MATH 342W / 650.4 Spring 2024 Homework #2

Laasya Indrakanti

Wednesday 28th 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.

The approach of the hedgehog makes you stick to one theory so all possible variables aren't captured whereas the fox approach allows for varying ideas and this results in more accurate models. Tetlock implies that fox analysis results in more accurate predictions in more complicated real life phenomena in politics and history as well.

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

He likes them because hedgehogs are straightforward as opposed to the complex and nuanced solutions given by fox. A lot of people must think that way because they like exact answers.

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

because more education introduces confirmation bias or overconfidence bias.

(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 allow you to make an informed decision based off of probabilities and they can decide how significant their result is rather than just interpreting a class.

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

Nearest neighbor

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

no because performance decreases past a certain age

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

Scouts can use their experience and take into account factors that PECOTA might not.

(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?

It's possible that at that time, the technology or sufficient data were not available.

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?

SVM hypothesis set is $\mathcal{H} = \{\mathbb{1}_{\vec{w}\cdot\vec{x}-b\geq 0} : \vec{w} \in \mathbb{R}^p, b \in \mathbb{R}\}$

The perceptron learning algorithm has a different hypothesis set of...

$\mathcal{H} = \{\mathbb{1}_{c\vec{w}\cdot\vec{x}\geq 0} : \vec{w} \in \mathbb{R}^{p+1}\}$

(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.

upper limit:

$y_i \in \{-1, 1\}$

$y_i \mathbb{1}_{\vec{w}\cdot\vec{x}-(b+1)\geq 0} = y_i$

$y = -1 \Rightarrow \vec{w}\cdot\vec{x_i} - (b+1) \geq 0 \Rightarrow \vec{w}\cdot\vec{x_i} - b \geq 1$

$y_i(\vec{w}\cdot\vec{x_i} - b) \geq 1$

lower limit:

$y_i \mathbb{1}_{\vec{w}\cdot\vec{x}-(b-1)\leq 0} = y_i$

$y = 1 \Rightarrow \vec{w}\cdot\vec{x_i} - (b-1) \leq 0 \Rightarrow \vec{w}\cdot\vec{x_i} - b \leq -1$

$\Rightarrow -(\vec{w}\cdot\vec{x_i} - b) \geq 1$

$\sum_{i=1}^{n} \max\{0, 1 - y_i(\vec{w}\cdot\vec{x_i} - b)\}$

(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.

$$\underset{\vec{w},b}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1-y_i(\vec{w}\cdot\vec{x_i}-b) + \lambda\|w\|^2)$$

Problem 3

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

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

The algorithm finds the k closest data points to an input, typically using Squared Euclidean Distance, to determine the class or the response of the input based on the most frequent class or the average response of the neighbors. k is a hyperparameter where the default is $k = \sqrt{n}$.

(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).

There should be zero error because the algorithm would assign the value of the nearest neighbor, which would fit perfectly and there would be no distance between the data and the output assignment.This is not a good estimate because the prediction is based off of a single data point that may not be accurate for all occurrences of that input, it may be an outlier.

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 $\mathcal{Y}$?

$\mathbb{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

$\mathcal{X}$ is the one dimensional space of predictors, and $\mathcal{Y}$ is the one dimensional space of responses.

(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.

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

$b_0 = \bar{y} - b_1\bar{x}$

$\hat{y} = b_0 + b_1 x$

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

$\Rightarrow$ when $x = \bar{x}$, $\hat{y} = \bar{y}$, so $\langle \bar{x}, \bar{y} \rangle$ exists on the 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 $\bar{y}$.

