

Review Exercises Chapter 10

- #2 (a) The individuals who scored 115 at age 18 scored exactly 1 SD_{18} above the mean back then. From our regression formula we know that $\#SD_{35} = r \times \#SD_{18} = 0.80 \times 1 = 0.80$. To see how many IQ-points at age 35 this is, we multiply the $\#SD_{35} \times SD_{35} = 0.8 \times 15 = 12$. Now the only thing that remains to be done is to add 12 to the mean IQ-Score of age 35, since 115 is above the mean IQ-Score back then. That is, $\mu_{35} + 12 = 100 + 12 = 112$. Combining the steps above in one single equation yields

$$\mu_{35} + \left(\frac{115 - \mu_{18}}{SD_{18}} \right) \times r \times SD_{35} = 100 + (1 \times 0.80) \times 15 = 100 + 12 = 112.$$

- (b) The computation here is exactly the same as in (a), since our estimate of the average score at age 35 is also our best guess given the information. Thus, we predict that her score will be 112 at age 35.
- #3 (a) Based on our information about the wives and the correlation coefficient ($r \approx .25$), our best guess of his wife's height is

$$\begin{aligned} \mu_{wives} + \left(\frac{72 - \mu_{husbands}}{2.7} \times r \right) \times SD_{wives} &= 63 + (1.48 \times 0.25) \times 2.5 \\ &= 63.93 \text{ inches} \end{aligned}$$

- (b) In this case, our best guess of his wife's height is

$$\begin{aligned} \mu_{wives} + \left(\frac{64 - \mu_{husbands}}{2.7} \times r \right) \times SD_{wives} &= 63 - (1.48 \times 0.25) \times 2.5 \\ &= 62.07 \text{ inches} \end{aligned}$$

- (c) A husband that is 68 inches tall has exactly the same height as the average husband in the data set (i.e., he is 0 SDs above/below the mean). Therefore, our best guess of his wife's height is

$$\begin{aligned} \mu_{wives} + (0 \times r) \times SD_{wives} &= 63 + (0 \times 0.25) \times 2.5 \\ &= 63 \text{ inches} \end{aligned}$$

- (d) If the height of the husband is unknown, then our best guess of his wife's height is $\mu_{wives} = 63$ inches.

- #4 (a) Based on the information given, our best guess of his wife’s education level is

$$\begin{aligned}\mu_{wives} + \left(\frac{18 - 12}{3} \times r\right) \times SD_{wives} &= 12 + (2 \times 0.5) \times 3 \\ &= 12 + 3 = 15 \text{ years}\end{aligned}$$

- (b) Here we know that the wife has an educational level 1 SD above the wife’s mean, so our best guess of her husband’s years of schooling is

$$\begin{aligned}\mu_{husbands} + (1 \times r) \times SD_{husbands} &= 12 + (1 \times 0.5) \times 3 \\ &= 12 + 1.5 = 13.5\end{aligned}$$

Even though it might appear unlikely that someone has “completed” half a year of schooling, our best prediction given the available information is 13.5 years.

- (c) If we assume that half of the men have married women with less years of schooling than they themselves received, while the other half married better educated women, we can easily see how both statements can be true even if the average level of education for both groups is the same.

- #7 None of the explanations given by the doctors are convincing. The observed pattern is most likely due to the regression effect, which leads us to expect that either unusually high or unusually low readings will on average converge back to the mean. The reasoning displayed by the doctors is thus an example of the so-called *regression fallacy*.

- #9 (a) Clearly, this student is in the left-tail of the distribution. In order to predict his/her likely rank on the final exam, we have to find the z-score for his/her midterm first. We know that at the student’s percentile rank 5% of all observations are located in the distribution’s left tail. Assuming that the exam scores are normally distributed, we thus look up the z-score for an area of 90% in the normal table of the book (90% because by symmetry there are also 5% in the upper tail), which yields a z-score of approximately 1.65. However, as we are interested in the left tail, $z = -1.65$. (In other words, the student scored 1.65 SDs below average on the midterm.) The regression method now predicts that

the student will be $r \times SD_{midterm} = (0.50) \times (-1.65) = -0.825$ standard deviations away from the mean on the final exam. Looking at the normal table in the book, this corresponds to an area of about 59%, i.e., 59% of all observations fall between 0.85 and -0.85 . Thus, about 20.5% of all observations are in each tail ($\frac{100-59}{2} = 20.5$). Therefore, we predict that the student whose percentile rank on the midterm was 5% will do considerably better on the final exam and will on average obtain a score corresponding to the 20.5th percentile.

