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:

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

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

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

actual = y_i, predicted by reg. line $= 3x_i + 7$

Problem 1. Pop Quiz

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.

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

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

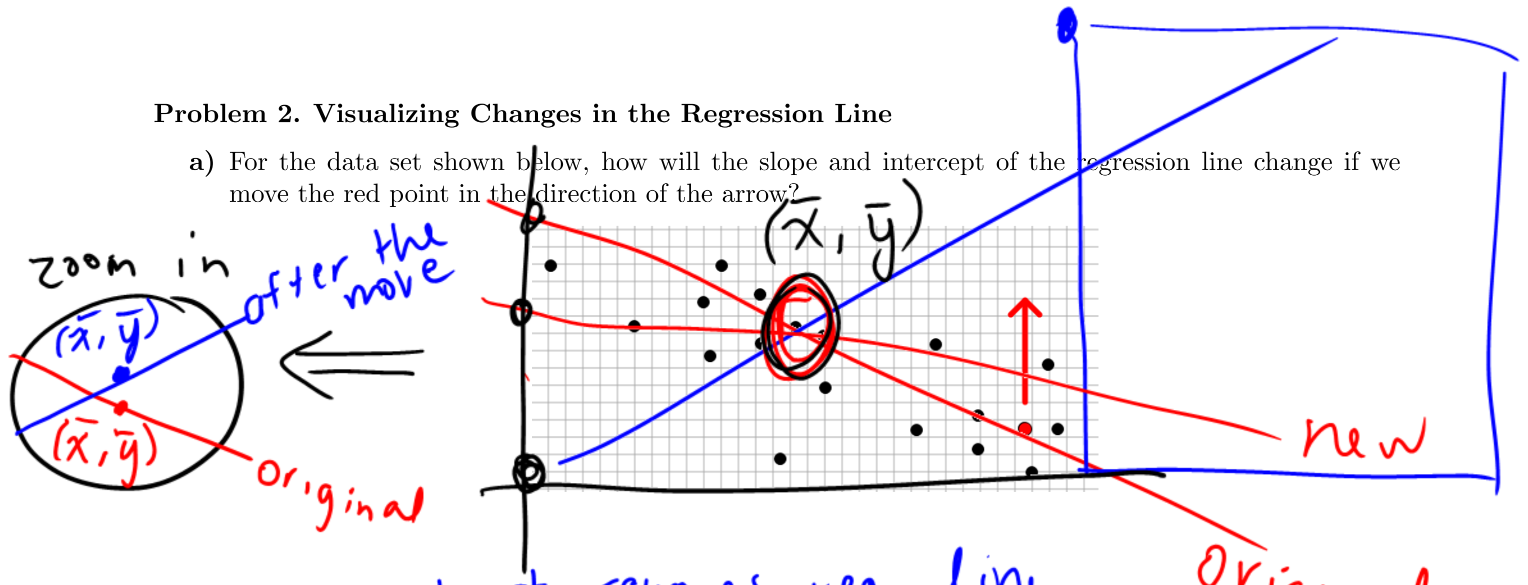
$$\frac{1}{n}\sum_{i=1}^{n}|3x_i + 7 - y_i|$$

could do: $y_i - (3x_i + 7)$ instead

but remember to distribute 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?

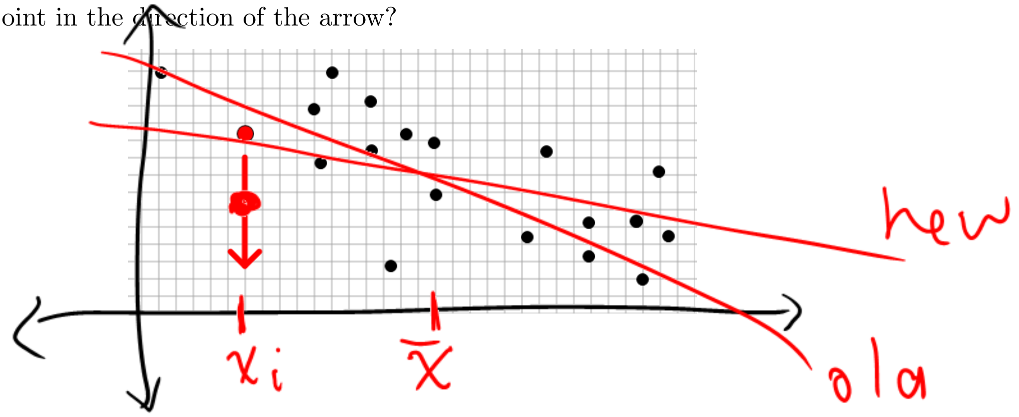


least squares reg. line very much influenced by outliers

slope increasing, int dec

(assuming data is in 1st quadrant)

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?



how to see this from formula

$$\text{slope} = w_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \underset{\substack{\uparrow \\ \text{in HW3}}}{=} \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Key thing:
Only one term changes (in numerator)

before: $(x_i - \bar{x})y_i$

↑ old y_i

after: $(x_i - \bar{x})(y_i - c)$

$= (x_i - \bar{x})y_i$

$\boxed{-\underbrace{(x_i - \bar{x})}_{\text{negative bc } x_i < \bar{x}}c}$

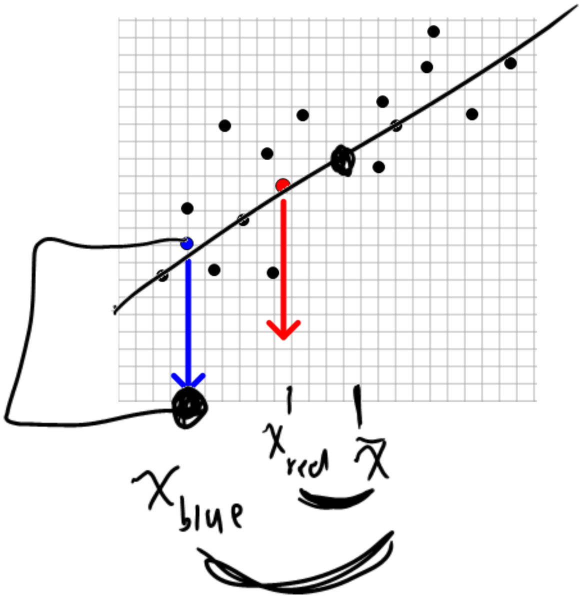
$\Rightarrow$ positive

slope increases bc the only term that changed got bigger

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?



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?

factor out 2

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

recognize from groupwork

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?

$$w_1 = \frac{\sum_{i=1}^{n}2(2x_i-2\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}4(2x_i-2\bar{x})^2}$$

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

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:

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

Round the parameters a and b to three decimal places.

nonlinear relationship between x and y

idea: manipulate this until it look like

linear in new variables →

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

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

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

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

$$\frac{x^2}{y} = a\,y + b$$

plays role of y

plays role of x

y	34			
x^2/y	136			

← think of it as x

← 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.