MATH 342W / 642 / RM742 Spring 2025 HW #2

Sergio E. Garcia Tapia

Friday 28th February, 2025

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.

A hedgehog will claim a small collection of features $x_1, \ldots, x_p$ accurately captures the information of the true drivers $z_1, \ldots, z_t$. For them, the relation $y = t(z_1, \ldots, z_t) = f(x_1, \ldots, x_p) + \delta$ will have a small misspecification δ. Meanwhile foxes will recognize that p should be much larger to be representative of the z's. Hedgehogs will be confident in the functional form of h^*, so that in the relation $f(x_1, \ldots, x_p)+\delta = h^*(x_1, \ldots, x_p)+\epsilon$, the value of ϵ will be very small. Foxes will be more reserved in this claim since they acknowledge f remains unknown, so they are sensitive to higher misspecification error.

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

Truman liked decisions that were absolute, or a low degree of doubt. Given the impact of decisions about the economy on people's lives, this makes sense; acting on predictions one is confident about is ideal. However, the reality is that no prediction is guaranteed to manifest. and it is this ambiguity that Truman was not content with. Many people are uncomfortable with ambiguity and only want to act on decisions they are 100% confident about. However, we have to accept that ignorance error δ will always be present.

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

It may be because we tend to become more specialized in our focus. For example, mathematics has many areas of research, and no one mathematician could hope to become an expert in all of them. We tend to pick a direction that we find significant

and focus on that. The deeper we engage ourselves, the more confident we may become in the significance of the information we study. We come to believe that there is not much more beyond the features $x_1, \ldots, x_p$ that we have seen, or believe that they are the most impactful, because we have actively shunned focus on much else. At the same time, we become less sensitive to dissent, and the possibility of errors. For example, even if accept that we may never know the functional form f, we may believe h^* is about the best we can do. Thus, in light of new information, we will insist on the information that agrees with this view is correct and reject what does not agree, perhaps dismissing it as noise.

(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.

"Vanilla" classifiers strive to be deterministic. That is, they strive to describe a phenomenon by making an unshakable statement such as $\hat{y} = g(x)$. In reality, there is much noise ϵ in the data, and having such a high degree of confidence in a classification ignores the effect of that noise. Probabilistic classifiers, on the other hand, embrace the noise and express prediction in terms of the likelihood that something will happen. It may be that the probability of the outcome $\hat{y}$ given x is overwhelmingly high, but nevertheless it accepts that $\hat{y}'$ or $\hat{y}''$ might occur instead.

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

It is most like the k-Nearest Neighbors (KNN) algorithm because it predicts how a baseball player will do by comparing it to a set of other players with similar statistics and histories. In fact on page 85, Silver says his algorithm performs nearest neighbor analysis.

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

No. Silver depicts both the quadratic-looking age curve as introduced by James which attempted to describe every player, as well as the noisy curve that attempted to account for different players. In both cases, the trend is not to only increase or only decrease.

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

The dialogue between Silver and Sanders reveals that Sanders believes a good scout is marked by their ability to gather information. This might mean both increasing n, the number of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, as well as increasing p, the number of features $x_1, \ldots, x_p$. The information may be qualitative as opposed to quantitative; it may be more like the mental tools that Silver identified in his conversation with Sanders rather than the Five Tools used by scout which are based on physical aspects. Such data is often hard to quantify and can therefore not always be made available to statistical systems such as PECOTA. A good scout will do better by taking all such information into account without unjustly assigning too heavy a weight to any of the features x_i.

(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?

The information measured by Pitch f/x has a more qualitative essence. It's difficult to assign a weight to a ball's horizontal or vertical movements in predicting the likelihood that a pitcher will strike out a batter.

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 $\mathcal{H}$ in the SVM is the set of hyperplanes in $\mathbb{R}^{p+1}$. Recall that for a fixed $\mathbf{w} \in \mathbb{R}^{p+1}$, a hyperplane is a set of the form

$$\{\mathbf{x} \in \mathbb{R}^{p+1} : \ \mathbf{w} \cdot \mathbf{x} = 0\}$$

Therefore, the set of candidate functions is a union over all such set, except the first entry of $\mathbf{x}$ must be 1:

$$\mathcal{H} = \bigcup_{\mathbf{w} \in \mathbb{R}^{p+1}} \{\mathbf{x} \in \{1\} \times \mathbb{R}^p : \mathbf{w} \cdot \mathbf{x} = 0\}$$

Here the feature vector has been extended to length $p+1$ by prepending a component with the value 1, and correspondingly extending $\mathbf{w}$ to length $p+1$ so that it also contains the bias term.

(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.

Suppose $\mathbb{D}$ is a linearly separable data set. Then there is a hyperplane that separates $\mathbb{R}^p$ into two parts, such that inputs $\mathbf{x}_i$ on one side all have a corresponding output $y_i = 1$, and inputs $\mathbf{x}_j$ on the other side have a corresponding output $y_j = -1$. Let ℓ be a given hyperplane, so that there is a fixed $\mathbf{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$ so that

$$\ell = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{w} \cdot \mathbf{x} - b = 0\}.$$

Let $\delta > 0$ be small enough that the hyperplanes ℓ_U and ℓ_L defined by shifting ℓ by $+\delta$ and $-\delta$, repectively. In particular,

$$\ell_U = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{w} \cdot \mathbf{x} - (b + \delta) = 0\}$$
$$\ell_L = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{w} \cdot \mathbf{x} - (b - \delta) = 0\}.$$

