MATH 342W / 650.4 Spring 2021 Homework #2

Professor Adam Kapelner

Due 11:59PM Sunday, March 7, 2021 by email

(this document last updated 7:08pm on Wednesday 3rd March, 2021)

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.* 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 and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: Meihe Liu

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.

To fit a model for a phenomenon y, a fox would take a lot of proxies x_i's and may try many different algorithms.
A hedgehog would concentrate on only a few of proxies x_i's so they could make bold predictions.

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

Harry Truman likes hedgehogs because he likes them having simple answers for complicated problems.

I think many people agree with their assumption.

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

(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 are better than vallina classifiers because probabilistic classifiers produce a better distribution of outcomes that represents the δ i.e the uncertainty of the real world.

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

k nearest neighbor (kNN)

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

No baseball performance as a function of age is not a lineal model. There are many other factors z_i may affect the baseball performance of a player.
The performance of a baseball player will hit a peak in a certain period of time and decline until the end of career.

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

Baseball scouts can do better than a prediction system because they have more information about the players (i.e factors z_i's for modeling) and use a hybrid approach so they would have much less δ (i.e uncertainties).

(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?

Because there wasn't enough data collected for the predict algorithm.

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?

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

It is different than the $\mathcal{H}$ used for $\mathcal{A}$ = perceptron learning because the SVM has an intercept b but perceptron doesn't.

(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.

① All $y=1$'s are above or equal to the upper line.

for $\forall\, y_i = 1$

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

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

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

② All $y=-1$'s are below or equal to the lower line

for $\forall\, y_i = -1$

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

$$\vec{w}\cdot\vec{x_i} - b \leq -1$$

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

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

(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.

$$H_i := \max\{0, y_i(\vec{w}\cdot\vec{x_i} - b)\}$$

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

$$\underset{\vec{w}, b}{\operatorname{argmin}} \left\{\frac{1}{n} SHE + \lambda\|\vec{w}\|^2\right\}$$

Problem 3

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

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

A model that predicts on a new x_* by looking through the training data and finding the "closest" $\vec{x_i}$ and returning its y_i as the predicted response value is called a "nearest neighbor" model. Further, a model that finds the k closest observations and return the mode of these k observations as the predicted response value is called "k nearest neighbors" (KNN)
k is a natural number parameter.

(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).

when $k=1$, the algorithm will choose a training sample which is closest to the test sample, in this case, X_i is compared to itself. Therefore there will be zero error.

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 $\mathcal{Y}$?

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\mathbb{D} = \{ < X, \vec{y} > \}$$

(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.

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

$$y = b_0 + b_1 x \Rightarrow \quad b_0 + b_1 \bar{x} = \bar{y} - b_1 \bar{x} + b_1 \bar{x} = \bar{y}$$

$\therefore$ $< \bar{x}, \bar{y} >$ is on this 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 $\bar{y}$.

$$\bar{\hat{y}} = \frac{\sum \hat{y}_i}{n} = \frac{\sum g(x_i)}{n} = \frac{\sum b_0 + b_1 x_i}{n} = \frac{\sum b_0 + b_1 \sum x_i}{n} = \frac{n}{n} b_0 + b_1 \bar{x} = \bar{y}$$

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

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

$$\bar{e}_i = \frac{\sum e_i}{n} = \frac{\sum y_i - \hat{y}_i}{n} = \frac{\sum y_i}{n} - \frac{\sum \hat{y}_i}{n} = \bar{y} - \bar{\hat{y}}_i$$

from (c) we know that $\bar{\hat{y}} = \bar{\hat{y}}_i$

$\therefore \bar{e}_i = \bar{y} - \bar{y} = 0$

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

RMSE indicates the absolute fit of the model to the data — how close the observed data points are to the model's predicted values. Whereas R^2 is a relative measure of fit

(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 $R^2 < 0$.

R^2 can be negative when $S_e^2 > S_y^2$ meaning the model is predicting worse than $g_0 = \bar{y}$

(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).

