MATH 390.4 / 650.2 Spring 2020 Homework #3

Pizon Shetu

Due noon Friday, March 13, 2020 under the door of KY604

(this document last updated 1:21am on Thursday 19th March, 2020)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read Chapters 3-6 of Silver's book. You should be googling and reading about all the concepts introduced in class online. This is your responsibility to supplement in-class with *your own* readings.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 7 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document *including this first page* and write in your answers. I do not accept homeworks which are *not* on this printout.

NAME: PIZON SHETU

Problem 1

These are questions about Silver's book, chapters 3-6. For all parts in this question, answer 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) [easy] What algorithm that we studied in class is PECOTA most similar to?

K Nearest Neighbors

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

No, it is difficult to rate players on age alone as the data is noisy, but for most players performances increases from early 20's to 30 and decline as they age.

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

They would have know more features, by that I mean more relevant independent features which might allow them to tune out the noise and make the data fit better.

(d) [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?

Because that type of data was extremely expensive and at that time, not a lot of people had access to such tools to monitor such data

(e) [difficult] Chapter 4 is all about predicting weather. Broadly speaking, what is the problem with weather predictions? Make sure you use the framework and notation from class. This is not an easy question and we will discuss in class. Do your best.

Being able to predict the weather with high accuracy and precision is nearly impossible, weather takes into multiple factors in reality, and we would need to be able to predict accurately for each of those factors making weather predictions a multi-level prediction where each predictions is predicated on the supporting factors to be correct for its predictions to be valid. In theoretical sense f is difficult to understand, we would need to able to properly calculate motion, moisture, and etc but this is difficult so we simplify and make approximations.

(f) [easy] Why does the weatherman lie about the chance of rain? And where should you go if you want honest forecasts?

Its easier to be wrong on bad news than to be right on good news. If there is a possibility of rain the weatherman will lie about the chances to make sure he has some knowledge of the occurrences but if he says there is no chance of rain even though it will be a sunny day he hedges and comes out good because people don't get upset about wrong news if its good news.

(g) [difficult] Chapter 5 is all about predicting earthquakes. Broadly speaking, what is the problem with earthquake predictions? It is *not* the same as the problem of predicting weather. Read page 162 a few times. Make sure you use the framework and notation from class.

Its difficult to measure actually data since most of it is under ground. Experts really only have access to past results. It is hard to predict having mostly results from the past and not being able to access current x's

(h) [easy] Silver has quite a whimsical explanation of overfitting on page 163 but it is really educational! What is the nonsense predictor in the model he describes?

Having the combination to 3 locks, also knowing the flaws of these locks but using this to lock pick other locks. This is over-fitting as combinations to 3 specific locks have no correlation to other locks nor does knowing the flaws of 3 specific locks.

(i) [easy] John von Neumann was credited with saying that "with four parameters I can fit an elephant and with five I can make him wiggle his trunk". What did he mean by that and what is the message to you, the budding data scientist?

Essentially a good model can be over-fitted merely by adding 1 unnecessary feature

(j) [difficult] Chapter 6 is all about predicting unemployment, an index of macroeconomic performance of a country. Broadly speaking, what is the problem with unemployment predictions? It is *not* the same as the problem of predicting weather or earthquakes. Make sure you use the framework and notation from class.

Unemployment predictions are often too late, there predictions become relevant only few months prior to the phenomena. Economic forecasters face 3 fundamental challenges according to Silver, "First, it is very hard to determine cause and effect from economic statistics alone. Second, the economy is always changing, so explanations of economic behavior that hold in one business cycle may not apply to future ones. And third, as bad as their forecasts have been, the data that economists have to work with isn't much good either." In other words f is extremely hard to find, and not x fit with the ever changing economy so f is also always changing thus making the prediction difficult to manufacture.

(k) [E.C.] Many times in this chapter Silver says something on the order of "you need to have theories about how things function in order to make good predictions." Do you agree? Discuss.