The vector $\mathbf{w}$ is normal to ℓ. Normalize it into a vector

$$\mathbf{w}_0 = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

Let $\mathbf{z} \neq \mathbf{0}$ be a normal vector residing on ℓ. Since $\mathbf{z}$ is normal to ℓ, it must be in $\text{span}(\mathbf{w}_0)$, so there is $\alpha \in \mathbb{R}$ such that $\alpha \neq 0$ and $\mathbf{z} = \alpha\mathbf{w}_0$. Since $\mathbf{z} \in \ell$, it also satisfies its Hesse normal form, so $\mathbf{w} \cdot \mathbf{z} - b = 0$. If we put these together, we have

$$\begin{aligned} \mathbf{w} \cdot \mathbf{z} - b &= 0 \\ \mathbf{w} \cdot (\alpha\mathbf{w}_0) - b &= 0 \\ (\|\mathbf{w}\|\mathbf{w}_0) \cdot (\alpha\mathbf{w}_0) - b &= 0 \\ \alpha &= \frac{b}{\|\mathbf{w}\|} \end{aligned}$$

since $\mathbf{w}_0 \cdot \mathbf{w}_0 = 1$ (it is a unit vector). Thus,

$$\mathbf{z} = \frac{b}{\|\mathbf{w}\|}\mathbf{w}_0$$

If we similarly define $\mathbf{z}_U$ to be a normal vector residing on ℓ_U and $\mathbf{z}_L$ to be a normal vector residing on ℓ_L, then an analogous computation shows that

$$\mathbf{z}_U = \frac{b+\delta}{\|\mathbf{w}\|}\mathbf{w}_0 \qquad\qquad = \frac{b-\delta}{\|\mathbf{w}\|}\mathbf{w}_0$$

Let m be the perpendicular distance between ℓ_L and ℓ_U. We want to maximize m, thereby determining the maximum-margin hyperplane. The value m is precisely given by

$$\begin{aligned} m &= \|\mathbf{z}_U - \mathbf{z}_L\| \\ &= \left\| \frac{b+\delta}{\|\mathbf{w}\|}\mathbf{w}_0 - \frac{b-\delta}{\|\mathbf{w}\|}\mathbf{w}_0 \right\| \\ &= \left\| \frac{2\delta}{\|\mathbf{w}\|}\mathbf{w}_0 \right\| \\ &= \frac{2\delta}{\|\mathbf{w}\|} \end{aligned}$$

since $\|\mathbf{w}_0\| = 1$. Therefore, to maximize m, we need to minimize $\|\mathbf{w}\|$. To obtain the objective function (our condition for minimization), recall that the Hesse normal form for ℓ is overdetermined, so for any $c \neq 0$, we have $c(\mathbf{w} \cdot \mathbf{x} - b) = 0$. If we pick $c = \frac{1}{\delta}$, so that we restrict ourselves to a fixed hyperplane, the equation for ℓ_U gives

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x} - (b+\delta) &= 0 \\ \frac{1}{\delta}\mathbf{w} \cdot \mathbf{x} - \frac{1}{\delta}(b+\delta) &= 0 \end{aligned}$$

$$\underbrace{\frac{\mathbf{w}}{\delta}}_{\mathbf{w}'} \cdot \mathbf{x} - \left(\underbrace{\frac{b}{\delta}}_{b'} + 1 \right) = 0$$
$$\mathbf{w}' \cdot \mathbf{x} - (b' + 1) = 0.$$

Likewise, for ℓ_L, we have $\mathbf{w}' \cdot \mathbf{x} - (b' - 1) = 0$. Now, going back to ℓ_U, we know that for points $\mathbf{x}_i$ above ℓ_U it holds that $y_i = 1$, so for such i it is true that

$$\begin{aligned} \mathbf{w}' \cdot \mathbf{x}_i - (b' + 1) &\geq 0 \\ \mathbf{w}' \cdot \mathbf{x}_i - b' &\geq 1 \\ y_i(\mathbf{w} \cdot \mathbf{w}_i - b') &\geq 1, \quad i \text{ such that } y_i = 1 \end{aligned}$$

since $y_i = 1$. Similarly, for all i such that $y_i = -1$ (those points below ℓ_L), we have

$$\begin{aligned} \mathbf{w}' \cdot \mathbf{x}_i - (b' - 1) &\leq 0 \\ \mathbf{w}' \cdot \mathbf{x}_i - b' &\leq -1 \\ -1(\mathbf{w} \cdot \mathbf{x}_i - b') &\geq (-1)(-1) \\ y_i(\mathbf{w} \cdot \mathbf{x}_i - b') &\geq 1, \quad i \text{ such that } y_i = -1 \end{aligned}$$

where I have multiplied by -1 thus reversing the inequality, and then made use of the fact that $y_i = -1$. In particularly, the same inequality holds for all i, so we may write for all points $(\mathbf{x}_i, y_i) \in \mathbb{D}$ that

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b') \geq 1, \quad \text{for all integers } 1 \leq i \leq n. \tag{1}$$

which is the cost function for SVM in the linearly separable case with $\mathcal{Y} = \{-1, 1\}$.

(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.

If $\mathbb{D}$ is not linearly separable and a point is on the "wrong side" of the hyperplane (it is missclassified), then it will fail to satisfy the objective function we derived in (c). Therefore, such a point $(\mathbf{x}_k, y_k)$ will instead satisfy the inequality

$$y_k(\mathbf{w}' \cdot \mathbf{x}_k - b) < 1$$

In this case, we can let $d_k > 0$ be number of units this expression is below 1. Then

$$\begin{aligned} y_k(\mathbf{w}' \cdot \mathbf{x}_k - b) &= 1 - d_k \\ d_k &= 1 - y_k(\mathbf{w} \cdot \mathbf{x}_k - b) \end{aligned} \tag{2}$$

Defining the hinge error $H_i := \max\{\{\}\, 0, d_i\}$ for all i, and the sum of hinge errors

$$SHE := \sum_{i=1}^{n} H_i$$

the cost function to minimize is

$$\frac{1}{n}SHE + \lambda\|\mathbf{w}'\|$$

where $\mathbf{w}'$ satisfies Equation (1) and the hinge errors $H_i = \{0, d_i\}$ are given by with d_i as in Equation (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 receives three inputs:

- A data set $\mathbb{D}$ with n pairs $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ in $\mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{Y}$ is a finite set.
- A sequence inputs L of the form $\mathbf{x}^* \in \mathcal{X}$.
- A distance function $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$.

The output of the algorithm is

- A sequence of responses $\mathbf{y}^* \in \mathcal{Y}$ corresponding to each $\mathbf{x}^* \in \mathcal{X}$ in the input sequence L.

