MATH 342W / 650.4 Spring 2024 Homework #2

Professor Adam Kapelner

Thursday 21st March, 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.

The distinction between Hedgehog and Fox methods refers to two different approaches to understanding or predicting an outcome, represented by variable y. The Hedgehog approach is rooted in the conviction that a single, overarching theory or model can illuminate the entirety of a phenomenon. In contrast, the Fox approach advocates for employing a collection of smaller, varied theories and concepts to make predictions about y. For each issue that arises, this method utilizes these diverse small-scale ideas to address it.

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

Harry Truman is well-known for his wish for a "one-handed economist," showcasing his preference for clear-cut, straightforward advice rather than detailed, conditional analyses. He was not in favor of the cautious and nuanced approach often associated with "foxes" within his team, who were hesitant to provide quick, definitive answers due to their awareness of their own limitations in knowledge. This illustrates Truman's inclination towards a more hedgehog-like approach to decision-making, valuing certainty and simplicity. Such an approach is attractive for its directness and decisiveness, particularly in situations where the stakes are high or rapid decisions are necessary.

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

The concept that a higher level of education could result in less precise forecasts is akin to the phenomenon of "overfitting." This suggests that the more educated a person *becomes, the more likely they are to apply all of their acquired knowledge and experiences in creating models and solving problems. Such an approach can lead to overly complex solutions that may not work well when applied to new, unseen data.*

(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.

Probabilistic classifiers offer outputs in the form of probabilities, providing a richer set of information compared to classifiers that only yield class labels. This attribute grants them greater flexibility, as they not only indicate the most likely class but also convey the level of certainty or uncertainty associated with each classification.

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

PECOTA is similar to the k-Nearest Neighbors (KNN) algorithm we learned in class.

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

Baseball performance as a function of age $y = f(age) + \epsilon$ *is not strictly linear* $\not\equiv f_{linear}$.

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

Baseball scouts can outperform systems like PECOTA because they use their experience and gut feelings to judge players. They look at things like how hard a player works and their ability to lead, which numbers can't always show. Scouts also consider how well a player would fit into a team's specific style and environment, something that stats and predictions might not capture.

(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?

In Nate Silver's book, the Pitch f/x data, while offering valuable insights into player performance, had not yet replaced traditional scouting. It served more as a complement, providing numerical data alongside the scouts' experiential evaluations. The reason Pitch f/x data wasn't fully utilized for predicting future success in baseball was not due to its lack of potential but rather because of limited access to comprehensive data sets. This restricted the ability to harness its full predictive power.

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 hypothesis set H for SVM algorithm

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

(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"?

K nearest neighbors algorithm works in the premise that similar things exist near others. With this you select an arbitrary K value of neighbors you will consider, then calculate the distance between the instance and its neighbors, depending on the data you would use are Euclidean, Manhattan, or Hamming distance. From this you assign labels to all the data and the mode of them is the returned answer. In this case K is considered a hyper-parameter as the fact that the value of K can have massive consequences on the accuracy of the model.

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

(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).

The error should be zero because stating that there is only 1 nearest neighbor it means its closest neighbor its itself. The distance between the data and itself would be zero. This doesnt mean that k=1 is a good estimate of future error when new data comes in, with this K the algorithm becomes sensitive to noise in the data due to overfitting.

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 $\mathcal{Y}$?

The matrix $\mathbb{D}$ has each row is a single point, and columns are: x_1 - the predictor variable and y is the response. $\mathcal{X}$ is a column vector full of xs. $\mathcal{Y}$ is a column vector of ys.

(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.

(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 $\bar{y}$.

(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.

(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?

The complex sums in the b_0 and b_1 calculations derived represent the inclusion of weights w_i for each y_i and x_i data point, indicating that every point is adjusted according to its weight. This adjustment is crucial in weighted least squares analysis, as it ensures the model captures the underlying pattern in the data more accurately. Unlike in ordinary least squares (OLS), where each data point contributes equally to the estimates of b_0 and b_1, in weighted least squares, the data points are scaled to reflect their relative importance, aligning with our objective to give more significance to certain observations in the model.