- (b) The approach here is the same as in (a). A 80th percentile rank gives us a $z = 0.85$ (60% of all observations fall between ± 0.85). We then predict a score $(0.50) \times (0.85) = 0.425$ SDs above the mean of the final, which translates to about 33.5% of all observations between 0.425 and -0.425 . Thus, this student's percentile rank on the final exam is approximately $\frac{33.5\%}{2} + 50\% = 66.75\%$.
- (c) A student whose percentile rank on the midterm was 50% was of course 0 SDs away from the mean on the midterm. Recall that the 50th percentile is by definition equal to the median and in the special case in which the data is normally distributed, it is also equal to the mean. We thus predict no change for this student's percentile rank on the final exam, i.e., our prediction is that this student will be located at the 50th percentile on the final exam.
- (d) Our best guess of a student's performance on the final with unknown midterm performance, is the average student's final score. Since the scatter diagram is football-shaped, we know that the grades are normally distributed with a single mode. Thus, the mean will be equal to the median, leading us to predict that this student will be located at the 50th percentile rank on the final.

Review Exercises Chapter 11

#3 To answer this question, recall that the r.m.s. error of the regression line of y on x is given by $\sqrt{1 - r^2} \times SD_y$. Thus:

(a)

$$r.m.s. \text{ error} = \sqrt{1 - 0.8^2} \times 2.5 = 0.6 \times 2.5 = 1.5 \text{ inches}$$

(b)

$$r.m.s. \text{ error} = \sqrt{1 - 0.8^2} \times 1.7 = 0.6 \times 1.7 = 1.02 \text{ inches}$$

- #4 (a) Recall that if the scatter diagram is football-shaped, the r.m.s. error of the regression line tells us how far certain points are above or below the regression line. We can then apply what we learned in chapter 5 with regard to the r.m.s. error and the regression line.

The $\frac{1}{3}$ of students, who's scores are furthest away from the predicted values (i.e., the regression line) can be found to about equal amounts above and below the regression line. Thus, the scores of $\frac{2}{3}$ of the students are clustered closer to the regression line in absolute distance. Knowing that the scatter diagram is football-shaped we can use a normal approximation to find the z-scores in between which 66.6% (i.e., $\frac{2}{3} \times 100$) of all observations fall: this is approximately ± 1 (in fact slightly less), as you recall from our empirical rule. Multiplying the r.m.s. error for the regression line (i.e., $\sqrt{1 - 0.6^2} \times 15 = 12$) by that number gives us the number of points away from the regression line within which $\frac{2}{3}$ of the students fall. Thus, for about $\frac{1}{3}$ of the students, the prediction for the final score was off by more than 12 points.

- (b) The students who scored 80 on the midterm are better than average. As a group they are expected to do better than average on the final exam – although there is a fare amount of spread as suggested by the correlation coefficient. Their average on the final can be estimated by the regression method: 80 is 1.2 SDs above average, so theses students will score about $r \times 1.2 = 0.6 \times 1.2 = 0.72$ SDs above average on the final exam. This is $0.72 \times 15 = 10.8$ points above average. So their predicted score is $55 + 10.8 = 65.8$.
- (c) Students, who scored 80 in the midterm are a smaller and more homogenous group. So, the SD of their final scores will be less than 15. How much less? Since the diagram is football-shaped, the scatter around the regression-line is about the same in each vertical strip and is given by the r.m.s. error for the regression line. The SD of the prediction in (b) is therefore

$$\sqrt{1 - r^2} \times SD_y = \sqrt{1 - 0.6^2} \times 15 = 12 \text{ points.}$$

Thus, the prediction is likely to be off by 12 points or so.

- #9 The data do not conclusively support the notion of a “sophomore slump.” What is most likely at work here is the regression effect, which

would lead us to predict that unusually high observations (in this case, a batting average of .285) will regress towards the mean in another round of measurements, regardless of players being “distracted” by TV appearances and the like.

#10 No, the regression method was not used to make the predictions. This can be seen by comparing the actual 1992 prices with the predicted 1993 prices. The first and probably most intuitive way to approach the problem, is to plot the two variables in a scatter diagram. If the predictions were derived by the regression method, you should be able to connect the five points by a straight line. This is clearly not the case.

The second way to think about this problem exploits the linear property of the regression method. If the regression method was used, then an increase of 2 in the 1992 price should be associated with an increase or decrease of x in the predicted 1993 price and due to the linear property an increase of 4 in the 1992 price should be associated with an increase or decrease of $2x$ in the predicted 1993 price. Looking at the data we can easily see that this is not the case: an increase of 2 from the second to the third observation in the 1992 price column is associated with an increase of 5 in the 1993 prediction column. However, the same increase from observation three to four in the 1992 price column is associated with a decrease of 1 in the 1993 prediction column.

#12 ¹ This question requires us to first predict what level of blood pressure the man under scrutiny is expected to have, given his level of education and then to establish a range of values that can be said to be typical for men at his educational level. Only if the predicted value falls outside this “normal” range can we conclude that his blood pressure level is unusual.

