MATH 342W / 650.4 Spring 2024 Homework #2

Osman Khan

Saturday 24th 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.

Hedgehogs have simple $\mathcal{H}$, which are not complex enough for the events about which forecasts are made, while foxes have more complex $\mathcal{H}$. In addition, foxes have models that have done a better job at learning from $\mathbb{D}$, and also at model validation. Hedgehogs tend to spend more time thinking from a theoretical point of view, and almost ignorning $\mathbb{D}$.

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

Harry Truman liked hedgehogs because they would have given him a clear answer/line of action. I think a lot of people are not like hedgehogs (or foxes) because not everyone have strong feelings and positions about things. I think most people have a goal of keeping as many people happy as possible, and are worried having a strong position might offend some.

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

With more education, one tends to become more of an expert in one specific field, and have an increase in hubris, which would cloud that person's judgement in regards to model making for forecasts. It tends to confirm their biases.

(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.

It is more realistic to say a range of things can happen, rather than saying a specific thing would happen. For e.g., saying there is a high chance of 4-7 mm of rain happening tomorrow v. there is a high chance of 5.5 mm of rain happening tomorrow.

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

PECOTA is most similar to the nearest neighbor model.

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

According to James, baseball performance as a function of age was a quadratic model with a peak at 27. I agree with that as most athletes tend to peak at around 27, and their performanes do follow the quadratic trajectory of getting better until 27, and then getting worse afterwards. There surely are exceptions to this, as some outliers do exist, i.e. Justin Verlander of the Astros.

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

Baseball scouts are able to see the soft skills that PECOTA cannot see. This is more clearly seen in minor-leagues (when players are younger). A good pitcher, with low speeds, might be getting lots of outs in the minor-league as the batters aren't good enough, but a good scout would know that pitcher would get thrashed in the major league.

(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?

Pitch f/x was not available at all minor-league stadiums. Incidentally, it was replaced by Trackman in 2017.

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?

$\mathcal{H} = \{\mathbb{1}_{\vec{w}.\vec{x}\geq 0 = w_0 + w_1 x_1 + w_2 x_2 \geq 0} : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}, w_2 \in \mathbb{R} = \vec{w} \in \mathbb{R}^3\}$.

No, it is the same Hypothesis Set (all lines: $\mathcal{H} = \{\mathbb{1}_{x_2 \leq a + bx,} : a \in \mathbb{R}, b \in \mathbb{R}\}$), just reparametrized. We reduce the dimension of the w vector and state the intercept separately as b, to get:

$\mathcal{H} = \mathbb{1}_{\vec{w}.\vec{x} - b \geq 0} : \vec{w} \in \mathbb{R}^p, b \in \mathbb{R}\}$

(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.

I) All y=1's are above or equal to L_U:
$\forall i' s s.t. y_i = 1, \vec{w}\vec{x_i} - (b+1) \geq 0 \Rightarrow \vec{w}\vec{x_i} - b \geq 1 \Rightarrow y_i(\vec{w}\vec{x_i} - b) \geq 1$
II) All y=0's are below or equal to L_L:
$\forall i' s s.t. y_i = -1, \vec{w}\vec{x_i} - (b-1) \leq 0 \Rightarrow \vec{w}\vec{x_i} - b \leq -1 \Rightarrow y_i(\vec{w}\vec{x_i} - b) \geq 1$

(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.

THE := $\sum_{i=1}^{n} \max\{0, 1 - y_i(\vec{w}\vec{x_i} - b)\}$,
H_i =$\max\{0, 1 - y_i(\vec{w}\vec{x_i} - b)\}$,
Vapnik Objective Function:
$\underset{argmin}{\vec{w}, b} \{\frac{1}{n} THE + \lambda ||\vec{w}||^2\} \Rightarrow g = \mathcal{A}(\mathbb{D}, \mathcal{H}, \lambda)$

Problem 3

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

(a) [easy] Describe how the algorithm works. Is k a "hyperparameter"?

KNN looks at K nearest neighbors of a datapoint, and returns the mode of those K nearest neighbors as the prediction for a new data point. K is indeed a hyperparameter.