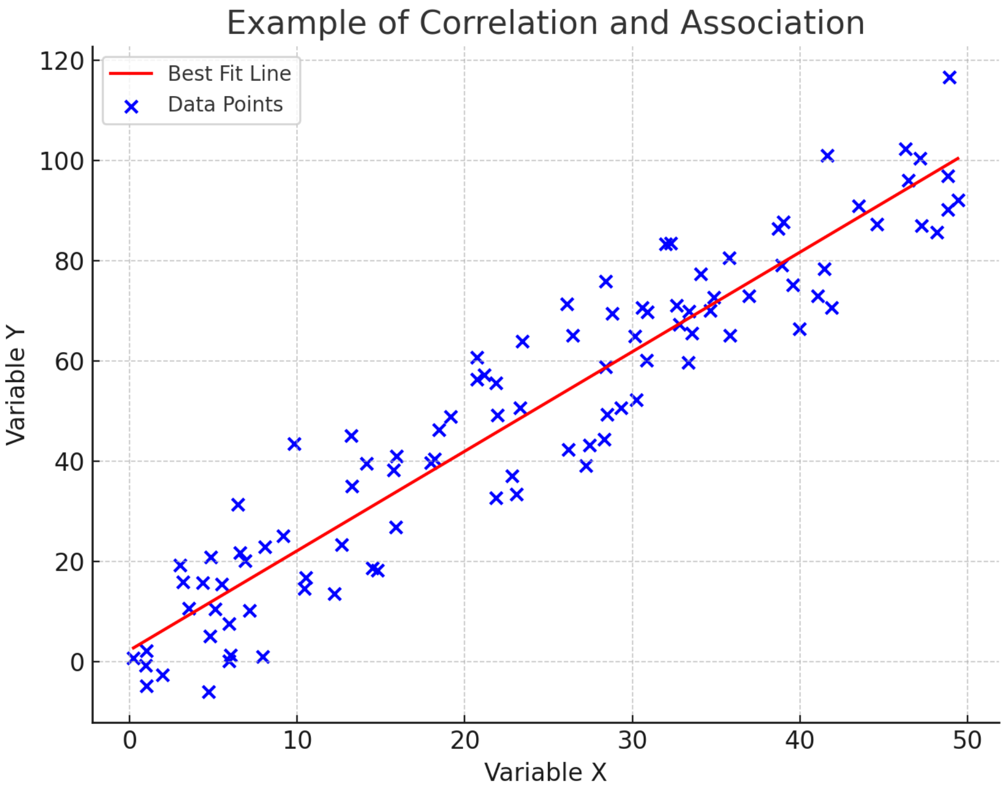
(i) [E.C.] In class we talked about $x_{raw} \in \{\text{red}, \text{green}\}$ and the OLS model was the sample average of the inputted x. Imagine if you have the additional constraint that x_{raw} is ordinal e.g. $x_{raw} \in \{\text{low}, \text{high}\}$ and you were forced to have a model where $g(\text{low})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

Problem 5

These are questions about association and correlation.

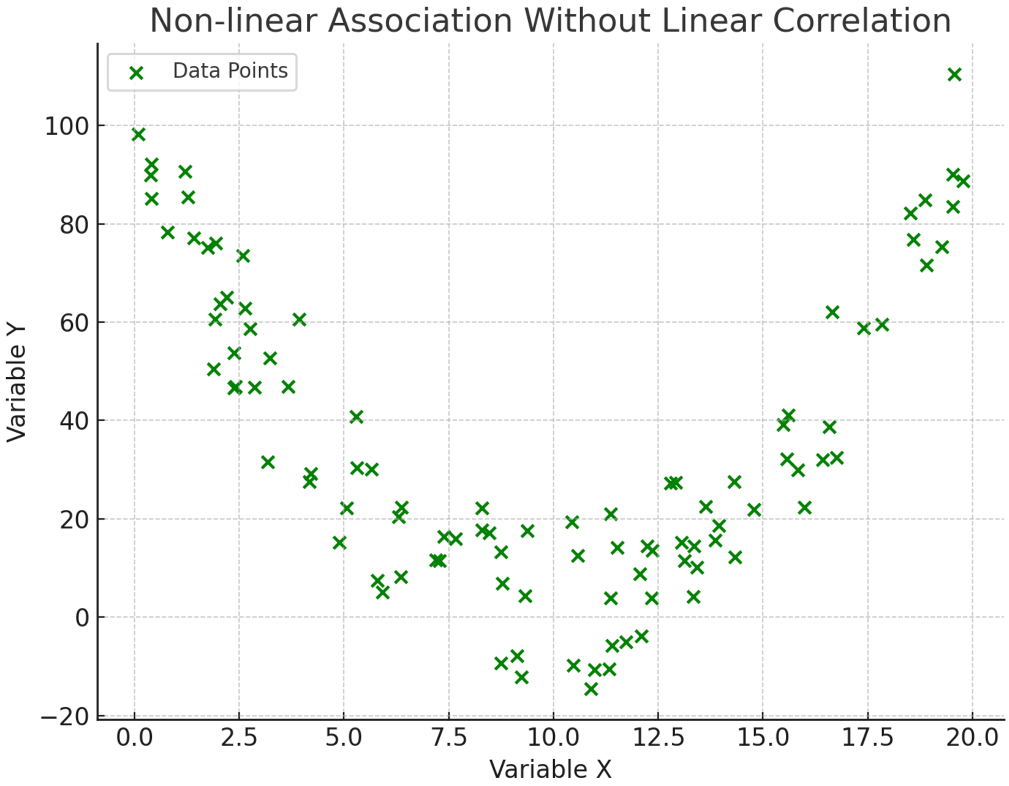
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.

