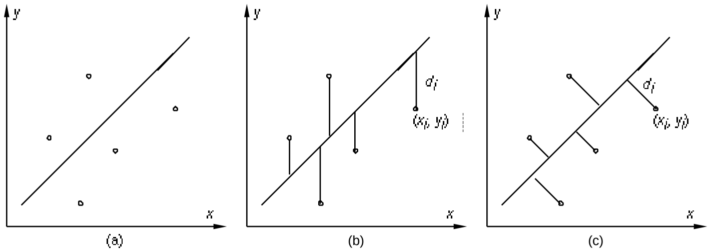
MATH 164, Spring 2001 **Due Date:**____________ **Name(s):** ________________________

Honors Project 1: Least Squares

The Analytic Approach

In many science and engineering applications, one is often given a set of data points $\{(x_i, y_i), i = 1, ..., N\}$ in R^2, and interested in finding the line which "best fits this data" (see Fig. (a) below). One solution to this problem is provided by *classical least squares*: finding the line for which the sum of the squares of the distances d_i from the data points to the line in the direction of the dependent axis is a minimum (see Fig. (b) below). Another solution to this problem is provided by *invariant least squares*: finding the line for which the sum of the squares of the perpendicular distances d_i from the data points to the line is a minimum (see Fig. (c) below). In general, these lines are different.

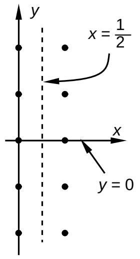


In MATH 261, it will follow (as an application of optimization of functions of two variables) that the line $y = mx + b$ which best approximates the data in the sense of classical least squares is given by

$$m = \frac{N \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{N \sum x_i^2 - \left(\sum x_i\right)^2} \quad \text{and} \quad b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N \sum x_i^2 - \left(\sum x_i\right)^2}.$$

In MATH 351, we will discuss how to find the line which best approximates the data in the sense of invariant least squares.

The classical least squares line is easier to compute than the invariant least squares line, which is certainly a strength of classical least squares. On the other hand, the classical least squares line depends on which variable is assumed to be independent and which is assumed to be dependent, and this is a weakness. In particular, classical least squares assumes the line which best approximates data has an equation of the form $y = mx + b$, and for a set of data points such as $\{(0, 0), (0, \pm 1), (0, \pm 2), (1, 0), (1, \pm 1), (1, \pm 2)\}$, it will not necessarily give the "best" answer. (The line whose equation is $x = 1/2$ — which is the invariant least squares line of fit — appears to fit the data better than the line whose equation is $y = 0$ — which is the classical least squares line of fit.)



Exercise 1: Find the constants m and b that define the line $y = mx + b$ that best matches the following data in the sense of classical least squares, to the nearest 2 decimal places.

x	1.0	1.2	1.5	1.8	2.0	2.3	2.4	2.6	3.4	3.7	3.8	3.9
y	2.7	4.0	3.6	4.2	4.5	5.1	5.4	5.7	7.2	7.8	7.8	6.0

Exercise 2: Repeat Exercise 1 (use the same data set, but now) finding the line $x = ny + d$ that best matches the data in the sense of classical least squares.

Exercise 3: Rewrite the equation $x = ny + d$ in the form $y = mx + b$. (You should get an equation that is *different* than the one you got in Exercise 1!)

In many science and engineering applications, one is also often given a set of data points $\{(x_i, y_i), i = 1, ..., N\}$ in R^2, and interested in finding a curve other than a line which "best fits this data". For example, you might want to fit a curve of the form $y = Ae^{kx}$ to such a set of data points. (This might happen if you were studying population growth or radioactive decay problems, for example.) The difficulty in trying to mimic the MATH 261 computations that we used to arrive at the classical least squares solution to our problem, is that the systems of equations we would have to solve is much more complicated; in fact it is too complicated to be of practical value. Instead of approaching our problem this way, it is much easier to try to transform the data and the equation to which data is to be fit. For example, to fit a curve of the form $y = Ae^{kx}$ to the data $\{(x_i, y_i), i = 1, ..., N\}$, we might take logarithms, transforming our equation into $\ln y = kx + \ln A$, and our data into $\{(x_i, \ln y_i), i = 1, ..., N\}$. By using the classical least squares solution to the problem of fitting a line $Y = mx + b$ to this "new" data set, we would find the $m = k$ and $b = \ln A$ — or $k = m$ and $A = e^b$ — for which $y = Ae^{kx}$ best fits our original data set $\{(x_i, y_i), i = 1, ..., N\}$.

Exercise 4: Find the constants A and k that define the curve of the form $y = Ae^{kx}$ that best matches the following data, to the nearest 2 decimal places.

x	1.0	1.4	1.8	2.5	2.9	3.2	3.6	3.9	4.0	4.2	4.4	4.5
y	0.9	1.1	1.3	1.7	2.0	2.3	2.8	3.1	3.3	3.6	3.9	4.0

One hidden advantage to transforming data by taking logarithms (as we did above) is that taking logarithms tends to temper the size of data: if the y_i-values in a data set are much larger or much smaller than the corresponding x_i-values, then the logs of the y_i-values may be more commensurate with the corresponding x_i-values.