The KNN algorithm repeats the following steps for each input $\mathbf{x}^*$. It computes the k "closest" points to $\mathbf{x}^*$ by computing the distances $d(\mathbf{x}^*, \mathbf{x}_j)$ for all $\mathbf{x}_j$ belonging to a pair in $\mathbb{D}$. Suppose the indices of the k closest points are $I = \{i_1, \ldots, i_k\} \subseteq \{1, 2, \ldots, n\}$. To be precise, this means

$$d(\mathbf{x}^*, \mathbf{x}_l) \leq d(\mathbf{x}^*, \mathbf{x}_j), \quad \forall l \in I, \quad \forall j \notin I.$$

The heuristic of the algorithm is to predict the response for $\mathbf{x}^*$ to be similar to the response of these inputs. It assigns to $\mathbf{x}^*$ the value $\text{Mode}[y_{i_1}, \ldots, y_{i_k}]$.

(b) [difficult] [MA] Assuming $\mathcal{A} = \text{KNN}$, 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).

It will be zero. If we are predicting a response for $\mathbf{x}^* \in \mathcal{X}$, and if there is $(\mathbf{x}_i, y_i) \in \mathbb{D}$ such that $\mathbf{x}^* = \mathbf{x}_i$ for some integer $1 \leq j \leq n$, then $d(\mathbf{x}^*, \mathbf{x}_i) = 0$, which implies $\mathbf{x}^* = \mathbf{x}_i$. Therefore, KNN would predict $g(\mathbf{x}^*) = y_i$. But recall that our data set $\mathbb{D}$ consists of the points observed and therefore produced by the phenomenon, so $f(\mathbf{x}^*) = f(\mathbf{x}_i) = y_i$.

The fact that g matches f on a given input means that there is zero error, since the error is given by

$$e^* = y^* - g(\mathbf{x}^*) = f(\mathbf{x}^*) - g(\mathbf{x}^*) = y_i - y_i = 0$$

It is not a good estimate of future error. It may be that the response y_i for the input $\mathbf{x}_i$ is exceptional or an outlier, and is thus not representative of what the response should be for similar inputs or even $\mathbf{x}_i$. For example, our features do not capture all of the truly causal information implied by the true drivers $z_1, \ldots, z_t$, so even though $\mathbf{x}_i = \mathbf{x}^*$, it may be that an unknown characteristic differs and the output should differ after all. With $k = 1$, our model is sensitive to this type of situation.

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 set is a collection of points $\mathbb{D}$ in $\mathbb{R}^2$, like a scatterplot. Here, $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$.

(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.

Proof. Suppose we are given a data set $\mathbb{D}$ with n points (x_i, y_i), where $1 \leq i \leq n$. The OLS algorithm computes $\mathbf{b} = \begin{bmatrix} b_1 & b_2 \end{bmatrix}^\top$ whose entries are given by

$$b_0 = \bar{y} - b_1 \bar{x} \tag{3}$$

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \tag{4}$$

$$\tag{5}$$

and makes predictions with the function

$$y^* = g(x^*) = b_0 + b_1 x^*$$

If $x^* = \bar{x}$, then

$$\begin{aligned} g(\bar{x}) &= b_0 + b_1 \bar{x} \\ &= (\bar{y} - b_1 \bar{x}) + b_1 \bar{x}, \quad \text{by Equation 3} \\ &= \bar{y} \end{aligned}$$

Hence, the point $(\bar{x}, \bar{y})$ is on the least square regression line. □

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

Proof. Using the setting of part (b), the residuals are given by

$$e_i = y_i - g(x_i), \quad \forall_i \in \mathbb{N} : 1 \leq i \leq n$$

The average residual is given by $\frac{1}{n}\sum_{i=1}^{n} e_i$, so

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} e_i &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 \cdot 1 - b_1 x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \cdot \frac{1}{n}\sum_{i=1}^{n}(1) - b_1 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \bar{y} - b_0 \cdot \frac{1}{n} \cdot n - b_1 \cdot \bar{x} \\
&= \underbrace{(\bar{y} - b_1 \bar{x})}_{\text{Equation 3}} - b_0 \\
&= b_0 - b_0 \\
&= 0
\end{aligned}$$

□

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

It's possible to have $R^2 \approx 1$, and yet, we may be predicting in a context where due to the scale of the values in question, the $RMSE$ would be large in an absolute sense even if it is relatively small.

(e) [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 $R^2 < 0$.

Fix $\gamma \neq 0$ and $n > 1$, and let $\mathbb{D}$ be a set of n points, namely, $(i, \gamma) \in \mathbb{D}$ for $i < n$, and $(n, 0) \in \mathbb{D}$. Let $g(x) = 0$ be our linear model. First, note the average of the response is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n} \cdot \sum_{i=1}^{n-1} \gamma + \frac{1}{n} \cdot 0 = \frac{n-1}{n}\gamma$$

Thus the null model g_0 predicts $g_0(x) = \frac{n-1}{n}\gamma$. Now the SST is given by

$$SST = SSE_0$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} e_{0,i}^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n-1} (\gamma - \frac{n-1}{n}\gamma)^2 + (0 - \frac{n-1}{n}\gamma)^2 \\
&= (n-1) \cdot \frac{\gamma^2}{n^2} + \frac{(n-1)^2}{n^2}\gamma^2 \\
&= \frac{(n-1)^2 + (n-1)}{n^2}\gamma^2
\end{aligned}
$$

Then for g

$$SSE_g = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n-1} (\gamma - 0)^2 + (0 - 0)^2 = (n-1)\gamma^2$$

and

$$
\begin{aligned}
R^2 &= 1 - \frac{SSE}{SST} \\
&= 1 - \frac{(n-1)\gamma^2}{\frac{(n-1)^2+(n-1)}{n^2}\gamma^2} \\
&= 1 - \frac{1}{\frac{n-1}{n^2}} \\
&= 1 - \frac{n^2}{n-1}
\end{aligned}
$$

Notice that for $n > 1$, we have $n^2 > n - 1$, and hence $R^2 < 0$.

(f) [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).

