MATH 390.4 / 650.2 Spring 2018 Homework #3t

Darshan Patel

Thursday 22^{nd} March, 2018

Problem 1

These are questions about Silver's book, chapter 2.

(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? 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.). Connecting this to the modeling framework should really make you think about what Tetlock's observation means for political and historical phenomena.

The difference between the approaches of the hedgehog and the fox is that the fox demonstrated the ability to make predictions using a multitude of approaches toward a problem, whereas the hedgehog only does no better than random chance. In notation form, the hedgehog's approach is $\bar{x}$ while the fox's approach is $\mathcal{A}(\mathcal{D}, \mathcal{H})$.

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

Harry Truman liked hedgehogs because they gave simple one answers whereas foxes couldn't give a concise answer. A lot of people don't think like hedgehogs because they usually create predictions from data and not from a random guess.

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

The more education that one acquires, the less accurate its' predictions become due to the accumulation of bias. With the information it receives, it can change them to confirm its bias. It can add more noise to all the information at hand and can disturb the ability to make an accurate prediction.

(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 vanilla classifiers because they don't return a clear image of the predictions but rather the chance of something occurring. Rather than saying which value is predicted, it allows for probability to dictate the chance of different outcomes occurring.

Problem 2

These are questions about Finlay's book, chapter 2-4. We will hold off on chapter 1 until we cover probability estimation after midterm 2.

(a) [easy] What term did we use in class for "behavioral (outcome) data"?

phenomenon.

(b) [easy] Write about some reasons why data scientists implement models that are subpar in predictive performance (p27).

Models that are subpar in predictive performance will be implemented sometimes to respect business requirements and constraints. It is also common for models to be simple for the regular laymen to understand and implement.

(c) [easy] In the first wine example, what is the outcome metric and what kind of supervised learning was employed?

The outcome metric is a classification model, a table of values showcasing responses and the score distribution amongst people who had different scores. It was employed using a decision tree algorithm.

(d) [easy] In the second wine example, what is the outcome metric and kind of supervised learning was employed?

The outcome metric is a regression model, a table that predicts the gross profit collected given the score. It was also employed using a decision tree algorithm.

(e) [easy] In the third chapter, why is it that some organizations cannot use predictive modeling to improve their business?

Some businesses cannot use predictive modeling to improve their business because it is not accepted by the entire organization. Not everyone from the senior manager to front line staff is comfortable acting upon the decisions made by the predictive model.

(f) [easy] In the bankruptcy case, what is the problem with merely using g to obtain a $\hat{y}$ without any other information from the model?

By not using other information from the model, a proper assessment of the model and outcome cannot be created. People will use their own judgement if they are confused about what's in the model and whether it is useful or not.

(g) [easy] Chapter 3 talks about using the model with human judgment. Under what circumstances is this beneficial? 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.).

It is beneficial to use models with human judgment when you have experience or biases. This is referred to as e, the difference between the true result you believe in which differs from the result made by the model.

(h) [difficult] In Chapter 4 Finaly makes an interesting observation based on his experience in data science. He says most predictive models have $p \leq 30$. Why do you think this is? Discuss.

Most predictive models have $p \leq 30$. This is a good constraint to have. If you have too many variables to measure a phenomenon, you can run into different situations where one variable will have an impact on another and are thus not independent. Perhaps you will see that there is no clear conclusion that can be made from more than 30 factors about a phenomenon. Each variable may not be independent of another or play little role if any.

(i) [easy] He says there is "almost always other data that could be acquired ... [which] doesn't always come for free". The "data" he is talking about here specifically means "more predictors" i.e. increasing p. In what cases would someone be willing to pay for this data?

Someone would be willing to pay for "more predictors" if it can be proved that the predictors can create a clear model of the phenomenon. If the value of the data can be determined to be a lot, then it will be purchased. A situation of this is when a social media network sells your data to an advertising company to determine when to show which ads to users to bring in the most profit, "clicks", to the advertising company.