Problem 2

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

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

$$\boldsymbol{c}^\top A\boldsymbol{c} = \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ \dots \\ c_n \end{bmatrix}$$

$$= \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} \cdot \begin{bmatrix} c_1a_{11} + c_2a_{12} + \dots + c_na_{1n} \\ c_1a_{21} + c_2a_{22} + \dots + c_na_{2n} \\ \dots \\ c_1a_{n1} + c_2a_{n2} + \dots + c_na_{nn} \end{bmatrix}$$

$$= c_1\big(c_1a_{11} + c_2a_{12} + \dots + c_na_{1n}\big) + \dots + c_n\big(c_1a_{21} + c_2a_{22} + \dots + c_na_{2n}\big)$$

where

$$= \sum_{i=1}^n c_i \left(\sum_{j=1}^n c_j a_{ij} \right) \quad \text{and} \quad \frac{\partial}{\partial \boldsymbol{c}_i} \left[\boldsymbol{c}^\top A\boldsymbol{c}\right] = \frac{\partial}{\partial \boldsymbol{c}_i} \left[\sum_{i,j=1}^n c_i c_j a_{ij}\right]$$

where

$$= \sum_{j=1}^n c_j a_{ij} + c_j a_{ji}$$

(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.

$$SSE = \sum \big(\vec{y}_i - \vec{\hat{\boldsymbol{y}}}\big)^2 = \big(\vec{y} - \vec{\hat{\boldsymbol{y}}}\big)^T\big(\vec{y} - \vec{\hat{\boldsymbol{y}}}\big) = \big(\vec{y}^T - \vec{\hat{\boldsymbol{y}}}^T\big)\big(\vec{y} - \vec{\hat{\boldsymbol{y}}}\big)$$

$$= \vec{y}^T\vec{y} - \vec{y}^T\vec{\hat{\boldsymbol{y}}} - \vec{\hat{\boldsymbol{y}}}^T\vec{y} + \vec{\hat{\boldsymbol{y}}}^T\vec{\hat{\boldsymbol{y}}}$$

$$= \vec{y}^T\vec{y} - 2\vec{\hat{\boldsymbol{y}}}^T\vec{y} + \vec{\hat{\boldsymbol{y}}}^T\vec{\hat{\boldsymbol{y}}}$$

$$= \vec{y}^T\vec{y} - 2\big(X\vec{b}\big)^T\vec{y} + \big(X\vec{b}\big)^T\big(X\vec{b}\big)$$

$$= \vec{y}^T\vec{y} - 2\vec{b}^TX^T\vec{y} + \vec{b}^TX^TX\vec{b}$$

$$\tfrac{\partial}{\partial\vec{b}}\big[\vec{y}^T\vec{y} - 2\vec{b}^TX^T\vec{y} + \vec{b}^TX^TX\vec{b}\big] = -2X^T\vec{y} + 2X^TX\vec{b} = 0$$

$$\big(X^TX\big)^{-1}\big(X^TX\big)\vec{b} = \big(X^TX\big)^{-1}X^T\vec{y}$$

$$\vec{b} = \big(X^TX\big)^{-1}X^T\vec{y}$$

(c) [harder] Consider the case where $p = 1$. Show that the solution for $\boldsymbol{b}$ you just derived is the same solution that we proved for simple regression in Lecture 8. 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 $b_1 = r\frac{s_y}{s_x}$.

