Josh Ulrich

U0805610

Statistics and Probability: Estimating Body Weight with Weighted Least Squares Method

 As the world becomes more interconnected with the advancement of civilization and technology, it is increasingly apparent how related many topics are. This is especially true in mathematics, as many topics apply knowledge gained from various other mathematical genres. Linear algebra is no exception, and applies knowledge from other levels of math, and can be applied to compute many different problems more efficiently. One area that can be greatly benefitted by the applications of linear algebra is the generating and computation of large amounts of data. This is very valuable for many statistics research projects. One of the ways that this can help is by allowing the researcher to modify the data that has been collected quickly and effectively. For example, if a researcher performed a large survey of the height and weight of random people, they may wish to make a regression model in order to predict the relationship and effect of height on weight. However, they may discover that their subjects responded with slightly inaccurate data. This is one of the risks of having self-reported data, as each subject can choose how to respond. However, it is important to use a correct method for adjusting the data, rather than arbitrarily adding a few pounds to each subject’s reported weight. In this project, we assume each subject underestimated their weight by 2-4%. First, the calculation of the least squares regression line of the original data must be done. X corresponds to each subject’s reported height in inches, and y corresponds to each subject’s reported weight in pounds.

x=(77,72,64,73,69,64,72,67,65,73,74,73,75,66,74,80,63,68,53,63,71,62,77,63,73,73,72,67,74,70,70,71,74,69) for all 34 subjects. The y-values correspond with x-values, in order. y=(240,230,120,175,150,180,175,170,112,215,200,185,160,112,195,125,114,200,110,115,155,100,340,150,145,160,195,180,170,180,190,150,330,150).

This is a plot of the height and weight of all the subjects, using the aforementioned values. 

Now, a least squares regression line must be found for the data. This can be done by using the method found in sections 6.5 and 6.6 of Lay’s textbook, *Linear Algebra and Its Applications, 5E (2015).* This is done by taking the height data, x, and writing it as a matrix, X, then doing the same for the weight data, y. Then, the relationship is described by the equation XB = Y, where B is the matrix composed of β0 and β1, which are the y-intercept and slope of the regression line, respectively. In order to solve for B, we must multiply both sides by X transpose, thus changing X into XTX, which is a square matrix. This allows us to find the inverse of XTX, and by multiplying both sides by this inverse, B is left alone, described by the equation B = (XTX)-1 \* (XTY). Calculating this equation gives B = (-215.95, 5.58) which represent β0 and β1, respectively. Plugging these values of B into the regression line gives the equation y = -215.95 +5.58x. Displayed with the plot, this is the regression line.



This allows us to estimate any subject’s weight when given a height. Using a given height, 6’2”, or 74 inches (my height), gives us y = -215.95 + 5.58(74) = 196.97 pounds. However, we suspect that each subject underestimates their weight by 2-4%, which would cause this estimation to be an underestimate. One way to find the new value would be to multiply the result (196.97) by 1.03, taking 3% as the average underestimation percentage. This gives a new value of 202.88 pounds. However, simpler than finding the new value for each result after it has been calculated, we can create a new regression line by applying the adjustment to the weights from our sample data. The new y-values are y\* = (247.20, 236.90, 123.60, 180.25, 154.50, 185.40, 180.25, 175.10, 115.36, 221.45, 206.00, 190.55, 164.80, 115.36, 200.85, 128.75, 117.42, 206.00, 113.30, 118.45, 159.65, 103.00, 350.20, 154.50, 149.35, 164.80, 200.85, 185.40, 175.10, 185.40, 195.70, 154.50, 339.90, 154.50). From this, we can reapply the solution for B, but using y\* instead of y. This gives us B = (XTX)-1 \* (XTY\*). The new values of B are (-225.52, 5.787). Thus, comparing side by side the two regression lines:

y = -215.95 +5.58x

y\* = -225.52 +5.787x

Obviously, there are differences between both equations, but it is paramount to observe how the two equations look when fitted with our data. The blue line represents the original y estimate regression line, and the red line represents the y\* estimate.



Using the same 74 inches as an example height, y = -215.95 +5.58(74) = 196.97 pounds, and y\* = -225.52 +5.787 (74) = 202.718 pounds. By solving the difference in estimates

d = (Weighted estimate)/(Original estimate)

We find d = 202.718/196.97= 1.029 or approximately 1.03. Our weighted model is effective at mitigating the 3% underestimate of weight by the subjects. This demonstrates how useful linear algebra methods can be for analyzing statistical data. This weighted regression line will allow researchers in the health fields to make more accurate inferences about their patients knowing only their height. This can be useful in computing average height to weight relationships that people may use to self-determine their fitness. Overall, just as was mentioned in the introduction to the topic, by combining the knowledge of multiple fields of study in mathematics, we were able to generate a more accurate model for predicting the weight of a subject given their height.