CSE 103

Homework #7
Fall 2019

Due: Monday, November 18, 2019 at 11:00PM on Gradescope

1 Directions

You may work with one other student. If working with a partner, **submit only one submission per pair** : one partner uploads the submission and adds the other partner to the Gradescope submission. You can post public questions about the assignment to Piazza, discuss the questions and their answers with at most one other student, and ask questions in office hours

Your answers have to be typeset, not handwritten. This is for two reasons: (a) to reduce ambiguity of the answers, and (b) to be kind to the TA's eyesight. We recommend you use latex, but you can also use word-processors that support mathematical formulas. More directions are available here: `https://tinyurl.com/y2gv9bn9`.

You will submit this assignment via Gradescope (`https://www.gradescope.com`) in the assignment called "Homework 7". You can submit each question as many times as you like. You should solve the problems and ask questions about them offline first, then try submitting once you are confident in your answers.

No late submissions are accepted.

2 Problems

1. (24 points) In a certain class, midterm scores average out to 50 with an SD of 15, as do scores on the final. The correlation between midterm scores and final scores is about 0.50. Estimate the average final score for the students whose midterm scores were:

   (a) 75

   (b) 30

   (c) 60

   Solution

$$y - y_0 = \frac{r * SD_y}{SD_x}(x - x_0)$$

$x_0 = 50, y_0 = 50, r = 0.5, SD_y = 15, SD_x = 15$

$$y = 0.5x + 25$$

   (a) 75

$$y = 0.5(75) + 25$$

   The average final is 62.5

   (b) 30

$$y = 0.5(30) + 25$$

   The average final is 40

   (c) 60

$$y = 0.5(55) + 25$$

   The average final is 55

2. (36 points) For the men age 18 and over in HANES5,
average height $\approx$ 69 inches, SD $\approx$ 3 inches
average weight $\approx$ 190 pounds, SD $\approx$ 42 pounds
$r \approx 0.41$. Estimate the average weight of the men whose heights were:

   (a) 69 inches

   (b) 66 inches

   (c) 24 inches

   (d) 0 inches

   Solution

$$y - y_0 = \frac{r * SD_y}{SD_x}(x - x_0)$$

$y_0 = 190, r = 0.41, SD_y = 42, SD_x = 3$

$$y = 0.42 * 14(x - 69) + 190$$

   (a) 69 inches
   The average weight is 190

   (b) 66 inches
   The average weight is 173

   (c) 24 inches
   The average weight is -68

   (d) 0 inches
   The average weight is -206

   Comment on your answers to (c) and (d).

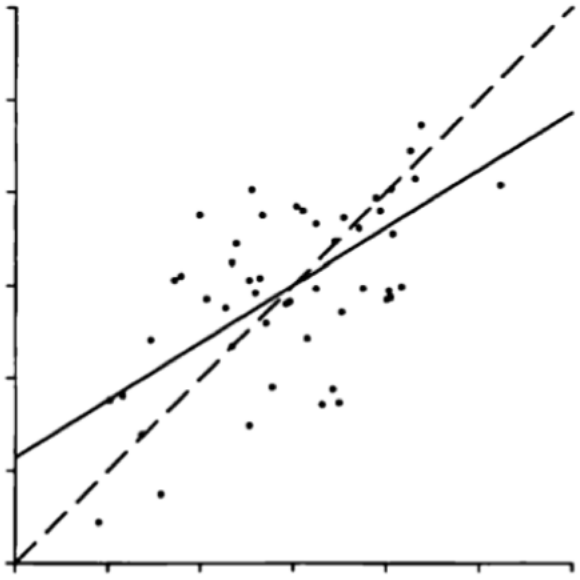
   Height of 24 inches is way below average, height of 0 is impossible. Regression is unable to provide an estimate in those two cases. That is why we got negative value.

3. (8 points) As part of their training, air force pilots make two practice landings with instructors, and are rated on performance. The instructors discuss the ratings with the pilots after each landing. Statistical analysis shows that pilots who make poor landings the first time tend to do better the second time. Conversely, pilots who make good landings the first time tend to do worse the second time.
   Conclusion: Criticism helps the pilots while praise makes them do worse. As a result, instructors were ordered to criticize all landings, good or bad.
   Was this conclusion warranted by the facts? Answer yes or no, and explain briefly.

