MATH 390.4 / 650.2 Spring 2018 Homework #3t

Rebecca Strauss

Friday 23rd 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.

Hedgehog would (mistakenly) be searching for the 'absolute truth', $y = t(z_1, .., z_n)$, since they are "order-seekeing". They probably don't account for δ ignorance error, since they are "stubborn"-mistakes are blamed on bad luck. Hedgehogs overfit their models, by incorparting new x^* - "Stalwart"
Foxes are "empirical"-uses better observations for their historical records$\mathcal{D}$. "Adatable"-try different $\mathcal{H}$ and $\mathcal{A}$. Foxes know they are looking for h^*, (aware of all types of eror)"tolerant of complexity"

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

Harry Truman liked hedgehogs more because they're able to given a defintive answer with confidence, while foxes return ranges and porbabilites. A lot of people would prefer to listen to hedgehogs because of their confidence and "size" of prediction.

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

(d) [easy] Why are probabilistic classifiers (i.e. algorithms that output functions that return probabilities) better than vanilla classifiers (i.e. algorithms that only return the class label)? We will move in this direction in class soon.

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 (outome) data"?

Response data

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

Data scientist might implent a model that is subpar in predictive performance because it is simple and easy to explain. Even though the data scientist builds the model, someone without a statistical background might be using it for their job. In this case, it won't matter how well the model can predict if the user can't comprehend what the predictions mean.

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

In finlays first example the outcome metric was the response rate of people buying wine and they employed classification.

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

In the second wine problem, the outcome metric was gross profit on a case of wine, given that someone responded to the original campaign.REGREEION

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

Some organizations cannot implent predictive models to improve bussiness because they are unwilling to fully commit to the "model lifestyle"-implementing predictive anayltics for the first time requires a lot of money, IT power, and change. (the problem isnt building the model it's using the model.) Additionally, when a new model is implemented, the people using the model don't understand what the outcome metric means and will subsititute their own opinion in. Also, if I'm a generic worker at some company, and I'm given a model to "help" with my job, if I see the model working succesfully, I may misreport the results so I don't loss my job.

(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?

In the bankruptcy case, the bank forgot to take into account the real world limitations of acting on their prediction g. On average, for every 50 reviews the bank had to conduct, only 1 would voluntary foreclose. Even though the model was returning accurate predictions, each review process took about 2.5 hours. Hence, implementing the model on g alone is not cost effective.

(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.).

HUman judgement is beneficial to model-based predictions when the human can bring some outside information that is relevant to the predictions but that is not inferrable from the data. In the wine example, our $\mathcal{D}$ has a large p consisting of contact details, age, income, martial status... The classification model breaks up age into 3 classes: $\leq 34, 35-54, \geq 55$. Without applying any human judgement to this model, people can end up on our wine campaign list who are not old enough to legally buy wine. Wastefull

(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.

(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?

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

Finlay finds primary behaviors to be the most important data type in predictive modeling. Primary behaviors refer to past behaviors that are similiar to the behavior you are trying to model. This is important data because if you've done something once, you are likely to do it again. If you are someone who borrows money from your friends and always "forgets" to pay them back, the next time you borrow money, you probably wont pay it back.

(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 friutful for predictive ability?

You wouldn't gain any useful information because there's about a 90% of your spouse being +-5 years away from your age(spouses age is a function of your age, usually)

(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 is finite. It's lifespan has come to an end when "the relationships that were found between the predictor data and the outcome data when the model was orginially constructed no longer apply"

(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.).

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

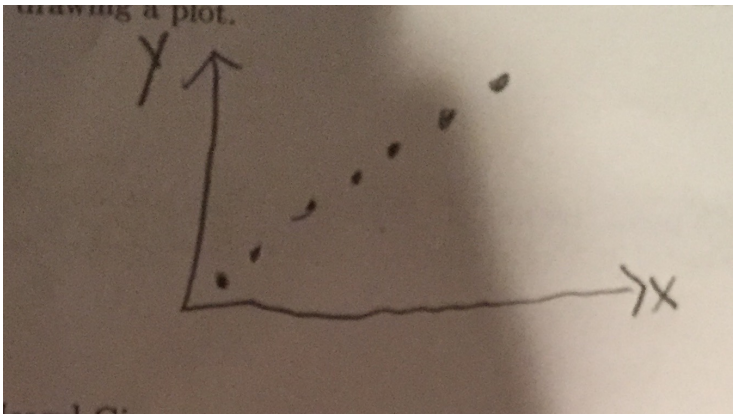
Finlays opinion regarding the hype about "big data"-huge benefits from using massive amounts of data, is a myth. Finlay says, in his own experince, once you go above a sample size of 10,000, the benefits you gain from adding more samples is marginal. Increasing from 10,000 to 100,000 can give you a 1-3% increase in predictive ability. However, the extra storage capacity and processing power needed to ruN predictions on a sample this large, might not make adding these extra samples worth it.

