MATH 390.4 / 650.2 Spring 2020 Homework #3

Tyron Samaroo

Saturday 9th May, 2020

Problem 1

These are questions about Silver's book, chapters 3-6. For all parts in this question, answer using notation from class (i.e. $t, f, g, h^*, \delta, \epsilon, e, t, z_1, \ldots, z_t, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}$, $x_{1\cdot}, \ldots, x_{n\cdot}$, etc.

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

PECOTA seems to be most similar to KNN(k nearest neighbors) based on his premises of comparing players by points and finding the difference.

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

Baseball performance as a function of age is not a linear model since some players can perform really well early on and better than more experience players.

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

Baseball scouts were able to do better than a prediction system like PECOTA because scouts use a hybrid approach. They have access to more information than statistics alone.

(d) [harder] Why hasn't anyone (at the time of the writing of Silver's book) taken advantage of Pitch f/x data to predict future success?

They had other tools like radar guns but didn't have completely functional three-dimensional cameras. Also Pitch f/x seems expensive and Moneyball could do the same.

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

The problem with weather prediction is that it can add a lot of dimensions and make it very difficult and messy since its a dynamic system. There is also many inaccuracies in the data $\mathbb{D}$

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

Weatherman lie about the chance of rain because the public won't believe then either way. Weatherman would rather get chance rain wrong because people wont curse them hence economic incentive. If you want honest forecast you should go to The Weather Channel since they can use all of the government's raw data.

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

We do not know the true function t with its causal inputs z's. We also don't have enough data in X and knowledge of features x_i to predict earthquakes especially ones with higher magnitude. We end up building a model that over-fits and will then have huge errors and bad prediction.

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

The nonsense predictor in model Silver describe is someone who assumes that since the locks with certain color unlocks with a certain code then all locks with same color will unlock with the same code. This is not true since other locks will not unlock with the same code and have the same color. So this is over fitting. You assume you learned something but in fact you did not.

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

Adding more features to the model can create noise and not do what is intended to do. We end up fitting noise than the signal and perform badly in real world.

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

It's much harder to find something that identifies the signal; variables that are leading indicators in one economic cycle will not be in the next. There are other decisions such political decisions you have to worry about other than economic ones. Its too difficult to understand the true function with its causal inputs and to even estimate it.

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

Problem 2

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

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

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

$$\frac{\partial}{\partial \boldsymbol{c}}\left[\boldsymbol{c}^\top A\boldsymbol{c}\right] = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{c}}\left[A\vec{\boldsymbol{c}_1}\right] \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{c}}\left[A\vec{\boldsymbol{c}_n}\right] \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{c}}\left[A\vec{\boldsymbol{c}_1}\right] \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{c}}\left[A\vec{\boldsymbol{c}_n}\right] \end{bmatrix} = 2\begin{bmatrix} \vec{a_1} \\ \vdots \\ \vec{a_n} \end{bmatrix}\vec{\boldsymbol{c}} = 2A\vec{\boldsymbol{c}}$$

(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.

Proof. We know that $\vec{\hat{y}} = X\vec{w}$

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

Then we can do

$$\begin{aligned} \frac{\partial}{\partial \vec{w}}[SSE] &= \frac{\partial}{\partial \vec{w}}\left[\vec{y}^T\vec{y} - 2\vec{w}^T X^T\vec{y} + \vec{w}^T X^T\vec{w}\right] \\ &= \frac{\partial}{\partial \vec{w}}\left[\vec{y}^T\vec{y}\right] - 2\frac{\partial}{\partial \vec{w}}\left[\vec{w}^T X^T\vec{y}\right] + \frac{\partial}{\partial \vec{w}}\left[\vec{w}^T X^T X\vec{w}\right] \\ &= \vec{0} - 2X^T\vec{y} + 2X^T X\vec{w} \\ X^T\vec{y} &= X^T X\vec{w} \\ \vec{w} = \vec{b} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\vec{y} \end{aligned}$$

□

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

When p=1 we use the lost function to fit the b's. We can use the SSE or sum of squared error formula and do some manipulation.

$$\hat{y} = b_0 + b_1(x_i)$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Substitute $\hat{y}$

$$= \sum_{i=1}^{n} y_i - (b_0 + b_1 x_i)^2$$

$$= \sum_{i=1}^{n} y_i^2 + b_0^2 + b_1^2 x_i^2 - 2y_i b_0 - 2y_i b_1 x_i + 2b_0 b_1 x_i$$

$$= \sum y_i^2 + nb_0^2 + b_1^2 \sum x_i^2 - 2n\bar{y}b_0 - 2b_1 \sum x_i y_i + 2b_{0b1} n\bar{x}$$

Then we take partial of b_0

$$\frac{\partial}{\partial \vec{b_0}} [SSE] = 2nb_0 - 2n\bar{y}_2 b_1 n\bar{x} = 0$$

