MATH 390.4 Spring 2020 Homework #3

Tziporah Horowitz

Monday 16th March, 2020

1 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 is most similar to k nearest neighbors.

(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 because it is parabolic, with different players peaking at different ages.

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

Scouts did better than PECOTA because they used a hybrid approach. They used both human judgement and statistics and therefore had more data available to them.

(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?

At the time, they did have complete 3-dimensional recordings of everything on the baseball field.

(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 predicting weather is that it is too chaotic. The weather behaves like a dynamic system, so f cannot be linear. Slight inaccuracies in prior data ($\mathbb{D}$) cause much larger inaccuracies in the predictions. In addition, to predict accurately, you have to increase the number of dimensions; however, if you increase the number of dimensions, you're left with exponentially more equations to solve.

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

Weathermen lie about the chance of rain for economic incentives. It give them room to for fewer complaints when it does rain in a low probability case. To get more "honest" forecasts, you should turn to The National Weather Service.

(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 cannot truly measure X so it becomes too noisy and the models are left with too much residual error, e.

(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?

Using the color of a lock to match it to known combinations. It is overly specific.

(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?

Neumann meant that when you model noise, it creates illustrations of reality that are inaccurate. It's important to recognize the difference between necessary features and noise.

(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.

Economic concepts are hard to predict because they rely on economic policy; once we begin to measure $\mathbb{D}$, its behavior begins to change. Similar to weather systems, economic data are chaotic. Nevertheless, meteorologists have a stronger understanding of the atmosphere and can therefore make more accurate forecasts, while economists cannot.

(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.

In order to make good predictions, you need a good X. Since there is no way of knowing the true causal inputs, the definition of X is subjective to your knowledge about y. So, the more you know about how y functions, the better your judgement will be about choosing X.

2 Multivariate Linear Model Fitting Using LS

(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.

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{c}}\left[\boldsymbol{c}^\top A \boldsymbol{c}\right] &= \frac{\partial}{\partial \boldsymbol{c}}\left[\boldsymbol{c}^\top \begin{bmatrix} \boldsymbol{a}_{1\cdot} \\ \vdots \\ \boldsymbol{a}_{n\cdot} \end{bmatrix} \boldsymbol{c}\right] \\
&= \frac{\partial}{\partial \boldsymbol{c}}\left[\boldsymbol{c}^\top \begin{bmatrix} \boldsymbol{a}_{1\cdot}\boldsymbol{c} \\ \vdots \\ \boldsymbol{a}_{n\cdot}\boldsymbol{c} \end{bmatrix}\right] \\
&= \frac{\partial}{\partial \boldsymbol{c}}\left[c_1 \boldsymbol{a}_{1\cdot}\boldsymbol{c} + \ldots + c_n \boldsymbol{a}_{n\cdot}\boldsymbol{c}\right] \\
&= \frac{\partial}{\partial \boldsymbol{c}}\left[c_1(a_{11}c_1 + \ldots + a_{1n}c_n) + \ldots + c_n(a_{n1}c_1 + \ldots + a_{nn}c_n)\right] \\
&= \frac{\partial}{\partial \boldsymbol{c}}\left[(c_1^2 a_{11} + c_1 a_{12} c_2 + \ldots + c_1 a_{1n} c_n) + \ldots + (c_n a_{n1} c_1 + \ldots + c_n^2 a_{nn})\right] \\
&\quad \Rightarrow \frac{\partial}{\partial c_1} = (a_{1\cdot} + a_{\cdot 1})\boldsymbol{c} \\
&= \begin{bmatrix} (a_{1\cdot} + a_{\cdot 1})\boldsymbol{c} \\ \vdots \\ (a_{n\cdot} + a_{\cdot n})\boldsymbol{c} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{a}_{1\cdot} \\ \vdots \\ \boldsymbol{a}_{n\cdot} \end{bmatrix} \boldsymbol{c} + \begin{bmatrix} \boldsymbol{a}_{\cdot 1} \\ \vdots \\ \boldsymbol{a}_{\cdot n} \end{bmatrix} \boldsymbol{c}
\end{aligned}
$$

(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.

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}}\left[\text{SSE}\right] &= \frac{\partial}{\partial \boldsymbol{w}}\left[\boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{w}^\top X^\top \boldsymbol{y} + \boldsymbol{w}^\top X^\top X \boldsymbol{w}\right] \\
&= \mathbf{0}_3 - 2X^\top \boldsymbol{y} + 2X^\top X \boldsymbol{w} \overset{\text{set}}{=} 0
\end{aligned}
$$

$$
\begin{aligned}
X^\top X \boldsymbol{w} &= X^\top \boldsymbol{y} \\
\boldsymbol{w} &= (X^\top X)^{-1} X^\top \boldsymbol{y}
\end{aligned}
$$