(j) [easy] Table 4 lists "data types" about what type of observations?

categorical observations.

(k) [easy] What type of data does he find in his experience to be the most important to predictive modeling? Why do you think this is so?

The most important type of data in predictive modeling, according to Finaly, is data about primary behavior. This should be the most important because your behavior can dictate your future behaviors. It is common to predict something that you already do.

(l) [easy] If $x_{\cdot 17}$ was age and $x_{\cdot 18}$ is age of spouse, what is the most likely reason why adding $x_{\cdot 18}$ to $\mathbb{D}$ not be fruitful for predictive ability?

Adding $x_{\cdot 18}$ to $\mathbb{D}$ would not be fruitful because it could be partially correlated to $x_{\cdot 17}$ and thus not add any new information.

(m) [difficult] What is the lifespan of a predictive model? Why does it not last forever? 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.).

The lifespan of a predictive model depend among different companies. Models in healthcare or credit risk have a lifespan of several years whereas ones for marketing only last for a few months. The lifespan of a predictive model does not last forever because responses can change over time due to new ideas, likings, medical advancements, economic situations, etc. What this means is that $t(z_1, \ldots, z_t)$ is always changing.

(n) [difficult] What does “large enough to representative of the full population” (p80) mean? 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.).

The quote “large enough to representative of the full population” means to use a use a sample that resembles the population. We want the $\mathcal{D}$ to have all the information we need to create a model that can be represent the entire $\langle \mathcal{X}, \mathcal{Y} \rangle$.

(o) [easy] Is there a hype about “big data” i.e. including millions of observations instead of a few thousand? Discuss Finlay’s opinion.

There is a hype about “big data”. Finlay believes that you shouldn’t use big data if not all of it is directly applicable to what you want to predict and model. In addition, having more data to evaluate and more predictor variables, hence, can require more computing power than you have.

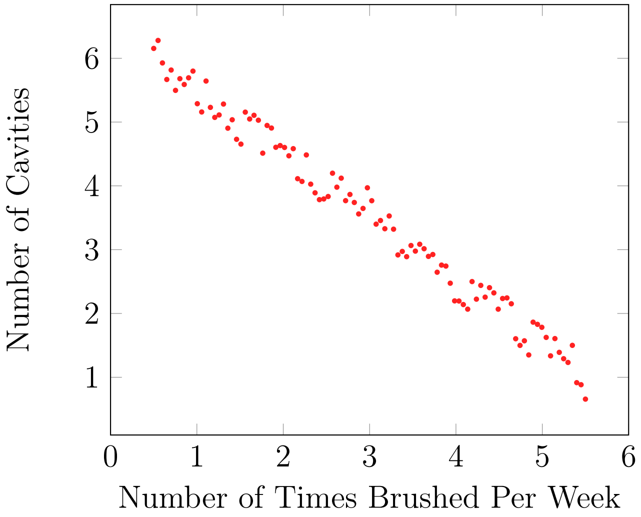
(p) [easy] What is Finlay’s solution to “overfitting” (p84)?

Finlay’s solution to “overfitting” is is to use large samples so that all points are accounted for rather than just a small amount.

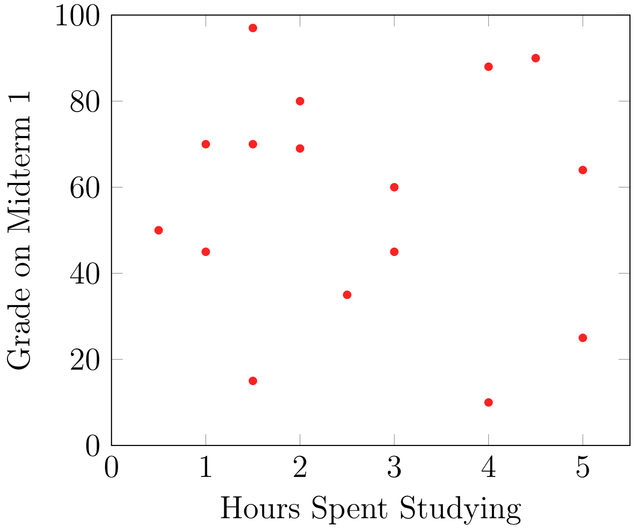
Problem 3

These are questions about association and correlation.