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

For KNN, we will not be predicting data that have never occurred before. The prediction will always be a subset of the historical data. This means our $\mathcal{H}$ could be a 'multisubset' of $\mathbb{D}$.

(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).

Errors are measured by y-yhat. But, if K=1, then there is only one value/classification of that point, and it will be identical to yhat. This is not a good estimate of future error as the new data does not necessarily have to belong to the same classification of y/yhat.

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 $\mathcal{Y}$?

$\mathbb{D}$ looks like $< \vec{x}, \vec{y} >$, where both $\mathcal{X}$ & $\mathcal{Y}$ are column vectors.

(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.

For $\mathcal{Y} = R$ or $\mathcal{Y} \subset R$, $p = 1$, our candidate sets are linear models. Then, for OLS, $g(x) = b_0 + b_1 x$, where we derived $b_1 = r\frac{S_y}{S_x}, b_0 = \bar{y} - r\frac{S_y}{S_x}\bar{x}$. Now, if we plug in $\bar{x}$ into $g(x)$ we get: $\bar{y} - r\frac{S_y}{S_x}\bar{x} + r\frac{S_y}{S_x}\bar{x} = \bar{y}$.■

(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 $\bar{y}$.

As average is taking the sum and dividing by the number, we find that the 'average' prediction is: $\frac{\sum_{i=1}^{n}(\hat{y}_i)}{n} = \frac{\sum_{i=1}^{n}(b_0 + b_1\hat{x}_i)}{n} = \frac{nb_0}{n} + \frac{nb_1\sum_{i=1}^{n}(\hat{x}_i)}{n} = b_0 + b_1\bar{x} = \bar{y}$.■

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

$e = y - \hat{\boldsymbol{y}}$. We know the average of y is $\bar{y}$, and from 4.d we know that the average prediction $\hat{y}_i = \bar{y} \therefore e_i = \bar{y} - \hat{y}_i = \bar{y} - \bar{y} = 0$.■

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

RMSE is in the same unit as Y, and gives us a 95% confidence interval of containing the true Y. On the other hand, R^2 just gives us the relationship with the null model. If worse than the null model, R^2 is negative.

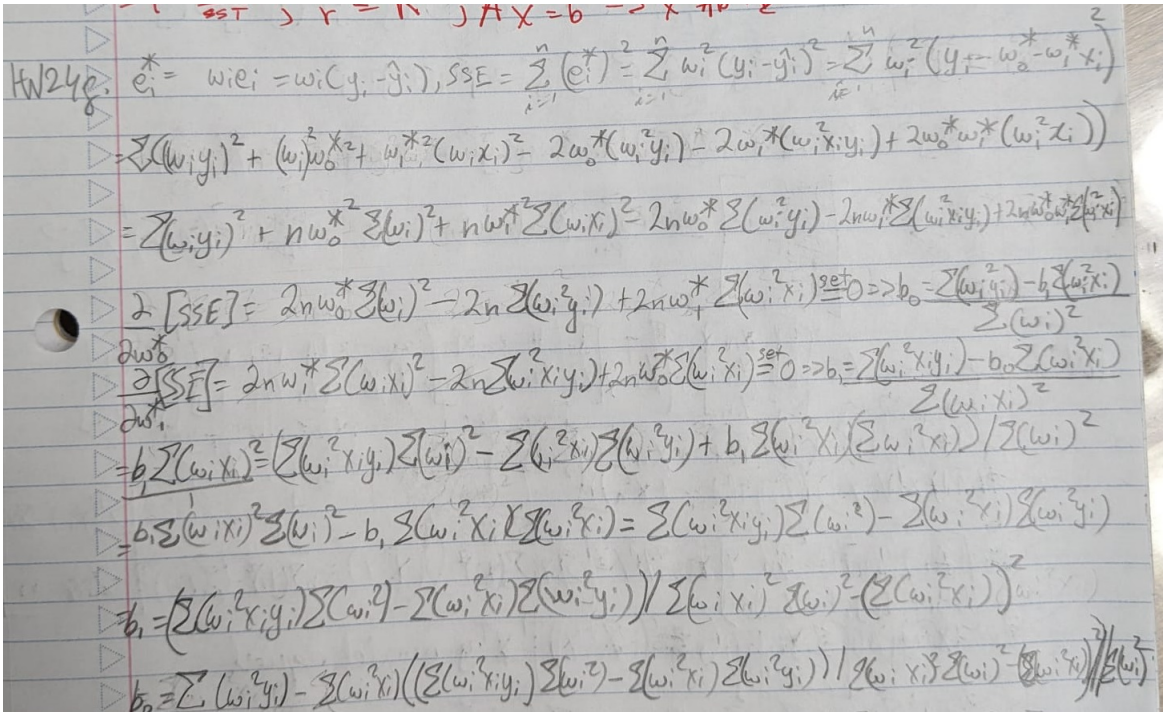
(f) [harder] R^2 is commonly interpreted as "proportion of the variance explained by the model" and proportions are constrained to the interval $[0, 1]$. While it is true that $R^2 \leq 1$ for all models, it is not true that $R^2 \geq 0$ for all models. Construct an explicit example $\mathbb{D}$ and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