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

Then we take partial of b_1

$$\frac{\partial}{\partial \vec{b_1}} [SSE] = 2b_1 \sum x_i^2 - 2\sum x_i y_i + 2b_0 n\bar{x}$$

Substitute in b_0

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

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

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

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

$$\mathbf{b_1} = \frac{(n-1)S_x y}{(n-1)S_x^2}$$

$$= r\frac{S_y}{S_x}$$

Now since $b_1 = r\dfrac{S_y}{S_x}$ *and* $b_0 = \bar{y} - b_1\bar{x}$

we substitute in b_0 *to get* $b_0 = \bar{y} - r\dfrac{S_y}{S_x}$

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

If X is rank deficient we can't solve for $\boldsymbol{b}$ because it will not be invertable. In order to solve this problem you had to remove anything that is linearly dependent in X

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

$\boldsymbol{X}^T\boldsymbol{X}$ has the dimension of $\mathbb{R}^p + 1xp + 1$ When $\boldsymbol{X}^T\boldsymbol{X}$ is full rank it is invertable and rank of X is the same, p + 1. All the columns are all linear independent so the rank of $\boldsymbol{X}^T\boldsymbol{X}$ and X are both p + 1

(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.

(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.

$$
\begin{aligned}
R^2 &= \frac{\text{SSR}}{\text{SST}} \\
&= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \\
&= \frac{\sum \hat{y}_i^2 - n\bar{y}^2}{\sum(y_i - \bar{y})^2} \\
&\textit{Subsitute in for } \hat{y} = b_0 + b_1 x_i \\
&= \frac{\sum(b_0 + b_1 x_i)^2 - n\bar{y}^2}{(n-1)s_y^2} \\
&= \frac{\sum(b_0^2 + 2b_0 b_1 x_i + b_1^2 x_i^2) - n\bar{y}^2}{(n-1)s_y^2} \\
&= \frac{nb_0^2 + 2b_0 b_1 \sum x_i + b_1^2 \sum x_i^2 - n\bar{y}^2}{(n-1)s_y^2} \\
&= \frac{(\bar{y} - b_1\bar{x})^2 n + 2(\bar{y} - b_1\bar{x})b_1 n\bar{x} + b_1^2 \sum x_i^2 - n\bar{y}^2}{(n-1)s_y^2} \\
&= \frac{b_1^2 \sum x_i^2 - b_1^2 \bar{x}^2 n}{(n-1)s_y^2} \\
&= \frac{b_1^2(\sum x_i^2 - \bar{x}^2 n)}{(n-1)s_y^2}
\end{aligned}
$$

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

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

Since $\bar{y}$ is the average of all the y_i

$$\bar{y} = \frac{1}{n} \sum y_i$$

y_i can be $b_0 + b_1 x_1 + b_2 x_2 ... b_{pxp}$ where each x is offset by some b

$$= \frac{1}{n} \sum b_0 + b_1 x_1 + ... + b_{pxp}$$
$$= \frac{1}{n} \sum b_0 + \frac{1}{n} \sum b1_{x1} + ... + \frac{1}{n} \sum b_p x_p$$
$$= b_0 + b_1 \bar{x_1} + ... + b_p \bar{x_p}$$

Now b_0 will be some constant vector $\vec{1}$.

So the $b's$ are apart of some function g that gives

$$\bar{y} = g([1\ \bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_p])$$

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

$$\bar{e} = \frac{1}{n} \sum (y_i - \hat{y_i})$$
$$= \frac{1}{n} \sum y_i - \frac{1}{n} \sum b_0 + b_1 x_1 + ... + b_{pxp}$$

Multiplying each term by $\frac{1}{n}$ would make the $x's$ become average$\bar{x}'s$

$$= \frac{1}{n} \sum y_i - b_0 + b_1 \bar{x_1} + ... + b_p \bar{x_p}$$

From part h we can see that $\bar{y} = b_0 + b_1 \bar{x_1} + ... + b_p \bar{x_p}$

$$= \bar{y} - \bar{y}$$
$$= 0$$

(j) [difficult] If you model $\boldsymbol{y}$ with one categorical nominal variable that has levels A, B, C, prove that the OLS estimates look like $\bar{y}_A$ if $x = A$, $\bar{y}_B$ if $x = B$ and $\bar{y}_C$ if $x = C$. You can choose to use an intercept or not. Likely without is easier.