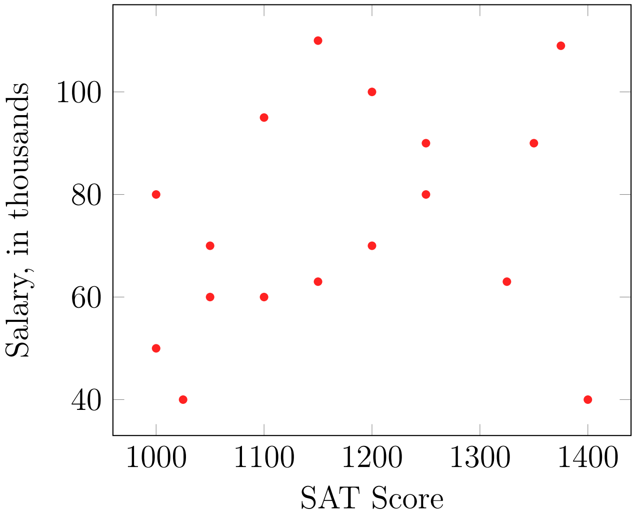
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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.

Two variables can be correlated but not associated due to sheer coincidence. For example, suppose as ice cream sales increased, the rate of drowning deaths increased sharply. Therefore ice cream consumption causes drowning. These two variables are positively correlated but have nothing to do with each other. It could be that ice cream sales are high during summer months as well as water activities. But drowning deaths are only caused by more exposure to water based activities, not ice cream.

Problem 4

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.

First find $A\boldsymbol{c}$,

$$A\boldsymbol{c} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \dots + a_{1n}c_n \\ a_{21}c_1 + a_{22}c_2 + \dots + a_{2n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2}c_2 + \dots + a_{nn}c_n \end{bmatrix}$$

Then

$$\begin{aligned} \boldsymbol{c}^\top A\boldsymbol{c} &= \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \dots + a_{1n}c_n \\ a_{21}c_1 + a_{22}c_2 + \dots + a_{2n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2}c_2 + \dots + a_{nn}c_n \end{bmatrix} \\ &= c_1(a_{11}c_1 + a_{12}c_2 + \dots + a_{1n}c_n) \\ &+ c_2(a_{21}c_1 + a_{22}c_2 + \dots + a_{2n}c_n) \\ &+ \dots + c_n(a_{n1}c_1 + a_{n2}c_2 + \dots + a_{nn}c_n) \end{aligned}$$

Note that

$$\frac{\partial}{\partial c_1}\left[\boldsymbol{c}^\top A\boldsymbol{c}\right] = a_{11}c_1 + (a_{11}c_1 + a_{12}c_2 + \dots + a_{1n}c_n) + a_{21}c_2 + a_{31}c_3 + \dots + a_{n1}c_n$$

Let's do the next row to note the pattern

$$\frac{\partial}{\partial c_2}\left[\boldsymbol{c}^\top A\boldsymbol{c}\right] = a_{12}c_1 + a_{22}c_2 + (a_{21}c_1 + a_{22}c_2 + \dots + a_{2n}c_n) + a_{32}c_3 + \dots + a_{n2}c_n$$

Since A is not symmetric, we cannot clean this up and can only write it as a form of summations

$$\begin{aligned} \frac{\partial}{\partial c_1}\left[\boldsymbol{c}^\top A\boldsymbol{c}\right] &= \sum a_{1i}c_i + \sum a_{i1}c_i = \sum (a_{1i} + a_{i1})c_1 \\ \frac{\partial}{\partial c_2}\left[\boldsymbol{c}^\top A\boldsymbol{c}\right] &= \sum a_{2i}c_i + \sum a_{i2}c_i = \sum (a_{2i} + a_{i2})c_2 \end{aligned}$$

Therefore

$$\frac{\partial}{\partial \boldsymbol{c}}\left[\boldsymbol{c}^\top Ac\right] = \begin{bmatrix} \sum(a_{1i} + a_{i1})c_1 \\ \sum(a_{2i} + a_{i2})c_2 \\ \vdots \\ \sum(a_{ni} + a_{in})c_n \end{bmatrix}$$

(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.

