouping the data points by themselves. You obtain a 0 error because the model predicts each data point will be grouped with itself as it's the closest one. This is not a good estimate because it's not entirely true that you obtained a model that can't make mistakes. It will perform poorly on unseen data and will often misclassify data points into incorrect classes.

Problem 4

These are questions about the linear model with $p = 1$.

(a) [easy] What does $\mathbb{D}$ look like in the linear model with $p = 1$? What is $\mathcal{X}$? What is $\mathcal{Y}$?

Answer:

When 'p' = 1 in linear models it means that there is only one feature being used to explain the response or the target variable. This then assumes the relationship between feature and target is linear. The model then assumes that there is a straigh-line relationship and the goal is to find the best fitting line that describes their relationship. For example in a regression problem the number of rooms may determine house price or in a classification problem that line may separate data by class.

(b) [easy] Consider the line fit using the ordinary least squares (OLS) algorithm. Prove that the point $< \bar{x}, \bar{y} >$ is on this line. Use the formulas we derived in class.

Formula for Linear Regression Line:

$$Y = b_0 + b_1 X$$

$$b_0 (y\text{-intercept}) = \bar{y} - b_1 \bar{x}$$

$$b_1 (\text{slope}) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

If we substitute $\bar{x}$ for X

$$Y = b_0 + b_1 \bar{x}$$

↳ substitute

$$Y = (\bar{y} - b_1 \bar{x}) + b_1 \bar{x}$$

$$\boxed{Y = \bar{y}}$$

Therefore when $X = \bar{x}$ & $y = \bar{y}$ they both will lie on the OLS regression line.

(c) [harder] Consider the line fit using OLS. Prove that the average prediction $\hat{y}_i := g(x_i)$ for $x_i \in \mathbb{D}$ is $\bar{y}$.

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

$$\hat{b}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

To show the average of the predicted values $\hat{y}_i$ over all $x_i \in D$ is $\bar{y}$

$$\bar{\hat{y}} = \frac{1}{n} \sum_i^n \hat{y}_i \quad \text{convert}$$

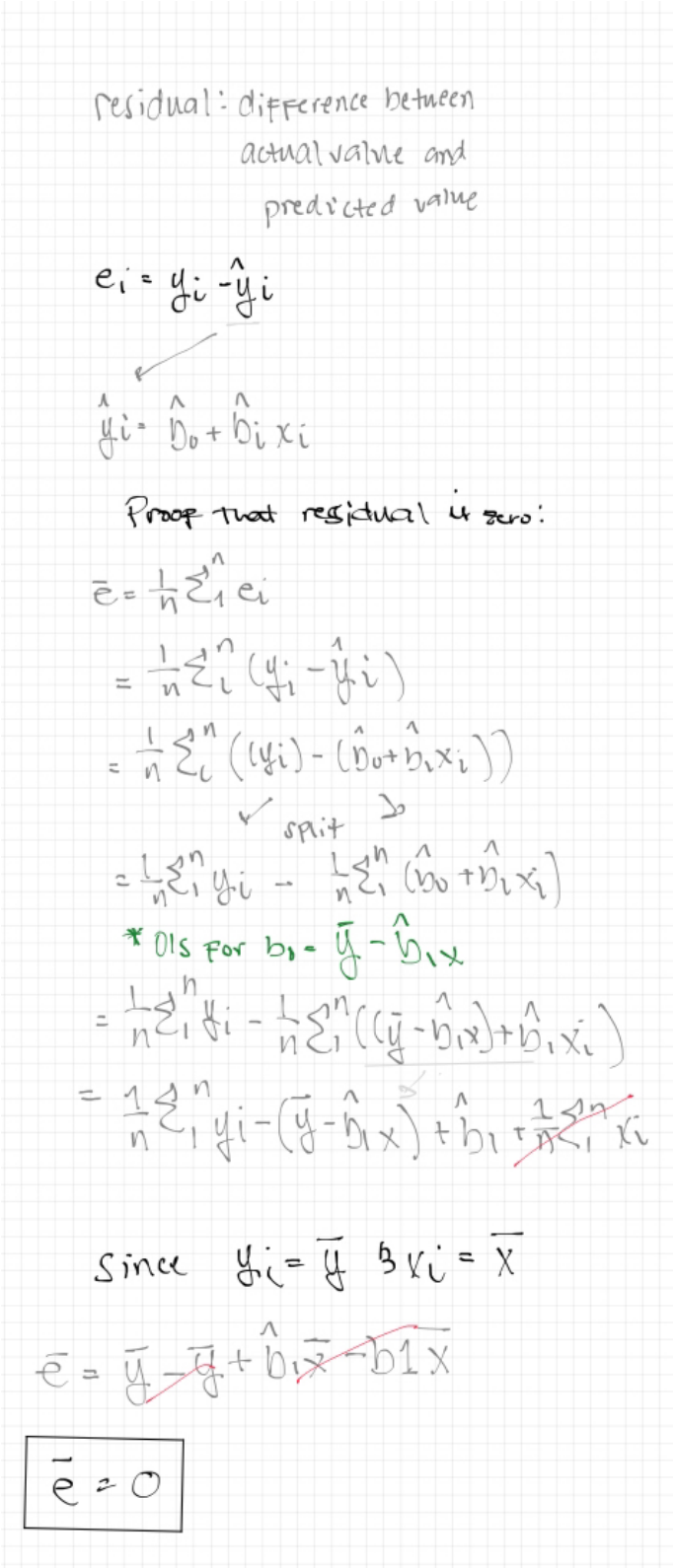
$$\bar{\hat{y}} = \frac{1}{n} \sum_i^n \left(\hat{b}_0 + \hat{b}_1 x_i \right)$$