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

$$
\begin{aligned}
\boldsymbol{b} &= (X^\top X)^{-1} X^\top \boldsymbol{y} \\
&= \left(\begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \\
&= \left(\begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\
&= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\
&= \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum x_i y_i \end{bmatrix} \\
&= \frac{1}{n(\sum x_i^2 - n\bar{x}^2)} \begin{bmatrix} \sum x_i^2 \sum y_i - n\bar{x}\sum x_i y_i \\ -n\bar{x}\sum y_i + n\sum x_i y_i \end{bmatrix} \\
&= \begin{bmatrix} \bar{y} - \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}\bar{x} \\ \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \end{bmatrix} \\
&= \begin{bmatrix} \bar{y} - r\frac{s_y}{s_x}\bar{x} \\ r\frac{s_y}{s_x} \end{bmatrix}
\end{aligned}
$$

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

If X is rank deficient, $X^\top X$ will not be invertable, so to solve for $\boldsymbol{b}$ you should remove the linearly dependent columns of X.

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

Definition: $\text{rank}\,[X] = \dim[X] - N[X]$

(a) Let $X \in N[A]$,

$$
\begin{aligned}
AX = \mathbf{0} &\Rightarrow A^\top AX = \mathbf{0} \\
&\Rightarrow X \in N[A^\top A]
\end{aligned}
$$

so, $N[A] \subseteq N[A^\top A]$.

(b) Let $X \in N[A^\top A]$,

$$\begin{aligned} A^\top AX &= \mathbf{0} \\ X^\top A^\top AX &= \mathbf{0} \\ (AX)^\top AX &= \mathbf{0} \Rightarrow AX = \mathbf{0} \\ &\Rightarrow X \in N[A] \end{aligned}$$

so, $N[A^\top A] \subseteq N[A]$.

By (a) and (b), $N[A] = N[A^\top A]$ and $\dim[A] = \dim[A^\top A]$. Therefore, $\text{rank}\,[X] = \text{rank}\left[X^\top X\right]$.

(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.

$$\begin{aligned} \text{SSE} &= (\boldsymbol{y} - X\boldsymbol{b})^\top C(\boldsymbol{y} - X\boldsymbol{b}) \\ &= (\boldsymbol{y}^\top C^{\frac{1}{2}} - \boldsymbol{b}^\top X^\top C^{\frac{1}{2}})(C^{\frac{1}{2}}\boldsymbol{y} - C^{\frac{1}{2}}X\boldsymbol{b}) \\ &= \boldsymbol{y}^\top C\boldsymbol{y} - \boldsymbol{y}^\top CX\boldsymbol{b} - \boldsymbol{b}^\top X^\top C\boldsymbol{y} + \boldsymbol{b}^\top X^\top CX\boldsymbol{b} \\ \frac{\partial}{\partial \boldsymbol{b}}\,[\text{SSE}] &= \mathbf{0}_{p+1} - 2X^\top C\boldsymbol{y} + 2X^\top CX\boldsymbol{b} \overset{\text{set}}{=} \mathbf{0}_{p+1} \end{aligned}$$

$$\begin{aligned} X^\top CX\boldsymbol{b} &= X^\top C\boldsymbol{y} \\ \boldsymbol{b} &= (X^\top CX)^{-1}X^\top C\boldsymbol{y} \end{aligned}$$

(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.

$$\begin{aligned} R^2 &= \frac{\text{SSR}}{\text{SST}} \\ &= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \\ &= \frac{\sum \hat{y}_i^2 - n\bar{y}^2}{\sum(y_i - \bar{y})^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum (b_0 + b_1 x_i)^2 - n\bar{y}^2}{(n-1)s_y^2} \\
&= \frac{\sum (b_0^2 + 2b_0 b_1 x_i + b_1^2 x_i^2) - n\bar{y}^2}{(n-1)s_y^2} \\
&= \frac{nb_0^2 + 2b_0 b_1 \sum x_i + b_1^2 \sum x_i^2 - n\bar{y}^2}{(n-1)s_y^2} \\
&= \frac{(\bar{y} - b_1\bar{x})^2 n + 2(\bar{y} - b_1\bar{x}) b_1 n\bar{x} + b_1^2 \sum x_i^2 - n\bar{y}^2}{(n-1)s_y^2} \\
&= \frac{b_1^2 \sum x_i^2 - b_1^2 \bar{x}^2 n}{(n-1)s_y^2} \\
&= \frac{b_1^2 (\sum x_i^2 - \bar{x}^2 n)}{(n-1)s_y^2} \\
&= \frac{r^2 \frac{s_y^2}{s_x^2} \sum (x_i - \bar{x})^2}{(n-1)s_y^2} \\
&= r^2
\end{aligned}
$$