First let's find a closed form for SSE:

$$\begin{aligned}
SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= (\vec{y} - \vec{\hat{y}})^\top(\vec{y} - \vec{\hat{y}}) = (\vec{y}^\top - \vec{\hat{y}}^\top)(\vec{y} - \vec{\hat{y}}) \\
&= \vec{y}^\top\vec{y} - \vec{\hat{y}}^\top\vec{y} - \vec{y}^\top\vec{\hat{y}} + \vec{\hat{y}}^\top\vec{\hat{y}} \\
&= \vec{y}^\top\vec{y} - 2\vec{\hat{y}}^\top\vec{y} + \vec{\hat{y}}^\top\vec{\hat{y}} \\
&= \vec{y}^\top\vec{y} - 2(X\vec{w})^\top\vec{y} + (X\vec{w})^\top(X\vec{w}) \\
&= \vec{y}^\top\vec{y} - 2\vec{w}^\top X^\top\vec{y} + \vec{w}^\top X^\top X\vec{w}
\end{aligned}$$

Integrate this with respect to $\vec{w}$ to attain $\boldsymbol{b}$:

$$\begin{aligned}
\frac{\partial}{\partial\vec{w}}SSE &= \frac{\partial}{\partial\vec{w}}(\vec{y}^\top\vec{y} - 2\vec{w}^\top X^\top\vec{y} + \vec{w}^\top X^\top X\vec{w}) \\
&= \frac{\partial}{\partial\vec{w}}(\vec{y}^\top\vec{y}) - 2\frac{\partial}{\partial\vec{w}}(\vec{w}^\top X^\top\vec{y}) + \frac{\partial}{\partial\vec{w}}(\vec{w}^\top X^\top X\vec{w}) \\
&= \vec{0} - 2X^\top\vec{y} + 2X^\top X\vec{w} \overset{\text{set}}{=} 0 \\
X^\top\vec{y} &= X^\top X\vec{w} \\
\boldsymbol{w} = \boldsymbol{b} &= (X^\top X)^{-1}X^\top\vec{y}
\end{aligned}$$

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

Let $X \in \mathbb{R}^{n\times(1+1)} = \mathbb{R}^{n\times 2}$. This looks like $\begin{bmatrix} 1_n & \vec{x} \end{bmatrix}$. Additionally, $\vec{y}$ is a column of n values of ys. Take tthe above solution one piece at a time. First let's compute $X^\top X$. This is

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

The inverse of $X^\top X$ is

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

Before finding $\boldsymbol{b}$, first find $X^\top \vec{y}$, which is

$$X^\top \vec{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Thus

$$\begin{aligned} \boldsymbol{b} &= (X^\top X)^{-1}(X^\top \vec{y}) \\ &= \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum(x_i-\bar{x})^2} \\ -\frac{\bar{x}}{\sum(x_i-\bar{x})^2} & \frac{1}{\sum(x_i-\bar{x})^2} \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sum y_i}{n} + \frac{\sum y_i \bar{x}^2}{\sum(x_i-\bar{x})^2} - \frac{\sum x_i y_i \bar{x}}{\sum(x_i-\bar{x})^2} \\ -\frac{\sum y_i \bar{x}}{\sum(x_i-\bar{x})^2} + \frac{x_i y_i}{\sum(x_i-\bar{x})^2} \end{bmatrix} \\ &= \begin{bmatrix} \bar{y} - b_1 \bar{x} \\ \frac{\sum y_i (x_i - \bar{x})}{\sum(x_i-\bar{x})^2} \end{bmatrix} \\ &= \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \end{aligned}$$

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

If X is rank deficient, then there is duplicated data where one or more columns are a linear combination of the other columns in X. To solve for $\boldsymbol{b}$, remove the duplicated columns so that all the columns are linearly independent.

