Homework 2

Ng Tze Kean
Student number: 721370290002

March 28, 2024

STAT 4710J, Spring 2024

Homework #2

Properties of Simple Linear Regression

1. (4 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) (2 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.

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

(b) (1 point) Using your result from part (a), prove that $\bar{y} = \bar{\hat{y}}$.

Answer.

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

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

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

$$\bar{y} = \bar{\hat{y}}$$

(c) (1 point) Prove that $(\bar{x}, \bar{y})$ is on the simple linear regression line.

Answer.

$$\hat{y} = \theta_0 + \theta_1 x$$

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

Let $\hat{y}$ be equal to $\bar{y}$, then

$$x = \bar{x}$$

Geometric Perspective of Least Squares

2. (4 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) (2 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.

$$\begin{aligned}\sum_{n}^{i=1} e_i &= 1^T e \\ &= 1(\mathbb{Y} - \hat{\mathbb{Y}})^T \\ &= 1(\mathbb{Y} - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y})^T\end{aligned}$$

Since we know that the first column of our design matrix is 1, we let $1 = \mathbb{X}e$ where e is a column vector with all zeroes except the first row. We substitute the following

$$\begin{aligned}&= e^T\mathbb{X}^T(\mathbb{Y} - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y})^T \\ &= e^T(\mathbb{X}^T\mathbb{Y} - \mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y})^T \\ &= e^T(\mathbb{X}^T\mathbb{Y} - \mathbb{X}^T\mathbb{Y})^T \\ &= 0\end{aligned}$$

(b) (1 point) 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 $e = \mathbb{Y} - \mathbb{X}\hat{\theta}$ is orthogonal to the *span*$(\mathbb{X})$ then the dot product of $\vec{x} \cdot e$ will be zero.

(c) (1 point) Explain why the predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

Answer. The predicted response $\hat{\mathbb{Y}}$ lies on the *span*$(\mathbb{X})$. While e lies on shortest line between $\mathbb{Y}$ and $\hat{\mathbb{Y}}$, this line is orthogonal to *span*$(\mathbb{X})$, making both $\mathbb{Y}$ and e 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. (2 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.

$$\begin{aligned}
\frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2 &= 0 \\
\frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2 &= 0 \\
\frac{1}{n}\sum_{i=1}^{n}(2\gamma x_i^2 - 2y_i x_i) &= 0 \\
\sum_{i=1}^{n}(2\gamma x_i^2 - 2y_i x_i) &= 0 \\
\sum_{i=1}^{n}(\gamma x_i^2 - y_i x_i) &= 0 \\
\sum_{i=1}^{n}\gamma x_i^2 - \sum_{i=1}^{n} y_i x_i &= 0 \\
\gamma &= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

4. (4 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) (1 point) $\sum_{i=1}^{n} e_i = 0$.

Answer. We cannot guarantee that the equation holds true.

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

By substituting our answers from part (3) we can see that the residuals cannot guaranteed to be equal to 0.

(b) (1 point) The column vector $\vec{x}$ and the residual vector e are orthogonal.

Answer. Still true. The shortest distance from $\mathbb{Y}$ to $\hat{\mathbb{Y}}$ is the orthorgonal projection of $\mathbb{Y}$ on the $span(x)$ is the orthorgonal distance. Since e is the residual vector, the difference between $\mathbb{Y}$ and $\hat{\mathbb{Y}}$, it must be also orthogonal to the $span(x)$, consequently the column vector $\hat{(x)}$

(c) (1 point) The predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

Answer. Still true. $\hat{\mathbb{Y}}$ lies on the $span(x)$, following from (b), then both must be orthogonal as well.

(d) (1 point) $(\bar{x}, \bar{y})$ is on the regression line.

Answer. We cannot guarantee it holds true. Without the intercept term the equation simplifies and from question (3) we found the term that minimizes the residual. However, the expression does not contain $\bar{x}, \bar{y}$ removing the constrain of the line needing to pass through the means.

MSE "Minimizer"

5. (6 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) (1 point) 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 ____ ($x_i/y_i/\gamma$).

   Answer. quadratic, γ

   (b) (1 point) 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.

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

   Since the 2nd derivative is positive, we have proven that g_i is convex.

(c) (1 point) 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. Given a convex function, the values before the critical point must be monotonically decreasing, and after the critical point, the function must be monotonically increasing. Thus, obtaining the minimum point for g.

(d) (2 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. (1 point) 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. Let $f(x) = g(x) + h(x)$ then we have the following, for some $f(x)$

$$\begin{aligned} f(cx_1 + (1-c)x_2) &= g(cx_1 + (1-c)x_2) + h(cx_1 + (1-c)x_2) \\ &\leq cg(x_1) + (1-c)g(x_2) + ch(x_1) + (1-c)h(x_2) \\ &\leq cf(x_1) + (1-c)f(x_2) \end{aligned}$$

We thus have shown that the sum of 2 convex function will also be a convex function

ii. (1 point) 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. Based could use induction to think of this. We can think of $g(x)+h(x)$ as another convex function $g'(x)$. We can iteratively repeat the summation for the next step n times following the same proof. Since the next step is still a convex function, we will obtain a convex function at the end of the iteration.

(e) (1 point) 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 guaranteed that the solution we find will minimize the MSE.

Answer. Using the intuition from d(ii), we know that the MSE equation for a single iteration is convex as proven in (b). Then, The summation of convex functions must still be convex and thus the MSE is a convex function. Using the 2nd derivation will guarantee that the solution is minimum following the explanation in (c)

Congratulations! You have finished Homework 2!