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum w_i (y_i - b_0 - b_1 x_i)^2$$

$$= \sum w_i y_i^2 + b_0^2 \sum w_i + b_1^2 \sum w_i x_i - 2b_0 \sum w_i y_i - 2b_1 \sum w_i x_i y_i + 2b_0 b_1 \sum w_i x_i$$

$$\frac{\partial SSE}{\partial b_0} = 2b_0 \sum w_i - 2 \sum w_i y_i + 2b_1 \sum w_i x_i$$

set to $0 \rightarrow \quad 0 = b_0 \sum w_i - \sum w_i y_i + b_1 \sum w_i x_i$

$$b_0 \sum w_i = \sum w_i y_i - b_1 \sum w_i x_i$$

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

$$\frac{\partial SSE}{\partial b_1} = 2b_1 \sum w_i x_i - 2 \sum w_i x_i y_i + 2b_0 \sum w_i x_i$$

Set $0 \rightarrow \quad 0 = b_1 \sum w_i x_i - \sum w_i x_i y_i + b_0 \sum w_i x_i$

$$b_1 \sum w_i x_i = \sum w_i x_i y_i - \left(\frac{\sum w_i y_i - b_1 \sum w_i x_i}{\sum w_i} \right) \sum w_i x_i$$

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

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

(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?

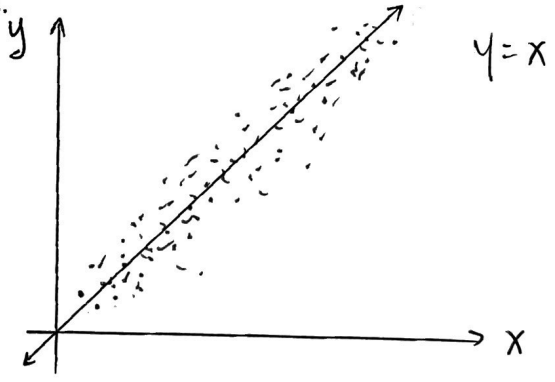
we are using the same idea deriving b_0 and b_1 as we did for OLS except now we have weight associate with the errors. This makes sense because it can adjust the slope and intercept of our model.

(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.

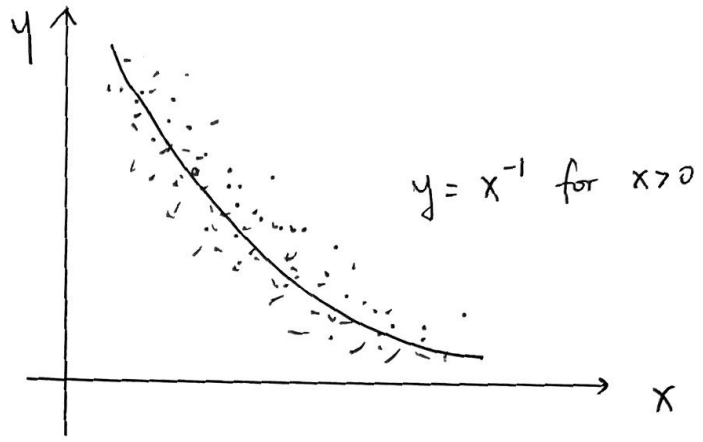
Problem 5

These are questions about association and correlation.

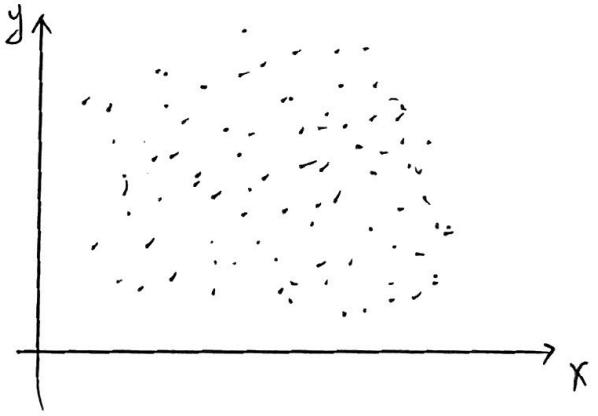
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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, to be correlated it has to be associated.

Problem 6

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

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

$$[c^T A c] = [c_1 \cdots c_n] \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = [(a_{11}c_1 + \cdots + a_{n1}c_n) \cdots (a_{1n}c_1 + \cdots + a_{nn}c_n)] \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$$

$$= \left[\sum_{i=1}^n a_{i1}c_i \cdots \sum_{i=1}^n a_{in}c_i\right] \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = c_1 \sum_{i=1}^n a_{i1}c_i + \cdots + c_n \sum_{i=1}^n a_{in}c_i$$

$$= \sum_{j=1}^n c_j \sum_{i=1}^n a_{ij}c_i = \sum_{j=1}^n \sum_{i=1}^n a_{ij}c_i c_j$$

$$\frac{\partial}{\partial c_p}[c^T A c] = \frac{\partial c^T A c}{\partial c_p} = \frac{\partial}{\partial c_p}\left(\sum_{j=1}^n \sum_{i=1}^n a_{ij}c_i c_j\right) = \frac{\partial}{\partial c_p}\left(c_1 \sum_{i=1}^n a_{i1}c_i + \cdots + c_p \sum_{i=1}^n a_{ip}c_i + \cdots + c_n \sum_{i=1}^n a_{in}c_i\right)$$