(p) [easy] What is Finlay's solution to "overfitting" (p84)?
Finlays solution to overfitting is to increase the sample size

Problem 3

These are questions about association and correlation.

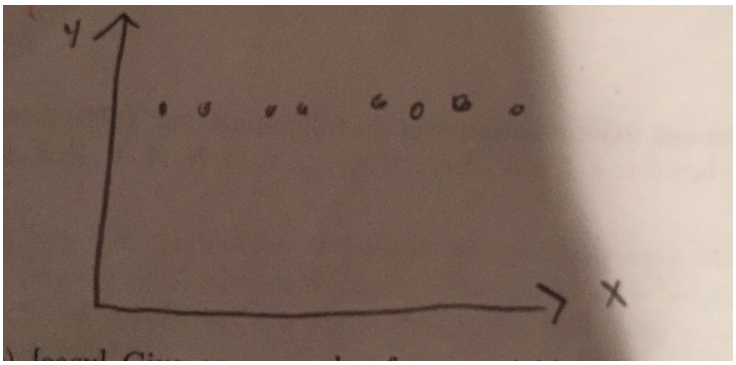
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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.
No. Correlation is a type of association.



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.

$$A\vec{c} = \begin{bmatrix} a_{11}c_1 + a_{12}c_2 \cdots + a_{1n}c_n \\ a_{21}c_1 + a_{22}c_2 \cdots + a_{2n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2}c_2 \cdots + a_{nn}c_n \end{bmatrix} \in \mathbb{R}^{n\times 1}$$

Then

$$\vec{c}^T(A\vec{c}) = c_1(a_{11}c_1+a_{12}c_2\cdots+a_{1n}c_n)+c_2(a_{21}c_1+a_{22}c_2\cdots+a_{2n}c_n)\cdots+c_n(a_{n1}c_1+a_{n2}c_2\cdots+a_{nn}c_n)$$

Take derivative with respect to c_1

$$\begin{aligned}\frac{\partial}{\partial c_1} &= (2a_{11}c_1 + a_{21}c_2 \cdots + a_{1n}c_n) + a_{21}c_2 + a_{31}c_3 \cdots + a_{n1}c_n \\ &= 2a_{11}c_1 + c_2(a_{12}a_{21}) + c_3(a_{13}a_{31}) \cdots + c_n(a_{1n}a_{n1})\end{aligned}$$

Take derivative with respect to c_2

$$\frac{\partial}{\partial c_2} = (2a_{22}c_2 + a_{21}c_1 \cdots + a_{2n}c_n) + a_{12}c_1 + a_{32}c_3 \cdots + a_{n2}c_n$$

(b) [easy] Given matrix $X \in \mathbb{R}^{n \times (p+1)}$, full rank and first column consisting of the $\mathbf{1}_n$ vector, rederive the least squares solution $\boldsymbol{b}$ (the vector of coefficients in the linear model shipped in the prediction function g). No need to rederive the facts about vector derivatives.

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

which we can rewrite as

$$(\vec{y} - \vec{\hat{y}})^T (\vec{y} - \vec{\hat{y}}) = (\vec{y}^T - \vec{\hat{y}}^T)(\vec{y} - \vec{\hat{y}})$$

Foil

$$\vec{y}^T\vec{y} - \vec{y}^T\vec{\hat{y}} - \vec{\hat{y}}^T\vec{y} + \vec{\hat{y}}^T\vec{\hat{y}}$$

Replace $\vec{y}^T\vec{\hat{y}} = \vec{y}\vec{\hat{y}}^T$

$$\vec{y}^T\vec{y} - 2\vec{y}\vec{\hat{y}}^T + \vec{\hat{y}}^T\vec{\hat{y}}$$

Replace $\vec{\hat{y}} = X\vec{w}$

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

Take partials

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

$$\frac{\partial}{\partial \vec{w}}[SSE] = \vec{0}_{p+1} - 2X^T\vec{y} + 2X^TX\vec{w}$$

Set $\frac{\partial}{\partial \vec{w}}[SSE] = 0$, factor out the 2

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

Since X has full rank $p+1$, we know that X^TX is invertible, and Let $\vec{w} = \vec{b}$

