Correlation & Best Fit Lines

1. True/false practice:

   (a) Let X and Y be random variables with finite means and variances. Then $\text{Cov}[10X, 10Y] = \text{Cov}[X, Y]$.

   False. We have that

   $\text{Cov}[10X, 10Y] = E[10X \cdot 10Y] - E[10X]E[10Y] = E[100XY] - 10E[X] \cdot 10E[Y] = 100(E[XY] - E[X]E[Y]$

   One motivation for introducing the notion of correlation, rather than just relying on sample covariances, is that the sample standard deviations in the denominator make it so that larger values of the measurements don't immediately cause larger values of the correlation (correlation is always between -1 and 1, while covariance gets bigger the bigger the data are).

   (b) The least squares line of best fit is the line which minimizes the sum of the distances from the observed data points (x_i, y_i) to the line.

   False. The line of best fit is the line which minimizes the sum of the *squares* of the distances from the observed data points (x_i, y_i) to the line. There are other notions of best-fit lines, but least-squares has a number of good theoretical properties that make it a very common choice.

   (c) Assume that we have two random variables X and Y that have a relationship of the form $Y = a + bX + N(0, \sigma)$, where $N(0, \sigma)$ is a normal random variable with mean 0 representing the noise. Then if we do a number of samples and compute the sample correlation coefficient r, this r will be our maximum-likelihood estimate for b.

   False. The maximum-likelihood estimate for b will be $r \cdot \frac{\sigma_y}{\sigma_x}$, where σ_y and σ_x are the sample standard deviations of the observed Y-values and X-values, respectively.

2. **(cp. HW 36 #3)** We are testing if a four-sided die is fair by rolling it 40 times. For the data in the second and third columns below, find

   - the correlation coefficient
   - the angle between the two data streams
   - the best-fitting line.

value on die	observed	expected
1	9	10
2	8	10
3	11	10
4	12	10

We denote the observed values by x_i and the expected values by y_i. Because we have that $\sigma_y = E[Y^2] - E[Y]^2 = 100 - 10^2 = 0$, the formula $r = \frac{\sum_{i=1}^{4}(x_i-\overline{x})(y_i-\overline{y})}{4\sigma_x\sigma_y}$ will be an indeterminate form of the form $\frac{0}{0}$. There won't be any way to use a couple of our other formulas either. We can find the correlation coefficient by interpreting it as the cosine of the angle between the two data streams, however.

The two data streams can be viewed as the vectors $\langle 9, 8, 11, 12\rangle$ and $\langle 10, 10, 10, 10\rangle$. We can find the cosine of the angle between these data streams by taking a dot product and dividing by the product of the lengths of the two vectors:

$$\begin{aligned}\cos\theta &= \frac{\langle 9, 8, 11, 12\rangle \cdot \langle 10, 10, 10, 10\rangle}{||\langle 9, 8, 11, 12\rangle|| \cdot ||\langle 10, 10, 10, 10\rangle||} \\ &= \frac{400}{\sqrt{410}\sqrt{400}} \\ &= \sqrt{\frac{400}{410}} \approx .988.\end{aligned}$$

This is the value of the correlation coefficient we were looking for. $\cos^{-1}\left(\sqrt{\frac{400}{410}}\right) \approx 0.157 \approx 8.96$ degrees.

The slope of the best fit line $y = \beta_0 + \beta_1 x$ will be $\beta_1 = r \cdot \frac{\sigma_y}{\sigma_x} = 0$, and we can solve for β_0 by using the formual $\beta_1\overline{x} + \beta_0 = \overline{y}$ to see that $\beta_0 = \overline{y} = 10$. Our line of best fit is thus $y = 10$.