$$= c_1 a_{p1} + \cdots + \left(\sum_{i=1}^n a_{ip}c_i + c_p a_{pp}\right) + \cdots + c_n a_{pn} = \sum_{j=1}^n a_{pj}c_j + \sum_{i=1}^n a_{ip}c_i$$

$$= (p^{th} \text{ row of } A)\, c + (p^{th} \text{ row of } A)^T c = [(p^{th} \text{ row of } A) + (p^{th} \text{ row of } A)^T]\, c$$

$$\Rightarrow \frac{\partial}{\partial c}[c^T A c] = (A + A^T)\, c$$

(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.

$$SSE = \vec{y}^T \vec{y} - 2\vec{w}^T X^T \vec{y} + \vec{w}^T X^T X \vec{w}$$

$$\frac{\partial SSE}{\partial \vec{w}} = \frac{\partial}{\partial \vec{w}}[\vec{y}^T \vec{y}] - 2\frac{\partial}{\partial \vec{w}}[\vec{w}(X^T \vec{y})] + \frac{\partial}{\partial \vec{w}}[\vec{w}^T (X^T X)\vec{w}]$$

$$= -2X^T y + \frac{\partial}{\partial \vec{w}}[\vec{w}^T (\vec{X}^T X)\vec{w}]$$

$$= -2X^T y + 2X^T X \vec{w}$$

set $\vec{0}_{p+1}$ and solve for b

$$\Rightarrow -X^T y + X^T X \vec{w} = 0$$

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

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

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

6.C) $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}$ $n \times 2$ $\quad X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}$ $2 \times n$

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

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

$$(X^T X)^{-1} X^T = \frac{1}{n\sum(x_i^2) - (\sum x_i)^2} \begin{bmatrix} \sum(x_i^2) & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

$$= \frac{1}{n\sum(x_i^2) - (\sum x_i)^2} \begin{bmatrix} \sum(x_i^2) - x_1\sum x_i & \cdots & \sum(x_i^2) - x_n\sum x_i \\ -\sum x_i + n x_1 & \cdots & -\sum x_i + n x_n \end{bmatrix}$$

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

$$= \frac{1}{n\sum(x_i^2) - (\sum x_i)^2} \begin{bmatrix} y_1(\sum(x_i^2) - x_1\sum x_i) + \cdots + y_n(\sum(x_i^2) - x_n\sum x_i) & \leftarrow b_0 \\ y_1(-\sum x_i + n x_1) + \cdots + y_n(-\sum x_i + n x_n) & \leftarrow b_1 \end{bmatrix}$$

$$b_1 = \frac{-y_1\sum x_i + n x_1 y_1 - y_2\sum x_i + n x_2 y_2 - \cdots - y_n\sum x_i + n x_n y_n}{n\sum(x_i^2) - n^2\bar{x}^2}$$

$$= \frac{-y_1\sum x_i - y_2\sum x_i - \cdots - y_n\sum x_i + n x_1 y_1 + n x_2 y_2 + \cdots + n x_n y_n}{n(\sum(x_i^2) - n\bar{x}^2)}$$

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

$$= \frac{(\sum x_i y_i - n\bar{x}\bar{y})/(n-1)}{(\sum(x_i^2) - n\bar{x}^2)/(n-1)} = \frac{S_{x,y}}{S_x^2} = \boxed{\frac{r S_y}{S_x}}$$

$$b_0 = \frac{y_1(\sum(x_i^2) - x_1\sum x_i) + \cdots + y_n(\sum(x_i^2) - x_n\sum x_i)}{n\sum(x_i^2) - (\sum x_i)^2}$$

$$= \frac{y_1\sum(x_i^2) - x_1y_1\sum x_i + \cdots + y_n\sum(x_i^2) - x_ny_n\sum x_i}{n\sum(x_i^2) - n^2\bar{x}^2}$$

$$= \frac{y_1\sum(x_i^2) + \cdots y_n\sum(x_i^2) - (x_1y_1\sum x_i + \cdots + x_ny_n\sum x_i)}{n(\sum(x_i^2) - n\bar{x}^2)}$$

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

$\therefore \frac{S_{x,y}}{S_x^2} = \frac{\sum x_iy_i - n\bar{x}\bar{y}}{\sum(x_i^2) - n\bar{x}^2}$ so we can use S_x^2 for denominator b/c we're looking for $S_{x,y}$ as numerator also.

