MATH 390.4 / 650.3 Spring 2018 Homework #3t

Adina Bechhofer

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.

While the hedgehogs pick and choose the x's that will generate predictions that fit with their world view, the foxes use all of the x's available to them.

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

Hedgehogs seem so much more confident in their predictions. Truman wanted a "one handed economist" because he disliked the fox's tendency to weigh things on both hands. Many people like hedgehogs because they seem confident, they're more likely to make bold predictions which are biased.

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

The more pieces of information a hedgehog acquires, the more oppurtunity he/ she has to manipulate and twist them to fit a biased 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.

A probabilistic classifier returns a probability of an event occurring rather than a binary prediction. If the model is good, over the long run, the event will occur p% of the times. This model is much more informative.

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"?

Response variable, dependent variable, outcome, output.

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

Although predictive power is important, the model must comply to other requirements posed by the business implementing it. For example, a business could require that the model will be simple and explicable, so it can be understood by non experts.

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

The outcome metric was classification into categories of likelihood to respond to advertising and purchase wine. The learning was done by fitting a decision tree with 11 nodes, each one representing a different likelihood to respond to advertising.

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

The outcome was a continuous variable that predicts profit from contacting a customer. the learning was done by regressing on the existing data of which wine people from different categories ended up buying.

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

1. Ignorance that leads to lack of motivation to implement new techniques.
2. Poor management information, and weak governance. The higher ups aren't checking to see that decisions are made using the model,and insufficient controls are put in place to ensure that score-based decisions are executed correctly.
3. Having to hire and fire workers.

(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 foreclosure example, using the model wasn't profitable. This is because hours of financial consulting were spent with customers who were likely to foreclose as predicted by the model, while only one of 500 went for voluntary foreclosure.

(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 judgment seeks to find the reason or cause to things, while data driven models correlate variables. When human judgment get closer to identifying more x's that are close to the z's it's beneficial to employ that.

(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.

Finaly believes that most models can have only about 30 linearly dependent meaningful features. It is correct that piling up meaningless data will increase the R^2 of the model. However, those features don't actually explain $\boldsymbol{y}$.

(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?

If it significantly improves the performance of the model.

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

Behavioral features and associations with other people to predict burglaries.

(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?

Primary behavior. This is because people rarely change. The way a person acted in the past is a strong predictor of the way he or she will continue to behave.

(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?

Although not perfectly correlated, the age of a person and the age of their spouse is highly correlated. Thus, adding column 18 to the model, won't add much 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.).

$\boldsymbol{y} = t(z_1(t), \ldots z_n(t))$

The z's depend on time. As long as the x's gathered are still close to the z's, g will be close enough to t, and the model will be usable. However, after long enough, new x's will be needed to train the model.

(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.).

Essentially, we need a $\mathbb{D} =< X, \boldsymbol{y} >$ that's large enough to capture the behaver of individuals inside the dataset. However, we don't want it to be large to the point where it becomes very hard to analyze the behavior.

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

Since data storage became much less costly, people have been storing loads of useless data. A large portion of the data stored should just be discarded.

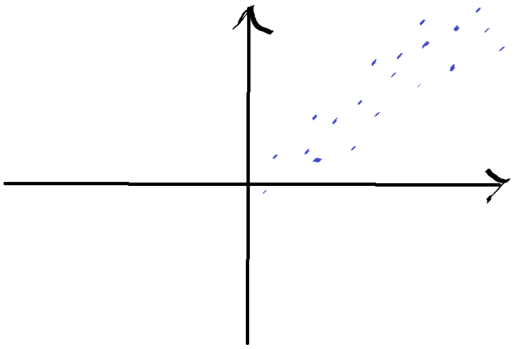
(p) [easy] What is Finlay's solution to "overfitting" (p84)?

Use a larger sample of data.

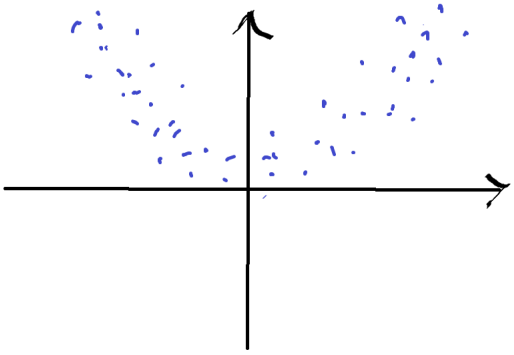
Problem 3

These are questions about association and correlation.