This would require us to have a terrible example, whereby our model is way of the data i.e. the null model has better performance.

We will let $\mathcal{D} =< \vec{X}, \vec{Y} >$ where $\vec{X} =< 7, 17, 27, 37 >$ while $\vec{Y} =< 37, 27, 17, 7 >$. We will have $w_0 = -100$ and $w_1 = -1$, which will give terrible errors.

(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).

When we reparametrized the Hypothesis space of line, we attached weights to features (except w_0 which was the intercept/bias term). Later on, while deriving least squared regression, at some point, we took the derivative w.r.t w_0 & w_1. The weights given in this question, for Weighted Least Squares Regression, also has weights $[w_1, w_2, ...w_n]$. We will denote the earlier weights as w_0^* & w_1^* to avoid confusion.



(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?

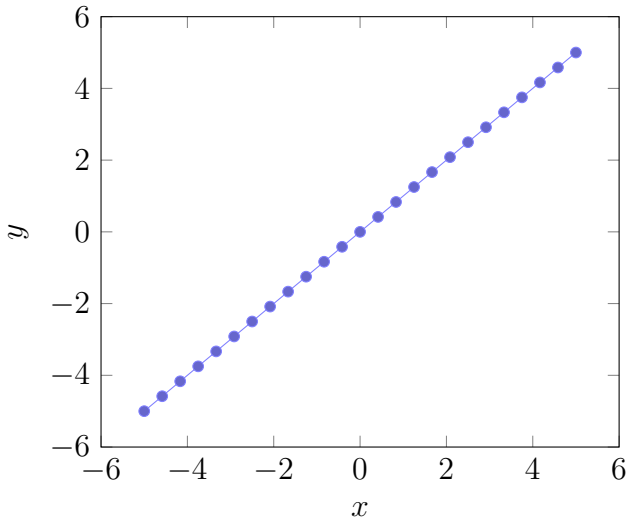
Yes, it makes sense, as our goal is to have the appropriate/estimate weight/impact of each error on our prediction.

(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.

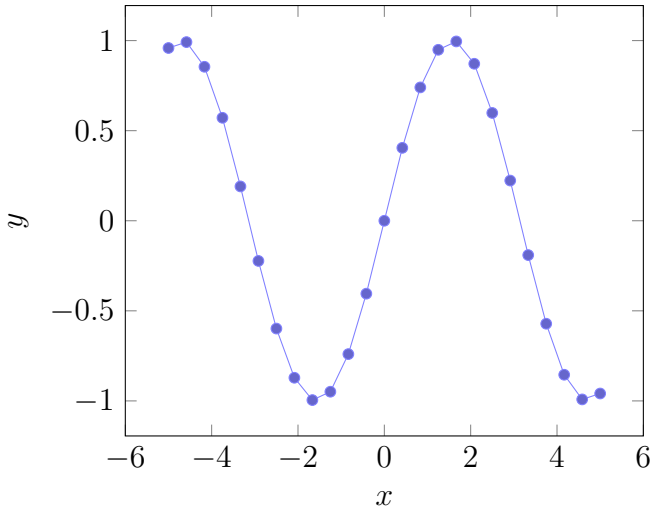
Problem 5

These are questions about association and correlation.