$$\bar{\hat{y}} = \frac{1}{n} \sum_i^n \left(\bar{y} - b_1 \bar{x} + b_1 x_i \right)$$

$$\bar{\hat{y}} = \bar{y} - b_1 \bar{x} + b_1 \frac{1}{n} \sum_i^n x_i$$

$= \bar{\hat{y}} = \bar{y}$, which proves that the average predicted values of $\hat{y}_i$ is equal to the mean of $\bar{y}$

(d) [harder] Consider the line fit using OLS. Prove that the average residual e_i is 0 over $\mathbb{D}$.

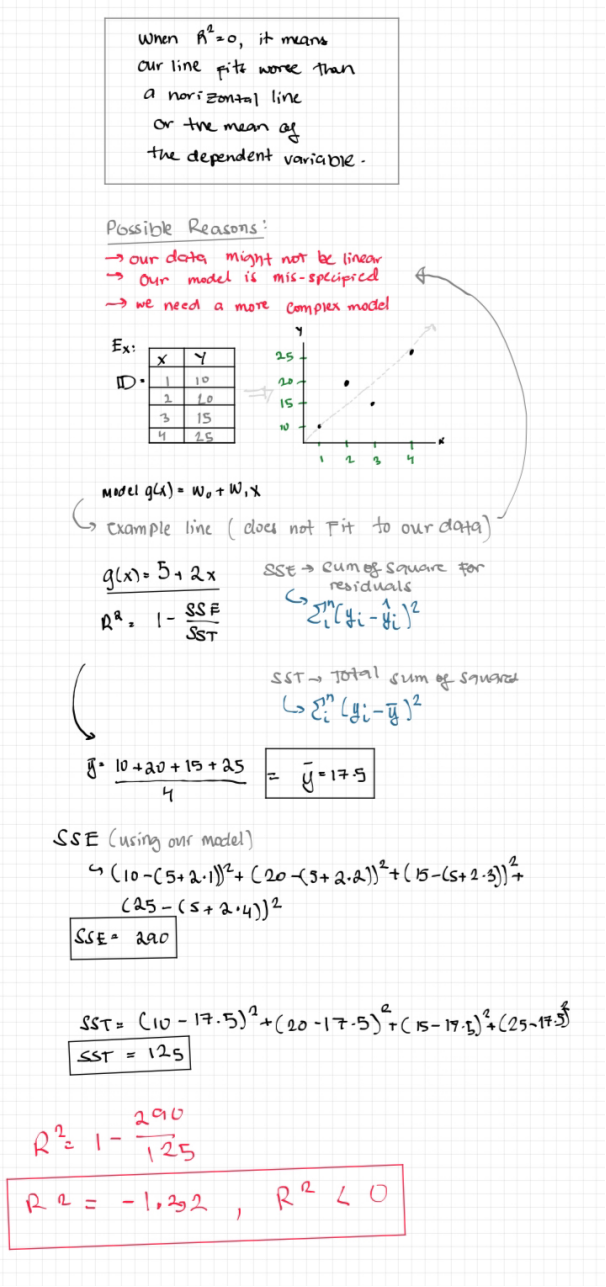


(e) [harder] Why is the RMSE usually a better indicator of predictive performance than R^2? Discuss in English.

Answer:

For instance, when attempting to predict a car's price based on features such as the number of doors, engine size, etc., RMSE provides a value in the same unit as the response. This tells us, on average, how much our predictions deviate from the actual prices. If the RMSE is \$1,000, it indicates that the average error in the model's predicted price is \$1,000. In contrast, R^2 will return a statistic like 0.80 or 80%, suggesting that this proportion of variability in car prices is accounted for by the factors in the model. The remaining 20% might be due to factors not included in the model. In this example, RMSE is more informative because it returns a value in the same unit as the response, making it easier to understand the extent of our deviations. On the other hand, R^2 yields a dimensionless statistic that explains the proportion of variance in the dependent variable but does not directly indicate the magnitude of the errors.

