Worksheet 27

Correlation & Best Fit Lines

1. True/false practice:

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

   (b) The 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.

   (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.

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

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.

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?

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.