3. **(modified Stewart/Day 11.3.19)** We have the following before-and-after data from a drug trial:

Before	After
7.4	3.7
5.1	2.6
6.9	3.4
7.2	3.6
1.4	0.7
4.3	2.1

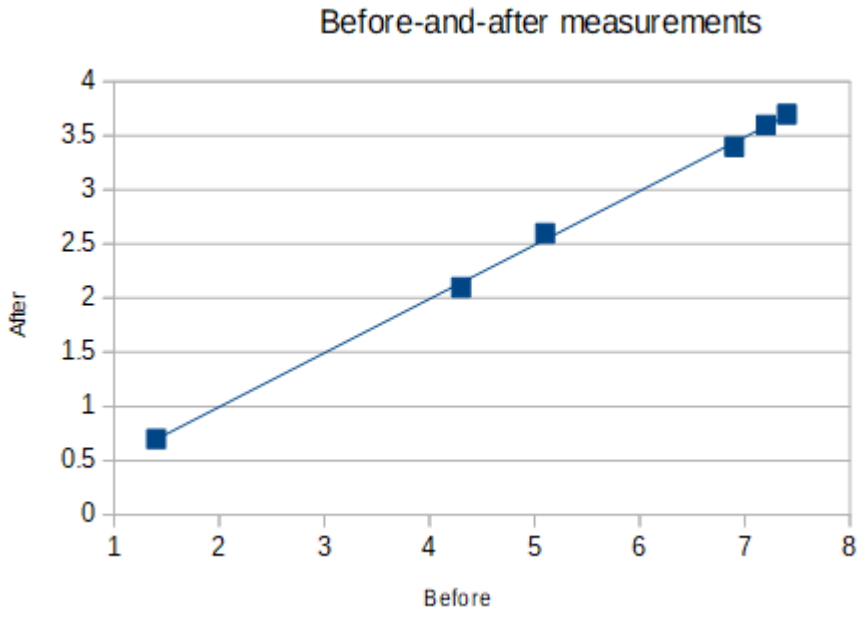
What is the sign of the correlation coefficient? What does that mean? Calculate the line of best fit and plot the points along with the best fit line.

We call the "before" values x and the "after" values y. We calculate the correlation coefficient to be approximately 0.999 using the formula $r = \frac{\sum_{i=1}^{6}(x_i-\overline{x})(y_i-\overline{y})}{6\sigma_x\sigma_y}$; this is positive, which means that higher x values correspond to higher y values (equivalently, that there is a positive relationship between x and y).

Calculating the standard deviations, we find $\sigma_y \approx 1.054$ and $\sigma_x \approx 2.113$, so we find that the slope β_1 of the best fit line is approximately 0.4986.

The intercept β_0 will be given by $\overline{y} = \beta_1\overline{x} + \beta_0$; we find that β_0 is approximately -0.0011, so our best fit line is approximately $y = 0.4986x - 0.0011$.

A picture of the points and the best-fit line (which, as we expect since r is close to 1, is very close to going through all the points) is shown below:



4. **(original)** We suspect that the price of oil influences the price of the stock of a plastic manufacturer. We have the following data:

oil price	stock price
100	19
80	32
60	41
40	48

What is the sign of the correlation coefficient? What does that mean? Calculate the line of best fit.

Suppose you believe the price of oil will be 90 a year from now. What do you expect the stock price of the plastic manufacturer to be a year from now?

We let y denote the company's stock price and x the oil price.

We use the formula $r = \frac{\sum_{i=1}^{4}(x_i-\overline{x})(y_i-\overline{y})}{4\sigma_x\sigma_y}$ compute a correlation of about -0.990; this negative correlation means that as oil prices go up, stock prices for the plastic manufacturer go down (i.e. there is an inverse relationship).

We use the formula $\beta_1 = r\frac{\sigma_y}{\sigma_x}$ to find a slope of -0.480.

We use the formula $\overline{y} = \beta_0 + \beta_1\overline{x}$ to gind an intercept of 68.60, so our best-fit line is $y = 68.60 - 0.48x$.

We can plug $x = 90$ into our best-fit line to estimate what the stock price will be a year from now when the price of oil is 90. We have $y = 68.60 - 0.48 \cdot 90 = 25.4$, so we get an estimate of 25.4.

Note that *interpolating*, that is, using our best-fit line to guess the y-values for x-values between data points we used in developing our best-fit line, as we did here tends to be much safer than *extrapolating* (using the line to guess y-values for x-values outside the range we used in developing our best-fit line); if we had tried to use this best-fit line to predict the stock price at an oil price of 200, for example, we would have gotten a negative stock price, which doesn't make much sense.

Acknowledgments

Problems inspired by the HW36 problem set.

Problems labeled Stewart/Day from Day, Troy and Stewart, James. *Biocalculus: Calculus, Probability, and Statistics for the Life Sciences.* Cengage Learning, 2019.