Data 100, Fall 2020

Homework #5

Due Date: Thursday, October 22th at 11:59 PM

Total Points: 26

Submission Instructions

You must submit this assignment to Gradescope by **Thursday, October 22th, at 11:59 PM**. While Gradescope accepts late submissions, you will not receive **any** credit for a late submission if you do not have prior accommodations (e.g. DSP).

You can work on this assignment in any way you like.

- One way is to download this PDF, print it out, and write directly on these pages (we've provided enough space for you to do so). Alternatively, if you have a tablet, you could save this PDF and write directly on it.
- Another way is to use some form of LaTeX. Overleaf is a great tool.
- You could also write your answers on a blank sheet of paper.

Regardless of what method you choose, the end result needs to end up on Gradescope, as a PDF. If you wrote something on physical paper (like options 1 and 3 above), you will need to use a scanning application (e.g. CamScanner) in order to submit your work.

When submitting on Gradescope, you **must** assign pages to each question correctly (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our tutors. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. If you have any questions about the submission process, please don't hesitate to ask on Piazza.

Collaborators

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others please include their names at the top of your submission.

Properties of Simple Linear Regression

1. (7 points) In Lecture 12, we spent a great deal of time talking about simple linear regression, which you also saw in Data 8. To briefly summarize, the simple linear regression model assumes that given a single observation x, our predicted response for this observation is $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$.

   In Lecture 12, we saw that the $\hat{\theta}_0$ and $\hat{\theta}_1$ 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) (3 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 $\sum_{i=1}^{n} e_i = 0$.

Since $e_i = y_i - \hat{y}_i$, thus:
$e_i = y_i - \left(\bar{y} + r\frac{\sigma_y}{\sigma_x}(x_i - \bar{x})\right)$

Considering the sum of residuals, $\sum_{i=1}^{n} e_i$:
$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}\left[y_i - \left(\bar{y} + r\frac{\sigma_y}{\sigma_x}(x_i - \bar{x})\right)\right]$
$= \sum_{i=1}^{n} y_i - \left\{\sum_{i=1}^{n}\left(\bar{y} + r\frac{\sigma_y}{\sigma_x}(x_i - \bar{x})\right)\right\}$
$= \sum_{i=1}^{n} y_i - \left\{\sum_{i=1}^{n}\bar{y} + r\frac{\sigma_y}{\sigma_x}\sum_{i=1}^{n} x_i - r\frac{\sigma_y}{\sigma_x}\sum_{i=1}^{n}\bar{x}\right\}$

Since $\bar{x}$ and $\bar{y}$ do not involve i, their summations can be written as $n\bar{x}$ and $n\bar{y}$ respectively.
$= \sum_{i=1}^{n} y_i - \left\{n\bar{y} + r\frac{\sigma_y}{\sigma_x}\sum_{i=1}^{n} x_i - r\frac{\sigma_y}{\sigma_x} n\bar{x}\right\}$

By definition $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
$= \sum_{i=1}^{n} y_i - \left\{n\cdot\frac{1}{n}\sum_{i=1}^{n} y_i + r\frac{\sigma_y}{\sigma_x}\sum_{i=1}^{n} x_i - r\frac{\sigma_y}{\sigma_x} n\cdot\frac{1}{n}\sum_{i=1}^{n} x_i\right\}$
$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i - r\frac{\sigma_y}{\sigma_x}\sum_{i=1}^{n} x_i + r\frac{\sigma_y}{\sigma_x}\sum_{i=1}^{n} x_i$
$= 0$
Thus, $\sum_{i=1}^{n} e_i = 0$, as proved above. □

   (b) (2 points) Using your result from part a, prove that $\bar{y} = \bar{\hat{y}}$.

Since $e_i = y_i - \hat{y}_i$, and this is true for all i, this must be true for their summations as well.
Thus: $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i)$
$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{y}_i$

Multiplying by $\frac{1}{n}$ on either side of the equation:
$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{y}_i\right)$
$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$
Since $\sum_{i=1}^{n} e_i = 0$ from part (a):

$\frac{1}{n}\cdot 0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$
Since $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$, by definition, then:
$0 = \bar{y} - \bar{\hat{y}}$
Thus $\bar{y} = \bar{\hat{y}}$ proved. □

   (c) (2 points) Prove that $(\bar{x}, \bar{y})$ is on the simple linear regression line.

We know:
$\hat{y} = \bar{y} + r\sigma_y\frac{(x - \bar{x})}{\sigma_x}$

When $x = \bar{x}$
$\hat{y} = \bar{y} + r\sigma_y\frac{(\bar{x} - \bar{x})}{\sigma_x}$
$\hat{y} = \bar{y} + r\sigma_y\cdot\frac{0}{\sigma_x}$

$\hat{y} = \bar{y} + 0$
Thus $\hat{y} = \bar{y}$.