$$\vec{b} = \left(X^T X\right)^{-1} X^T \vec{y}$$

$$X \in \mathbb{R}^{n\times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \cdot \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$\sum x_i = n\bar{x}$$

$$\left(X^T X\right)^{-1} = \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

$$\left(X^T X\right)^{-1}\left(X^T \vec{y}\right) = \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$$

$$= \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \cdot \begin{bmatrix} n\bar{y} \\ \sum y_i x_i \end{bmatrix}$$

$$= \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} n\bar{y}\sum x_i^2 - n\bar{\sum} y_i x_i \\ -n^2\bar{x}\bar{y} + n\sum y_i x_i \end{bmatrix}$$

$$b_0 = \frac{\bar{y}\left(\sum x_i^2 - n\bar{x}^2\right) - \bar{x}\left(\sum y_i x_i - n\bar{x}\bar{y}\right)}{\sum x_i^2 - n\bar{x}^2}$$

$$b_1 = \frac{\sum y_i x_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

(d) [easy] If X is rank deficient, how can you solve for $\boldsymbol{b}$? Explain in English. Make X full-rank by making all columns linearly independent

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

$\text{rank}\,[x]$ = Dim of whole space - $N(X)$

Prove that $N(A) = N(A^T A)$:
$N(A) \subset N(A^T A)$
$x \in N(A)$
$Ax = 0$
$A^T Ax = A^T 0 = 0 \qquad \Rightarrow \qquad x \in N(A^T A)$
so $N(A) \subset N(A^T A)$
$N(A^T A) \subset N(A)$

$x \in N(A^TA)$
$A^TAx = 0$
$x^TA^TAx = x^T0 = 0$
$(Ax)^T(Ax) = 0$
$||Ax||^2 = 0$
$Ax = 0$
so $N(A^TA) \subset N(A)$

Therefore, $N(X) = N(X^TX)$ and rank $[X]$ = rank $\left[X^TX\right]$

(f) [difficult] Given matrix $X \in \mathbb{R}^{n\times(p+1)}$, full rank and first column consisting of the $\mathbf{1}_n$ vector, now consider cost multiples ("weights") $c_1, c_2, \ldots, c_n$ for each mistake e_i. As an example, previously the mistake for the 17th observation was $e_{17} := y_{17} - \hat{y}_{17}$ but now it would be $e_{17} := c_{17}(y_{17} - \hat{y}_{17})$. Derive the weighted least squares solution $\boldsymbol{b}$. No need to rederive the facts about vector derivatives. Hints: (1) show that SSE is a quadratic form with the matrix C in the middle (2) Split this matrix up into two pieces i.e. $C = C^{\frac{1}{2}}C^{\frac{1}{2}}$, distribute and then foil (3) note that a scalar value equals its own transpose and (4) use the vector derivative formulas.

$$\begin{aligned} SSE &= \left(\boldsymbol{y} - X\vec{b}\right)^\top C\left(\boldsymbol{y} - X\vec{b}\right) \\ &= \left(\boldsymbol{y}^\top C\boldsymbol{y} - \boldsymbol{y}^\top CX\vec{b} - \vec{b}^\top X^\top C\boldsymbol{y} + \vec{b}^\top X^\top CXb\right) \end{aligned}$$

$$\frac{\partial SSE}{\partial \vec{b}} \;=\; -2X^\top C\boldsymbol{y} + 2X^\top CX\vec{b} \;=\; 0$$

$$\vec{b} = \left(X^\top CX\right)^{-1}X^\top C\boldsymbol{y}$$

(g) [difficult] 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.

$$b_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$
$$b_0 = \bar{y} - b_1\bar{x}$$
$$\hat{\boldsymbol{y}}_i = b_0 + b_1x_i \;=\; \bar{y} - \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right]\bar{x} \;+\; \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right]x_i$$

$$\begin{aligned} R^2 &= \frac{SSR}{SST} \\ &= \frac{\sum(\hat{\boldsymbol{y}}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \\ &= \frac{\sum\left(\bar{y} - \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right]\bar{x} \;+\; \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right]x_i - \bar{y}\right)^2}{\sum\left(y_i - \bar{y}\right)^2} \end{aligned}$$

$$= \frac{\sum \left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right)^2 (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}$$

$$= \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}$$

$$= \frac{\left[Cov(x, y) \right]^2}{s_x^2 s_y^2}$$

$$= r^2$$

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

Recall that

$$\bar{y} = b_0 + b_1 \bar{x}_1 \ + \ b_2 \bar{x}_2 \ + \ \ldots \ + \ b_{p+1} \bar{x}_{p+1}$$

