ECE 4710J, Spring 2022

Homework #4

Properties of Simple Linear Regression

1. (10 points) In lecture, we spent a great deal of time talking about simple linear regression. To briefly summarize, the simple linear regression model assumes that given a single observation x, our predicted response for this observation is $\hat{y} = \theta_0 + \theta_1 x$. (Note: In this problem we write (θ_0, θ_1) instead of (a, b) to more closely mirror the multiple linear regression model notation.)

   We saw that the $\theta_0 = \hat{\theta_0}$ and $\theta_1 = \hat{\theta_1}$ that minimize the average L_2 loss for the simple linear regression model are:

$$\hat{\theta_0} = \bar{y} - \hat{\theta_1}\bar{x}$$
$$\hat{\theta_1} = r\frac{\sigma_y}{\sigma_x}$$

   Or, rearranging terms, our predictions $\hat{y}$ are:

$$\hat{y} = \bar{y} + r\sigma_y \frac{x - \bar{x}}{\sigma_x}$$

   (a) (4 points) As we saw in lecture, a residual e_i is defined to be the difference between a true response y_i and predicted response $\hat{y}_i$. Specifically, $e_i = y_i - \hat{y}_i$. Note that there are n data points, and each data point is denoted by (x_i, y_i).

   Prove, using the equation for $\hat{y}$ above, that $\sum_{i=1}^{n} e_i = 0$ (meaning the sum of the residuals is zero).

Answer. According to the formula:

$$\hat{b} = r\frac{\sigma_y}{\sigma_x}$$
$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$
$$e_i = y_i - \hat{y}_i$$

and

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

We have

$$\begin{aligned}
e_i &= y_i - \hat{y}_i \\
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (y_i - \bar{y}_i) \\
&= \sum_{i=1}^{n} \left(y_i - \bar{y} - r\sigma_y \frac{x_i - \bar{x}}{\sigma_x} \right) \\
&= \sum_{i=1}^{n} y_i - n\bar{y} - r\sigma_y \frac{1}{\sigma_x} \sum_{i=1}^{n} x_i + \frac{r \times \frac{\sigma_y}{\sigma_x} \cdot \bar{x}}{1} \\
&= \sum_{i=1}^{n} y_i - n \times \frac{1}{n} \sum_{i=1}^{n} y_i - r\sigma_y \frac{1}{\sigma_x} \cdot \left(\sum_{i=1}^{n} x_i - \frac{1}{n} \times n \times \sum_{i=1}^{n} x_i \right). \\
&= 0
\end{aligned}$$

(b) (3 points) Using your result from part (a), prove that $\bar{y} = \bar{\hat{y}}$.

Answer. Since we already have the answers in part a

$$\begin{aligned}
&\overline{\widehat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \frac{1}{n} \sum_{i=1}^{n} \left(\bar{y} + r\sigma_y \frac{x_i - \bar{x}}{\sigma_x} \right) \\
&= \frac{1}{n} \times n \times \bar{y} + \frac{1}{n} \sum_{i=1}^{n} \left(r\sigma_y \frac{x_i}{\bar{\sigma}_x} - r\sigma_y \frac{\overline{x}}{\sigma_x} \right) \\
&= \bar{y}
\end{aligned}$$

(c) (3 points) Prove that $(\bar{x}, \bar{y})$ is on the simple linear regression line.

Answer. It is on the line

$$\overline{\widehat{y}} = \bar{y} + r\sigma_y \frac{\overline{\widehat{x}} - \bar{x}}{\sigma_x}$$

Geometric Perspective of Least Squares

2. (10 points) We also viewed both the simple linear regression model and the multiple linear regression model through linear algebra. The key geometric insight was that if we train a model on some design matrix $\mathbb{X}$ and true response vector $\mathbb{Y}$, our predicted response $\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$ is the vector in span($\mathbb{X}$) that is closest to $\mathbb{Y}$ ($\hat{\mathbb{Y}}$ is the orthogonal projection of $\mathbb{Y}$ onto the span($\mathbb{X}$)).