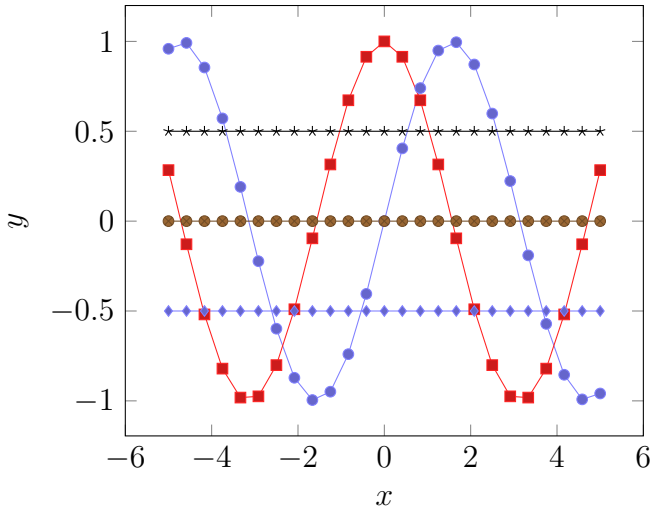
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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.



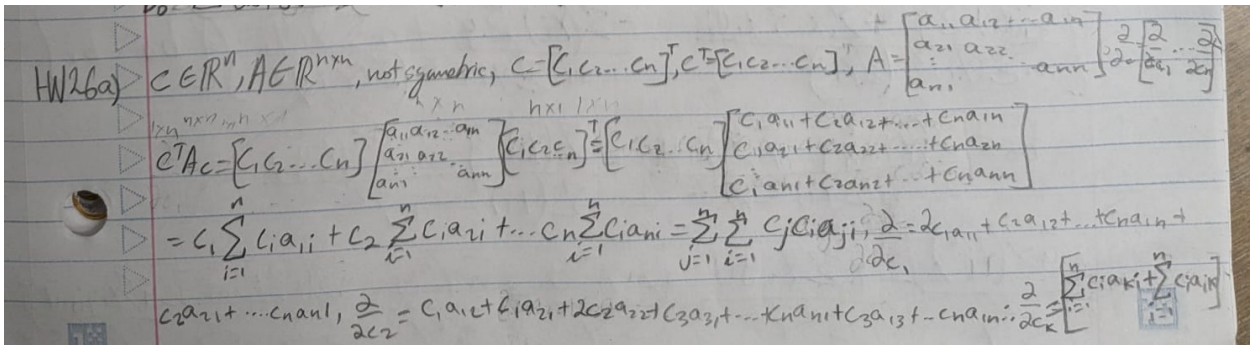
(d) [easy] Can two variables be correlated but not associated? Explain.

No, because correlation is a subset of association, as correlation is linear association.