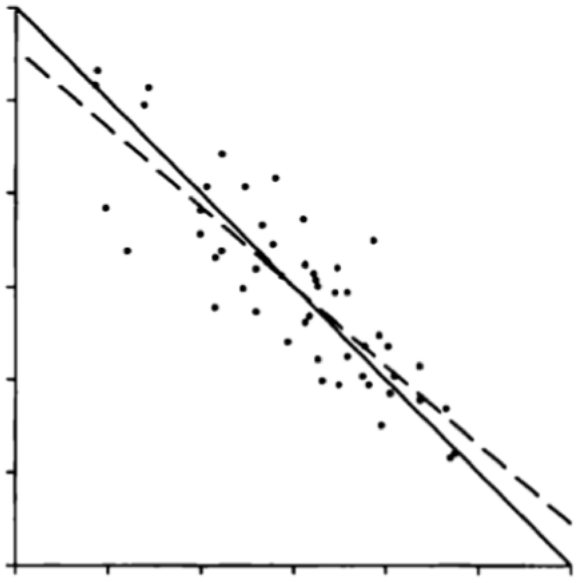
   Solution

   No, This is an example of regression fallacy. It fails to account for various factors. One example: pilots didn't do well the first time, received some feedback and try harder. those who did well, might not try as hard. We don't know whether or not the improvement of "bad" pilot is higher than the the worsening of "good" pilot

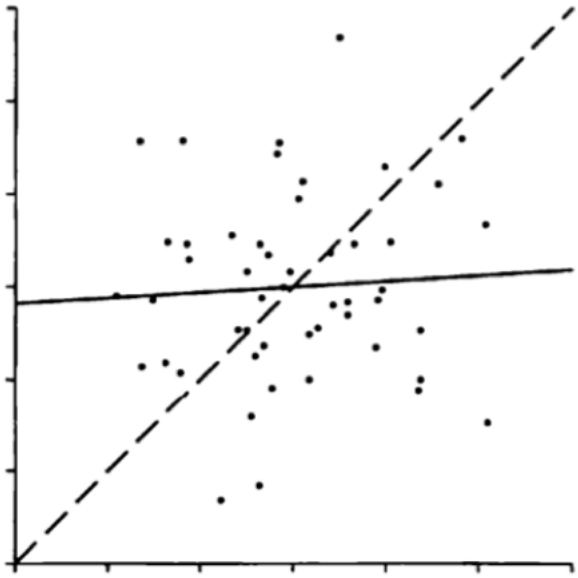
4. (12 points) Below are four scatter diagrams, each with a solid line and a dashed line. For each diagram, say which is the SD line and which is the regression line for y on x.



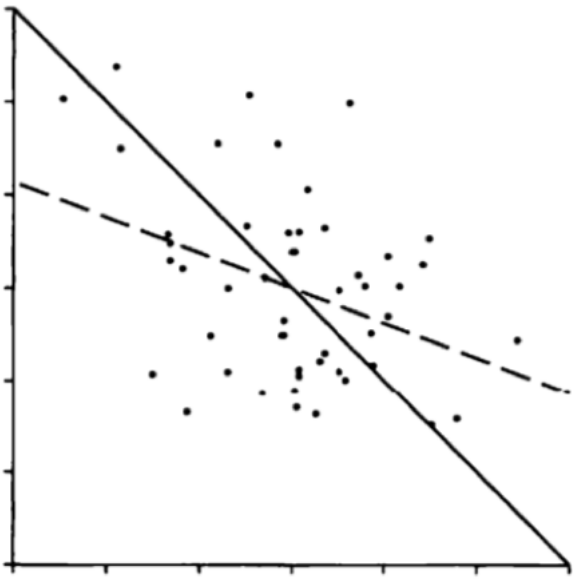
(a) SD line: Dashed
Regression line: Solid
Reason: Slope of regression is calculated using slope of SD. Correlation coefficient is between 0 and 1, slope of regression has to be less than or equal to slope of SD



(b) SD line: Solid
Regression line: Dashed
Reason:Slope of regression is calculated using slope of SD. Correlation coefficient is between 0 and 1, slope of regression has to be less than or equal to slope of SD



(c) SD line: Dashed
Regression line: Solid
Reason:Slope of regression is calculated using slope of SD. Correlation coefficient is between 0 and 1, slope of regression has to be less than or equal to slope of SD



(d) SD line: Solid
Regression line: Dashed
Reason:Slope of regression is calculated using slope of SD. Correlation coefficient is between 0 and 1, slope of regression has to be less than or equal to slope of SD

5. (20 points) You are given a set of points: $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. You need to find the second order polynomial that minimizes MSE error given by:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i + cx_i^2))^2$$

Write a set of linear equations from which you can determine a, b, and c.
(Hint: take the partial derivatives and then simplify).

Solution

$$\frac{\partial}{\partial a}MSE(a,b,c) = \frac{1}{n}\sum_{i=1}^{n}(-2)(y_i - (a + bx_i + cx_i^2))$$

$$\frac{\partial}{\partial b}MSE(a,b,c) = \frac{1}{n}\sum_{i=1}^{n}(-2x_i)(y_i - (a + bx_i + cx_i^2))$$

$$\frac{\partial}{\partial c}MSE(a,b,c) = \frac{1}{n}\sum_{i=1}^{n}(-2x_i^2)(y_i - (a + bx_i + cx_i^2))$$

$$\begin{cases} \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)) = 0 \\ \sum_{i=1}^{n}(y_i x_i - ax_i - bx_i^2 - cx_i^3)) = 0 \\ \sum_{i=1}^{n}(y_i x_i^2 - ax_i^2 - bx_i^3 - cx_i^4)) = 0 \end{cases}$$