

Exercise 10- Distance Decay and Tuberculosis Rates (10 pts)

The role of distance in the frequency of disease is an important, but often overlooked, factor. The literature is replete with research documenting distance decay effects associated with point source pollution, but very little research has been done regarding infectious diseases and distance decay. To that end, this exercise will guide you through an examination of tuberculosis in the United States and the role that distance has in the frequency of this disease.

First we need a little background in how distance might influence tuberculosis. There has been a recent upsurge in drug-resistant tuberculosis in the United States over the last decade. Research suggests that this new form of tuberculosis has been imported to the East Coast from Latin America and Southeast Asia (1,2). As infected populations move to new areas for work or other reasons, it would be expected that the tuberculosis rates would mirror the path of emigration, and that the highest rates would be found at the “source” areas (e.g. East Coast). It would also be expected that tuberculosis rates would decrease with distance from the “source” areas, since shorter distance moves are much more common than longer distance moves. Therefore, if we were to gather data for tuberculosis rates for selected cities at increasing distances from the East Coast, we should be able to determine the rate of change in tuberculosis rates with increasing distance. Luckily, these data are available, as are the means of analyzing them.

In order to determine the rate of change in tuberculosis rates with increasing distance we need to fit a distance-decay function to the tuberculosis data. Often these functions are a negative exponential, meaning that the disease rate decreases and an increasing rate. However, fitting a curve to this type of function is difficult and tedious. Has luck would have it, our TB data appear to be linear, which makes fitting a line to the function much easier but equally as tedious.

The equation for fitting a simple line to a function is $y = a + bx$. Yes, it is a regression or “best fit” line, and it is very useful in predicting values of y based on the values of x . However, determining the parameters a and b requires some doing. The first step is to determine the *Sum of the Squared Deviations from the Mean* (or SS), which is a measure of the total amount of x and y variation from the means of x and y within the data set. The equation for this looks imposing but it really is not:

$$SS = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

Where X is the distance measurement and n is the sample size. To more clearly illustrate this process let’s use an example.

Distance(X)	Rate (Y)	$\bar{x} = 30$	$\bar{y} = 7.18$	$\sum x = 150$	$\sum y = 35.9$
10	9.2				
20	7.8				
30	7.1				
40	6.3				
50	5.5				
				$\sum X_i^2 = 10^2 + 20^2 + 30^2 + 40^2 + 50^2 = 5500$	
				$(\sum X)^2 = (10 + 20 + 30 + 40 + 50)^2 = 22500$	
				Therefore:	$SS = 5500 - \frac{22500}{5} = 1000$

1: McKenna MT, McCray E, Onorato IM, Castro KG. [The epidemiology of tuberculosis among foreign-born persons in the United States, 1986 to 1993](#). N Engl J Med 1995;332:1071-6.

2: Enarson DA, Wang JS, Dirks JM. [The incidence of active tuberculosis in a large urban area](#). Am J Epidemiol 1989;129:1268-76.

Next we will need to calculate the *Sum of the Crossproducts* (or SC) of the deviations from the mean. This parameter denotes the deviation of a value of x (or y) from all other values of x (or y), and is a measure of the dispersion of the data. This equation also appears formidable, but when broken down into its constituent parts is very straightforward:

$$SC = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$$

So in our example:

$$\sum X_i Y_i = 988$$

$$(\sum X_i)(\sum Y_i) = 150 \times 35.9 = 5385$$

$$\text{Therefore: } SC = 988 - \frac{5385}{5} = -89$$

Note that the parameter is negative. This means that the values of y decrease with increasing values of x ... distance decay. We can then combine both equations to give us the b parameter:

$$b = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

You should note that the numerator (top part) is the *Sum of the Crossproducts* and denominator (bottom part) of this equation is the *Sum of the Squared Deviations from the Mean*. The equation give us the average amount of change (either positive or negative) in x for every unit change in y . Therefore, in our continuing example:

$$b = \frac{-89}{1000} = -0.089$$

It can be show mathematically that the mean of both x and y always fall on the line of best fit. Therefore, to calculate the y-intercept or a we simply need to rearrange our regression equation as:

$$a = \bar{y} - b\bar{x}$$

So in our example:

$$a = 7.18 - (-0.089)30 = 9.85$$

This results in the completed regression equation:

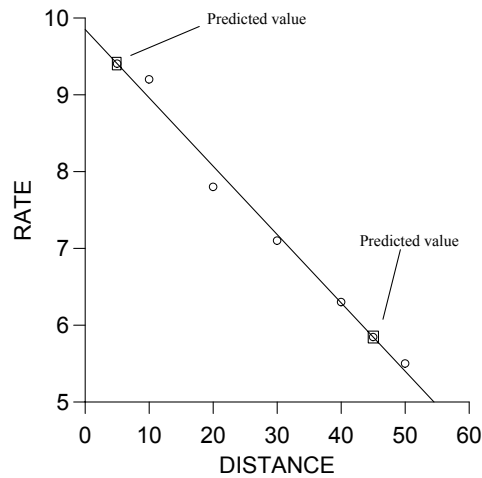
$$y = 9.85 + (-0.089)x$$

Now it is a simple matter of plugging various values of x (or distance) to predict values of y (disease rate). So now let's predict what the disease rate in our example would be at 5 miles and 45 miles.

$$y = 9.85 + (-0.089)5 = 9.405$$

$$y = 9.85 + (-0.089)45 = 5.845$$

Let's see how we did. Below is a plot of the original data set and our two predictions.



Not bad... The usefulness of this tool should be apparent. We are now able to predict disease rates at different distances from the "source" area. This will allow researchers (and Geo 532 students) to determine the health care needs for cities and towns in advance, and notify health care officials and facilities of the expected number of tuberculosis cases.

The Exercise

Using the method detailed above, please determine the line of best fit (regression equation) for the data below. Then use that equation to predict the number of tuberculosis cases for the distances listed in the table. Please show your work were specified and **PLEASE** be neat.

<i>Tuberculosis Data</i>	
Distance	TB Rate
332	18.14
379	12.83
502	7.83
603	11.05
694	6.94
710	7.72
779	2.16
837	2.64
844	3.86
879	3.90
950	0.92

$n =$ _____

$\sum x =$ _____

$\sum y =$ _____

$\bar{x} =$ _____

$\bar{y} =$ _____

$b =$ _____

$a =$ _____

Please fill out the following equations:

$SS =$

$SC =$

Regression Equation : _____

Please predict the tuberculosis rates for the following distances. Make sure to write in the equation that you used.

Distance (Kilometers)	Equation	Predicted TB Rate
350		
435		
572		
660		
821		
913		