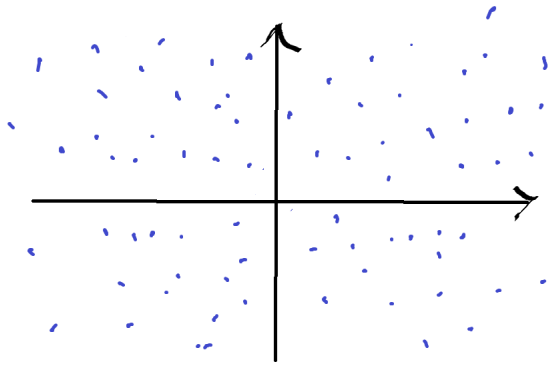
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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 subset of association which refers to a linear relationship between variables. Variables that aren't associated, have no relationship.

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.

$$\boldsymbol{c}^\top A\boldsymbol{c} = \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ \dots \\ c_n \end{bmatrix}$$

$$= \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} \cdot \begin{bmatrix} c_1a_{11} + c_2a_{12} + \dots + c_na_{1n} \\ c_1a_{21} + c_2a_{22} + \dots + c_na_{2n} \\ \dots \\ c_1a_{n1} + c_2a_{n2} + \dots + c_na_{nn} \end{bmatrix}$$

$$= c_1\left(c_1a_{11} + c_2a_{12} + \dots + c_na_{1n}\right) + c_2\left(c_1a_{21} + c_2a_{22} + \dots + c_na_{2n}\right) + \dots + c_n\left(c_1a_{21} + c_2a_{22} + \dots + c_na_{2n}\right)$$

$$= \sum_{i=1}^n c_i\left(\sum_{j=1}^n c_ja_{ij}\right)$$

$$\frac{\partial}{\partial \boldsymbol{c}_i}\left[\boldsymbol{c}^\top A\boldsymbol{c}\right] = \frac{\partial}{\partial \boldsymbol{c}_i}\left[\sum_{i,j=1}^n c_ic_ja_{ij}\right]$$

$$= \sum_{j=1}^n c_ja_{ij} + c_ja_{ji}$$

(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.

$$SSE = \sum \left(y_i - \hat{\vec{y}}\right)^2 = \left(\vec{y} - \hat{\vec{y}}\right)^T \left(\vec{y} - \hat{\vec{y}}\right) = \left(\vec{y}^T - \hat{\vec{y}}^T\right)\left(\vec{y} - \hat{\vec{y}}\right)$$

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

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

$$= \vec{y}^T\vec{y} - 2\left(X\vec{b}\right)^T\vec{y} + \left(X\vec{b}\right)^T\left(X\vec{b}\right)$$

$$= \vec{y}^T\vec{y} - 2\vec{b}^T X^T\vec{y} + \vec{b}^T X^T X\vec{b}$$

$$\frac{\partial}{\partial\vec{b}}\left[\vec{y}^T\vec{y} - 2\vec{b}^T X^T\vec{y} + \vec{b}^T X^T X\vec{b}\right] = -2X^T\vec{y} + 2X^T X\vec{b} = 0$$

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

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

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

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

$$X \in \mathbb{R}^{n\times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix}$$

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

$$\sum x_i = n\bar{x}$$

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

$$\left(X^T X\right)^{-1}\left(X^T\vec{y}\right) = \frac{1}{n\sum x_i^2 - n^2\bar{x}^2}\begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$$

$$= \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \cdot \begin{bmatrix} n\bar{y} \\ \sum y_i x_i \end{bmatrix}$$

$$= \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} n\bar{y}\sum x_i^2 - n\bar{\sum} y_i x_i \\ -n^2\bar{x}\bar{y} + n\sum y_i x_i \end{bmatrix}$$

$$b_0 = \frac{\bar{y}\left(\sum x_i^2 - n\bar{x}^2\right) - \bar{x}\left(\sum y_i x_i - n\bar{x}\bar{y}\right)}{\sum x_i^2 - n\bar{x}^2}$$

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

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

Eliminate linearly dependent feature columns in the data matrix, until X is full rank.