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

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

$$b_1 = r\frac{s_y}{s_x} = \frac{(n-1)S_{xy}}{(n-1)S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{1}$$

$$b_0 = \bar{y} - r\frac{s_y}{s_x}\bar{x} \tag{2}$$

From the previous question, we know that $\vec{b} = (X^TX)^{-1}X^T\vec{y}$ and since $p = 1$, $X \in \mathbb{R}^{n\times 2}$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_n \end{bmatrix} X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

using (note:$\sum_{i=1}^{n} y_i = n\bar{y}$)

$$X^T\vec{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

Now mutiply(note:$\sum_{i=1}^{n} x_i = n\bar{x}$)

$$X^TX = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Since X has full rank we can take the inverse of it(note: $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$)

$$\det\left[X^TX\right] = n\sum_{i=1}^{n} x_i^2 - (n\bar{x})^2 = n(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = nS_{xx}$$

The inverse is then

$$(X^TX)^{-1} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} = \frac{1}{S_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

Now we have

$$\vec{b} = (X^TX)^{-1}X^T\vec{y} = \frac{1}{S_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \frac{1}{S_{xx}} \begin{bmatrix} \frac{1}{n}(\sum_{i=1}^{n} x_i^2)(n\bar{y}) - \bar{x}\sum_{i=1}^{n} x_i y_i \\ -\bar{x}n\bar{y} + \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$\vec{b} = \frac{1}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \begin{bmatrix} \bar{y}\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \end{bmatrix}$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = r\frac{s_y}{s_x} \tag{3}$$

EquatioN 3 matches equation 1 from above
Now sub b_1 into the equation for b_0

$$b_0 = \frac{\bar{y}\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \bar{y} - \bar{x}r\frac{s_y}{s_x} \tag{4}$$

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

Since X is rank deficient, b is not in the column space of X. Instead of solving exactly for b, we could find the element in column space of X that is "closest" to b, which is its' projection. So we are looking for the hat matrix that gives us $H\vec{y} = Proj_b(X)$

(e) [difficult] Prove rank $[X]$ = 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.

(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.

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

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?

(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?

Let $\text{colsp}[X] = [X_1, X_2 \cdots, X_{p+1}] \in \mathbb{R}^{n \times p+1}$
Notes

- (1) $\boldsymbol{y} = \text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] + \vec{e}$
  $\Rightarrow \vec{e} = \boldsymbol{y} - \text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}]$
- (2)$\vec{e} \cdot \text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = \vec{0}$ and $\vec{e} \cdot \text{colsp}[X] = \vec{0}$

Since $\boldsymbol{y}$ is being orthogonally projected onto the column space of X, we know that $\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] \in \text{colsp}[X]$. In other words, $\exists$ a scalar $\vec{w}$ s.t

$$\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = w_1 X_1 + w_2 X_2 \cdots w_{p+1} X_{p+1} = X\vec{w} \tag{5}$$

By note 1 $\vec{e} = \boldsymbol{y} - X\vec{w}$ and by note 2 $X^T(\boldsymbol{y} - X\vec{w}) = \vec{0}$

$$\begin{aligned} X^T(\boldsymbol{y} - X\vec{w}) &= \vec{0} \\ X^T\boldsymbol{y} - X^T X\vec{w} &= \vec{0} \\ X^T\boldsymbol{y} &= X^T X\vec{w} \end{aligned}$$

Since X has full rank $p+1$, $X^T X$ is also of full rank which means we can take the inverse firecetly

$$(X^T X)^{-1} X^T \boldsymbol{y} = \vec{w} \tag{6}$$

Now plug $\vec{w}$ into equation 1 and we get

$$\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = X(X^T X)^{-1} X^T \boldsymbol{y} \tag{7}$$

Which 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.

Each residual, e_i, is calculated by $e_i = y_i - \hat{y}_i$ where $\hat{y}_i = X_i w_i$. Hence $\vec{e}$ is dependent on X. Running a regression on dependent variables will return a smaller error rate than if you ran the regression on independent variables. So, yes, this process will yield a better model on iteration 2.

(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.

Calculate $Q^TQ = \begin{bmatrix} \leftarrow q_1 \rightarrow \\ \leftarrow q_2 \rightarrow \\ \vdots \\ \leftarrow q_{p+1} \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ q_1 & q_2 & \cdots & q_{p+1} \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \ddots & & & \\ 0 & 0 & \cdots & 1 \end{bmatrix}$ So $Q^TQ = I_{p+1}$. Linear algebra rules tell us that if we have a vector times its' own transpose that equals the identity matrix, the transpose of that matrix must be equal to the inverse. In basic math, if we have $xy = 1$, without loss of generality, $x = y^{-1}$. Here we just have the vector version of this relationship. Hence $Q^T = Q^{-1}$

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

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

(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.

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

(i) [E.C.] Prove that rank $[H]$ = tr $[H]$.