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{matrix} A \\ B \\ C \end{matrix}\begin{bmatrix} 1 & \dots & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 & \dots & 1 \end{bmatrix} \begin{matrix} A & B & C \\ \end{matrix} \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix}$$

This only leaves the diagonal

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

Now the inverse $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$

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

Now the $\boldsymbol{X}^T y$

$$\boldsymbol{X}^T y = \begin{matrix} A \\ B \\ C \end{matrix}\begin{bmatrix} 1 & \dots & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum y\mathbb{1}_{A=1} \\ \sum y\mathbb{1}_{B=1} \\ \sum y\mathbb{1}_{C=1} \end{bmatrix}$$

Now put $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ *and* $\boldsymbol{X}^T y$ *together*

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T y = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix} \begin{bmatrix} \sum y\mathbb{1}_{A=1} \\ \sum y\mathbb{1}_{B=1} \\ \sum y\mathbb{1}_{C=1} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n_A}\sum y\mathbb{1}_{A=1} \\ \frac{1}{n_B}\sum y\mathbb{1}_{B=1} \\ \frac{1}{n_C}\sum y\mathbb{1}_{C=1} \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$

Problem 3

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

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

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

I_n is an orthogonal project is it is both symmetric and idempotent. If ${I_n}^T = \mathrm{I}_n$ then it is symmetric. If $I_n I_n = I_n$ then it is idempotent.

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

I_n projects onto the subspace R^N

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

The resulting model with have $p+1$ degrees of freedom.

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

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

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

$$\boldsymbol{Q}^\top = \boldsymbol{Q}^{-1} => \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{Q}^{-1}\boldsymbol{Q} = I_n$$

$$\boldsymbol{Q}^\top\boldsymbol{Q} = \begin{bmatrix} q_1^\top \\ \vdots \\ q_n^\top \end{bmatrix} \begin{bmatrix} q_1 & \dots & q_n \end{bmatrix}$$

$$= \begin{bmatrix} q_1^\top q_1 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & q_n^\top q_n \end{bmatrix} = I_n$$

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

$$\text{Proj}_v[a] = \frac{\vec{v}\vec{v}^T}{||v||^2 a} = H\vec{a}$$

If $\text{Proj}_v[a]$ *is in the column space of* $\boldsymbol{V}$ *where* $V = \{\vec{v_1}\dots\vec{v_k}\} \in \mathbb{R}^{n\times k}$

$\text{Proj}_v[a]$ $\in$ *the column space of* V *then* $\text{Proj}_v[a] = w_1\vec{v_1} + ... + w_k\vec{v_k} = V\vec{w}$

Now if V *is orthogonal*

$$\begin{aligned}
\text{Proj}_v[a] &= \text{Proj}_{v_1}[a] + \text{Proj}_{v_2}[a] \\
&= \left(\frac{\vec{v_1}\vec{v_1}^T}{||v_1||^2} + \frac{\vec{v_2}\vec{v_2}^T}{||v_2||^2}\right)\vec{a}
\end{aligned}$$

If V *is orthonormal length* $= 1$ *then*

$$\begin{aligned}
\text{Proj}_v[a] &= \left(\vec{v_1}\vec{v_1}^T + \vec{v_2}\vec{v_2}^T\right)\vec{a} \\
&= \left(\begin{bmatrix} v_1{}_1^2 v_1 & v_{11}v_{12} & \dots & v_{11}v_{1n} \\ & \ddots & & \vdots \\ \vdots & & \ddots & \\ v_{1n}v_1 & \dots & & v_1^2 v_n \end{bmatrix} + \dots + \begin{bmatrix} v_{n1}^2 v_1 & v_{n1}v_{n2} & \dots & v_{n1}v_{nn} \\ & \ddots & & \vdots \\ \vdots & & \ddots & \\ v_{nn}v_{n1} & \dots & & v_n^2 v_n \end{bmatrix}\right) \\
&= \left(\begin{bmatrix} \vec{v_1} & \vec{v_2}\dots\vec{v_n} \end{bmatrix}\begin{bmatrix} \vec{v_1}^T \\ \vdots \\ \vec{v_n}^T \end{bmatrix}\right) \\
&= \boldsymbol{Q}\boldsymbol{Q}^\top
\end{aligned}$$

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

$$\begin{aligned}
\text{Proj}_Q[\boldsymbol{a}] &= \boldsymbol{Q}(\boldsymbol{Q}^\top\boldsymbol{Q})^{-1}\boldsymbol{Q}^\top\boldsymbol{a} \\
&= \boldsymbol{Q}I_n\boldsymbol{Q}^\top\boldsymbol{a} \\
&= \boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{a} \\
&= \sum q_i q_i^\top\boldsymbol{a} \\
&= \sum \text{Proj}_{q_i}[\boldsymbol{a}]
\end{aligned}$$

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

$SST = SSR + SSE$

Adding another column does not change the formula since it its another y value "i"

SSR is sum of the squared regression or how far the prediction is from the $\bar{y}$ and SSE is sum of

$$\begin{aligned} SST &= (SSR + i) + (SSE - k) \\ &= SSR + SSE + i - i \\ &= SST \end{aligned}$$

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

Problem 4

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

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

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

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

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

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

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

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

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

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

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

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

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

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