and

$$\hat{\boldsymbol{y}}^* = g([1\ \bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_p]$$

since

$$\hat{\boldsymbol{y}}^* = b_0 + b_1 \bar{x}_1 \ + \ b_2 \bar{x}_2 \ + \ \ldots \ + \ b_{p+1} \bar{x}_{p+1}$$

$$\hat{\boldsymbol{y}}^* = \frac{1}{n} \sum b_0 + \frac{1}{n} \sum b_1 x_{i1} \ + \ \frac{1}{n} \sum b_2 x_{i2} \ + \ \ldots \ + \ \frac{1}{n} \sum b_{p+1} x_{ip+1}$$

and we know that

$$\hat{\boldsymbol{y}}^* = \frac{1}{n} \sum \hat{\boldsymbol{y}}_i \text{ since } \hat{\boldsymbol{y}}_i = b_0 + b_1 x_{i1} \ + \ b_2 x_{i2} \ + \ \ldots \ + \ b_{ip+1} x_{ip+1}$$

Also recall $\frac{1}{n} \sum \left(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i \right) \ = \ \frac{1}{n} \sum e_i \ = \ 0$
So $\frac{1}{n} \sum \hat{\boldsymbol{y}}_i = \frac{1}{n} \sum \boldsymbol{y}_i = \bar{y}$

Hence,

$$\hat{\boldsymbol{y}}^* = g([1\ \bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_p] = \bar{y}$$

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

(j) [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.

$$\vec{x} = \begin{bmatrix} A \\ B \\ A \\ C \\ \ldots \end{bmatrix} \ , \ \mathbb{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ . & , & , \end{bmatrix} , \ \vec{b} = \left(X^T X \right)^{-1} X^T \vec{y}$$

$$(X^TX) = \begin{bmatrix} 1 & 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ . & , & , \end{bmatrix} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

$$(X^TX)^{-1} = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix}, (X^T\vec{y} = \begin{bmatrix} 1 & 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=A} Y_i \\ \sum_{i=B} Y_i \\ \sum_{i=C} Y_i \end{bmatrix}$$

$$\vec{b} = (X^TX)^{-1}(X^T\vec{y} = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix} \begin{bmatrix} \sum_{i=A} Y_i \\ \sum_{i=B} Y_i \\ \sum_{i=C} Y_i \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$

Problem 3

These are questions related to the concept of orthogonal projection, QR decomposition and its relationship with least squares linear modeling.

(a) [difficult] [MA] Prove that if a square matrix is both symmetric and idempotent then it must be an orthogonal projection matrix.

(b) [easy] Prove that I_n is an orthogonal projection matrix $\forall n$.

Assume I_n is an orthogonal projection matrix then I_n is symmetrical and idempotent

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \text{ then } I_n^T = I_n \text{ meaning } \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

And finally the idempotency $I_nI_n = I_n$

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(c) [easy] What subspace does I_n project onto?

It projects onto the colspace of I_n

(d) [easy] Consider least squares linear regression using a design matrix X with rank $p+1$. What are the degrees of freedom in the resulting model? What does this mean?

Since there are p+1 degrees of freedom thus $\hat{\boldsymbol{y}} = w_0 + w_1 x_1 + \cdots + w_p x_p$ would have p+1 weight parameters which can be adjusted

(e) [harder] If you are orthogonally projecting the vector $\boldsymbol{y}$ onto the column space of X which is of rank $p+1$, derive the formula for $\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}]$. Is this the same as in OLS?