First, predict the expected blood pressure for a man with 18 years of education. 18 years of education is $\frac{5}{3}$ SDs above the average education level. Using the information provided in the question, we can now predict the man’s blood pressure level:

¹The text above is for the third edition. The problem in the fourth edition is slightly different. The man’s education is $7/3$ SD’s beyond the average education. Therefore, we expect his blood pressure, on average, to be $119 + (7/3) \times 11 \times -0.1 = 116.4$. Also, *r.m.s. error* ≈ 11 . As in the third edition, this man’s blood pressure is not a rare event given the data collected.

$$124 + r \times \frac{5}{3} \times SD_{\text{blood pressure}} = 124 + (-0.1) \times \frac{5}{3} \times 14 = 121.67\text{mm}$$

Next, we have to find the r.m.s. error for this prediction. This is

$$r.m.s. \text{ error} = \sqrt{1 - (-0.1)^2} \times 14 = 13.93\text{mm}.$$

Now we are finally in a position to judge whether a blood pressure level of 123mm is a bit on the high side for the man with 18 years of education. Based on our forecast of 121.67mm and the fact that about 68% of the actual blood pressure level values are expected to be within a range of $\pm 13.93\text{mm}$ around this value, 123mm does *not* appear to be on the high side. It seems to be a rather typical value.

Review Exercises Chapter 12

#1 This question is most easily answered by first computing the slope of the regression line. To do so, we can use the formula from page 204 in FPP:

$$\text{slope} = \frac{r \times SD_{\text{final}}}{SD_{\text{midterm}}} = \frac{0.60 \times 20}{10} = 1.2$$

Next, we can use this information to find the intercept. We know that the regression line has to go through the point of averages, which allows us to write

$$\text{mean}_{\text{final}} = \text{intercept} + 1.2 \times \text{mean}_{\text{midterm}} = \text{intercept} + 84$$

$$\Rightarrow \text{intercept} = 55 - 84 = -29.$$

All parts of the regression equation have now been identified. Thus, the regression equation is

$$\widehat{\text{final score}} = -29 + 1.2 \times \text{midterm score}$$

#3 ² If someone puts on 20 pounds, we do *not* predict that he will get taller by 0.9 inches, since weight obviously does not *cause* changes

²The above is for the third edition, the fourth edition changes the numbers but has the same interpretation. We don't know anything about how much that man would grow (probably none at all), but we know about people, on average, who are twenty pounds taller.

in height. It is only associated with height. What the slope of the regression line means in this case is that we predict that all those who are 20 pounds heavier than this person will on average be about 0.9 inches taller.

- #5 ³ False, the second investigator has not found the regression equation predicting the husband's income from the wife's income. Using the formula for the slope coefficient, we know that

$$\begin{aligned}\widehat{\text{husband's income}} &= \beta \times \text{wife's income} + \alpha \\ \Rightarrow \widehat{\text{husband's income}} &= \frac{24,000}{15,000} \times 0.2 \times \text{wife's income} + \alpha\end{aligned}$$

and the above β is $\frac{24,000}{15,000} \times 0.2 = .32 \neq 8$.

- #7 Recall that a straight line is uniquely defined by two points in the (x,y)-plane, i.e., there are no two distinct straight lines that can go through the same two points in the plane. Thus, if we want to compute the slope of the regression line from the information given in this problem, we need to find two points based on that information. We know that the regression line will always go through the point of averages (\bar{x}, \bar{y}) , and as we are given the average values of both beer and pizza consumption, we know that one point on the regression line is (4,4). The information provided also tells us that the intercept is +2, i.e., for an x-value of 0, we have a y-value of 2. Thus, the second point is (0,2). These two points suffice for the calculation of the slope of the regression line, which turns out to be 0.5, as

$$\text{slope} = \frac{\text{rise}}{\text{run}} = \frac{4 - 2}{4 - 0} = \frac{2}{4} = 0.5$$

- #9 ⁴ What is described here is an approximation to what is done in the regression method. If we were to decrease the size of the income groups up to the point where each dollar (or even cent) amount falls into a single group, we would end up with points forming the regression line. Thus, the slope of that line is an approximation to the slope of the regression

³The above is for the third edition. Although the numbers in the question change, the proof is the same. Simply calculating the regression using $\frac{SD_y}{SD_x}r$ coefficient shows that the second researcher has not found the regression equation predicting the husband's income from the wife's income. That is, $\frac{39,000}{26,000} \times 0.25 = .375 \neq 6$.

⁴The above is for the third edition. The fourth edition is slightly different. Here, $\text{slope} = \frac{15}{45,000} * 0.5 = \frac{1}{6,000}$.

line, so we can use our known formula for the regression slope, which is

$$\text{slope} = \frac{r \times SD_{IQ}}{SD_{Income}} = \frac{0.50 \times 15}{\$15,000} = \frac{1}{2,000}$$