(f) [harder] R^2 is commonly interpreted as "proportion of the variance explained by the model" and proportions are constrained to the interval $[0, 1]$. While it is true that $R^2 \leq 1$ for all models, it is not true that $R^2 \geq 0$ for all models. Construct an explicit example $\mathbb{D}$ and create a linear model $g(x) = w_0 + w_1 x$ whose $R^2 < 0$.



(g) [difficult] You are given $\mathbb{D}$ with n training points $< x_i, y_i >$ but now you are also given a set of weights $[w_1\ w_2\ \ldots\ w_n]$ which indicate how costly the error is for each of the i points. Rederive the least squares estimates b_0 and b_1 under this situation. Note that these estimates are called the *weighted least squares regression* estimates. This variant $\mathcal{A}$ on OLS has a number of practical uses, especially in Economics. No need to simplify

your answers like I did in class (i.e. you can leave in ugly sums).

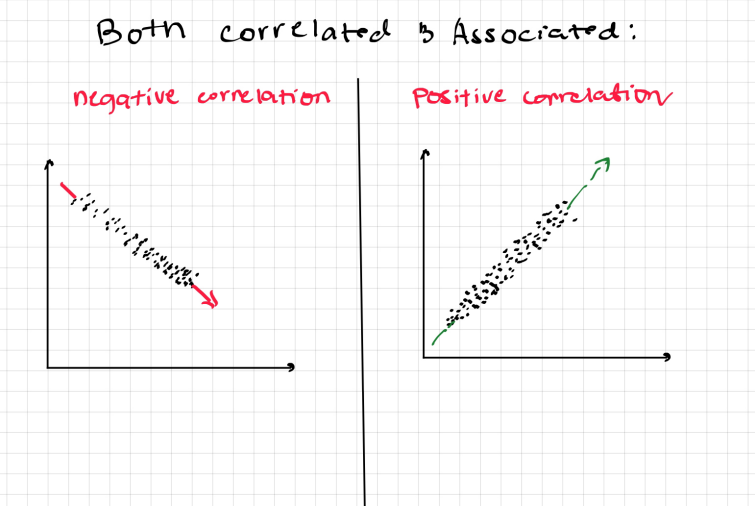
(h) [harder] Interpret the ugly sums in the b_0 and b_1 you derived above and compare them to the b_0 and b_1 estimates in OLS. Does it make sense each term should be altered in this matter given your goal in the weighted least squares?

(i) [E.C.] In class we talked about $x_{raw} \in \{\text{red}, \text{green}\}$ and the OLS model was the sample average of the inputted x. Imagine if you have the additional constraint that x_{raw} is ordinal e.g. $x_{raw} \in \{\text{low}, \text{high}\}$ and you were forced to have a model where $g(\text{low}) \leq g(\text{high})$. Write about an algorithm $\mathcal{A}$ that can solve this problem.

