Least Squares Linear Regression

ILAP 2 – MA205 **Time allotted: 7 hours per cadet**
Due in class, 16 November 1999

Authors: MAJs Weate, Kobylski, Dr. Rickey D/Math
MAJ Cross D/Physics

On the very first day of the nineteenth century, the astronomer Giuseppe Piazzi observed a star that he had never seen before. He was able to observe it for several more nights over 41 days and was amazed to see that it was moving in a direction opposite to the other stars. He lost track of it because of bad weather and could not find it again, so he published his observations in the hope that someone could compute where to look. The problem stumped astronomers. But a young mathematician, a mathematician who had devised the method of least squares a few years earlier, was able to find it using this technique and assuming that the orbit was an ellipse (others of the time assumed a circular orbit in order to make the computations reasonable). Carl Friedrich Gauss (1777-1855) predicted where to look and the star was found immediately. The star was actually the first asteroid ever discovered. The discovery of Ceres made Gauss famous. In this project you will learn about the method of least squares, a method used by scientists of all disciplines as a fundamental means of approximation and prediction.

Karl Friedrich Gauss

In your PH201 labs this semester, you have often used this technique of linear regression. For each experiment, you collected data points in the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where the y value was some measured observation based upon a fixed observation of an x value. For example, in Laboratory 3 you collected the times it took for a ball to drop from 10 different heights. Next, you hypothesized that there was an underlying linear relationship between the recorded x values and corresponding y values, and then used the EXCEL spreadsheet program to find a "best fit" line through these data points.

The method of finding the line that best fits the collected data is known as linear regression. As a MA205 student at the 40 lesson mark, you have all the tools needed to understand the mathematics behind this linear regression that you have been conducting with great faith in your physics labs. The purpose of this project is to understand the math behind this very useful and ubiquitous technique.

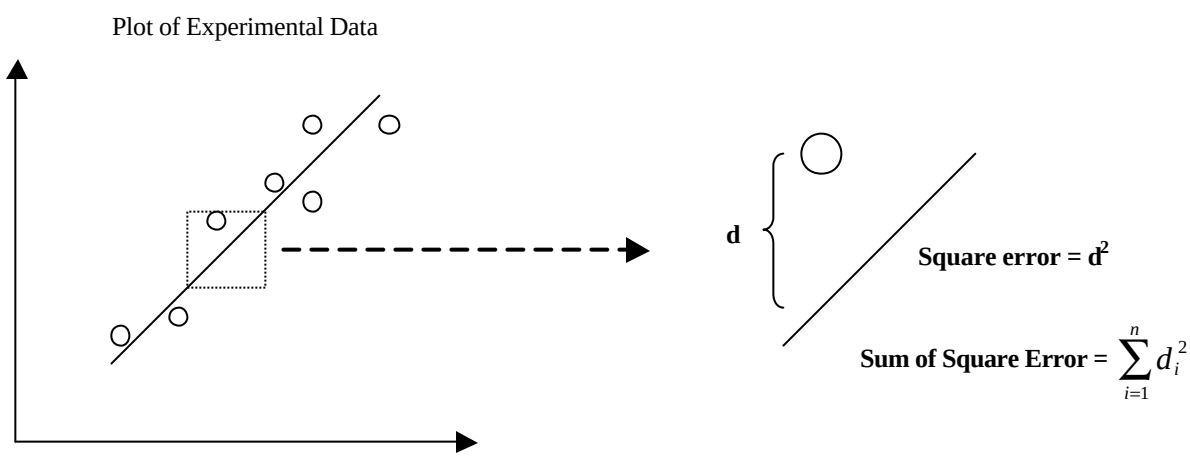

Figure 1. Square Error

In order to find the line that "best fits" our given data points, we need to first have an idea of what we mean by "best fit." How will you measure the fit of a model line $\boldsymbol{y = mx + b}$ through the data? One common, and certainly not exclusive way of measuring this closeness of fit is to add up the square error for each data point. The error of a point can be defined as the difference in y value of a given data point and the y value of the model line for any given x value. The square error is just this error squared. Figure 1 above illustrates this idea. The objective then is to find the constants $\boldsymbol{m}$ and $\boldsymbol{b}$ for a line $\boldsymbol{y = mx + b}$ that give the line the minimum possible sum of square error.

