DSC 40A - Extra Practice Session 2

Wednesday, January 26, 2022

Recall that the least squares solutions to the problem of fitting a straight line, $h(x) = w_1 x + w_0$, to the data (x_i, y_i) are:

$$\boxed{w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$w_0 = \bar{y} - w_1\bar{x}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Problem 1. Pop Quiz

actual: y_1 predicted by line: $3x_1 + 7$

Consider the data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and the line $y = 3x + 7$.

a) Without looking at any notes, write down the expression for the mean squared error of this line on the data set.

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{3x_i + 7 - y_i}_{\text{error}}\Big)^2$$

b) Without looking at any notes, write down the expression for the mean absolute error of this line on the data set.

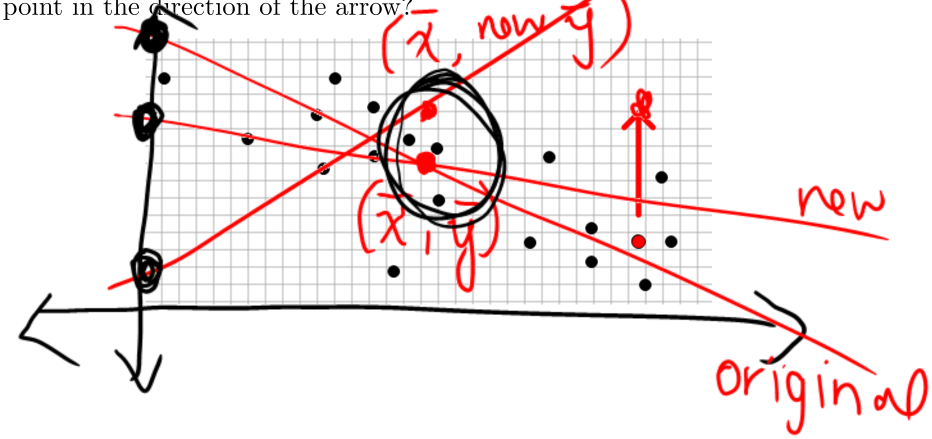
$$\frac{1}{n}\sum_{i=1}^{n}|3x_i + 7 - y_i|$$

note: could also write $(y_i - (3x_i + 7))$

but remember to distribute the minus sign

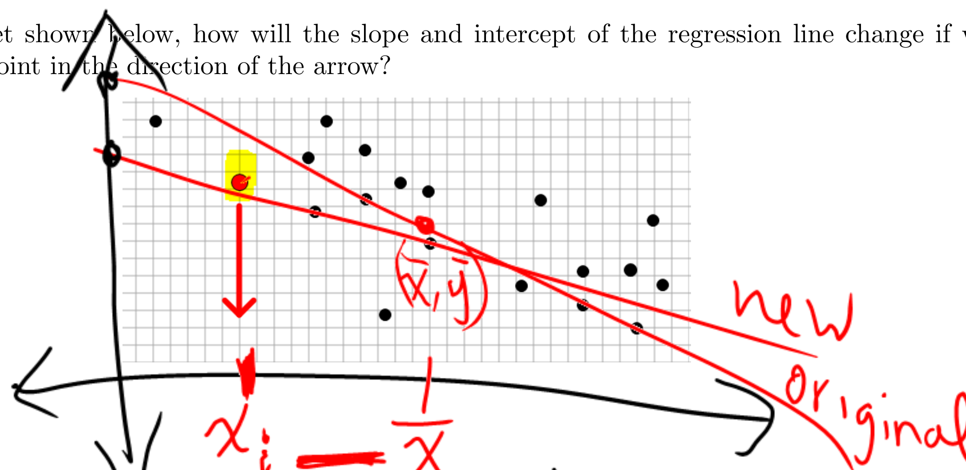
Problem 2. Visualizing Changes in the Regression Line

a) For the data set shown below, how will the slope and intercept of the regression line change if we move the red point in the direction of the arrow?



Only changing one y-value

b) For the data set shown below, how will the slope and intercept of the regression line change if we move the red point in the direction of the arrow?



