Homework 2 Solutions

by the respectable Asa Levi

p134 # 1 (10 pts)

- (a) Too low! - in this plot the point of averages is around $(60, 60)$, way lower than $(100, 100)$.
- (b) Too small! - in this plot the SD is much smaller than 15. It looks like all the x values are within 20 of 100.
- (c) Too big! - here it looks like the point of averages is around $(90, 90)$, the correlation is too high, and also the SD on both axes is too large.
- (d) Just right :)

p135 # 3 (5 pts)

The correlation would be 1. Then the relationship between the two variables would be perfectly linear as we would have:

$$\text{wife's height} = 0.92(\text{husband's height})$$

Notice that the slope of the line of best fit (which is defined by the equation above) is not the correlation, because we are not in standard units!

p154 # 5 (10 pts)

If the correlation were one then the points would lie on a line. Any two points determine a line, so we can use the first two points in each question to get our equation for a line and then solve. To get our equation for a line, we will use the two points to compute the slope and then give the point-slope form of the line.

Part (a)

Our two points are $(1, 1)$ and $(2, 3)$. So we get:

$$\text{slope} = \frac{3-1}{2-1} = 2$$

And our line is defined by the equation:

$$y - 1 = 2(x - 1)$$

Which we can rewrite as:

$$y = 2x - 1$$

Now we can see if our third point is on the line. The third point is also $(2, 3)$, so we know it is, but just to check we plug $x = 2$ into our equation and check that we get 3: $2(2) - 1 = 3$ woohoo! So the point $(2, 3)$ is on our line. Now, our final point has $x = 4$. Plugging this in we get: $2(4) - 1 = 7$, so we need to fill a 7 into the blank.

Part (b)

Using the same logic as last time we get the equation for the line to be:

$$y = 2x - 1$$

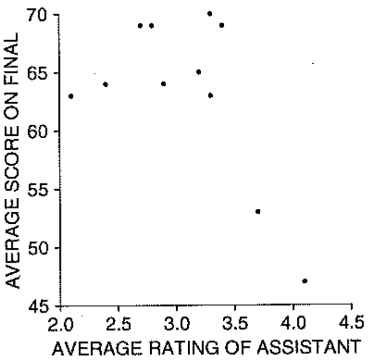
This is the same, since the two points we used are the same. Now when we plug $x = 3$ in we get $2(3) - 1 = 5$. This is a problem though, because $(3, 4)$ is in our data set. Thus the three points given are not colinear, so there is no value we could put in the blank to make r be 1.

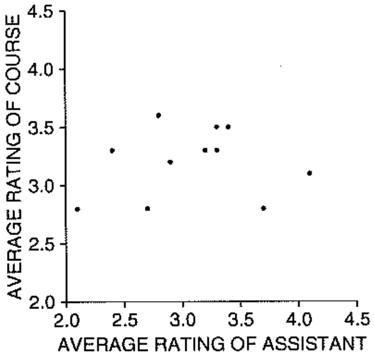
p155 # 8 (5 pts)

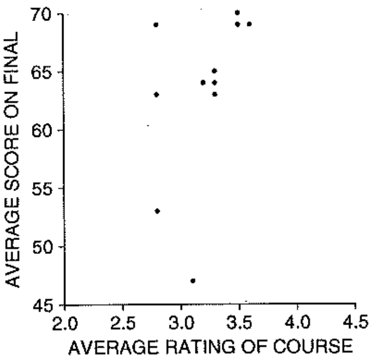
We have $r = -0.2$ which means that as age goes up education level among women goes down in this sample. This is because education levels have been going up over time and younger people were born later, thus are more educated. If every year you asked the same group of women for their educational level you would expect to see a non-linear relationship, with education increasing with age generally, but peaking at a certain age.

p155 # 9 (15 pts)

The scatter plots are as follows:







Variable 1	Variable 2	r
Average Score on Final	Average Rating of Assistant	-0.57
Average Rating of Course	Average Rating of Assistant	0.12
Average Score on Final	Average Rating of Course	0.46

Part (a)

False. For these two variables $r \approx -0.57$, which means that as assistant rating increased score on final decreased. I guess that means I should be meaner :P

Part (b)

True. For these two variables $r \approx 0.12$, which indicates that there is no significant relationship between assistant rating and course rating.

Part (c)

False. For these two variables $r \approx 0.46$ which means that as the course rating increased so did the average final score.

p176 # 2 (10 pts)

Part (a)

There are two ways to present the computations, although they are both the same thing. The equation for the regression line is:

$$(y - AVG(Y)) = r * \frac{SD(Y)}{SD(X)}(x - AVG(X))$$

In our case this is:

$$(y - 100) = 0.8 * \frac{15}{15}(x - 100)$$

We are trying to estimate the average y value in the vertical strip above $x = 115$. This is just where the regression line passes through this strip, so we plug $x = 115$ into the equation above and get $y = 112$. Thus we estimate the average score at age 35 for people who scored 115 at age 18 to be 112. Regression to the mean makes us stupider :P
Another way to think about it is to notice that 115 is 1 SD above AVG(X), so following the regression line, when we move 1 SD(X) over we rise r*SD(Y) above AVG(Y). This means that we expect our score to be $100 + (0.8) * 15 = 112$. This is the way that the book does it, but it is just a reformulation of the equation for the regression line.