   In the simple linear regression case, our optimal vector θ is $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]^T$, and our design matrix is

$$\mathbb{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} | & | \\ \mathbb{1} & \vec{x} \\ | & | \end{bmatrix}$$

   This means we can write our predicted response vector as $\hat{\mathbb{Y}} = \mathbb{X}\begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} = \hat{\theta}_0\mathbb{1} + \hat{\theta}_1\vec{x}$.

   Note, in this problem, $\vec{x}$ refers to the n-length vector $[x_1, x_2, ..., x_n]^T$. In other words, it is a feature, not an observation.

   For this problem, assume we are working with the **simple linear regression model**, though the properties we establish here hold for any linear regression model that contains an intercept term.

   (a) (4 points) Using the geometric properties from lecture, prove that $\sum_{i=1}^{n} e_i = 0$.

   Hint: Recall, we define the residual vector as $e = \mathbb{Y} - \hat{\mathbb{Y}}$, and $e = [e_1, e_2, ..., e_n]^T$.

Answer.

$$x = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & 1 \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & \bar{x} \\ 1 & 1 \end{bmatrix} + y = \theta_0 1 + \theta_1 x$$

and

$$\begin{aligned} R(\theta) &= \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, \hat{y}_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_i x_i\right)^2 \\ &= 0 \end{aligned}$$

Finally,

$$e_i = y_i - \hat{y}_i = y_i - \hat{\theta}_0 - \hat{\theta}_i x_i$$
$$\sum_{i=1}^{n} e_i = 0$$

(b) (3 points) Explain why the vector $\vec{x}$ (as defined in the problem) and the residual vector e are orthogonal. *Hint: Two vectors are orthogonal if their dot product is 0.*

Answer. Since we have

$$R(\theta) = \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\theta\|_2^2$$
$$\hat{\theta} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}$$

We need to prove that

$$\vec{e} \cdot \vec{x} = 0$$

$\therefore$ Least squares estimators must satisfy

$$\begin{aligned}\left.\frac{\partial S}{\partial \theta}\right|_{\hat{\theta}} &= -2x'y + 2x' \times \hat{\theta} = 0\\ &\Rightarrow x' \times \hat{\theta} = x'y\\ &\Rightarrow x'y - x' \times \hat{\theta} = 0\\ &\Rightarrow x'(y - x\hat{\theta}) = 0\\ &\Rightarrow x'e = 0\end{aligned}$$

vector x the residual vector e are orthogonal

(c) (3 points) Explain why the predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

Answer. We can use the answers in part a and part b, we have

$$\sum_{i=1}^{n} e_i = 0$$

and

$$\vec{e} \cdot \vec{x} = 0$$
$$\begin{aligned}\sum_{i=1}^{n}\left[e_i \cdot \left(\hat{\theta}_0 + \hat{\theta}_1 x_i\right)\right] &= \vec{e} \cdot \left(\hat{\theta}_0 + \hat{\theta}_1 \vec{x}\right)\\ &= \vec{e} \cdot \vec{Y} = 0\end{aligned}$$

they are orthogonal.

Properties of a Linear Model With No Constant Term

Suppose that we don't include an intercept term in our model. That is, our model is now

$$\hat{y} = \gamma x,$$

where γ is the single parameter for our model that we need to optimize. (In this equation, x is a scalar, corresponding to a single observation.)

As usual, we are looking to find the value $\hat{\gamma}$ that minimizes the average L_2 loss (mean squared error) across our observed data $\{(x_i, y_i)\}, i = 1, \ldots, n$:

$$R(\gamma) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2$$

The normal equations derived in lecture no longer hold. In this problem, we'll derive a solution to this simpler model. We'll see that the least squares estimate of the slope in this model differs from the simple linear regression model, and will also explore whether or not our properties from the previous problem still hold.