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

Let X has rank and let $x \in N(X)$ where $N(X)$ is the null space of X. The null space of X is such that $xX = 0$. If $\text{rank}\left[X\right] = \text{rank}\left[X^\top X\right]$, then x is in the null space of $X^\top X$, or $x \in N(X^\top X)$. Hence $N(X) \subseteq N(X^\top X)$. Now suppose $x \in N(X^\top X)$, the null space of $X^\top X$. Then

$$\begin{aligned} X^\top X x &= 0 \\ x^\top X^\top X x &= 0 \\ (Xx)^\top (Xx) &= 0 \\ Xx &= 0 \end{aligned}$$

Here we find that x is in the null space of X, or $x \in N(X)$. Therefore $N(X^\top X) \subseteq N(X)$. Furthermore,

$$\begin{aligned} N(X^\top X) &= N(X) \\ \dim[N(X^\top X)] &= \dim[N(X)] \\ \text{rank}\left[X^\top X\right] &= \text{rank}\left[X\right] \end{aligned}$$

(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.

First find a closed from for SSE:

$$\begin{aligned}
SSE &= \sum c_i (y_i - \hat{y}_i)^2 \\
&= (\vec{y} - \vec{\hat{y}})^\top C (\vec{y} - \vec{\hat{y}}) \\
&= (\vec{y}^\top - \vec{\hat{y}}^\top) C (\vec{y} - \vec{\hat{y}}) \\
&= (\vec{y}^\top - \vec{\hat{y}}^\top) C^{\frac{1}{2}} C^{\frac{1}{2}} (\vec{y} - \vec{\hat{y}}) \\
&= (\vec{y}^\top C^{\frac{1}{2}} - \vec{\hat{y}}^\top C^{\frac{1}{2}})(C^{\frac{1}{2}} \vec{y} - C^{\frac{1}{2}} \vec{\hat{y}}) \\
&= \vec{y}^\top C^{\frac{1}{2}} C^{\frac{1}{2}} \vec{y} - \vec{\hat{y}}^\top C^{\frac{1}{2}} C^{\frac{1}{2}} \vec{y} - \vec{\hat{y}}^\top C^{\frac{1}{2}} C^{\frac{1}{2}} \vec{y} + \vec{\hat{y}}^\top C^{\frac{1}{2}} C^{\frac{1}{2}} \vec{\hat{y}} \\
&= \vec{y}^\top C \vec{y} - 2 \vec{\hat{y}}^\top C \vec{y} + \vec{\hat{y}}^\top C \vec{\hat{y}} \\
&= \vec{y}^\top C \vec{y} - 2 (X\vec{w})^\top C \vec{y} + (X\vec{w})^\top C (X\vec{w}) \\
&= \vec{y}^\top C \vec{y} - 2 \vec{w}^\top X^\top C \vec{y} + \vec{w}^\top X^\top C X \vec{w}
\end{aligned}$$

Integrate this with respect to $\vec{w}$ to attain $\boldsymbol{b}$:

$$\begin{aligned}
\frac{\partial}{\partial \vec{w}} SSE &= \frac{\partial}{\partial \vec{w}} (\vec{y}^\top C \vec{y} - 2 \vec{w}^\top X^\top C \vec{y} + \vec{w}^\top X^\top C X \vec{w}) \\
&= \frac{\partial}{\partial \vec{w}} (\vec{y}^\top C \vec{y}) - 2 \frac{\partial}{\partial \vec{w}} (\vec{w}^\top X^\top C \vec{y}) + \frac{\partial}{\partial \vec{w}} (\vec{w}^\top X^\top C X \vec{w}) \\
&= \vec{0} - 2 X^\top C \vec{y} + 2 X^\top C X \vec{w} \overset{\text{set}}{=} 0 \\
X^\top C X \vec{w} &= X^\top C \vec{y} \\
\boldsymbol{b} = \vec{w} &= (X^\top C X)^{-1} X^\top C \vec{y}
\end{aligned}$$

(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.