Given the errors $e_i = y_i - \hat{y}_i$, weighting it amounts to multiplication by w_i. Whereas before we had to minimize the SSE (sum of squared errors), this time we need to minimize the weighted sum of squared errors:

$$SSE_\alpha = \sum_{i=1}^{n} w_i (e_i)^2 = \sum_{i=1}^{n} w_i [y_i - \hat{y}_i]^2$$

In the context of OLS, our prediction is given by $\hat{y}_i = \alpha_0 + \alpha_1 x_i$, for $\begin{bmatrix}\alpha_0 & \alpha_1\end{bmatrix}^\top \in \mathbb{R}^2$, so we can write the equation above as

$$SSE_\alpha = \sum_{i=1}^{n} w_i[y_i - \alpha_0 - \alpha_1 x_i]^2$$

The least square estimates are obtained by minimizing SSE_α, which by taking the partial derivative with respect to both α_0 and α_1 and then setting the to 0, yielding $\alpha_0 = b_0$ and $\alpha_1 = b_1$. We begin with α_0:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \alpha_0}[SSE_\alpha] \\
&= \sum_{i=1}^{n} -2w_i \cdot [y_i - b_0 - b_1 x_i] \\
&= \sum_{i=1}^{n} w_i y_i - b_0 \sum_{i=1}^{n} w_i - b_1 \sum_{i=1}^{n} w_i x_i
\end{aligned}$$

Next, we do the same for α_1:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \alpha_1}[SSE_{\mathbf{w}}] \\
&= \sum_{i=1}^{n} -2w_i x_i [y_i - b_0 - b_1 x_i] \\
&= \sum_{i=1}^{n} w_i x_i y_i - b_0 \sum_{i=1}^{n} w_i x_i - b_1 \sum_{i=1}^{n} w_i x_i^2
\end{aligned}$$

We can now combine these equations into a system of equations:

$$\begin{aligned}
\left(\sum_{i=1}^{n} w_i\right) b_0 + \left(\sum_{i=1}^{n} w_i x_i\right) b_1 &= \sum_{i=1}^{n} w_i y_i \\
\left(\sum_{i=1}^{n} w_i x_i\right) b_0 + \left(\sum_{i=1}^{n} w_i x_i^2\right) b_1 &= \sum_{i=1}^{n} w_i x_i y_i
\end{aligned}$$

We can write it as a matrix equation $A\mathbf{b} = \mathbf{d}$:

$$\begin{bmatrix} q & r \\ r & s \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix}$$

We can solve it by using the inverse and matrix multiplication:

$$\mathbf{b} = A^{-1}\mathbf{d}$$

$$= \frac{1}{\det[A]} \begin{bmatrix} s & -r \\ -r & q \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$
$$= \frac{1}{qs - r^2} \begin{bmatrix} su - rv \\ qv - ru \end{bmatrix}$$

Plugging in the corresponding values we see that

$$b_0 = \frac{\left(\sum_{i=1}^n w_i x_i^2\right)\left(\sum_{i=1}^n w_i y_i\right) - \left(\sum_{i=1}^n w_i x_i\right)\left(\sum_{i=1}^n w_i x_i y_i\right)}{\left(\sum_{i=1}^n w_i\right)\left(\sum_{i=1}^n w_i x_i^2\right) - \left(\sum_{i=1}^n w_i x_i\right)}$$
$$b_1 = \frac{\left(\sum_{i=1}^n w_i\right)\left(\sum_{i=1}^n w_i x_i y_i\right) - \left(\sum_{i=1}^n w_i x_i\right)\left(\sum_{i=1}^n w_i y_i\right)}{\left(\sum_{i=1}^n w_i\right)\left(\sum_{i=1}^n w_i x_i^2\right) - \left(\sum_{i=1}^n w_i x_i\right)^2}$$

(g) [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?

Note that if all errors are equally costly(i.e., $w_i = \frac{1}{n}$ for all i), then we recover the b_0 and b_1 estimates in OLS. Notice that

$$\begin{aligned} b_1 &= \frac{\left(\sum_{i=1}^n w_i\right)\left(\sum_{i=1}^n w_i x_i y_i\right) - \left(\sum_{i=1}^n w_i x_i\right)\left(\sum_{i=1}^n w_i y_i\right)}{\left(\sum_{i=1}^n w_i\right)\left(\sum_{i=1}^n w_i x_i^2\right) - \left(\sum_{i=1}^n w_i x_i\right)^2} \\ &= \frac{\left(\sum_{i=1}^n \frac{1}{n}\right)\left(\sum_{i=1}^n \frac{1}{n} x_i y_i\right) - \left(\sum_{i=1}^n \frac{1}{n} x_i\right)\left(\sum_{i=1}^n \frac{1}{n} y_i\right)}{\left(\sum_{i=1}^n \frac{1}{n}\right)\left(\sum_{i=1}^n \frac{1}{n} x_i^2\right) - \left(\sum_{i=1}^n \frac{1}{n} x_i\right)^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \frac{1}{n^2}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^n x_i\right)^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}(n\bar{x})(n\bar{y})}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(n\bar{x})^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \end{aligned}$$

Similarly,

$$\begin{aligned} b_0 &= \frac{\frac{1}{n}\sum_{i=1}^n x_i^2 \cdot \frac{1}{n}\sum_{i=1}^n y_i - \frac{1}{n}\sum_{i=1}^n x_i \cdot \frac{1}{n}\sum_{i=1}^n x_i y_i}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^n x_i\right)^2} \\ &= \frac{\bar{y}\sum_{i=1}^n x_i^2 - \bar{x} \cdot \sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\ &= \frac{\bar{y}\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} - \bar{x} \cdot \left(b_1 - \frac{n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) \\ &= \frac{\bar{y} \cdot \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} - b_1\bar{x} \\ &= \bar{y} - b_1\bar{x} \end{aligned}$$

Victory.