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

rank $[x]$ = Dim of whole space - $N(X)$

Prove that $N(A) = N(A^T A)$:
$N(A) \subset N(A^T A)$
$x \in N(A)$
$Ax = 0$
$A^T Ax = A^T 0 = 0 \quad \Rightarrow \quad x \in N(A^T A)$
so $N(A) \subset N(A^T A)$
$N(A^T A) \subset N(A)$
$x \in N(A^T A)$
$A^T Ax = 0$
$x^T A^T ax = x^T 0 = 0$
$(Ax)^T (Ax) = 0$
$||Ax||^2 = 0$
$Ax = 0$
so $N(A^T A) \subset N(A)$

Therefore, $N(X) = N(X^T X)$ and rank $[X]$ = rank $\left[X^T X\right]$

(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.

$$SSE = \left(\boldsymbol{y} - X\vec{b}\right)^\top C\left(\boldsymbol{y} - X\vec{b}\right)$$
$$= \left(\boldsymbol{y}^\top C\boldsymbol{y} - \boldsymbol{y}^\top CX\vec{b} - \vec{b}^\top X^\top C\boldsymbol{y} + \vec{b}^\top X^\top CXb\right)$$

$$\frac{\partial SSE}{\partial \vec{b}} \;=\; -2X^\top C\boldsymbol{y} + 2X^\top CX\vec{b} \;=\; 0$$

$$\vec{b} = \left(X^\top CX\right)^{-1}X^\top C\boldsymbol{y}$$

(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.

$b_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$
$b_0 = \bar{y} - b_1\bar{x}$
$\hat{\boldsymbol{y}}_i = b_0 + b_1x_i \;=\; \bar{y} - \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right]\bar{x} \;+\; \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right]x_i$

$$R^2 = \frac{SSR}{SST}$$
$$= \frac{\sum(\hat{\boldsymbol{y}}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$
$$= \frac{\sum\left(\bar{y} - \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right]\bar{x} \;+\; \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right]x_i - \bar{y}\right)^2}{\sum\left(y_i - \bar{y}\right)^2}$$
$$= \frac{\sum\left(\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right)^2\left(x_i - \bar{x}\right)^2}{\sum\left(y_i - \bar{y}\right)^2}$$
$$= \frac{\left[\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]^2}{\sum\left(x_i - \bar{x}\right)^2\sum\left(y_i - \bar{y}\right)^2}$$
$$= \frac{\left[Cov(x,y)\right]^2}{s_x^2 s_y^2}$$
$$= r^2$$

(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.

Show that $\begin{bmatrix}1 & \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_{p+1}\end{bmatrix}\vec{b} = \bar{y}$

$$\hat{\boldsymbol{y}}^* = b_0 + b_1\bar{x}_1 \;+\; b_2\bar{x}_2 \;+\; \ldots \;+\; b_{p+1}\bar{x}_{p+1}$$
$$\hat{\boldsymbol{y}}^* = \frac{1}{n}\sum b_0 + \frac{1}{n}\sum b_1x_{i1} \;+\; \frac{1}{n}\sum b_2x_{i2} \;+\; \ldots \;+\; \frac{1}{n}\sum b_{p+1}x_{ip+1}$$

Since $\hat{\boldsymbol{y}}_i = b_0 + b_1 x_{i1} \ + \ b_2 x_{i2} \ + \ \ldots \ + \ b_{ip+1} x_{ip+1}$

$$\hat{\boldsymbol{y}}^* = \frac{1}{n} \sum \hat{\boldsymbol{y}}_i$$

As proven on previous homework, $\frac{1}{n} \sum \left(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i \right) \ = \ \frac{1}{n} \sum e_i \ = \ 0$
Therefore, $\frac{1}{n} \sum \hat{\boldsymbol{y}}_i = \frac{1}{n} \sum \boldsymbol{y}_i = \bar{y}$
Thus,

$$\hat{\boldsymbol{y}}^* = \bar{y} = b_0 + b_1 \bar{x}_1 \ + \ b_2 \bar{x}_2 \ + \ \ldots \ + \ b_{p+1} \bar{x}_{p+1}$$

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?

There are p+1 degrees of freedom. This means that the resulting model: $\hat{\boldsymbol{y}} = w_0 + w_1 x_1 + \cdots + w_p x_p$ has p+1 weight parameters that can be adjusted.

(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?