3. (5 points) Use calculus to find the minimizing $\hat{\gamma}$. That is, prove that

$$\hat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Note: This is the slope of our regression line, analogous to $\hat{\theta}_1$ from our simple linear regression model.

Answer. Since we have

$$R(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \gamma x_i\right)^2$$

and we know that it should satisfy that

$$\frac{\partial R(\gamma)}{\partial \gamma} = 0$$

Then,

$$\frac{1}{n}\sum_{i=1}^{n} 2\left(y_i - \gamma x_i\right) \cdot \left(-x_i\right) = 0$$

which means

$$\sum_{i=1}^{n} \gamma x_i^2 - x_i y_i = 0$$

and

$$\hat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2}$$

4. (10 points) For our new simplified model, our design matrix $\mathbb{X}$ is:

$$\mathbb{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ \vec{x} \\ | \end{bmatrix}.$$

Therefore our predicted response vector $\hat{\mathbb{Y}}$ can be expressed as $\hat{\mathbb{Y}} = \hat{\gamma}\vec{x}$. ($\vec{x}$ here is defined the same way it was in Question 2.)

Earlier in this homework, we established several properties that held true for the simple linear regression model that contained an intercept term. For each of the following four properties, state whether or not they still hold true even when there isn't an intercept term. Be sure to justify your answer.

(a) (2 points) $\sum\limits_{i=1}^{n} e_i = 0$.

Answer.

False.

Since we have the formula

$$e_i = y_i - \gamma \cdot x_i$$

and in problem 3 we know

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(y_i - \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} x_i \right)$$

It is not necessarily 0 because we cannot simply delete x_i or y_i here since there exists Σ .

(b) (3 points) The column vector $\vec{x}$ and the residual vector e are orthogonal.

Answer.

True.

The same as problem 2

$$\begin{aligned}
\vec{e} \cdot \vec{x} &= \sum_{i=1}^{n} (y_i - \hat{\gamma} x_i) \cdot x_i \\
&= \sum_{i=1}^{n} \left(y_i x_i - \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} x_i^2 \right) \\
&= 0.
\end{aligned}$$

(c) (3 points) The predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

Answer.

True.
We can use the answers in part a and part b, we have

$$\sum_{i=1}^{n} e_i = 0$$

and

$$\vec{e} \cdot \vec{x} = 0$$

$$\begin{aligned}
\sum_{i=1}^{n} [e_i \cdot (\gamma x_i)] &= \vec{e} \cdot (\gamma \vec{x}) \\
&= \vec{e} \cdot \vec{Y} = 0
\end{aligned}$$

they are orthogonal.

(d) (2 points) $(\bar{x}, \bar{y})$ is on the regression line.

Answer.

True.
Since we have $\overline{\widehat{y}} = \gamma \overline{\widehat{x}}$
It also fits the regression model and it is on the line

MSE "Minimizer"

5. (15 points) Recall from calculus that given some function $g(x)$, the x you get from solving $\frac{dg(x)}{dx} = 0$ is called a *critical point* of g – this means it could be a minimizer or a maximizer for g. In this question, we will explore some basic properties and build some intuition on why, for certain loss functions such as squared L_2 loss, the critical point of the empirical risk function (defined as average loss on the observed data) will always be the minimizer.

   Given some linear model $f(x) = \gamma x$ for some real scalar γ, we can write the empirical risk of the model f given the observed data $\{x_i, y_i\}, i = 1, \ldots, n$ as the average L_2 loss, also known as mean squared error (MSE):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2.$$

   (a) (2 points) Let's break the function above into individual terms. Complete the following sentence by filling in the blanks using one of the options in the parenthesis following each of the blanks:

   The mean squared error can be viewed as a sum of n ____ (linear/quadratic/logarithmic/exponential) terms, each of which can be treated as a function of ____ (x_i/y_i/γ).