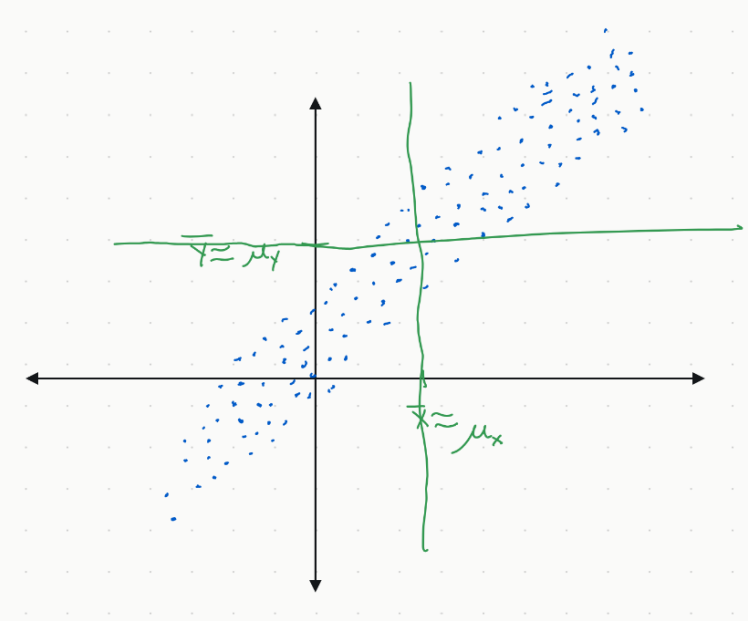

Figure 1: Exercise 5(a): Associated and correlated

(h) [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.

See Figure 1.

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

See Figure 2.

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

See Figure 3.

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

No. If the variables are correlated, then as one increases, the other decreases (or increases). This suggests a relationship between the variables, and hence an association.

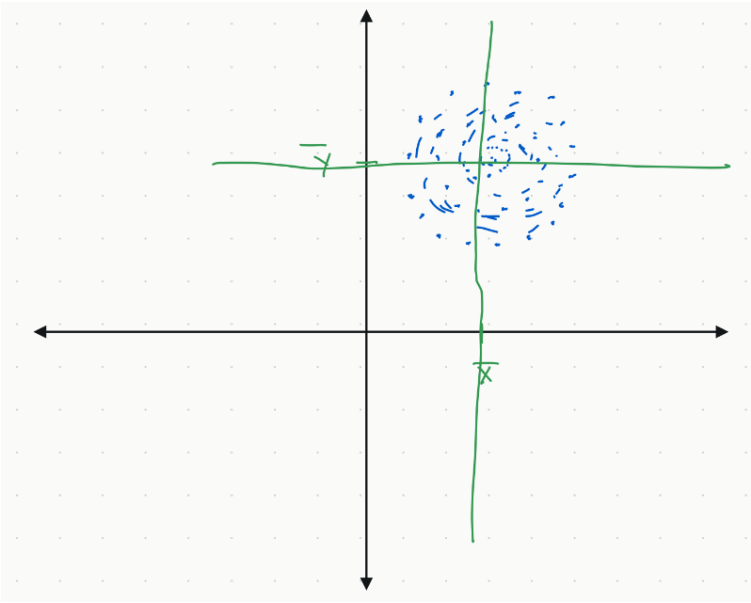

Figure 2: Exercise 5(b): Associated but not correlated

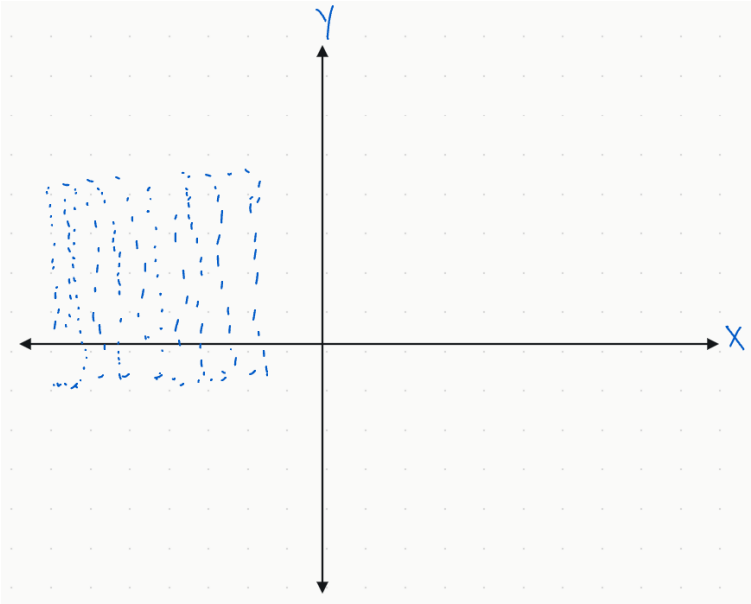

Figure 3: Exercise 5(c): Not associated and not correlated.

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.

Note that by the definition of matrix multiplication, the kth entry of $A\mathbf{c}$ is given by

$$(A\mathbf{c})_k = \sum_{j=1}^{n} a_{k,j} c_j$$

Then

$$\begin{aligned}\mathbf{c}^\top A\mathbf{c} &= \sum_{k=1}^{n} c_k \sum_{j=1}^{n} a_{k,j} c_j \\ &= \sum_{k=1}^{n}\sum_{j=1}^{n} a_{k,j} c_k c_j\end{aligned}$$

To compute the vector derivative, we can focus on the partial derivative with respect to the c_i

$$\begin{aligned}\frac{\partial}{\partial c_i}[\mathbf{c}^\top A\mathbf{c}] &= \frac{\partial}{\partial c_i}\left[\sum_{k=1}^{n}\sum_{j=1}^{n} a_{k,j} c_k c_j\right] \\ &= \sum_{k=1}^{n}\sum_{j=1}^{n} a_{k,j}\frac{\partial}{\partial c_i}[c_k c_j] \\ &= \sum_{k=1}^{n}\sum_{j=1}^{n} a_{k,j}\left[c_j\frac{\partial}{\partial c_i}[c_k] + c_k\frac{\partial}{\partial c_i}[c_j]\right] \\ &= \sum_{k=1}^{n}\sum_{j=1}^{n} a_{k,j} c_j\frac{\partial}{\partial c_i}[c_k] + \sum_{k=1}^{n}\sum_{j=1}^{n} a_{k,j} c_k\frac{\partial}{\partial c_i}[c_j] \\ &= \sum_{j=1} a_{i,j} c_j + \sum_{k=1}^{n} a_{k,i} c_k\end{aligned}$$

