Linear Regression / Least Squares

Consider the following data taken from a 2006 U.S. Environmental Protection Agency report that summarizes municipal solid waste in the United States. The full report is available at http://www.epa.gov/msw/pubs/ex-sum05.pdf. The report presents Table ES-1 as shown below:

Table ES-1
GENERATION, MATERIALS RECOVERY, COMPOSTING, COMBUSTION WITH ENERGY RECOVERY, AND DISCARDS OF MUNICIPAL SOLID WASTE, 1960 - 2005
(in millions of tons)

Activity	1960	1970	1980	1990	2000	2003	2004	2005
Generation	88.1	121.1	151.6	205.2	237.6	240.4	247.3	245.7
Recovery for recycling	5.6	8.0	14.5	29.0	52.7	55.8	57.2	58.4
Recovery for composting*	Neg.	Neg.	Neg.	4.2	16.5	19.1	20.5	20.6
Total materials recovery	5.6	8.0	14.5	33.2	69.1	74.9	77.7	79.0
Combustion with energy recovery†	0.0	0.4	2.7	29.7	33.7	33.7	34.1	33.4
Discards to landfill, other disposal‡	82.5	112.7	134.4	142.3	134.8	131.9	135.5	133.3

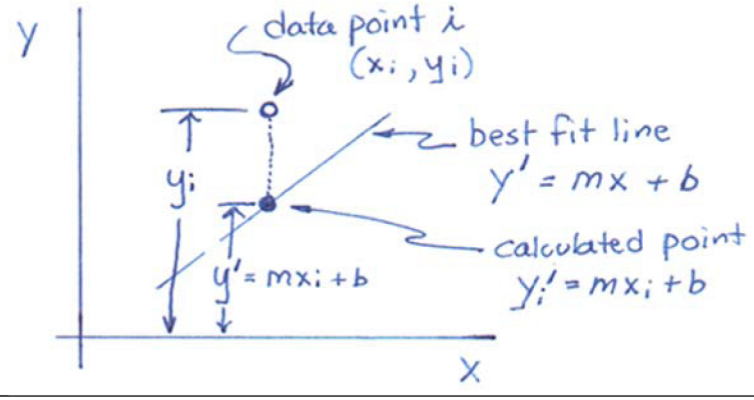
* Composting of yard trimmings, food scraps and other MSW organic material. Does not include backyard composting.
† Includes combustion of MSW in mass burn or refuse-derived fuel form, and combustion with energy recovery of source separated materials in MSW (e.g., wood pallets and tire-derived fuel).
‡ Discards after recovery minus combustion with energy recovery. Discards include combustion without energy recovery. Details may not add to totals due to rounding.

Class Problem: Using linear regression, determine the best fit line of "recovery for recycling" versus "year" for 2000, 2003, 2004 and 2005. Use Excel.

(a) Compute the slope m and the y-intercept b.
(b) Plot the four raw data points versus the fit. Use markers only (with no lines) for the raw data and lines only (no markers) for the fit.

Thus, the best fit line is _____ Don't erase your spreadsheet – we will use it later. Now, we can see that the fit above is "pretty good" since it passes very close to the points. But, how do we define the degree of goodness?????

Coefficient of Determination (r^2):



The distance between the data point and the best fit line is . . .

Since some errors are + and some -, we need to square the error terms . . .

If we simply reported the *error*^{2} term above, the number would vary in <u>size</u> depending on the problem being solved. We would like a number that varies between 0 (for a poor fit) to 1 (for a perfect fit). So, we "normalize" the error term as follows . . .

$r^2 =$

Here, $\bar{y}$ is the average value of y_i

This normalizes r^2

A more convenient form is

$$r^2 = \left[\frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}} \right]^2$$

Class Problem: Find r^2 for the problem worked at the beginning of class:

How do we interpret the result $\rightarrow r^2$?

if it's $\emptyset$: No apparent relationship between x & y

if it's 1 : x determines y 100%

...or the variation in y is wholly due to x

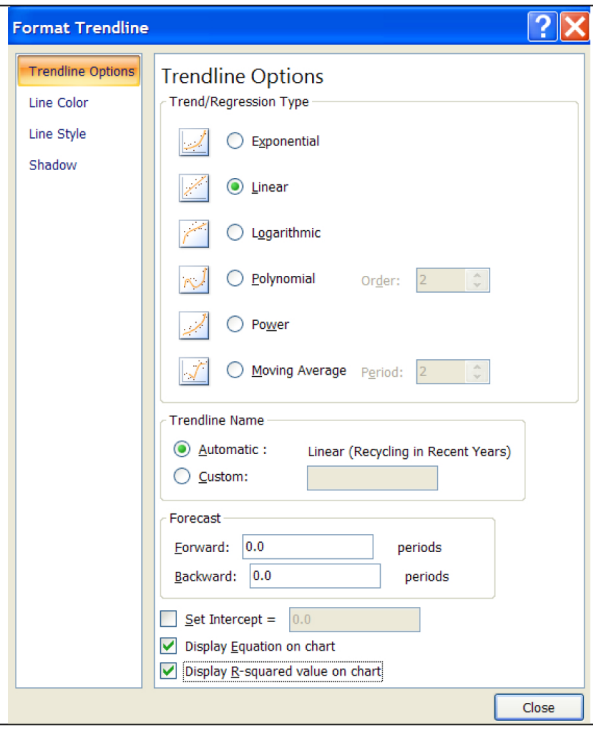
...or y depends on x and there are no other variables that affect y

For the class problem, the year is responsible for _____ of the millions of tons recycled between 2000 and 2005. Something else leads to the _____ variation in recycling from the fit.

Repeating in Excel: Use Excel to do the hard work for us.

Steps:

1. Enter x and y data
2. Create a scatter plot
3. Right click on the markers and select “Add Trendline”
4. Select “Linear”, “Display Equation on chart,” and “Display R-squared value on chart



EXAM TIP: When studying, be sure that you can solve problems like the one in class today BY HAND and USING EXCEL.