Suppose $y = b_0 + b_1 x$. Then $r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$. Note that $b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$. and $b_0 = \bar{y} - b_1\bar{x}$. Then

$$\begin{aligned}
\sum(\hat{y}_i - \bar{y})^2 &= \sum(b_0 + b_1 x_i - \bar{y})^2 \\
&= \sum(\hat{y} - b_1\bar{x} + b_1 x_i - \bar{y})^2 \\
&= b_1^2 \sum(x_i - \bar{x})^2 \\
&= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})^2}{\sum(x_i - \bar{x})^2} \sum(x_i - \bar{x})^2 \\
&= \frac{[\sum(x_i - \bar{x})(y_i - \bar{y})]^2}{\sum(x_i - \bar{x})^2}
\end{aligned}$$

Therefore

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

(h) [harder] Prove that the point $\langle 1, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p, \bar{y}\rangle$ is a point on the least squares linear solution.

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} b_0 + b_1 x_{11} + \cdots + b_p x_{1p} \\ b_0 + b_1 x_{21} + \cdots + b_p x_{2p} \\ \vdots \\ b_0 + b_1 x_{p1} + \cdots + b_p x_{pp} \end{bmatrix} = X\boldsymbol{b}$$

Then

$$\begin{aligned}
\sum_{i=1}^{p} y_i &= \sum_{i=1}^{p} (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}) \\
\frac{\sum y_i}{n} &= \frac{\sum(b_0 + b_1 x_{i1} + \cdots + b_p x_{ip})}{n} \\
\frac{\sum y_i}{n} &= \frac{\sum b_0}{n} + \frac{\sum b_1 x_{i1}}{n} + \cdots + \frac{\sum b_p x_{ip}}{n} \\
\frac{\sum y_i}{n} &= \frac{n b_0}{n} + \frac{n b_1 \bar{x}_1}{n} + \cdots + \frac{n b_p \bar{x}_p}{n} \\
\bar{y} &= b_0 + b_1 \bar{x}_1 + \cdots + b_p \bar{x}_p
\end{aligned}$$

Problem 5

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

(a) [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?

The degrees of freedom in the resulting model is $p+1$. It means that there are $p+1$ independent variables to pick to describe $\boldsymbol{y}$.

(b) [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 the least squares solution?

Project $\boldsymbol{y}$ onto the column space of X.

$$\begin{aligned} X\boldsymbol{b} &= \boldsymbol{y} \\ \vec{x}_1^\top(X\boldsymbol{b}-\boldsymbol{y}) &= 0 \\ \vec{x}_2^\top(X\boldsymbol{b}-\boldsymbol{y}) &= 0 \\ \vec{x}_n^\top(X\boldsymbol{b}-\boldsymbol{y}) &= 0 \end{aligned}$$

This means

$$\begin{aligned} X^\top(X\boldsymbol{b}-\boldsymbol{y}) &= 0 \\ X^\top X\boldsymbol{b} &= X^\top\boldsymbol{y} \\ \boldsymbol{b} &= (X^\top X)^{-1}X^\top\boldsymbol{y} \\ X\boldsymbol{b} &= X(X^\top X)^{-1}X^T\boldsymbol{y} \\ \text{Proj}_{\text{colsp}[X]}\,[\boldsymbol{y}] &= X(X^\top X)^{-1}X^\top\boldsymbol{y} \end{aligned}$$

This result is the same as the least squares solution.

(c) [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 $\boldsymbol{0}_n$ than the first $\boldsymbol{e}$? That is, wouldn't this yield a better model on iteration #2? Yes/no and explain.

