Name: *Caleb McWhorter — Solutions*
MATH 108
Spring 2022
Written HW 10: Due 04/27

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."
–John Tukey

Problem 1. (10pt) Find the least square regression line, along with the r and r^2 value, for the dataset $\{(-1,-2),(1,3),(2,2),(4,5)\}$. Show all your work.

Solution. We have 4 points so that $n = 4$. First, we compute the x and y averages—$\overline{x}$ and $\overline{y}$, respectively.

$$\overline{x} = \frac{\sum x_i}{n} = \frac{-1+1+2+4}{4} = \frac{6}{4} = 1.5$$
$$\overline{y} = \frac{\sum y_i}{n} = \frac{-2+3+2+5}{4} = \frac{8}{4} = 2.00$$

Now we compute s_x, s_y, r: Then we have

x	y	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$
-1	-2	-2.5	6.25	-4	16
1	3	-0.5	0.25	1	1
2	2	0.5	0.25	0	0
4	5	2.5	6.25	3	9
	Total:		13.00		26

$$s_x^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2 = \frac{1}{4-1}\cdot 13.00 = 4.3333 \Longrightarrow s_x = \sqrt{4.3333} = 2.0817$$
$$s_y^2 = \frac{1}{n-1}\sum (y_i - \overline{y})^2 = \frac{1}{4-1}\cdot 26 = 8.6667 \Longrightarrow s_y = \sqrt{8.6667} = 2.9439$$

Now we also compute the r value:

x	y	$x_i - \overline{x}$	$\frac{x_i-\overline{x}}{s_x}$	$y_i - \overline{y}$	$\frac{y_i-\overline{y}}{s_y}$	$\frac{x_i-\overline{x}}{s_x} \cdot \frac{y_i-\overline{y}}{s_y}$
1	1	-2.5	-1.2009	-4	-1.3587	1.6317
1	0	-0.5	-0.2402	1	0.3397	-0.0816
2	3	0.5	0.2402	0	0.0000	0.0000
3	4	2.5	1.2009	3	1.0191	1.2238
	Total:					2.7739

$$r = \frac{1}{n-1}\sum \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) = \frac{1}{4-1}\cdot 2.7739 = 0.9246$$

Therefore, $r^2 = 0.8549$. Finally, we can compute our regression coefficients:

$$b_1 = r\,\frac{s_y}{s_x} = 0.9246 \cdot \frac{2.9439}{2.0817} = 1.308 \quad \text{and} \quad b_0 = \overline{y} - b_1\overline{x} = 2.00 - 1.308 \cdot 1.5 = 0.038$$

Therefore, as $\widehat{y} = b_1 x + b_0$, we know $\widehat{y} = 1.308x + 0.038$.

Problem 2. (10pt) Given the following information below, find the least square regression line. Show all your work.

$$\begin{aligned} n &= 10 \\ \overline{x} &= 19.59, \quad s_x^2 = 2.9225 \\ \overline{y} &= 15.63, \quad s_y^2 = 5.2407 \\ R &= 0.8733 \end{aligned}$$

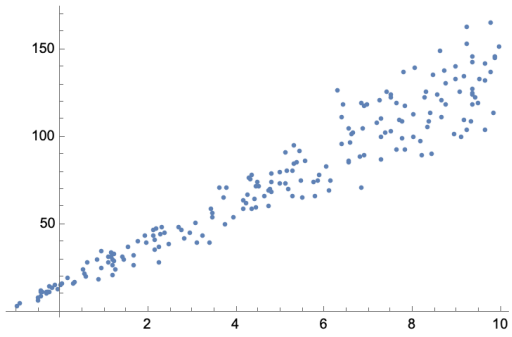
Solution. Because $s_x^2 = 2.9225$ and $s_y^2 = 5.2407$, we know that $s_x = \sqrt{2.9225} = 1.7095$ and $\sigma_y = \sqrt{5.2407} = 2.2893$. But then...