This means that for $x = \bar{x}$, the predicted y value $(\hat{y})$ is $\bar{y}$.
Thus $(\bar{x}, \bar{y})$ lies on the simple regression line. □

Geometric Perspective of Least Squares

2. (7 points) In Lecture 13, we viewed both the simple linear regression model and the multiple linear regression model through the lens of 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 $\mathbb{Y}$.

   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 $\hat{\mathbb{Y}} = \mathbb{X}\begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix}$, and also as $\hat{\mathbb{Y}} = \hat{\theta}_0\mathbb{1} + \hat{\theta}_1\vec{x}$.

   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) (3 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* $e = [e_1, e_2, ..., e_n]^T$.

Since $\vec{e} = Y - \hat{Y}$,

$\vec{e} = Y - (\hat{\theta}_0 + \hat{\theta}_1\vec{x})$

From lecture, we know that the residual vector is orthogonal to span(X). This is because the residual vector is minimized i.e. the true values are closest to the predicted values when the distance between them is the shortest possible distance, i.e the orthogonal distance.

Since $(1, \ldots, 1)^T$ is in span(X), $(1, \ldots, 1)^T$ is orthogonal to the residual vector, $\vec{e}$. By the definition of orthogonality, this means:

$$\vec{e} \cdot \vec{1} = 0$$

$$(e_1, \ldots, e_n)^T \underbrace{(1, \ldots, 1)^T}_{n \text{ 1s}} = 0$$

$$e_1 + e_2 + \ldots + e_n = 0$$

Thus $\sum_{i=1}^{n} e_i = 0$. □

(b) (2 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.*

From lecture, we know that the residual vector is orthogonal to span (X). This implies that $\vec{e}$ is orthogonal to the span of vectors within X i.e $\vec{e}$ is orthogonal to span$(\vec{1}, \vec{x})$.

Furthermore, consider: $(x_1, \dots, x_n)^T = (1, \dots, 1)^T \cdot (x_1, \dots, x_n)^T$. This implies $\vec{x} = \vec{1} \cdot \vec{x}$, which means $x \in$ span $(\vec{1}, \vec{x})$. Since $\vec{e}$ is orthogonal to span $(\vec{1}, \vec{x})$ and $x \in$ span $(\vec{1}, \vec{x})$, this implies that $\vec{x}$ and $\vec{e}$ are orthogonal. □

(c) (2 points) Explain why the predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

We know $\hat{y} = \hat{\theta}_0 \cdot \vec{1} + \hat{\theta}_1 \cdot \vec{x}$.

Since $\hat{y}$ is a linear combination of vectors $\vec{1}$ and $\vec{x}$, this means that $\hat{y} \in$ span $(\vec{1}, \vec{x})$.

We also already know that e is orthogonal to $\vec{1}$ (from part (a)) and we also know that e is orthogonal to $\vec{x}$ (from part (b)). This implies that $\vec{e}$ is orthogonal to span $(\vec{1}, \vec{x})$.

Since $\hat{y} \in$ span $(\vec{1}, \vec{x})$ and $\vec{e}$ is orthogonal to span$(\vec{1}, \vec{x})$, this implies that $\vec{e}$ and $\hat{y}$ 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 simply $\hat{y} = \gamma x$, 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 squared loss ("empirical risk") 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. (4 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.

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

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

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

To find minimizing γ, let $R'(\gamma) = 0$

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

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

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

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

Thus, $\boxed{\hat{\gamma} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}}$. □

4. (8 points) For our new simplified model, our design matrix $\mathbb{X}$ is

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

And so 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) $\sum_{i=1}^{n} e_i = 0$.

If $\vec{1}$ is not a column of the feature matrix then the following is not necessarily true: $\vec{1} \cdot \vec{e} = 0$. If we were to apply the same logic as in (2a) for the above design matrix, Then $\vec{x} \cdot \vec{e} = 0$ i.e $\sum_{i=1}^{n} x_i e_i = 0$ which does not necessarily mean $\sum_{i=1}^{n} e_i = 0$.

Thus, FALSE.

(b) (2 points) The column vector $\vec{x}$ and the residual vector e are orthogonal.

Since the residual vector is orthogonal to span(X), and $\vec{x}$ lies in span(X), thus $\vec{e}$ is orthogonal to $\vec{x}$.

TRUE

(c) (2 points) The predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

We know $\hat{\mathbb{Y}} = \hat{\gamma}\vec{x}$. Since $\hat{\mathbb{Y}}$ is a linear combination of $\vec{x}$, thus $\hat{\mathbb{Y}} \in \text{span}(\vec{x})$

From (b), we also know that $\vec{e}$ is orthogonal to $\vec{x}$. Thus, $\hat{\mathbb{Y}}$ and $\vec{e}$ are orthogonal. TRUE

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

In (1c) we did the following proof:

We know:

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

When $x = \bar{x}$

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

$$\hat{y} = \bar{y} + r\sigma_y \cdot \frac{0}{\sigma_x}$$

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

However, in our case, we have no intercept term thus $\hat{y} \neq \bar{y}$.

Thus $(\bar{x}, \bar{y})$ is not on the regression line.

FALSE.