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

$$
\begin{aligned}
\bar{y} &= \frac{1}{n} \sum y_i \\
&= \frac{1}{n} \sum (b_0 + b_1 x_{1_i} + \ldots + b_p x_{p_i} + e_i) \\
&= \frac{1}{n} \sum b_0 + \frac{1}{n} \sum b_1 x_{1_i} + \ldots + \frac{1}{n} \sum b_p x_{p_i} + \frac{1}{n} \sum e_i \\
&= b_0 + b_1 \bar{x}_1 + \ldots + b_p \bar{x}_p + \bar{e} \\
&= b_0 + b_1 \bar{x}_1 + \ldots + b_p \bar{x}_p \\
&= g([1\ \bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_p])
\end{aligned}
$$

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

$$
\begin{aligned}
\bar{e} &= \frac{1}{n} \sum (y_i - \hat{y}_i) \\
&= \frac{1}{n} \sum y_i - \frac{1}{n} \sum (b_0 + b_1 x_{1_i} + \ldots + b_p x_{p_i}) \\
&= \bar{y} - (b_0 + b_1 \bar{x}_1 + \ldots + b_p \bar{x}_p) \\
&= \bar{y} - \bar{y} \\
&= 0
\end{aligned}
$$

(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.

Show $\boldsymbol{b} = (X^\top X)^{-1} X^\top \boldsymbol{y}$ is $\bar{y}_A$ if $x = A$, $\bar{y}_B$ if $x = B$ and $\bar{y}_C$ if $x = C$.

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

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

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

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

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

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

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

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

3 Orthogonal Projection and QR Decomposition

(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 $\forall n$.

(a) Symmetry: $I_n^\top = I_n$

(b) Idempotency: $I_n I_n = I_n$

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

$\mathbb{R}^n$

(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?

There are $p+1$ degrees of freedom, i.e. $p+1$ independent parameters.

(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?

Yes,

$$\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = X\boldsymbol{w}$$

$$\Rightarrow \begin{cases} \boldsymbol{x}_1^\top(\boldsymbol{y} - X\boldsymbol{w}) = 0 \\ \vdots \\ \boldsymbol{x}_n^\top(\boldsymbol{y} - X\boldsymbol{w}) = 0 \end{cases}$$

$$\begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} (\boldsymbol{y} - X\boldsymbol{w}) = \mathbf{0}_n$$

$$X^\top(\boldsymbol{y} - X\boldsymbol{w}) = \mathbf{0}_n$$

$$X^\top\boldsymbol{y} = X^\top X\boldsymbol{w}$$

$$\boldsymbol{w} = (X^\top X)^{-1}X^\top\boldsymbol{y}$$

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

No, the projection mapping yields the least squares error. Any further projection onto the colsp $[X]$ would only yield the same result.

(g) [harder] Prove that $\boldsymbol{Q}^\top = \boldsymbol{Q}^{-1}$ where $\boldsymbol{Q}$ is an orthonormal matrix such that colsp $[\boldsymbol{Q}]$ = 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.

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

$$\boldsymbol{Q}^\top\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{q}_1^\top \\ \vdots \\ \boldsymbol{q}_n^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_1 & \dots & \boldsymbol{q}_n \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{q}_1^\top\boldsymbol{q}_1 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \boldsymbol{q}_n^\top\boldsymbol{q}_n \end{bmatrix} = I_n$$

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

$$\begin{aligned}
\text{Proj}_V[\boldsymbol{a}] &= \boldsymbol{H}\boldsymbol{a}, \text{ Let } ||\boldsymbol{v}_1|| = \ldots = ||\boldsymbol{v}_d|| = 1 \\
\boldsymbol{H} &= \begin{bmatrix} \boldsymbol{v}_1\boldsymbol{v}_1^\top & \dots & \boldsymbol{v}_d\boldsymbol{v}_d^\top \end{bmatrix} \\
&= \begin{bmatrix} v_{11}\boldsymbol{v}_1 & \dots & v_{1n}\boldsymbol{v}_1 \end{bmatrix} + \begin{bmatrix} v_{21}\boldsymbol{v}_2 & \dots & v_{2n}\boldsymbol{v}_2 \end{bmatrix} + \ldots + \begin{bmatrix} v_{d1}\boldsymbol{v}_d & \dots & v_{dn}\boldsymbol{v}_d \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_d \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^\top \\ \vdots \\ \boldsymbol{v}_d^\top \end{bmatrix} \\
&= \boldsymbol{Q}\boldsymbol{Q}^\top
\end{aligned}$$

(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 $\boldsymbol{Q}$.

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

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

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

(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.

4 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 $\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\mathbf{0}_n, \sigma^2\boldsymbol{I}_n\right)$ 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 $\boldsymbol{U} \sim \mathcal{N}_n\left(_n, \boldsymbol{I}_n\right)$ independent of $\boldsymbol{V} \sim \mathcal{N}_n\left(_n, \boldsymbol{I}_n\right)$. Let θ be the r.v. model of the angle between $\boldsymbol{U}$ and $\boldsymbol{V}$. How is θ distributed?