MONT 106N – Identifying Patterns
Selected Solutions – Problem Set 5
November 16, 2009

Chapter 10 Review Exercises (p. 176 – 178)

4. (a) *Method 1*: If the husband has completed 18 years of schooling, the value of x is 2 SD's above the average of 12. The regression line predicts that the wife's education level will be $(r)(2) = (.5)(2) = 1$ SD above the average for the wives. Thus, the wife's educational level is predicted to be $12 + 3 = 15$ years.

Method 2: We can also solve problems like this by writing down the xy-equation of the regression line. In this case, it is $y - \text{ave}_y = \frac{r \cdot SD_y}{SD_x}(x - \text{ave}_x)$ or $y - 12 = \frac{(.5)(3)}{3}(x - 12)$, or $y = .5x + 6$. With $x = 18$, we get predicted value $y = (.5)(18) + 6 = 15$.

(b) This is similar to part (a), except that the roles of x and y are reversed in the above. A man with 15 years of schooling is 1 SD above the average, so the woman is predicted to be $(r)(1) = .5$ SD above the average for the women. This gives an educational level of $12 + (.5)(3) = 13.5$ years.

(c) This is possible since we are actually talking about *two different* regression lines here, one for predicting values of y = wife's education level from values of x = husband's education level, and the other for predicting values of x from values of y. (*Technical Note:* The first is the line $y = .5x + 6$ found in the solution of (a) by Method 2 above. The other would be the line $x = .5y + 6$ in this example. The general equation of the 2nd line would be $(x - \text{ave}_x) = \frac{r \cdot SD_x}{SD_y}(y - \text{ave}_y)$.) The two lines are located as in the figure from problem 6 on page 177. If we look at the intersection of the vertical line $x = 18$ with $y = .5x + 6$, then we get the point $(18, 15)$ from part (a). Then the horizontal line $y = 15$ gives $(13.5, 15)$ on the second line $x = .5y + 6$. Since the line for predicting x from y lies above the first line, this point is farther to the left. (*Draw the picture to make sure you understand this!*)

9. (a) A score at the 5th percentile means that the corresponding standard unit score $-z$ was located so that the area in the lower tail below $-z$ was 5%. By the symmetry of the normal curve, the upper tail above z was also 5%. So the area in the central region between $-z$ and z was 90%. From the normal curve area table, $z \doteq 1.65$. So the student's score was 1.65 SDs below average. On the final, the same student would be predicted to score $r \times 1.65 = .5 \times 1.65 \doteq .825$ SD's below average. Using the approximation .8 in our table gives $A(.8) = 57.63$, so area in the lower tail below $-.8$ is about $(100 - 57.63)/2 \doteq 42.3/2 \doteq 21.0$. The predicted percentile is approximately the 21st percentile.

(b) Similar to (a): A standard unit score z at the 80th percentile means the area in the upper tail to the right of z is 20%, and the central area from $-z$ to z is 60%. From the normal curve area table, this occurs at about $z = .85$. On the final, the predicted score would be $r \times .85 = .5 \times .85 = .425$ SD's above average. Using the approximation, .4 in the area table gives $A(.4) = 31.08\%$, so the area below $z = .4$ is about $31.08 + (100 - 31.08)/2 = 31.08 + 68.92/2 \doteq 65.54\%$.

(c) and (d) Both of these would be the 50th percentile.

Chapter 11 Review Problems (p. 198 – 201)

4. The general idea for this problem is that when the scatter diagram is football-shaped, the rms error for regression functions as an SD for the predicted values within a vertical strip.

(a) Recall that about 68 percent of a normally distributed data set is within 1 SD of the average. Hence we expect about 32 or roughly 1/3 of the data to be more than one SD from the average. This means that in this problem roughly 1/3 of the predicted final scores would be off by more than 1 rms error, or more than $\sqrt{1 - r^2} \times SD_y = \sqrt{1 - (.6)^2} \times 15 = .8 \times 15 = 12$ points.

(b) This is like the problems in the previous section, except that you must keep track of *which average and which SD to use* by whether you are working with the midterm scores or the final scores. A midterm score of 80 is $\frac{80-50}{25} = 1.2\ SD_x$'s above the average. The predicted final score is $r \times 1.2\ SD_y$'s above the average of the y's, so $55 + (.6)(1.2)(15) = 55 + 10.8 = 65.8$.

(c) 12 points = rms error (since SD is the expected size of the "experimental error").

12. To answer this question, of course we need to compute the predicted blood pressure for the man with 20 years of education. $x = 20$ is $7/3 = 2.33$ SD's above the average for education. The predicted blood pressure is

$$\text{ave}_y + (r \times SD_y)(2.33) = 119 + (-0.1 \times 11)(2.33) \doteq 116.4$$

(Note that we *do not* divide by $SD_x = 3$ again here, since that is already done in the computation of the 2.33 SD's above the average for x. The $\frac{r \times SD_y}{SD_x}$ is the slope of the regression line. If you use that equation, then "raw values" of x and y are used, not the corresponding standard unit values.) Thus, the blood pressure value of 118 for the man *is "a bit high"* (though of course a difference this small might not actually have any practical significance).