This would not yield a better model on iteration #2. After $\boldsymbol{y}$ is regressed using $\boldsymbol{X}$ to get $\hat{\boldsymbol{y}}$, it is projected onto the space orthogonal to X, or the residual space. Then regressing $\boldsymbol{e}$ using $\boldsymbol{X}$ will result in the same projection with no new residuals because it is already in the space orthogonal to $\boldsymbol{X}$. "$\boldsymbol{e}_{new}$" would not be closer to $\boldsymbol{0}_n$ but will be the same as $\boldsymbol{e}$.

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

If Q is an orthonormal matrix, then for any two vectors $q_i, q_j \in Q$, $q_i^\top q_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = i \end{cases}$.

This means that each vector is perpendicular to each other. Henceforth

$$Q^\top Q = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = I_n$$

If so, then

$$\begin{aligned} Q^\top Q &= I_n \\ (Q^\top Q)Q^{-1} &= I_n Q^{-1} \\ Q^\top &= Q^{-1} \end{aligned}$$

(e) [harder] Prove that the least squares projection $H = X\left(X^\top X\right)^{-1} X^\top$ is the same as $QQ^\top$.

Note that $H = X\left(X^\top X\right)^{-1} X^\top$. Let $X = Q$.

$$\vec{\hat{y}} = H\vec{y} = X(X^\top X)^{-1} X^\top \vec{y} = Q(Q^\top Q)^{-1} Q^\top \vec{y}$$

From the last problem, we see that $Q^\top Q = I_n$. Then

$$\vec{\hat{y}} = Q(Q^\top Q)^{-1} Q^\top \vec{y} = Q(I_n)^{-1} Q^\top \vec{y} = QQ^\top \vec{y}$$

Hence $X(X^\top X)^{-1} X^\top = QQ^\top$.

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

Divide the projection component wise

$$\begin{aligned} \text{Proj}_{\text{colsp}[X]}\,[\boldsymbol{y}] &= \text{Proj}_{\vec{x}_1}\,[\boldsymbol{y}] + \text{Proj}_{\vec{x}_2}\,[\boldsymbol{y}] + \dots + \text{Proj}_{\vec{x}_n}\,[\boldsymbol{y}] \\ &= \frac{\vec{x}_1 \vec{x}_1^\top}{||\vec{x}_1||^2}\boldsymbol{y} + \frac{\vec{x}_2 \vec{x}_2^\top}{||\vec{x}_2||^2}\boldsymbol{y} + \dots + \frac{\vec{x}_n \vec{x}_n^\top}{||\vec{x}_n||^2}\boldsymbol{y} \\ &= \left(\frac{\vec{x}_1 \vec{x}_1^\top}{||\vec{x}_1||^2} + \frac{\vec{x}_2 \vec{x}_2^\top}{||\vec{x}_2||^2} + \dots + \frac{\vec{x}_n \vec{x}_n^\top}{||\vec{x}_n||^2}\right)\boldsymbol{y} \end{aligned}$$