$$\Rightarrow \frac{\bar{y}\sum(x_i^2)}{S_x^2} - \frac{\bar{x}\sum x_iy_i}{S_x^2} + \frac{n\bar{x}^2\bar{y}}{S_x^2} - \frac{n\bar{x}^2\bar{y}}{S_x^2}$$

$$= \frac{\bar{y}\sum(x_i^2) - n\bar{x}^2\bar{y}}{S_x^2} - \frac{\bar{x}\sum x_iy_i + n\bar{x}^2\bar{y}}{S_x^2} = \bar{y}\frac{(\sum(x_i^2) - n\bar{x}^2)}{S_x^2} - \bar{x}\frac{(\sum x_iy_i - n\bar{x}\bar{y})}{S_x^2}$$

$$= \bar{y}\frac{S_x^2}{S_x^2} - \bar{x}\frac{S_{x,y}}{S_x^2} = \bar{y} - \frac{S_{x,y}}{S_x^2}\bar{x} = \boxed{\bar{y} - r\frac{s_y}{s_x}\bar{x}}$$

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

If X is rank deficient, then its columns are linearly dependent which means there are some columns that are multiplications of other columns. So we make X full rank and then solve for $b = (X^T X)^{-1} X^T \vec{y}$

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

Let A be a $n \times 1$ matrix and B be a $m \times n$ matrix

claim: $X^T X w = 0$ iff $Xw = 0$

proof: Suppose $Xw = 0$, by multiplying both sides by X^T. ∎

then we get $X^T X w = 0$

Suppose $X^T X w = 0$ and let $Y = Xw$

$X^T Y = 0$

$w^T X^T Y = 0$

$(Xw)^T Y = 0$

$Y^T Y = 0$

$\Sigma y_i = 0$

∴ all y_i's must be 0.

so Y is 0 and so $Xw = 0$ ∎

$\Rightarrow X^T X$ and X have the same null space

$\Rightarrow$ they have the same nullity

and by the nullity-rank thm we get

$n = \text{Nul}[X^T X] + \text{rank}[X^T X] = \text{Nul}[X] + \text{rank}[X]$

$\Rightarrow \boxed{\text{rank}[X] = \text{rank}[X^T X]}$ ∎

(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.

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

$$g(\vec{\bar{x}}_*) = \vec{\bar{y}}_* = \vec{\bar{x}}_* \vec{b}$$
$$= b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 + \cdots + b_p\bar{x}_p$$
$$= \frac{1}{n}\sum_{i=1}^{n} b_0 + \frac{1}{n}\sum_{i=1}^{n} b_1 x_{i1} + \frac{1}{n}\sum_{i=1}^{n} b_2 x_{i2} + \cdots + \frac{1}{n}\sum b_p x_{ip}$$
$$= \frac{1}{n}\sum \hat{y}_i$$
$$= \bar{y}$$

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

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_{i,1} - \cdots - b_k x_{i,k})^2$$
$$\frac{\partial SSE}{\partial b_0} = 2\sum (y_i - b_0 - b_1 x_{i,1} - \cdots - b_k x_{i,k})(-1)$$
$$= -2\sum e_i$$

Set $0 \Rightarrow$ $0 = -2\sum e_i$

$$0 = \sum e_i$$
$$0 = \frac{\sum e_i}{n}$$
$$\boxed{\bar{e} = 0}$$

(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.

$$\vec{x} = \begin{bmatrix} A \\ A \\ B \\ C \\ \vdots \\ \vdots \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix} \quad \vec{b} = (X^T X)^{-1} X^T \vec{y}$$

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

$$X^T \vec{y} = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=A} y_i \\ \sum_{i=B} y_i \\ \sum_{i=C} y_i \end{bmatrix} \Rightarrow \vec{b} = \begin{bmatrix} 1/n_A & 0 & 0 \\ 0 & 1/n_B & 0 \\ 0 & 0 & 1/n_C \end{bmatrix} \begin{bmatrix} \sum_{i=A} y_i \\ \sum_{i=B} y_i \\ \sum_{i=C} y_i \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$

$g(\vec{x}_*) = \vec{x}_* \vec{b}$ if $x = A$: $g([1\ 0\ 0]) = [1\ 0\ 0] \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} = \bar{y}_A$

if $x = B$: $g([0\ 1\ 0]) = [0\ 1\ 0] \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} = \bar{y}_B$

if $x = C$: $g([0\ 0\ 1]) = [0\ 0\ 1] \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} = \bar{y}_C$

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