Exercise 5: By taking logs of both x- and y-data values, find the constants A and p that define the curve of the form $y = Ax^p$ which best fits the following data, to the nearest 2 decimal places.

x	1	10	20	50	100	110
y	1.2	3.2	4.2	6.1	8.1	8.4

On the other hand, one disadvantage to transforming data is that data can often be transformed in more than one way, and the results obtained via different transformations may be different.

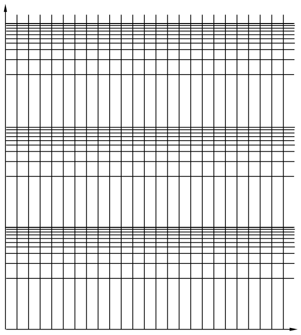
The Geometric Approach

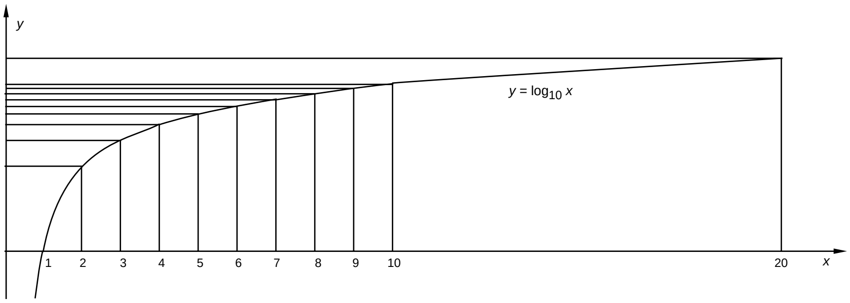
So far, the processes we have discussed have been analytical. In many applications, however, they are often studied geometrically.

Given a set of data points $\{(x_i, y_i), i = 1, ..., N\}$, finding the line $y = mx + b$ which best approximates it (in whatever sense you might please) is straightforward: the data is plotted, the line is estimated by eye, and — having drawn the line — the parameters that describe it are estimated by eye.

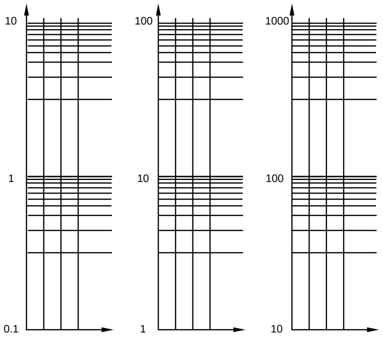
Exercise 6: By graphing, find the constants m and b that define the line $y = mx + b$ that best matches the data from Exercise 1 to the nearest 2 decimal places.

Finding the curve $y = Ae^{kx}$ that best fits a set of data points is a little trickier, however. To find the curve of the form $y = Ae^{kx}$ that best fits the set of data points $\{(x_i, y_i), i = 1, ..., N\}$, you might numerically transform the data into data of the form $\{(x_i, \ln y_i), i = 1, ..., N\}$, plot the data, estimate the line by eye, estimate the parameters m and b that define the line, and then numerically compute $k = m$ and $A = e^b$. Alternately, rather than taking the time and effort to transform the data, you might work on semi-log graph paper. A sample of such paper is to the right.



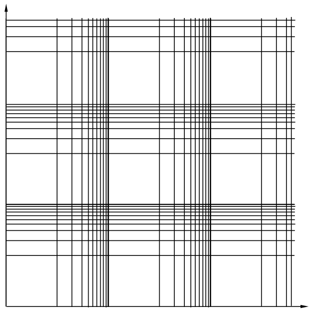


On semi-log graph paper the y-axis tic marks are rescaled by $\log_{10}$ (see the figure below) so that the units along the vertical axis can be labeled in any of the manners illustrated to the right. Using semi-log graph paper facilitates the plotting of y-values over a large range without the need for numerical conversions by hand. One of the things that makes semi-log graph paper so useful is the regularity in the irregularity of the spacing of the y-axis tic marks; this regularity is a consequence of the fact that $\log_{10} 10^n = n \log_{10} 10 = n$ (or, if you are thinking in terms of natural logs, $\ln e^n = n \ln e = n$.)



Exercise 7: By graphing on semi-log graph paper, find the constants A and k that define the line $y = Ae^{kx}$ that best matches the data from Exercise 4 to the nearest 2 decimal places.

In addition to semi-log graph paper — which facilitates the plotting of y-values over a large range without the need for numerical conversions by hand — there is log-log graph paper which facilitates the plotting of x- *and* y-values over a large range without the need for numerical conversions by hand.



Exercise 8: By graphing on log-log graph paper, find the constants A and p that define the line $y = Ax^p$ that best matches the data from Exercise 5 to the nearest 2 decimal places.

Exercise 9: Your answers to Exercises 7 and 8 are probably not the same. Why?