$x_i - \bar{x}$

inc in slope

dec in intercept

$$w_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

↑ in HW 3

original w_1 has a term

$$(x_j - \bar{x})y_j$$

new w_1 has as its term j:

$$(x_j - \bar{x})(y_j - c)$$

← move down, so $c > 0$

how much does w_1 change?

$$(x_j - \bar{x})(y_j - c)$$

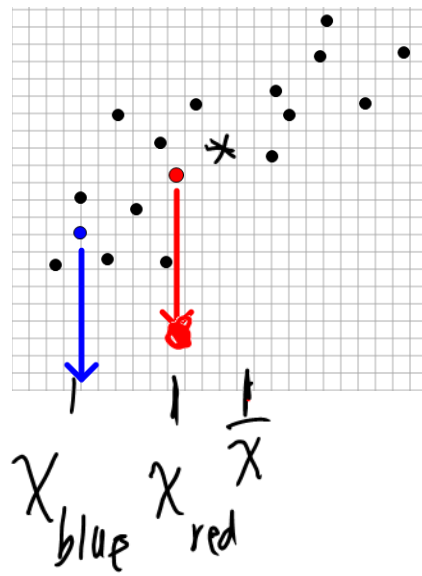
$$= \underbrace{(x_j - \bar{x})y_j}_{\text{old term}} \underbrace{\boxed{-\underbrace{(x_j - \bar{x})}_{\text{neg}}\underbrace{c}_{\text{pos}}}}_{\text{change}}$$

turns out positive

c) Compare two different possible changes to the data set shown below.

- Move the red point down c units.
- Move the blue point down c units.

Which move will change the slope of the regression line more? Why?



if change blue

$$\underline{(x_{blue} - \bar{x})}(\underbrace{y_{blue}}_{orig} - \underline{c})$$

if change red

$$\underline{(x_{red} - \bar{x})}(\underbrace{y_{red}}_{orig} - \underline{c})$$

d) Suppose we transform a data set of $\{(x_i, y_i)\}$ pairs by doubling each y-value, creating a transformed data set $\{(x_i, 2y_i)\}$. How does the slope of the regression line fit to the transformed data compare to the slope of the regression line fit to the original data? Can you prove your answer from the formula for the slope of the regression line?

- visually

- formula

$$w_1 = 2 \leftarrow \frac{\sum_{i=1}^{n} 2(x_i - \bar{x})(2y_i - 2\bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

saw this in groupwork 4

e) Suppose we transform a data set of $\{(x_i, y_i)\}$ pairs by doubling each x-value, creating a transformed data set $\{(2x_i, y_i)\}$. How does the slope of the regression line fit to the transformed data compare to the slope of the regression line fit to the original data? Can you prove your answer from the formula for the slope of the regression line?

- visually

- formula

$$w_1 = 2\frac{\sum (2x_i - 2\bar{x})(y_i - \bar{y})}{\sum (2x_i - 2\bar{x})^2}$$

$$4 \leftarrow \left(2(x_i - \bar{x})\right)^2$$

$$\frac{2}{4} = \frac{1}{2}$$

Problem 3. Nonlinear Function

x	68	70	72	72
y	34	20	18	27

For the data above, apply a suitable transformation then use linear regression to find the best fitting curve of the form:

$$\boxed{x = \sqrt{ay^2 + by}.} \leftarrow \text{nonlinear}$$

Round the parameters a and b to three decimal places.

idea: rewrite/manipulate until it look like

$$\underbrace{\boxed{x^2/y}}_{\text{variable}} = \text{constant} + \text{constant} * \underbrace{\boxed{y}}_{\text{variable}}$$

$$x = \sqrt{ay^2 + by}$$

$$x^2 = ay^2 + by$$

$$x^2 = y(\underline{ay + b})$$

plays role of y $\Rightarrow$ $\boxed{\frac{x^2}{y}} = a\boxed{y} + b$ $\Leftarrow$ plays role of x

y	34			
x^2/y				

$\leftarrow$ think of it as x

$\leftarrow$ think of it as y

Problem 4. Optimization Algorithm

In the supplementary Jupyter notebook (linked), write a Python function that takes as input an array of names and returns the longest name in the array, where longest means having the most individual characters. If multiple names are tied for the longest, you can select any one of them.