1. Take the data from one of your group member's PH201 Lab 5: Conservation of Energy experiment and explore the ideas behind linear regression in the following requirements:

a) Use the transformed velocity data from the experiment and find the "best" line $\boldsymbol{y = mx + b}$ through the data by finding the values of $\boldsymbol{m}$ and $\boldsymbol{b}$ by minimizing the sum of these square errors. (Hint: First calculate the difference between the y value of each data point and the value of the corresponding point on the "best fit" line. This difference will be expressed in terms of the unknowns $\boldsymbol{m}$ and $\boldsymbol{b}$. Now for each data point, square this difference term. Add up all of these square differences to get a function in terms of $\boldsymbol{m}$ and $\boldsymbol{b}$. Now minimize this function.) Once you determine $\boldsymbol{m}$ and $\boldsymbol{b}$, plot the data and the resulting linear regression on the same graph. Remember, that you must be able to show and explain your work in your presentation.

b) Given any set of n data points in the form (x_i, y_i), show that according to this method of minimizing the sum of square errors, the line of best fit is obtained when

$$m\sum x_i + bn = \sum y_i$$

$$m\sum x_i^2 + b\sum x_i = \sum x_i y_i$$

where the summations are taken over the entire data set. These equations are known as the *normal equations*.

c) From the *normal equations* above, how would you find the expressions for the unknown coefficients $\boldsymbol{m}$ and $\boldsymbol{b}$ below

$$\boldsymbol{b} = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} \qquad \boldsymbol{m} = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

where the summations are again taken over the entire data set? Do not actually do this; just explain how you could. These are the equations that programs like Excel or your calculator use to find the unknown linear coefficients.

d) Take the physics lab data from Part (a) above and use the equations for $\boldsymbol{m}$ and $\boldsymbol{b}$ above to again find the "best fit" line through these points. Compare and discuss any differences with your result from Part (a).

e) There are other ways to measure the "fitness" of a line. One method could be to define the "best fit" line by minimizing the sum of the errors instead of the sum of square errors. What problems occur when you try this approach?

f) The method of least squares assumes that there is some error in the y values. There could also be error in the actual x measurement. Using the linear model found above, if the error in measuring the x value is less than 0.001 meters, use the differential to express the corresponding error bound in the prediction for y when using this best fit line.

2. In your PH201 Lab 5 experiment, you transformed the collected velocity data by squaring each measured value.

a) Take the original velocity data collected and find the least square error line through these data points. Plot the points and the resulting line on the same graph.

b) The *residual* of a data point is the difference between the actual observed value and the corresponding value on the fitted line (the *d* value in the illustration above). Calculate the residuals for the linear regression on the untransformed data.

c) Calculate the residuals of the linear regression from Part 1 on the transformed data. The residuals are just the error terms for each point. Compare the two sets of residuals. Which linear regression is the better model of the data set?

d) Explain why you transformed the data in the physics experiment in the first place.

3. In the physics experiment, you used the linear regression line to calculate an approximation for the constant acceleration of gravity. The linear model obtained could also be used to predict the value of the velocity given the height of drop.

a) Use your linear model from Part (1) to predict the velocity when the height of drop in the experiment is exactly 0.5 meters. How close to the real value do you think you will be?

b) Explain how you think Gauss used linear regression to determine where to look for the strangely moving star?

Submission Requirements.

There is no formal write up requirement for this project. Each group will conduct a formal briefing for their instructor. The focus of the briefing will be to explain the method of least squares linear regression and its applications. Each group will submit a copy of any slides or other documents used in the briefing. In addition, each group will submit a copy of all work leading to the solution of each requirement above. Include any appropriate documentation. Your group's final grade will be based upon the briefing and the submissions listed above.