$\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = X\vec{w}$
$X^T(\boldsymbol{y} - X\vec{w}) = 0$ Because of orthogonality
$X^T\boldsymbol{y} - X^T X\vec{w} = 0$
$X^T\boldsymbol{y} = X^T X\vec{w}$
$\vec{w} = (X^T X)^{-1} X^T \boldsymbol{y}$
$\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = X\vec{w} = X(X^T X)^{-1} X^T \boldsymbol{y} = H\boldsymbol{y}$
Since this is the same as OLS, yes it is.

(f) [difficult] We saw that the perceptron is an *iterative algorithm*. This means that it goes through multiple iterations in order to converge to a closer and closer $\boldsymbol{w}$. Why not do the same with linear least squares regression? Consider the following. Regress $\boldsymbol{y}$ using $\boldsymbol{X}$ to get $\hat{\boldsymbol{y}}$. This generates residuals $\boldsymbol{e}$ (the leftover piece of $\boldsymbol{y}$ that wasn't explained by the regression's fit, $\hat{\boldsymbol{y}}$). Now try again! Regress $\boldsymbol{e}$ using $\boldsymbol{X}$ and then get new residuals $\boldsymbol{e}_{new}$. Would $\boldsymbol{e}_{new}$ be closer to $\mathbf{0}_n$ than the first $\boldsymbol{e}$? That is, wouldn't this yield a better model on iteration #2? Yes/no and explain.

$H = X(X^T X)^{-1} X^T$ and $H \cdot H = H$ Since the projection onto the X gives the least square error on the 1st iteration. Due to the idem-potency of the projection matrix, you would get the same thing over and over again.

(g) [harder] Prove that $\boldsymbol{Q}^\top = \boldsymbol{Q}^{-1}$ where $\boldsymbol{Q}$ is an orthonormal matrix such that $\text{colsp}[\boldsymbol{Q}] = \text{colsp}[\boldsymbol{X}]$ and $\boldsymbol{Q}$ and $\boldsymbol{X}$ are both matrices $\in \mathbb{R}^{n\times(p+1)}$. Hint: this is purely a linear algebra exercise.

$$\begin{bmatrix} \leftarrow & q_{\cdot 1} & \rightarrow \\ \leftarrow & q_{\cdot 2} & \rightarrow \\ & \cdots & \\ \leftarrow & q_{\cdot n} & \rightarrow \end{bmatrix} \cdot \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ q_{\cdot 1} & q_{\cdot 2} & \cdots & q_{\cdot n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} q_{\cdot 1}^\top q_{\cdot 1} & q_{\cdot 1}^\top q_{\cdot 2} & \cdots & q_{\cdot 1}^\top q_{\cdot n} \\ q_{\cdot 2}^\top q_{\cdot 1} & q_{\cdot 2}^\top q_{\cdot 2} & \cdots & q_{\cdot 2}^\top q_{\cdot n} \\ \cdots & \cdots & \cdots & \cdots \\ q_{\cdot n}^\top q_{\cdot 1} & q_{\cdot n}^\top q_{\cdot 2} & \cdots & q_{\cdot n}^\top q_{\cdot n} \end{bmatrix}$$

$q_{\cdot i}^\top q_{\cdot i} = ||q_{\cdot i}||^2$, and $q_{\cdot i}^\top q_{\cdot j} = 0$ when $i \neq j$ because of orthonormality of Q.

$$Q^\top Q = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{bmatrix} = I$$

(h) [harder] Prove that the least squares projection $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{Q}\boldsymbol{Q}^\top$.

Since $Q^\top Q = I$

$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = (QR)\Big((QR)^T QR\Big)^{-1}(QR)^T$$
$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T == QR\big(R^TQ^TQQ\big)^{-1}R^TQ^T$$
$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T == QR\big(R^TIR\big)^{-1}R^TQ^T$$
$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T == QR\big(R^TR\big)^{-1}R^TQ^T$$
$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T == QRR^{-1}(R^T)^{-1}R^TQ^T$$
$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T == QIIQ^T$$
$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{Q}\boldsymbol{Q}^\top.$$

(i) [harder] Prove that an orthogonal projection onto the colsp $[\boldsymbol{Q}]$ is the same as the sum of the projections onto each column of $\boldsymbol{Q}$.

