

Extra Practice for chapter 3:

Is cardiovascular fitness (as measured by time to exhaustion running on a treadmill) related to an athlete's performance in a 20-km ski cross country ski-race? The following data shows x = treadmill run time to exhaustion (in minutes) and y = 20-km ski time (in minutes).

1. Why should treadmill time be the explanatory variable? Explain.
2. Draw a scatterplot and interpret a scatterplot for the data.
3. Calculate the least squares line and graph it on the scatterplot.
4. Interpret the slope in the context of the problem.
5. Interpret the x - and y -intercepts in the context of the problem. Are these reasonable values in this context?
6. Find and interpret the value of the correlation coefficient.
7. If the time was measured in seconds, how would this value change?
8. If r is high, can we conclude that a change in treadmill time causes the ski time to change? Explain.
9. Calculate and interpret the residual for the first point in the data set.
10. Sketch the residual plot. What does it tell you?
11. Calculate and interpret the values of r^2 and s in the context of the problem.
12. If you were to use number of hours instead of number of minutes for the ski time, how would the values of r^2 change?
13. Predict the ski time for a runner who can last 8 minutes on the treadmill.
14. Suppose the observation (13, 59) was added to the data set. What effect will this have on the LSRL and the values of r , r^2 , s ?
15. Instead of (13, 59), suppose the new point was (13, 65). What effect will this have on the LSRL and the values of r , r^2 , s ?

Person	Treadmill Time	Ski Time
1	7.7	71
2	8.4	71.4
3	8.7	65
4	9	68.7
5	9.6	64.4
6	9.6	69.4
7	10	63
8	10.2	64.6
9	10.4	66.9
10	11	62.6
11	11.7	61.7

Answers to Chapter 5 Extra Practice

- Treadmill time is the explanatory variable since we are using treadmill time to predict the results of the ski race.
- The scatterplot shows that there is a moderate negative association between the treadmill time and ski time. The relationship appears to be linear.
- Predicted Ski Time = $88.795649 - 2.3335102$ Treadmill Time
- For each additional minute an athlete is able to run on the treadmill, the average time it will take to complete the ski race will decrease by 2.33 minutes.
- Y: For athletes that can last 0 minutes on the treadmill, the average ski time will be 88.80 minutes. This doesn't make sense. It isn't reasonable that an athlete couldn't run for any amount of time on the treadmill. Also, 0 is way below the other x-values in the data set, so we should be cautious making extrapolations. X: Athletes that can last 38.11 minutes on the treadmill will take an average of 0 minutes to finish the ski race. This is humanly impossible. Again, 38 minutes is way beyond the x-values in our data set, so we should be very cautious extrapolating.
- $r = -0.796$. The correlation coefficient tell us that there is a moderately strong linear relationship between treadmill time and ski time.
- It wouldn't change. The correlation coefficient is not dependent on units.
- No, both variable probably depend on general fitness. Besides, we would need an experiment to conclude cause and effect.
- $y - \hat{y} = 71 - 70.859 = 0.141$. The model's prediction was only 0.141 minutes too low for the first athlete (very close!).
- Since there is no pattern in the residual plot, this means that the linear model is appropriate for this data.
- r^2 : 63% of the variability in ski time can be explained by treadmill time (or, treadmill time accounts for 63% of the variability in ski time). s : When using treadmill time to predict ski time, the predictions will be about 2.19 minutes away from the actual times, on average.
- r^2 wouldn't change at all and s would be 60 times smaller (.0365 hours).
- Athletes that can last 8 minutes on a treadmill will complete the ski race in 70.16 minutes, on average.
- This point will be influential since it's x-value is bigger than the others. However, since it falls very close to the LSRL above, the LSRL will change very little (slope and y-intercept will be similar). However, the other measures will change. For example, s will go down, since the residual for this point will be smaller than average. The value of r will be closer to -1 and r^2 will be closer to 1 since this point gives extra evidence of a negative association (it follows the pattern of the scatterplot).
- This point will also be influential since it's x-value is bigger than the others. However, since it falls very far from the LSRL above, the LSRL will be affected (it will be flatter, so the y-intercept will be lower and the slope closer to 0). This also means that s will go up, since moving the line will make the typical residual get bigger. The value of r will be closer to 0 and r^2 will be closer to 0 since this point makes the data more scattered. (The data point falls away from the pattern of the scatterplot)