Note that X is orthonormal and so

$$\begin{aligned}
\text{Proj}_{\text{colsp}[X]}\left[\boldsymbol{y}\right] &= (\vec{x}_1\vec{x}_1^\top + \vec{x}_2\vec{x}_2^\top + \cdots + \vec{x}_n\vec{x}_n^\top)\boldsymbol{y} \\
&= \left(\begin{bmatrix} x_{11}^2 & x_{11}x_{12} & \dots & x_{11}x_{1n} \\ x_{12}x_{11} & x_{12}^2 & \dots & x_{12}x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n}x_{11} & x_{1n}x_{12} & \dots & x_{1n}^2 \end{bmatrix} + \begin{bmatrix} x_{21}^2 & x_{21}x_{22} & \dots & x_{21}x_{2n} \\ x_{22}x_{21} & x_{22}^2 & \dots & x_{22}x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2n}x_{21} & x_{2n}x_{22} & \dots & x_{2n}^2 \end{bmatrix}\right. \\
&\left. + \cdots + \begin{bmatrix} x_{n1}^2 & x_{n1}x_{n2} & \dots & x_{n1}x_{nn} \\ x_{n2}x_{n1} & x_{n2}^2 & \dots & x_{n2}x_{nn} \\ \vdots & \vdots & \ddots & \vdots \\ x_{nn}x_{n1} & x_{nn}x_{n2} & \dots & x_{nn}^2 \end{bmatrix}\right)\boldsymbol{y} \\
&= \left(\begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ \vec{x}_1 & \vec{x}_2 & \dots & \vec{x}_n \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix}\begin{bmatrix} \leftarrow & \vec{x}_1^\top & \rightarrow \\ \leftarrow & \vec{x}_2^\top & \rightarrow \\ \vdots & \vdots & \vdots \\ \leftarrow & \vec{x}_n^\top & \rightarrow \end{bmatrix}\right)\boldsymbol{y} \\
&= QQ^\top\boldsymbol{y} \\
&= \text{Proj}_{\text{colsp}[X]}\left[\boldsymbol{y}\right]
\end{aligned}$$

(g) [difficult] Trouble in paradise. Prove that the SSE of a multivariate linear least squares model always decreases (equivalently, R^2 always increases) upon the addition of a new independent predictor. Keep in mind this holds true even if this new predictor has no information about the true causal inputs to the phenomenon y.

Recall that

$$SSR = \sum \hat{y}_i - n\bar{y}^2 = \sum_{j=1}^{p} \left|\left|\text{Proj}_{\vec{x}_j}\left[\vec{y}\right]\right|\right|^2 - n\bar{y}^2$$

If a new predictor $\vec{x}_{new}$ is added to X, then

$$SSR_{new} = \left(\sum_{j=1}^{p} \left|\left|\text{Proj}_{\vec{x}_j}\left[\vec{y}\right]\right|\right|\right) + \underbrace{\left|\left|\text{Proj}_{\vec{x}_{new}}\left[\vec{y}\right]\right|\right|^2}_{>0} - n\bar{y}^2$$

Since the new term added is positive, this means that $SSR_{new} > SSR$ and since $SST = SSR + SSE$ is constant,

$$SSE_{new} < SSE$$

Therefore the SSE decreased. Furthermore,

$$\begin{gathered}
SSR_{new} > SSR \\
\frac{SSR_{new}}{SST} > \frac{SSR}{SST} \\
R^2_{new} > R^2
\end{gathered}$$

Thus R^2 always increases upon the addition of a new independent predictor.

(h) [harder] Why is this a bad thing? Explain in English.

This is a bad thing because if we keep adding new useless inputs, we will be able to attain $R^2 = 100\%$ where 100% of the variance will be explained. But all this variance will be explained by the help of random data points that has nothing to contribute to the phenomenon. The model will be overfit due to unhelpful data.

(i) [E.C.] Prove that $\text{rank}\left[H\right] = \text{tr}\left[H\right]$.

Preliminary step: Show $\text{tr}\left[AB\right] = \text{tr}\left[BA\right]$.

$$\begin{aligned} \text{tr}\left[AB\right] &= \sum_{i=1}^{n} (AB)_{ii} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} B_{ji} \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} B_{ji} A_{ij} \\ &= \sum_{j=1}^{n} (BA)_{jj} \\ &= \text{tr}\left[BA\right] \end{aligned}$$

Therefore, if $\text{rank}\left[H\right] = n + 1$, prove that $\text{tr}\left[H\right] = n + 1$.

$$\begin{aligned} \text{tr}\left[H\right] &= \text{tr}\left[X(X^\top X)^{-1} X^\top\right] \\ &= \text{tr}\left[(X^\top X)^{-1} X^\top X\right] \\ &= \text{tr}\left[Z^{-1} Z\right] \\ &= \text{tr}\left[I\right] \\ &= n + 1 \end{aligned}$$