If we do a projection on each column of $\boldsymbol{Q}$: $\text{Proj}_{q_{\cdot i}}[\vec{a}] = \frac{q_{\cdot i}q_{\cdot i}^\top}{||q_{\cdot i}||^2}\vec{a}$

Recall Q is orthonormal, $||q_{\cdot i}||^2 = 1$, and $\text{Proj}_{q_{\cdot i}}[\vec{a}] = q_{\cdot i}q_{\cdot i}^\top\vec{a}$
So

$$\sum_{i=1}^{p+1}\text{Proj}_{q_{\cdot i}}[\vec{a}] = \sum_{i=1}^{p+1} q_{\cdot i}q_{\cdot i}^\top\vec{a} \;=\; QQ^\top\vec{a}$$

Due to orthogonality of all the columns in $\boldsymbol{Q}$

(j) [easy] Prove that adding a new column to $\boldsymbol{X}$ results in SST remaining the same.

(k) [difficult] [MA] Prove that $\text{rank}[\boldsymbol{H}] = \text{tr}[\boldsymbol{H}]$. Hint: you will need to use facts about eigenvalues and the eigendecomposition of projection matrices that we learned in class.

Problem 4

All of these are extra credit. This is for students who want to get a taste of a first year linear model theory class at the graduate level. The prereq to do these problems is Math 368/621. Only attempt these if you have time!

In linear modeling, $\mathcal{H} = \{\boldsymbol{x}\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^{p+1}\}$ where $\boldsymbol{x} = [1\ x_1\ \ldots\ x_p]$, a row vector. Thus, there is a best function $h^*(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$ where $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \ldots\ \beta_p]^\top$, a column vector and $y = h^*(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \mathcal{E}$. Imagine that for all n observations in $\mathbb{D}$, the $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$ where $\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n(\boldsymbol{0}_n, \sigma^2\boldsymbol{I}_n)$ and $\boldsymbol{Y}$ is a random vector with dimension n modeling the responses of which $\boldsymbol{y}$ is a random realization. Assume σ^2 is known.

(a) [E.C.] Show that $\boldsymbol{Y} \sim \mathcal{N}_n\left(X\boldsymbol{\beta},\, \sigma^2 \boldsymbol{I}_n\right)$.

(b) [E.C.] Let $\boldsymbol{B} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, i.e. the r.v. that represents the OLS estimator of which $\boldsymbol{b}$ is one realization which changes based on the realizations of the error-vector r.v. $\boldsymbol{\mathcal{E}}$. Find the distribution of $\boldsymbol{B}$ and once this is done, its expectation and variance-covariance matrix. Do the entries in $\boldsymbol{B}$ have dependence?

(c) [E.C.] Find the distribution of $\hat{\boldsymbol{Y}}$, the vector r.v. of predictions.

(d) [E.C.] Find the distribution of $\boldsymbol{E}$, the vector r.v. of residuals.

(e) [E.C.] Find the distribution of SST.

(f) [E.C.] Find the distribution of SSE.

(g) [E.C.] Find the distribution of SSR.

(h) [E.C.] Find the distribution of R^2.

(i) [E.C.] Now let σ^2 be unknown. Use the MSE as its estimate. What is the distribution of $\boldsymbol{B}$ now?

(j) [E.C.] What is the distribution of MSE?

(k) [E.C.] What is the distribution of R^2?

(l) [E.C.] Let $\boldsymbol{U} \sim \mathcal{N}_n\left(\boldsymbol{0}_n,\, \boldsymbol{I}_n\right)$ independent of $\boldsymbol{V} \sim \mathcal{N}_n\left(\boldsymbol{0}_n,\, \boldsymbol{I}_n\right)$. Let θ be the r.v. model of the angle between $\boldsymbol{U}$ and $\boldsymbol{V}$. How is θ distributed?