Note that the evaluated to 0 whenever $i \neq k$ in the first sum, and whenever $i \neq j$ in the second sum. Now let's express this as a matrix multiplication:

$$\frac{\partial}{\partial c_i}[\mathbf{c}^\top A\mathbf{c}] = \sum_{j=1} a_{i,j} c_j + \sum_{k=1}^{n} a_{k,i} c_k$$

$$= A_{i,\cdot}\mathbf{c} + (A^\top)_{i,\cdot}\mathbf{c}$$

Here, $A_{i,\cdot}$ denotes the ith row of A, and $(A^\top)_{i,\cdot}$ denotes the ith row of $A^\top$. Now

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{c}}[\mathbf{c}^\top A\mathbf{c}] &= \begin{bmatrix} \frac{\partial}{\partial c_1}[\mathbf{c}^\top A\mathbf{c}] \\ \vdots \\ \frac{\partial}{\partial c_n}[\mathbf{c}^\top A\mathbf{c}] \end{bmatrix} \\
&= \begin{bmatrix} A_{1,\cdot}\mathbf{c} + (A^\top)_{1,\cdot}\mathbf{c} \\ \vdots \\ A_{n,\cdot}\mathbf{c} + (A^\top)_{n,\cdot}\mathbf{c} \end{bmatrix} \\
&= \begin{bmatrix} A_{1,\cdot}\mathbf{c} \\ \vdots \\ A_{n,\cdot}\mathbf{c} \end{bmatrix} + \begin{bmatrix} (A^\top)_{1,\cdot}\mathbf{c} \\ \vdots \\ (A^\top)_{n,\cdot}\mathbf{c} \end{bmatrix} \\
&= A\mathbf{c} + A^\top\mathbf{c}
\end{aligned}$$

In the special case where A is symmetric, we can see that we get $2A\mathbf{c}$.

(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.

We are assuming that the set of candidate functions $\mathcal{H}$ is the set of hyperplanes, so we have a prediction function g that can be used to predict $\hat{y}_i$ as follows:

$$\begin{aligned}
\hat{y}_i &= g(\mathbf{x}_i) \\
&= w_0 + w_1 x_{i,1} + \cdots + x_{i,p} \\
&= \begin{bmatrix} 1 & x_{i,1} & \cdots & x_{i,p} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} \\
&= \mathbf{x}_i^\top \mathbf{w}
\end{aligned}$$

Therefore, we can write the prediction vector as

$$\hat{\mathbf{y}} := \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{w} \\ \mathbf{x}_2^\top \mathbf{w} \\ \vdots \\ \mathbf{x}_n^\top \mathbf{w} \end{bmatrix} = X\mathbf{w}$$

The residual errors are given by $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, and hence the SSE is given by

$$\begin{aligned}
SSE &= \sum_{i=1}^{n} e_i \\
&= \mathbf{e}^\top \mathbf{e} \\
&= (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) \\
&= (\mathbf{y}^\top - \hat{\mathbf{y}}^\top)(\mathbf{y} - \hat{\mathbf{y}}) && \text{(Since } (A+B)^\top = A^\top + B^\top\text{)} \\
&= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \hat{\mathbf{y}} - \hat{\mathbf{y}}^\top \mathbf{y} + \hat{\mathbf{y}}^\top \hat{\mathbf{y}} \\
&= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \hat{\mathbf{y}} - (\mathbf{y}^\top \hat{\mathbf{y}})^\top + \hat{\mathbf{y}}^\top \hat{\mathbf{y}} && \text{(Since } (AB)^\top = B^\top A^\top\text{)} \\
&= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \hat{\mathbf{y}} + \hat{\mathbf{y}}^\top \hat{\mathbf{y}} && \text{(Since } \mathbf{y}^\top \hat{\mathbf{y}} \text{ is a scalar)} \\
&= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top X\mathbf{w} + (X\mathbf{w})^\top (X\mathbf{w}) \\
&= \mathbf{y}^\top \mathbf{y} - 2(X^\top \mathbf{y})^\top \mathbf{w} + \mathbf{w}^\top X^\top (X\mathbf{w}) && \text{(Since } (AB)^\top = B^\top A^\top\text{)}
\end{aligned}$$

To minimize this quantity, we differentiate with respect to $\mathbf{w}$:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}[SSE] &= \frac{\partial}{\partial \mathbf{w}} \left[\mathbf{y}^\top \mathbf{y} - 2(X^\top \mathbf{y}^\top)^\top \mathbf{w} + \mathbf{w}^\top X^\top (X\mathbf{w})\right] \\
&= \frac{\partial}{\partial \mathbf{w}}[\mathbf{y}^\top \mathbf{y}] - 2\frac{\partial}{\partial \mathbf{w}}[(X^\top \mathbf{y}^\top)^\top \mathbf{w}] + \frac{\partial}{\partial \mathbf{w}}[\mathbf{w}^\top X^\top (X\mathbf{w})] \\
&= \mathbf{0}_{p+1} - 2 \cdot X^\top \mathbf{y} + 2 \cdot X^\top X\mathbf{w}
\end{aligned}$$

When we set this quantity to $\mathbf{0}_{p+1}$, we arrive at a vector $\mathbf{w} = \mathbf{b}$ that minimizes the SSE:

$$\begin{aligned}
\mathbf{0}_{p+1} &= \mathbf{0}_{p+1} - 2 \cdot X^\top \mathbf{y} + 2 \cdot X^\top X\mathbf{b} \\
X^\top X\mathbf{b} &= X^\top \mathbf{y} \\
\mathbf{b} &= (X^\top X)^{-1} X^\top \mathbf{y}
\end{aligned}$$

provided that $X^\top X$ is invertible, or equivalently, that $\text{rank}[X^\top X] = p + 1$.