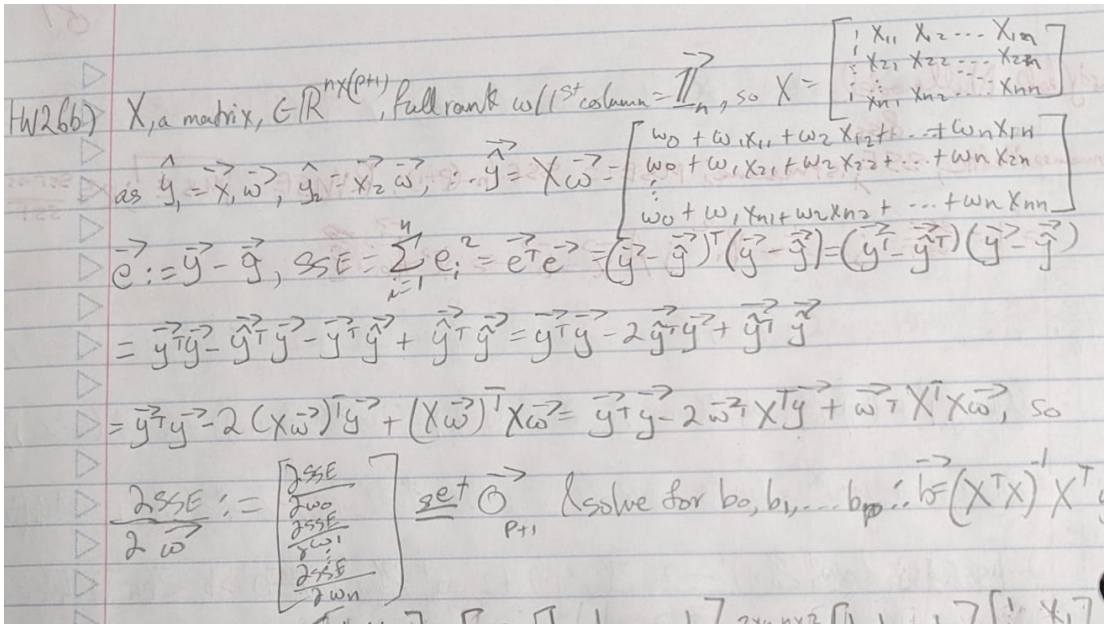
Problem 6

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

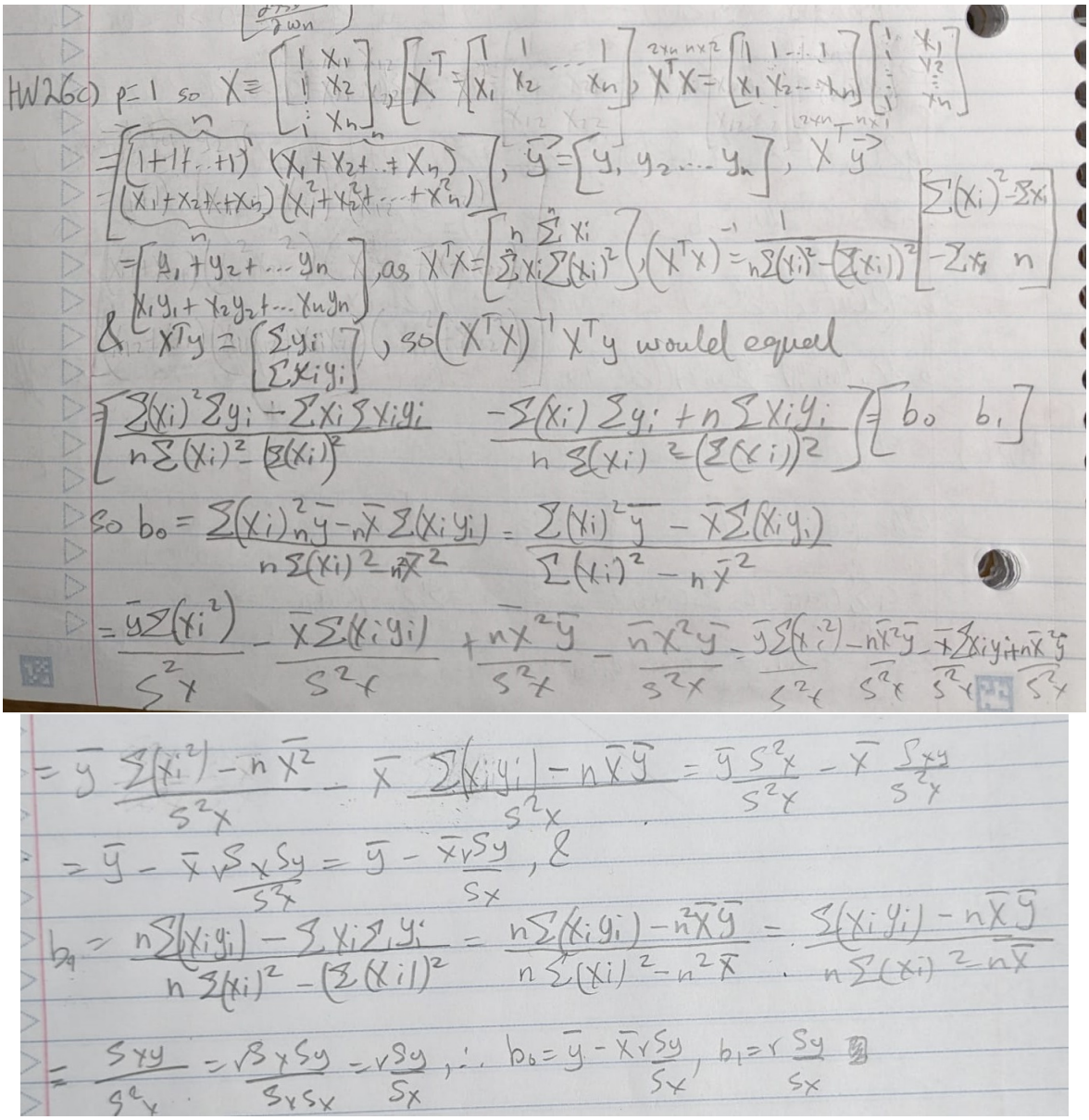
(a) [difficult] Derive $\frac{\partial}{\partial \boldsymbol{c}}\left[\boldsymbol{c}^\top A \boldsymbol{c}\right]$ where $\boldsymbol{c} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ but *not* symmetric. Get as far as you can.