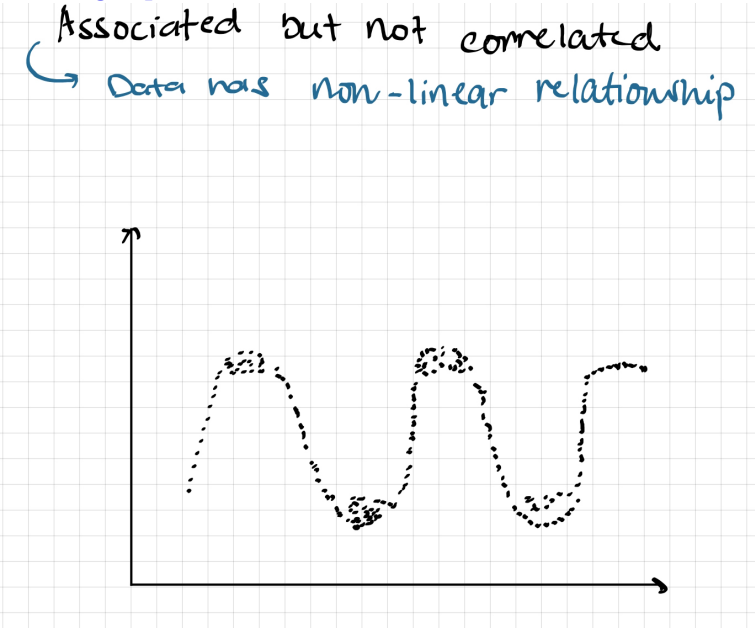
Problem 5

These are questions about association and correlation.

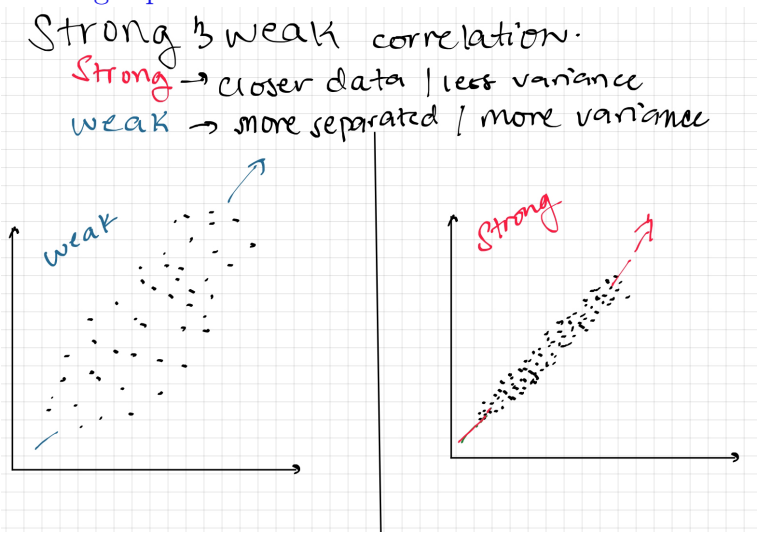
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(b) [easy] Give an example of two variables that are not correlated but are associated by drawing a plot.



(c) [easy] Give an example of two variables that are not correlated nor associated by drawing a plot.



(d) [easy] Can two variables be correlated but not associated? Explain.