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

If $p = 1$, then $X = \begin{bmatrix}\vec{\mathbf{1}}_n & \mathbf{x}_{\cdot,1}\end{bmatrix}$. We know that

$$\mathbf{b} = (X^\top X)^{-1} X^\top \mathbf{y}$$

Now

$$X^\top \mathbf{y} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$
$$= \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

Meanwhile,

$$X^\top X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$
$$= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$
$$= \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

Now we find the inverse of $X^\top X$:

$$X^\top X = \frac{1}{\det(X^\top X)} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$
$$= \frac{1}{n \cdot \sum_{i=1}^n x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & n \end{bmatrix}$$

Now we multiply:

$$\begin{aligned} \mathbf{b} &= (X^\top X)^{-1}(X^\top \mathbf{y}) \\ &= \frac{1}{n \cdot \sum_{i=1}^n x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\ &= \frac{1}{n \cdot \sum_{i=1}^n x_i^2 - n^2\bar{x}^2} \begin{bmatrix} n\bar{y} \sum_{i=1}^n x_i^2 - n\bar{x} \sum_{i=1}^n x_i y_i \\ -n^2\bar{x}\bar{y} + n \sum_{i=1}^n x_i y_i \end{bmatrix} \\ &= \frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \begin{bmatrix} \bar{y} \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i y_i \\ -n\bar{x}\bar{y} + \sum_{i=1}^n x_i y_i \end{bmatrix} \end{aligned}$$

Hence we see that

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \bar{x}^2}$$

as before, and

$$b_0 = \frac{\bar{y} \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

$$
\begin{aligned}
&= \frac{\bar{y}\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} - \bar{x} \cdot \left(b_1 - \frac{n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right) \\
&= \frac{\bar{y} \cdot (\sum_{i=1}^n x_i^2 - n\bar{x}^2)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} - b_1\bar{x} \\
&= \bar{y} - b_1\bar{x}
\end{aligned}
$$

Hence b_0 and b_1 are the same as when we did simple regression.

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

The fact that X is rank deficient implies that the columns of X is linearly independent, and hence, that some of the features are redundant. By eliminating the redundant features from each observation $\mathbf{x}_i$, we can build a new matrix X', which will have only the non-redundant features, and will be full rank.

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

Proof. Suppose $\mathbf{u} \in \text{null}(X)$. Then $X\mathbf{u} = \mathbf{0}_n$, and hence

$$(X^\top X)\mathbf{u} = X^\top(X\mathbf{u}) = X^\top(\mathbf{0}_n) = \mathbf{0}_{p+1}$$

Hence, $\text{null}(X) \subseteq \text{null}(X^\top X)$. On the other hand, if $\mathbf{u} \in \text{null}(X^\top X)$, then $X^\top X\mathbf{u} = \mathbf{0}_{p+1}$, so

$$
\begin{aligned}
X^\top X\mathbf{u} &= \mathbf{0}_{p+1} \\
\mathbf{u}^\top X^\top X\mathbf{u} &= 0 \\
(X\mathbf{u})^\top (X\mathbf{u}) &= 0 \\
\|X\mathbf{u}\|^2 &= 0
\end{aligned}
$$

The norm can only be zero if $X\mathbf{u} = \mathbf{0}_n$. Thus, we conclude that $\mathbf{u} \in \text{null}(X)$ and hence that $\text{null}(X^\top X) \subseteq \text{null}(X)$. With that, we can assert that $\dim \text{null}(X^\top X) = \dim \text{null}(X)$. Next, we use the Fundamental Theorem of Linear Maps (Rank-Nullity Theorem), applying it on both $X^\top X : \mathbb{R}^{p+1} \to \mathbb{R}^{p+1}$ and $X : \mathbb{R}^{p+1} \to \mathbb{R}^n$:

$$
\begin{aligned}
\dim \text{null}(X) + \dim \text{range}(X) &= \dim \mathbb{R}^{p+1} = p+1 \\
\dim \text{null}(X^\top X) + \dim \text{range}(X^\top X) &= \dim \mathbb{R}^{p+1} = p+1
\end{aligned}
$$

Subtracting the two equations and re-arranging, we arrive at $\dim \text{range}(X^\top X) = \dim \text{range}(X)$. Since the dimension of the range is precisely the rank, we are done. □

(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.

Proof.

$$\begin{aligned}
g(\begin{bmatrix}1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p\end{bmatrix}) &= \begin{bmatrix}1 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p\end{bmatrix}\mathbf{b} \\
&= \begin{bmatrix}\frac{1}{n}\sum_{i=1}^n 1 & \frac{1}{n}\sum_{i=1}^n x_{i,1} & \frac{1}{n}\sum_{i=1}^n x_{i,2} & \cdots & \frac{1}{n}\sum_{i=1}^n x_{i,p}\end{bmatrix}\mathbf{b} \\
&= \left(\frac{1}{n}\vec{\mathbf{1}}^\top X\right)\mathbf{b} \\
&= \left(\frac{1}{n}\vec{\mathbf{1}}^\top X\right)(X^\top X)^{-1}X^\top\mathbf{y} && \text{(by definition of } \mathbf{b}) \\
&= \frac{1}{n}\vec{\mathbf{1}}^\top H\mathbf{y} && \text{(by definition of } H) \\
&= \frac{1}{n}(H^\top\vec{\mathbf{1}})^\top\mathbf{y} && \text{(since } (AB)^\top B^\top A^\top) \\
&= \frac{1}{n}(H\vec{\mathbf{1}})^\top\mathbf{y} && (H = H^\top) \\
&= \frac{1}{n}(\vec{\mathbf{1}}^\top)\mathbf{y} && (\vec{\mathbf{1}} \in col[X]) \\
&= \bar{y}
\end{aligned}$$

□

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

Proof.

$$\begin{aligned}
\sum_{i=1}^n e_i &= \sum_{i=1}^n (y_i - \hat{y}_i) \\
&= \sum_{i=1}^n 1 \cdot (y_i - \hat{y}_i) \\
&= \vec{\mathbf{1}}_n^\top(\mathbf{y} - \hat{\mathbf{y}}) \\
&= \vec{\mathbf{1}}_n^\top(\mathbf{y} - H\mathbf{y}) \\
&= \vec{\mathbf{1}}_n^\top(I - H)\mathbf{y} \\
&= [(I - H)^\top\mathbf{1}]^\top\mathbf{y} \\
&= [(I - H)\vec{\mathbf{1}}_n]^\top\mathbf{y} && (I \text{ and } H \text{ are symmetric}) \\
&= [\mathbf{0}_n]^\top\mathbf{y} \\
&= 0
\end{aligned}$$

Note that $(I - H)$ is an orthogonal projection matrix onto $col[X]^\top$, the orthogonal complement of the column space of X. Since $\vec{\mathbf{1}}$ belongs to the column space of X, the orthogonal projection matrix $I - H$ maps it to $\mathbf{0}_n$. Now $\bar{e} = \frac{1}{n}\sum_{i=1}^n e_i = 0$. □

(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.