(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, then 1 or more of its columns are not linearly independent. This means they are not necessary. They most be removed so that we are left with columns that are linearly independent, and then we can solve for b.

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

$\rightarrow$ If $Xb = 0$ for some b, then $X^T Xb = 0$

$\leftarrow$ If $X^T Xb = 0$ for some b, then $b^T X^T Xb = 0 \Rightarrow Xb = 0 \therefore \text{rank}\,[X] = \text{rank}\,[X^T X]$

(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.

$r^2 = (\frac{S_{xy}}{S_x S_y})^2 = \frac{SST-SSE}{SST} = R^2$

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

As $g(\vec{x^*}) = \hat{\boldsymbol{y}}^* = b_0 + b_1 x^*{}_1 + \ldots + x^*{}_p$
then $g([1\ \bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_p]) = b_0 + b_1\bar{x}_1 + \ldots + \bar{x}_n$
which $= \frac{b_0}{n}\sum(1) + \frac{b_1}{n}\sum(\bar{x}_1) + \ldots + \frac{b_p}{n}\sum(\bar{x}_p)$
which $= \frac{1}{n}(b_0\sum(1) + b_1\sum(\bar{x}_1) + \ldots + \sum(b_p\sum(\bar{x}_p))$
which $= \frac{\frac{1}{n}}{\hat{\boldsymbol{y}}} = \frac{\hat{\boldsymbol{y}}}{n} = \underbrace{\overline{\hat{\boldsymbol{y}}} = \hat{\boldsymbol{y}}}_{from 4.c}$

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

$\bar{e} = \frac{1}{n}e_i$, as OLS is linear regression, the errors can either be positive (if the regression is above the data point) or negative (if the regression is below the data point). Add all these errors give us zero.

(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.

$\vec{y}$ is one categorical nominal variable that has levels A, B, C, so $\vec{x}$ would be a column vector of length n, while X would be an $n * 3$ matrix while X^T will be a $3 * n$ matrix. $\vec{b} = (X^T X)^{-1} X^T \vec{(y)}$, where $X^T X$ = a diagonal matrix of size three, which has nA, nB, & nC on its diagonal, so its inverse would have the reciprocal of those entries on the diagonal. $X^T\vec{y}$ would be a $3 * 1$ matrix with the entries being $\sum_{i=A} Yi, \sum_{i=B} Yi, \sum_{i=c} Yi \therefore \vec{b}$ = a column vector of length 3: $[\bar{y}_A\bar{y}_B\bar{y}_C]^T$. As $x = A = g([100]), x = B = g([010], x = C = g([001])$, their multiplication with $\vec{b}$ would yield $\bar{y}_A, \bar{y}_B, \& \bar{y}_C$ respectively.■

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

OLS is always at least as good as the null model. Therefore, even when there is an error on all points, the worst it can be is the null model, which gives and R^2 of 0 as $R^2 = \frac{SST-SSE}{SST}$, while when the model is super good, SSE is almost 0 and R^2 is close to 100%.