$\hat{y}_i = b_0 + b_1 x_i$

$b_0 = \bar{y} - b_1\bar{x}$

$\hat{y} = \bar{y} - b_1\bar{x} + b_1 x_i$

$\frac{1}{n}\sum_{i=1}^{n} \hat{y} = \frac{1}{n}\sum(\bar{y} - b_1\bar{x} + b_1 x_i)$

$\bar{\hat{y}} = \bar{y} - b_1\bar{x} + b_1\bar{x}$

$\bar{\hat{y}} = \bar{y}$

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

$e_i := y_i - \hat{y}_i \qquad \hat{y}_i = b_0 + b_1 x_i$

$SSE := \sum e_i^2 = \sum (y_i - b_0 - b_1 x_i)^2$

$\frac{\partial}{\partial b_0}[SSE] = 2\sum(y_i - b_0 - b_1 x_i) \stackrel{set}{=} 0$

$\sum y_i = n b_0 + b_1 \sum x_i \Rightarrow \sum y_i - \hat{y}_i = 0$

$\Rightarrow \sum e_i = 0$

$\Rightarrow \frac{1}{n}\sum e_i = 0$

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

R^2 is unitless and RMSE takes on the unit of the response so it is more comprehensible. Also, R^2 could be very close to 1 but still have high RMSE so the value of R^2 does not necessarily matter.

(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 $R^2 < 0$.

$\mathbb{D} = \{(1,2), (3,4), (5,6)\}$

$\bar{y} = \frac{2+4+6}{3} = 4$

$\hat{y} = 0$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$$= 1 - \frac{(2-0)^2 + (4-0)^2 + (6-0)^2}{(2-4)^2 + (4-4)^2 + (6-4)^2}$$

$$= 1 - \frac{4+16+36}{4+0+4} = 1 - \frac{56}{8} = 1 - 7 = -6$$

(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).

$$WLS := \sum w_i (y_i - (b_0 + b_1 x_i))^2$$

$$\frac{d}{db_0}[WLS] = \frac{d}{db_0}\left[\sum w_i (y_i - (b_0 + b_1 x_i))^2\right]$$

$$= 2\sum w_i (y_i - (b_0 + b_1 x_i)) \overset{set}{=} 0$$

$$\Rightarrow b_0 = \frac{\sum w_i y_i - b_1 \sum w_i x_i}{\sum w_i}$$

$$\frac{d}{db_1}[WLS] = -2\sum w_i x_i (y_i - (b_0 + b_1 x_i)) \overset{set}{=} 0$$

$$\Rightarrow b_1 = \frac{\sum w_i x_i y_i - \sum w_i x_i \left[\frac{\sum w_i y_i - b_1 \sum w_i x_i}{\sum w_i}\right]}{\sum w_i x_i^2}$$

(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?

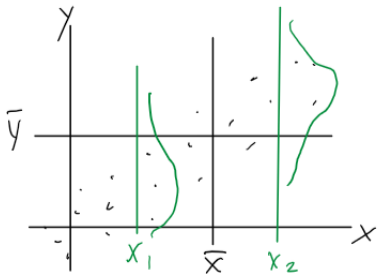
The sums are weighted sums, implying that there is an adjustment by weight to the statistics. It is different than OLS in that b_0 and b_1 are adjusted by weight. This makes sense because it would help reduce residuals.

(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.

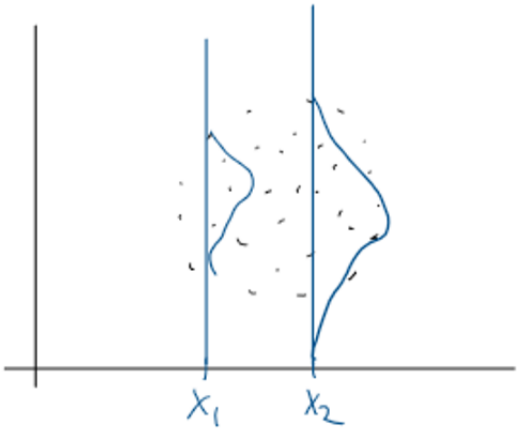
Problem 5

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, variables that are correlated are considered to be associated in linear relationships.

