Cancer Mortality

Purpose: In this lab we will explore the relation between proximity to a nuclear materials storage site and cancer mortality rates.

Preview: We will learn how to model data, fit it to a straight line, interpret the model and measure how good our model is.

Background: Larsen and Marx (1986) state "Since World War II, plutonium for use in atomic weapons has been produced at an Atomic Energy Commission facility in Hanford, Washington. One of the major safety problems encountered there has been the storage of radioactive wastes. Over the years, significant quantities of these substances, including strontium 90 and cesium 137—have leaked from their open-pit storage areas into the nearby Columbia River, which flows along the Washington-Oregon border, and eventually empties into the Pacific Ocean."

To measure the extent to which exposure to the radiation affected cancer mortality rates during 1959-1964, researchers calculated an *index of exposure*. The index takes into consideration the county's distance down the river from the plant and the distance from the river. More precisely, the index is formulated on the assumption that county or city exposure is directly proportional to river frontage and inversely proportional both to the distance from the Hanford site and the square of the county's (or city's) average depth away from the river (Smith and Moore, 1996). The following table gives the index of exposure and cancer mortalities (the average number of cancer deaths per 20,000 residents per year during the period 1959-1964).

County	Exposure	Mortality
Umatilla	2.49	147.1
Morrow	2.57	130.1
Gilliam	3.41	129.9
Sherman	1.25	113.5
Wasco	1.62	137.5
Hood River	3.83	162.3
Portland	11.64	207.5
Columbia	6.41	177.9
Clatsop	8.34	210.3

Sources

1. Fadeley, R.C. (1965). Oregon malignancy pattern physiographically related to Hanford, Washington, Radioisotope Storage, Journal of Environmental Health 27, 883-897.

2. Larsen, R.J., and Marx, M.L. (1986). An Introduction to Mathematical Statistics and its Applications, 2nd Edition. Prentice-Hall, Englewood Cliffs, New Jersey. Case Study 1.2.4.

3. David A. Smith and L.C. Moore, Calculus: Modeling and Application, Houghton Mifflin Co., Single-Variable Chapters, 1996.

4. http://www.statsci.org/data/general/hanford.html

For the questions below, unless the question in purely computational, answer using complete sentences. Your lab instructor will give you more details about the format of your report that you will hand in.

Part I: Modeling the Data

1. Open the spreadsheet for this lab from the class webpage, make a copy of it, and rename it with your group names.

2. Insert a scatter plot for columns B and C next to the data. Make sure you customize your chart to start at $(0, 0)$ so that it's easy to see where the line lies. Don't add a linear trend line just yet.

3. Why is it reasonable to model this data with a straight line?

4. What makes a good line to fit data? Write down a few criteria that characterize a good line.

5. What might be a reasonable slope and y-intercept for a line to fit this data? Insert your values in cells $B14$ and $B15$. Use these to compute data for your line in column D and add the line to your chart as you did in the training pre-lab sheet. You may want to change the vertical axis range to make sure you see the entire line. Does your line fit as well as you thought? If not, change the slope and y-intercept until it does. Describe what you did. Does your line fit your criteria from the previous part?

6. Insert a linear trend line for the data on your chart. How does this new line compare to your answer in 5?

Part II: Interpreting the Model (using your equation from Part I, Question 5)

7. What is the significance of the y-intercept in the equation? Is your answer reasonable, that is, does it make sense in the context of the situation?

8. What is the slope of your line? What is its significance in terms of the situation being modeled?

9. Estimate the mortality rate for a community with an index of exposure 7.0.

10. Estimate the index of exposure for a community with a mortality rate of 1000.

11. What numbers are reasonable values for the domain of this function? Explain.

Part III: A Measure of How Well a Line Fits the Data

In statistics, a common criterion for how well a line fits to a set of points is the sum of the squares of the vertical distances from the line to the data points.

12. What is an 'error' for a given line and a given data point? Why do we want to to square errors? Why do we want to sum them?

13. Is the following a correct formula for computing the sum of the squares of the vertical distances correct from a line $y = mx + b$ correct? If not, correct the formula.

$$\sum_{i=1}^{9}(y_i - (mx_i + b)),$$

where m is the slope of the line, b is the y-intercept, x_i is the index of exposure for the i^{th} county or city, and y_i is the cancer mortality for the i^{th} county or city.

14. In column E, insert the values predicted by your linear trend line (you may want to insert the slope and y-intercept values in $B18$ and $B19$). In columns F and G, compute the square differences for each of the two lines from Part I. At the bottom of each of those columns, sum the square differences to find the Sum of Square Errors for each of the lines. Compare your two answers. Do they make sense in comparison to one another? Explain.