An example of 2 variables that are both correlated and associated is x = temperature and y = ice cream sales.



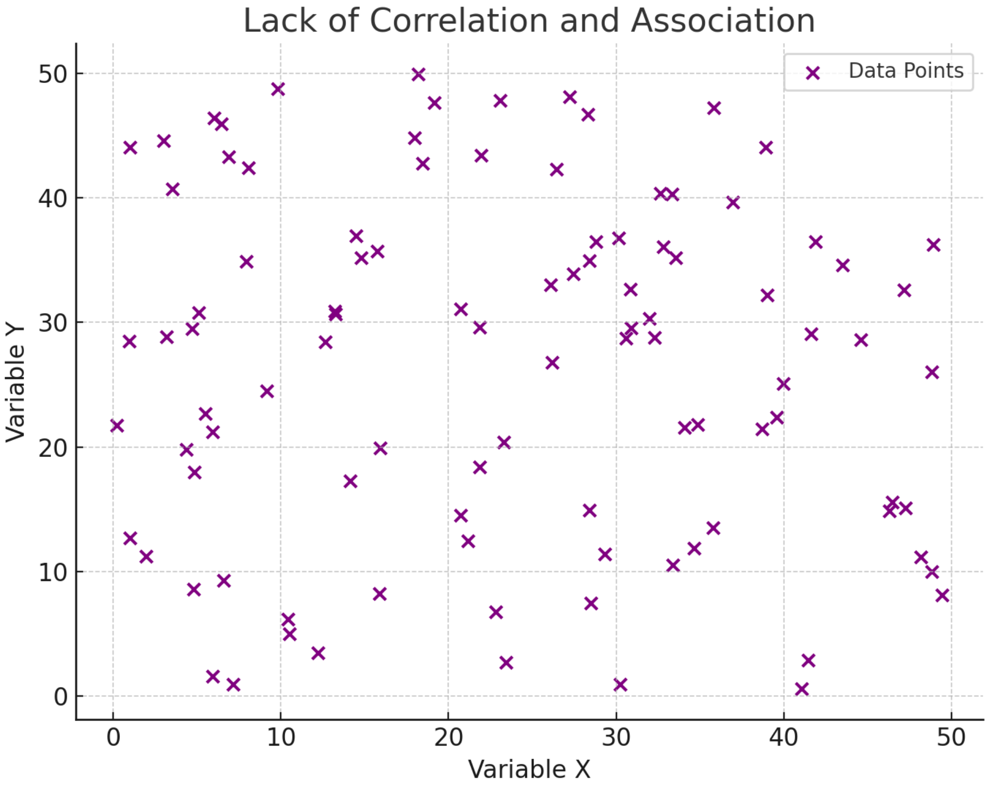
(b) [easy] Give an example of two variables that are not correlated but are associated by drawing a plot.

two variables that are not correlated but are associated, x = hours of work per day and y = productivity level.



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

An example of 2 variables that are neither correlated nor associated is where x = *train delays per hour, and* y = *number planes that left from JFK per hour*



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

Correlation and association describe relationships between variables, but in distinct ways. Correlation quantifies the linear relationship between two quantitative variables, indicating how one variable changes with another. Association, a broader term, refers to any relationship or dependency between variables, including linear and nonlinear patterns. Thus, if two variables are correlated, they are inherently associated, as correlation is a specific type of linear association. However, variables can be associated through various patterns without being linearly correlated, meaning the presence of correlation implies association, but association does not necessitate linear correlation.

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.

If X is rank deficient, meaning it lacks full rank, the solution for β can be problematic due to insufficient unique information. To address this issue, one can identify and eliminate redundant variables from the model. This action helps to alleviate multicollinearity, ensuring that each remaining variable in X contributes unique and valuable information to the determination of β. By doing so, we enhance the model's stability and the reliability of the β estimates.

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

(f) [harder] [MA] If $p = 1$, prove $r^2 = R^2$ i.e. the linear correlation is the same as proportion of sample variance explained in a least squares linear model.

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

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

(i) [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.

(j) [harder] [MA] Prove that the OLS model always has $R^2 \in [0, 1]$.