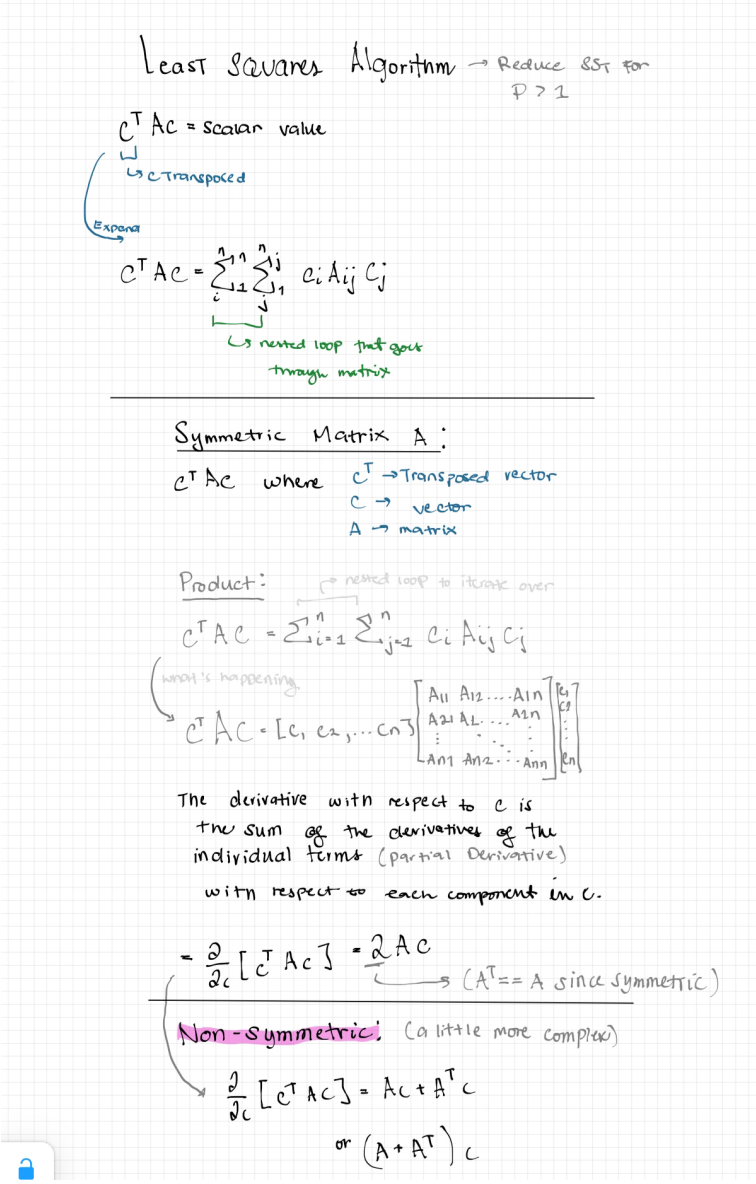
Answer:

Two variables can be either both correlated and associated, or just associated. It is impossible for variables to be correlated but not associated, as correlation is a specific type of association. When two variables have a linear relationship, they are correlated, but if they have a non-linear relationship, they are merely associated. In summary, if two variables are either correlated or associated, they are related in some way. However, to say that they have a specific type of relationship but are not associated is incorrect because stating that they are correlated inherently implies that they are associated.

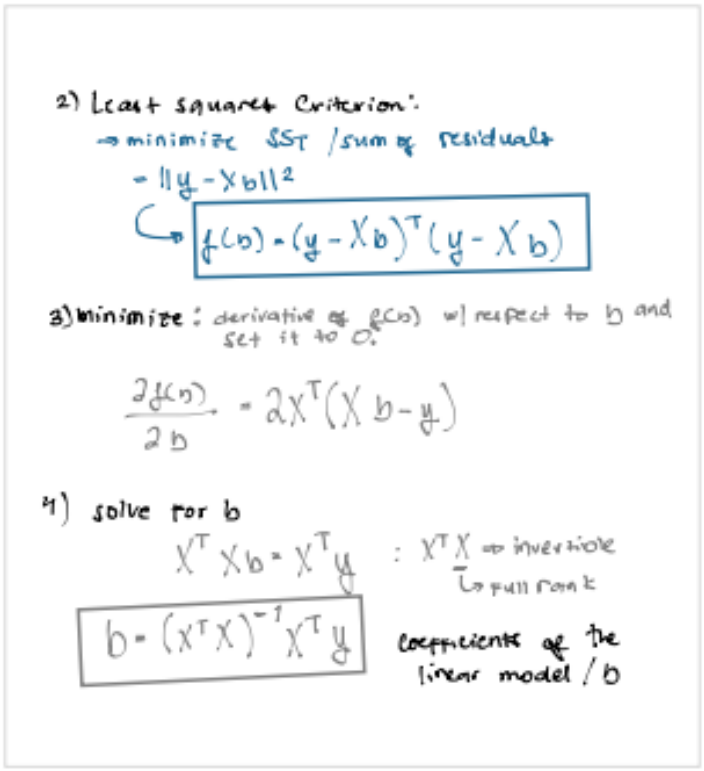
Problem 6

These are questions about multivariate linear model fitting using the least squares algorithm.