Part (b)

The best prediction for a score at age 35 for someone who scored 115 at age 18 would be the average of the scores at age 35 for people who scored 115 at age 18. We estimated this average in part (a) to be 112, so this is the best guess.

p176 # 3 (20 pts)

Again we have to compute the equation for the regression line and just plug in the different x values. Using the formula from the previous question we get that the equation for the regression line is:

$$y - 63 = 0.25 * \frac{2.5}{2.7}(x - 68)$$

Now in the following parts we just plug our x value into this equation.

Part (a)

Here plug $x = 72$ into the equation for the regression line:

$$y - 63 = 0.25 * \frac{2.5}{2.7}(72 - 68) \approx 63.93$$

Part (b)

Here plug $x = 64$ into the equation for the regression line:

$$y - 63 = 0.25 * \frac{2.5}{2.7}(64 - 68) \approx 62$$

Part (c)

Here plug $x = 68$ into the equation for the regression line:

$$y - 63 = 0.25 * \frac{2.5}{2.7}(68 - 68) = 63$$

Notice that this is just AVG(Y), since the regression line passes through the point of averages, and 68 is AVG(X).

Part (d)

Since we don't know what the height of the husband is in this case, the best guess is just AVG(Y) = 63.

p177 # 7 (5 pts)

At this point it seems like both doctors are wrong. There is not necessarily any effect due to anxiety or relaxation; regression to the mean would explain both of the effects they are observing.

p177 # 8 (5 pts)

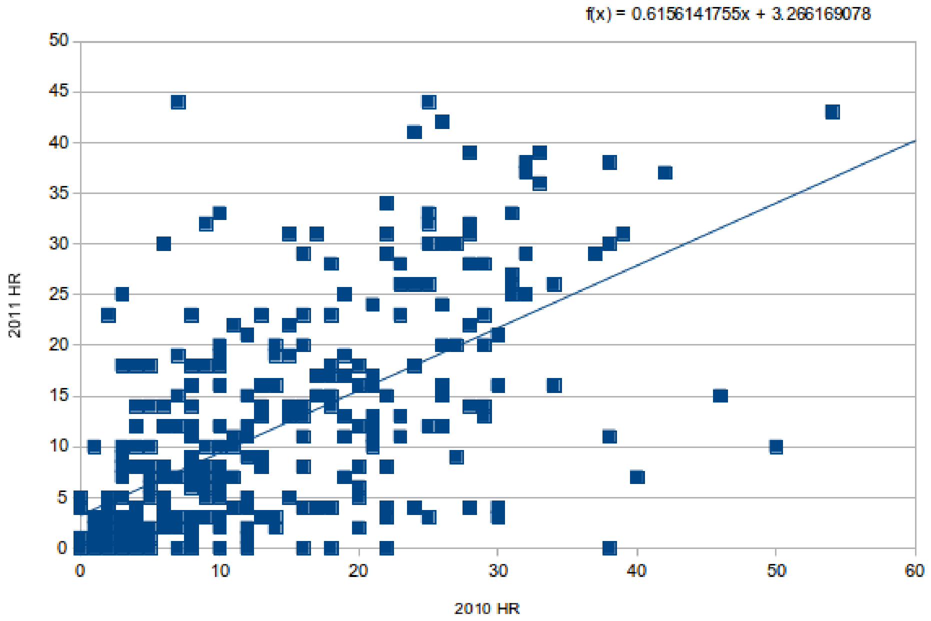
This extra data suggests that patients actually are more relaxed on the second reading. Regression would not change the average of the whole population, in fact regression leaves the average of the whole fixed. Although there is regression to the mean in each vertical strip, the positive gains to the left of the point of averages balance out the predicted losses to the right of the point of averages.

Worksheet 1 (5 pts)

The largest average was 2010, about 13.34 hr. The largest SD was 13.35 in 2006. I used SD$^+$; if you used a different SD function you might have gotten slightly different numbers. There has been no trend in the averages, while the SDs have been slowly decreasing.

Worksheet 2 (10 pts)

See scatter plot.



Worksheet 3 (5 pts)

The average correlation for predicting 3 years in the future is about 0.45. The average correlation coefficent is greater than 0.4 up to 5 years in the future. So you can predict 4 years in the future with correlation above 0.4. (It's fine if you said only 2 years also, since the correlation between 2008 and 2011 is about 0.399. But it makes more sense to consider the average correlation, because that is closer to the correlation you'd expect for random years.)

Worksheet 4 (10 pts)

You compute the 2012 predictions according to the formula for the regression line. In our case this is:

$$\text{2012 HR} = (\text{Avg 1 Year Correlation})\frac{\text{Avg HR SD}}{\text{2011 HR SD}}(\text{2011 HR} - \text{Avg 2011 HR})$$

In Excel you'd use this formula to create a new column with the predicted HR for each player in it. In that column you'd see that Carlos Beltran and Hunter Pierce's HR production are predicted to drop the most.

Worksheet 5 (10 pts)

Following the same method as in the last problem you can create a column with predicted 2011 HR from the 2010 data. Then you can compute the RMS error of these predictions by making a new column that has the difference of the predicted values and the actual values and taking its RMS error. Alternatively you could use the correlation coefficent for the 2010 and 2011 data and the formula RMS Error $= \text{SD(Y)}\sqrt{1-r^2}$. The RMS error for your prediction should be around 8.37. This might vary a little bit though depending on what SD you used, if you included 2011 in your average SD and average HR (it would be best not to, but I didn't specify either way) and how your computer rounds.

Worksheet 6 (Bonus)

Anything goes pretty much :) In class I presented some student's solutions that led to the best RMS error.