Proof. Without an intercept, the matrix X is

$$X = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{\cdot,A} & \mathbf{x}_{\cdot,B} & \mathbf{x}_{\cdot,C} \end{bmatrix}$$

The first column ($\mathbf{x}_{\cdot,A}$) corresponds to inputs where $x = A$, the second column ($\mathbf{x}_{\cdot,B}$) to inputs where $x = B$, and the third column ($\mathbf{x}_{\cdot,C}$)to inputs where $x = C$. The least square estimate is given by

$$\mathbf{b} = (X^\top X)^{-1} X^\top \mathbf{y}$$

We need to compute the vector, so we will do it in parts:

$$X^\top \mathbf{y} = \begin{bmatrix} \mathbf{x}_{\cdot,A}^\top \\ \mathbf{x}_{\cdot,B}^\top \\ \mathbf{x}_{\cdot,C}^\top \end{bmatrix} \mathbf{y} = \begin{bmatrix} \mathbf{x}_{\cdot,A}^\top \mathbf{y} \\ \mathbf{x}_{\cdot,B}^\top \mathbf{y} \\ \mathbf{x}_{\cdot,C}^\top \mathbf{y} \end{bmatrix} = \begin{bmatrix} \sum_{\{\, i | x_i = A\}} y_i \\ \sum_{\{\, i | x_i = B\}} y_i \\ \sum_{\{\, i | x_i = C\}} y_i \end{bmatrix}$$
$$= \begin{bmatrix} n_A \bar{y}_A \\ n_B \bar{y}_B \\ n_C \bar{y}_C \end{bmatrix}$$

Notice that since the vector $\mathbf{y}$ of responses corresponds to the inputs, the values line up to give, for example, $n_A \bar{y}_A$ (where n_A is the number of entries where $x = A$). Next we must compute $X^\top X$:

$$X^\top X = \begin{bmatrix} \mathbf{x}_{\cdot,A}^\top \mathbf{x}_{\cdot,A} & \mathbf{x}_{A}^\top \mathbf{x}_{\cdot,B} & \mathbf{x}_{A}^\top \mathbf{x}_{\cdot,C} \\ \mathbf{x}_{\cdot,B}^\top \mathbf{x}_{\cdot,A} & \mathbf{x}_{\cdot,B}^\top \mathbf{x}_{\cdot,B} & \mathbf{x}_{\cdot,B}^\top \mathbf{x}_{\cdot,C} \\ \mathbf{x}_{\cdot,C}^\top \mathbf{x}_{\cdot,A} & \mathbf{x}_{\cdot,C}^\top \mathbf{x}_{\cdot,B} & \mathbf{x}_{\cdot,C}^\top \mathbf{x}_{\cdot,C} \end{bmatrix} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

Since the matrix is diagonal, the inverse is easy to compute:

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

Now we can compute the least square estimate:

$$\begin{aligned}\mathbf{b} &= (X^\top X)^{-1} X^\top \mathbf{y} \\ &= \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix} \begin{bmatrix} n_A \bar{y}_A \\ n_B \bar{y}_B \\ n_C \bar{y}_C \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}\end{aligned}$$

The prediction is given by $g(\mathbf{x}) = \mathbf{x}^\top \mathbf{b}$, so we can proceed to compute the predictions of interest. Note that if $x = A$, then the input is $\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$. In particular, only one entry is 1 at a time. Finally,

$$\begin{aligned} g\left(\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top\right) &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \mathbf{b} \\ &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} \\ &= \bar{y}_A \end{aligned}$$

A similar calculations shows that g predicts $\bar{y}_B$ and $\bar{y}_C$ for $x = B$ and $x = C$, respectively. □

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

Proof. First compute the mean-control response:

$$\begin{aligned} \underset{col[X]}{\text{proj}} (\mathbf{y} - \bar{y} \cdot \vec{\mathbf{1}}_n) &= (\mathbf{y} - \bar{y} \cdot \vec{\mathbf{1}}_n) \\ &= H\mathbf{y} - \bar{y} H \vec{\mathbf{1}}_n \\ &= \hat{\mathbf{y}} - \bar{y} \cdot \vec{\mathbf{1}}_n \qquad \text{(since } \vec{\mathbf{1}} \in col[X]\text{)} \end{aligned}$$

Moreover,

$$(\mathbf{y} - \bar{y}\vec{\mathbf{1}}_n) - (\hat{\mathbf{y}} - \bar{y}\vec{\mathbf{1}}_n) = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}$$

In particular, since $\mathbf{e}$ is orthogonal to the column space of X, and $\hat{\mathbf{y}} - \bar{y}\vec{\mathbf{1}}_n$ belongs to the column space of X (it is a linear combination of such vectors), it follows that $\hat{\mathbf{y}} - \bar{y}\vec{\mathbf{1}}_n$ and $\mathbf{e}$ are orthogonal. Therefore, we can apply the Pythagorean Theorem to write

$$\begin{aligned} \|\mathbf{y} - \bar{y}\vec{\mathbf{1}}_n\|^2 &= \|\hat{\mathbf{y}} - \bar{y}\vec{\mathbf{1}}_n\|^2 + \|\mathbf{e}\|^2 \\ \sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 \\ SST &= SSR + SSE \end{aligned}$$

Now using trigonometry, we find that

$$\begin{aligned}
\cos(\theta)^2 &= \frac{\|\hat{\mathbf{y}} - \bar{y}\vec{\mathbf{1}}_n\|^2}{\|\mathbf{y} - \bar{y}\vec{\mathbf{1}}_n\|^2} \\
&= \frac{SST - SSE}{SST} \\
&= 1 - \frac{SSE}{SST} \\
&= R^2
\end{aligned}$$

Finally, since $\cos(\theta)^2 \in [0, 1]$, it follows that $R^2 \in [0, 1]$. □