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.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & \cdots & & a_{nn} \end{bmatrix} = \begin{bmatrix} \leftarrow a_{1\cdot} \rightarrow \\ \leftarrow a_{2\cdot} \rightarrow \\ \vdots \\ \leftarrow a_{n\cdot} \rightarrow \end{bmatrix} \qquad \vec{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

$n \times n$ $\qquad$ $n \times 1$

$$A\vec{c} = \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n \\ a_{21}c_1 + a_{22}c_2 + \cdots + a_{2n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2}c_2 + \cdots + a_{nn}c_n \end{bmatrix} = \begin{bmatrix} \vec{a}_{1\cdot}\,\vec{c} \\ \vec{a}_{2\cdot}\,\vec{c} \\ \vdots \\ \vec{a}_{n\cdot}\,\vec{c} \end{bmatrix}$$

$n \times 1$

Let $B = \underbrace{\vec{c}^{\,\top}}_{1\times n}\underbrace{A\vec{c}}_{n\times 1}$ (scalar) $= \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n \\ a_{21}c_1 + a_{22}c_2 + \cdots + a_{2n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2}c_2 + \cdots + a_{nn}c_n \end{bmatrix}$

$$= c_1\left(a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n\right) + c_2\left(a_{21}c_1 + a_{22}c_2 + \cdots + a_{2n}c_n\right) + \cdots + c_n\left(a_{n1}c_1 + a_{n2}c_2 + \cdots + a_{nn}c_n\right)$$

$$\frac{\partial}{\partial c_1}[B] = 2a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n + a_{21}c_2 + a_{31}c_3 + \cdots + a_{n1}c_n$$

(excludes $a_{11}c_1$; excludes $a_{11}c_1$)

$$= \underbrace{a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n}_{} + \underbrace{a_{11}c_1 + a_{21}c_2 + \cdots + a_{n1}c_n}_{} = a_{1\cdot}\vec{c} + a_{1}\vec{c}$$

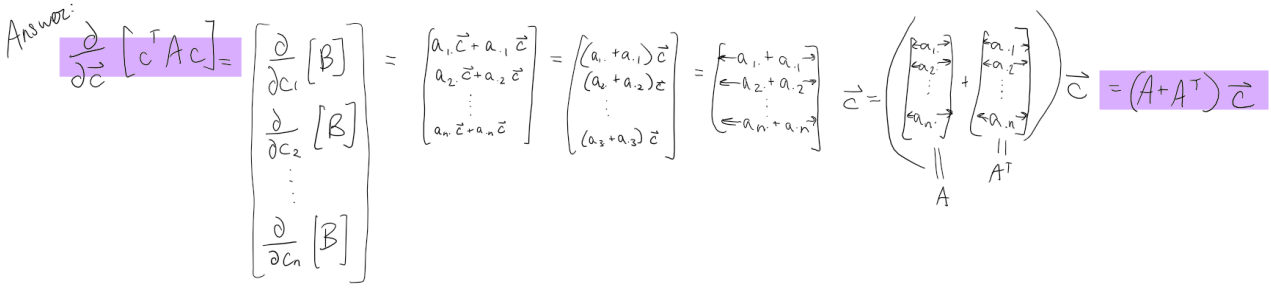
$$\frac{\partial}{\partial c_2}[B] = a_{12}c_1 + a_{21}c_1 + 2a_{22}c_2 + \cdots + a_{2n}c_n + a_{32}c_3 + a_{n2}c_4 + \cdots + a_{n2}c_n$$

$$= 2a_{22}c_2 + a_{21}c_1 + a_{23}c_3 + a_{24}c_4 + \cdots + a_{2n}c_n + a_{12}c_1 + a_{32}c_3 + a_{42}c_4 + \cdots + a_{n2}c_n$$