(a) [difficult] Derive $\frac{\partial}{\partial \boldsymbol{c}}\left[\boldsymbol{c}^\top A \boldsymbol{c}\right]$ where $\boldsymbol{c} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ but *not* symmetric. Get as far as you can.



(b) [easy] Given matrix $X \in \mathbb{R}^{n \times (p+1)}$, full rank and first column consisting of the $\mathbf{1}_n$ vector, rederive the least squares solution $\boldsymbol{b}$ (the vector of coefficients in the linear model shipped in the prediction function g). No need to rederive the facts about vector derivatives.



(c) [harder] Consider the case where $p = 1$. Show that the solution for $\boldsymbol{b}$ you just derived in (b) is the same solution that we proved for simple regression. That is, the first element of $\boldsymbol{b}$ is the same as $b_0 = \bar{y} - r\frac{s_y}{s_x}\bar{x}$ and the second element of $\boldsymbol{b}$ is $b_1 = r\frac{s_y}{s_x}$.

(d) [easy] If X is rank deficient, how can you solve for $\boldsymbol{b}$? Explain in English.

Answer:

If X is rank deficient, this implies that the features in X are not linearly independent. Therefore, we cannot use the normal equation $b = (X^\top X)^{-1}X^\top y$ because $X^\top X$ is not invertible, meaning it does not have a unique inverse. Techniques such as Ridge Regression and Pseudo-Inverse could be used to solve for b.

Citation:

n/a (2024, February 25).
"Rank Deficiency Solutions". Retrieved from https://chat.openai.com/c/e5a24cde-c381-4122-a7d5-11aef6e17fe8 (ChatGPT)

(e) [difficult] Prove $\text{rank}\,[X] = \text{rank}\left[X^\top X\right]$.

(f) [harder] Prove that $g([1\ \bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_p]) = \bar{y}$ in OLS.

The linear regression model could be expressed by,

$$\underline{y} = X\underline{b} + \epsilon$$

OLS solution for estimating b

$$b = (X^T X)^{-1} X^T y$$

Prediction

$$g(X) = Xb$$

Prove:

$$g([1\bar{x}_1, \bar{x}_2, \ldots \bar{x}_p]) = \bar{y}$$

1) Expression for prediction:

$$g([1\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)] = [1\bar{x}_1, \bar{x}_2, \ldots \bar{x}_p]b$$

2) Expand b

$$g = ([1\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]) = [1\bar{x}_1, \bar{x}_2, \ldots \bar{x}_p](X^T X)^{-1} X^T y$$

3) $\bar{X}$ is a row vector where each element is the mean of the corresponding column in x

$$\bar{X} = [1\bar{x}_1, \bar{x}_2, \ldots \bar{x}_p]$$

4) substitute

$$g(\bar{x}) = \bar{X}(X^T X)^{-1} X^T y$$

5) Property of OLS

- This implies

$$X^T(y - Xb) = 0 \;/\; X^T y = X^T X b$$

- Substituting $b = (X^T X)^{-1} X^T y$ into the orthogonality condition we find

$$X^T y = X^T X (X^T X)^{-1} X^T y$$

$$\rightarrow X^T y = X^T y$$

Therefore

$$g(\bar{x}) = \bar{X} b$$

"OLS Properties Proof". Retrieved from https://chat.openai.com/c/b0f4defd-66e4-45d0-bef7-98e25a4978e3 (ChatGPT)

(g) [harder] Prove that $\bar{e} = 0$ in OLS.

(h) [difficult] If you model $\boldsymbol{y}$ with one categorical nominal variable that has levels A, B, C, prove that the OLS estimates look like $\bar{y}_A$ if $x = A$, $\bar{y}_B$ if $x = B$ and $\bar{y}_C$ if $x = C$. You can choose to use an intercept or not. Likely without is easier.