$\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = X\vec{w}$
$X^T(\boldsymbol{y} - X\vec{w}) = 0$ Because of orthogonality
$X^T\boldsymbol{y} - X^T X\vec{w} = 0$
$X^T\boldsymbol{y} = X^T X\vec{w}$
$\vec{w} = (X^T X)^{-1} X^T \boldsymbol{y}$
$\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = X\vec{w} = X(X^T X)^{-1} X^T \boldsymbol{y} = H\boldsymbol{y}$

Yes. This 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.

No. The projection onto colsp$[X]$ gives the minimum square error on the first iteration. Going through that process again would yeild the same result because of the idempotency of the projection matrix.
$H = X(X^T X)^{-1} X^T$ and $H \cdot H = H$

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

Prove: $Q^\top Q = I$

$$\begin{bmatrix} \leftarrow & q_{\cdot 1} & \rightarrow \\ \leftarrow & q_{\cdot 2} & \rightarrow \\ & \dots & \\ \leftarrow & q_{\cdot n} & \rightarrow \end{bmatrix} \cdot \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ q_{\cdot 1} & q_{\cdot 2} & \dots & q_{\cdot n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} q_{\cdot 1}^\top q_{\cdot 1} & q_{\cdot 1}^\top q_{\cdot 2} & \dots & q_{\cdot 1}^\top q_{\cdot n} \\ q_{\cdot 2}^\top q_{\cdot 1} & q_{\cdot 2}^\top q_{\cdot 2} & \dots & q_{\cdot 2}^\top q_{\cdot n} \\ \dots & \dots & \dots & \dots \\ q_{\cdot n}^\top q_{\cdot 1} & q_{\cdot n}^\top q_{\cdot 2} & \dots & q_{\cdot n}^\top q_{\cdot n} \end{bmatrix}$$

$q_{\cdot i}^\top q_{\cdot i} = ||q_{\cdot i}||^2$, and $q_{\cdot i}^\top q_{\cdot j} = 0$ when $i \neq j$ because of orthonormality of Q.

$$Q^\top Q = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} = I$$

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

In previous part I proved: $Q^\top Q = I$

Show that $QQ^T = X\left(X^T X\right)^{-1} X^T$

$$\begin{aligned} &(QR)\Big((QR)^T QR\Big)^{-1}(QR)^T \\ &= QR\big(R^T Q^T QQ\big)^{-1} R^T Q^T \\ &= QR\big(R^T IR\big)^{-1} R^T Q^T \\ &= QR\big(R^T R\big)^{-1} R^T Q^T \\ &= QRR^{-1}(R^T)^{-1} R^T Q^T \\ &= QIIQ^T \\ &= QQ^T \end{aligned}$$

(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.

Projection onto each column of Q:

$\text{Proj}_{q_{\cdot i}}[\vec{a}] = \frac{q_{\cdot i} q_{\cdot i}^\top}{||q_{\cdot i}||^2}\vec{a}$

Since Q is orthonormal, $||q_{\cdot i}||^2 = 1$, and $\text{Proj}_{q_{\cdot i}}[\vec{a}] = q_{\cdot i} q_{\cdot i}^\top \vec{a}$

$$\sum_{i=1}^{p+1} \text{Proj}_{q_{\cdot i}}[\vec{a}] = \sum_{i=1}^{p+1} q_{\cdot i} q_{\cdot i}^{\top} \vec{a} \;=\; QQ^{\top}\vec{a}$$

Since all the columns in Q are orthogonal to each other.

(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.

As SSR increases $\Rightarrow$ SSE decreases. This is because SST = SSR + SSE, and SST is a measure of variance in $\boldsymbol{y}$; it doesn't depend on the x's or g.
Now, add a random x_{p+2} column to $\mathcal{X}$. rank $[X]$ = p+2.
$SSR_{new} = \sum_{i=1}^{n} \left(\hat{\boldsymbol{y}}_i - \bar{y}\right)^2 = \sum_{j=1}^{p+1} ||\text{Proj}_{q\cdot j}[\boldsymbol{y}]||^2 + ||\text{Proj}_{q\cdot p+2}[\boldsymbol{y}]||^2 \;\geq SSR_{old}$
Thus, $SSE_{new} \leq SSE_{old}$, and $R^2_{new} \geq R^2_{old}$

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

This means that any random, non orthogonal vector added to $\mathcal{X}$ would increase R^2. It makes it hard to determine whether or not the x's in the model actually explain $\boldsymbol{y}$.

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