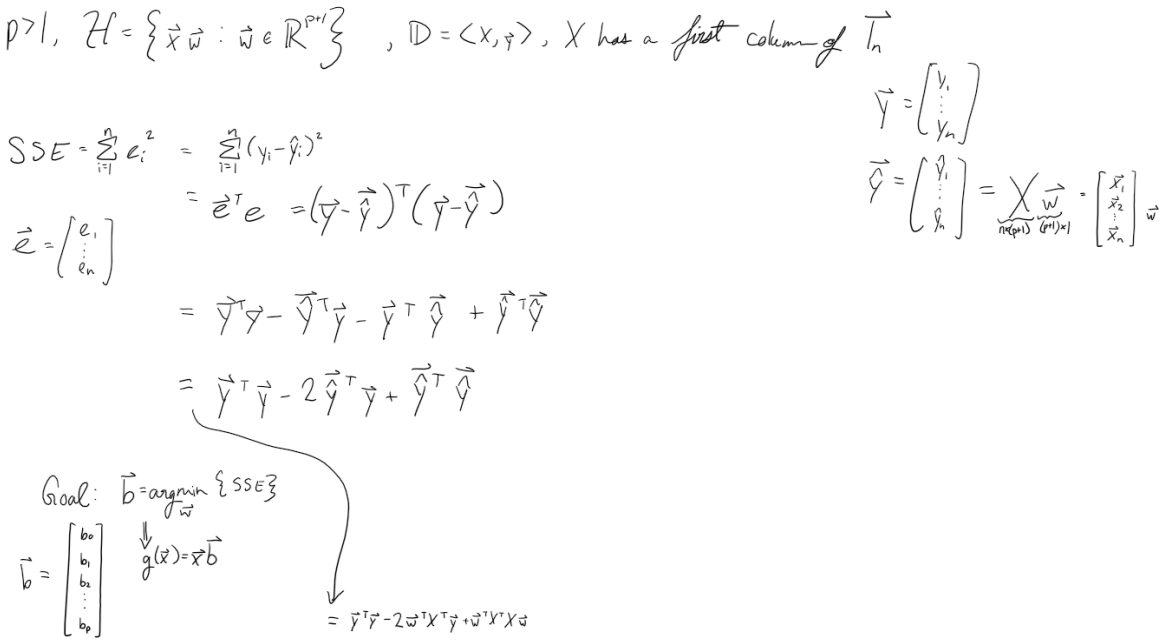
(excludes $a_{22}c_2$; excludes $a_{22}c_2$)

$$= \underbrace{a_{21}c_1 + a_{22}c_2 + \cdots + a_{2n}c_n}_{} + \underbrace{a_{12}c_1 + a_{22}c_2 + \cdots + a_{n2}c_n}_{} = a_{2\cdot}\vec{c} + a_{\cdot 2}\vec{c}$$

$$\frac{\partial}{\partial c_n}[B] = \underbrace{a_{n1}c_1 + a_{n2}c_2 + \cdots + a_{nn}c_n}_{} + \underbrace{a_{1n}c_1 + a_{2n}c_2 + \cdots + a_{nn}c_n}_{} = a_{n\cdot}\vec{c} + a_{\cdot n}\vec{c}$$



(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.



$$\frac{\partial}{\partial \vec{w}}\left[SSE\right] \overset{set}{=} \vec{0}_{p+1}$$

Rule 2

$$\Rightarrow \frac{\partial}{\partial \vec{w}}\left[\vec{y}^T\vec{y} - 2\vec{w}^T X^T\vec{y} + \vec{w}^T X^T X\vec{w}\right] = \underbrace{\vec{0}_{p+1}}_{\text{Rule 0}} - 2\underbrace{\frac{\partial}{\partial \vec{w}}\left[\vec{w}^T X^T\vec{y}\right]}_{\text{Rule 1}} + \underbrace{\frac{\partial}{\partial \vec{w}}\left[\vec{w}^T X^T X\vec{w}\right]}_{\text{Rule 3}}$$

$$= -2X^T\vec{y} + 2X^T X\vec{w} \overset{set}{=} \vec{0}_{p+1}$$

$$\Rightarrow X^T X\vec{w} = X^T\vec{y}$$

$$\Rightarrow \underbrace{(X^T X)^{-1}X^T X}_{I_{p+1}}\vec{w} = (X^T X)^{-1}X^T\vec{y}$$

$$\Rightarrow \boxed{\vec{w} = \vec{b} = (X^T X)^{-1}X^T\vec{y}}$$

(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 $b_1 = r\frac{s_y}{s_x}$.