   Answer. quadratic / γ

   (b) (4 points) Let's investigate one of the n functions in the summation in the MSE. Define $g_i(\gamma) = \frac{1}{n}(y_i - \gamma x_i)^2$ for $i = 1, \ldots, n$. Recall from calculus that we can use the 2nd derivative of a function to describe its curvature about a certain point (if it is facing concave up, down, or possibly a point of inflection). You can take the following as a fact: A function is convex if and only if the function's 2nd derivative is non-negative on its domain. Based on this property, verify that g_i is a **convex function**.

   Answer.

$$\frac{d}{d\gamma}g_i(\gamma) = \frac{1}{n}\frac{d}{d\gamma}\left(y_i - \gamma x_i\right)^2 = \frac{-2x_i}{n}\left(y_i - \gamma x_i\right)$$

$$\frac{d^2}{d\gamma^2}g_i(\gamma) = \frac{d}{d\gamma}\frac{d}{d\gamma}g_i(\gamma) = \frac{d}{d\gamma}\left(\frac{-2x_i}{n}\left(y_i - \gamma x_i\right)\right) = \frac{d}{d\gamma}\left(\frac{-2x_i y_i}{n} + \frac{2\gamma x_i^2}{n}\right) = \frac{2x_i^2}{n}$$

   Since x_i^2 is non-negative so that the second derivative is always non-negative, and $g_i(\gamma)$ is convex.

(c) (3 points) Briefly explain in words why given a convex function $g(x)$, the critical point we get by solving $\frac{dg(x)}{dx} = 0$ minimizes g. You can assume that $\frac{dg(x)}{dx}$ is a function of x (and not a constant).

Answer. The convex function $g(x)$ has a critical point it should be minimum. Here when $\frac{dg(x)}{dx} = 0$ it has a minimum.

(d) (4 points) Now that we have shown that each term in the summation of the MSE is a convex function, one might wonder if the entire summation is convex given that it is a sum of convex functions.

Let's look at the formal definition of a **convex function**. Algebraically speaking, a function $g(x)$ is convex if for any two points $(x_1, g(x_1))$ and $(x_2, g(x_2))$ on the function,

$$g(cx_1 + (1-c)x_2) \leq cg(x_1) + (1-c)g(x_2)$$

for any real constant $0 \leq c \leq 1$.

The above definition says that, given the plot of a convex function $g(x)$, if you connect 2 randomly chosen points on the function, the line segment will always lie on or above $g(x)$ (try this with the graph of $y = x^2$).

i. (2 points) Using the definition above, show that if $g(x)$ and $h(x)$ are both convex functions, their sum $g(x) + h(x)$ will also be a convex function.

Answer. According to the formula,

$$g\left(cx_1 + (1-c)x_2\right) \leq cg\left(x_1\right) + (1-c)g\left(x_2\right)$$

$$h\left(cx_1 + (1-c)x_2\right) \leq ch\left(x_1\right) + (1-c)h\left(x_2\right)$$

namely,

$$(g+h)\left(cx_1 + (1-c)x_2\right) \leq c(g+h)\left(x_1\right) + (1-c)(g+h)\left(x_2\right)$$

$g + h$ is convex

ii. (2 points) Based on what you have shown in the previous part, explain intuitively why the sum of n convex functions is still a convex function when $n > 2$.

Answer. We can let $g + h$ be a new function j, when we add new convex function p, it is still convex by induction that $g + h + p$ is also convex, the sum of n convex functions is still a convex function when $n > 2$

(e) (2 points) Finally, using the previous parts, explain why in our case that, when we solve for the critical point of the MSE by taking the gradient with respect to the parameter and setting the expression to 0, it is guranteed that the solution we find will minimize the MSE.

Answer. From previous part a, b, c, d we can know that the second derivative is always non-negative, when it is zero, the first derivative must be zero since it is convex and the MSE is minimized.

Congratulations! You have finished Homework 4!