$$b_1 = R\,\frac{s_y^2}{s_x^2} = 0.8733\,\frac{2.2893}{1.7095} = 1.1695$$

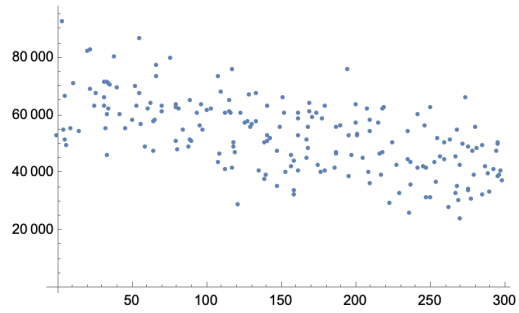
$$b_1 = \overline{y} - b_1\overline{x} = 15.63 - 1.1695 \cdot 19.59 = -7.2805$$

Therefore, as $\widehat{y} = b_1 x + b_0$, we know that $\widehat{y} = 1.1695x - 7.2805$.

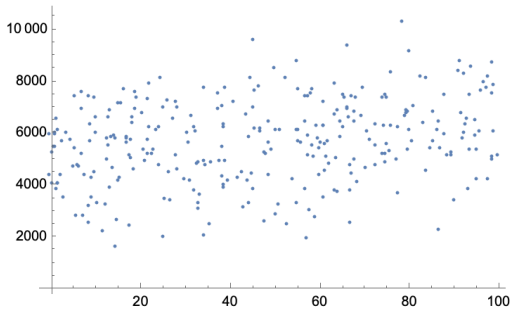
Problem 3. (10pt) Match each regression coefficient to its corresponding graph.



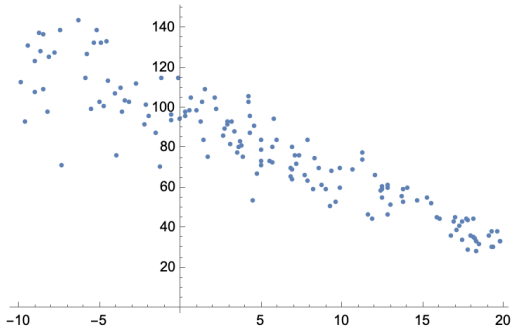
(a)



(b)



(c)



(d)

(i) __(d)__ : $R = -0.9197$

(ii) __(b)__ : $R = -0.6023$

(iii) __(c)__ : $R = 0.2527$

(iv) __(a)__ : $R = 0.9616$

Problem 4. (10pt) A researcher is trying to predict home run records. The researcher wants to predict what the next home run record will be. To model this, they will take the last ten home run records and label them as $r = 1, 2, \ldots, 10$, i.e. $r = 1$ is the tenth highest home run record, $r = 2$ is the ninth highest home run record, etc.. They fit a linear regression to this data and find a simple linear regression model of $h(r) = 20.21r + 559.3$.

(a) What are b_0 and b_1 for this linear regression?

(b) Predict the future home run record by finding $h(11)$.

(c) For $r = 8$, the home run record, $h(8)$, was known to be Babe Ruth's record of 714 home runs. Find the residual for this value given the model.

(d) The researcher finds an R^2 value of 0.9837. Is the linear model a good fit to the home run record data? Explain.

Solution.

(a) We know that a simple linear regression is of the form $\widehat{y} = b_1 x + b_0$. because we have $h(r) = 20.21r + 559.3$, we have $b_1 = 20.21$ and $b_0 = 559.3$.

(b) We have...

$$h(11) = 20.21(11) + 559.3 = 222.31 + 559.3 = 781.61$$

(c) We have...

$$h(8) = 20.21(8) + 559.3 = 161.68 + 559.3 = 720.98$$

Therefore, the residual is $e = y - \widehat{y} = 714 - 720.98 = -6.98$.

(d) Because $R^2 = 0.9837 > 0.85$, this is a 'good' fit for the data. Because $R^2 = 0.9837$, 98.37% of the variation in the home runs is explained by the model.