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

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}$$

$$b = (X^T X)^{-1}X^T\vec{y} = \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2}\underbrace{\begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}}_{2\times 2}\underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}}_{2\times n}\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n\times 1} \Rightarrow 2\times 1$$

$$= \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 - y_1 \sum x_i & \sum x_i^2 - x_2 \sum x_i & \dots & \sum x_i^2 - x_n \sum x_i \\ -\sum x_i + nx_1 & -\sum x_i + nx_2 & \dots & -\sum x_i + nx_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} y_1 \sum x_i^2 + y_2 \sum x_i^2 + \dots + y_n \sum x_i^2 - y_1 x_1 n\bar{x} - y_2 x_2 n\bar{x} - \dots - y_n x_n n\bar{x} \\ -n\bar{x} y_1 - n\bar{x} y_2 - \dots - n\bar{x} y_n + n x_1 y_1 + n x_2 y_2 + \dots + n x_n y_n \end{bmatrix}$$

$$\beta_1 = r\frac{s_y}{s_x} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \cdot \frac{\sqrt{\frac{1}{n-1}\sum (y_i - \bar{y})^2}}{\sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_1 = \frac{-n\sum_{i=1}^{n} \bar{x} y_i + n\sum_{i=1}^{n} x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - \sum \bar{x} y_i}{\sum x_i^2 - \frac{1}{n}\sum x_i \sum x_i} = \frac{\sum x_i y_i - \sum \bar{x} y_i}{\sum x_i^2 - \bar{x}\sum x_i} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$ ✓

$$b_0 = \frac{\sum_{i=1}^{n}\left(y_i\left(\sum x_i^2\right)\right) - \sum_{i=1}^{n}\left(y_i x_i \left(\sum x_i\right)\right)}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = \bar{y} - \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\bar{x}$$

??? womp womp

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

only consider unique columns

(e) [difficult] Prove rank $[X]$ = rank $[X^\top X]$.

if u is in the null space of $X^\top X$ so $X^\top X u = 0 \implies u^\top X^\top X u = u^\top 0 = 0$

if u is not in the null space of X, then $Xu = v \implies u^\top X^\top = v^\top$

$u^\top X^\top X u = 0 \implies v^\top v = 0 \implies v = Xu = 0$

$\therefore$ null$(X^\top X) \subseteq$ null(X)

the reverse isn't important so null$(X^\top X)$ = null(X).

Since Nullity of A + Rank of A = Total number of columns in A, rank$(X^\top X)$ = rank(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.

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

$$\bar{x} = \begin{bmatrix} 1 \\ \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

$$g(\bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \hat{\beta}_3 \bar{x}_3 + \ldots + \hat{\beta}_p \bar{x}_p$$

$$\hat{\beta}_0 = \bar{y} - \sum_{i=1}^{p} \hat{\beta}_i \bar{x}_i$$

$$g(\bar{x}) = \bar{y} - \sum_{i=1}^{p} \hat{\beta}_i \bar{x}_i + \sum_{i=1}^{p} \hat{\beta}_i \bar{x}_i = \bar{y}$$

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

$$e_i = y_i - \hat{y}_i$$

$$\bar{e} = \frac{1}{n} \sum (y_i - \hat{y}_i) = 0$$

$$\sum e_i = 0$$

or

$$\sum (y_i - \hat{y}_i) = 0$$

(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.

$$\vec{b} = (X^T X)^{-1} X^T \vec{y} \qquad X = \begin{bmatrix} \mathbb{1}_{i=A} & \mathbb{1}_{i=B} & \mathbb{1}_{i=C} \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{matrix} \\ \uparrow n_A \downarrow \\ \\ \uparrow n_B \downarrow \\ \\ \uparrow n_C \downarrow \end{matrix}$$

$$X^T X = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix}$$

$$(X^T X)^{-1} X^T \vec{y} = (X^T X)^{-1} \begin{bmatrix} \sum_{i:A} y_i \\ \sum_{i:B} y_i \\ \sum_{i:C} y_i \end{bmatrix} = \begin{bmatrix} \frac{\sum_A y_i}{n_A} \\ \frac{\sum_B y_i}{n_B} \\ \frac{\sum_C y_i}{n_C} \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$

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