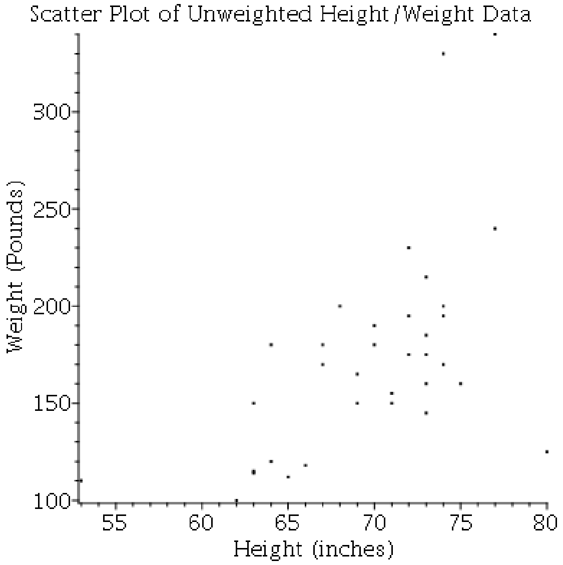
Veronica Dean-Perry
Maria Novozhenya

Using Weighted Least Squares to Model Data Accurately

Linear algebra has applications across many, if not all, mathematical topics. These days, every industry uses and generates data. Most of this data is either self-reported or entered into a platform by a human. Anytime a human is involved in a process, the possibility of error increases. There is great value is producing an estimate for this error and modeling an estimated correct entry. This is especially important in the field of Statistics, with most of the data coming from small samples. Luckily, there are procedures to account for the bias of self-reporting.

Looking at the height and weight data given in Lab 3, we can see that the data was self-reported. Below is a scatterplot with each point corresponding to a point within the data set.



In order to predict the expected weight for a given height, we must create a least squares regression model, which will create a linear equation for which we can plug in height values, given variable x, to calculate an expected weight, given variable y. The commonly used matrix equation Ax=b is often used to do so. This matrix equation can be rewritten as *X*ß=*y* where *X*, and *y* correspond to the following matrices.

$$X=\begin{bmatrix} 1 & 77 \\ 1 & 72 \\ 1 & 64 \\ 1 & 73 \\ 1 & 69 \\ 1 & 64 \\ 1 & 72 \\ 1 & 67 \\ 1 & 65 \\ 1 & 73 \\ 1 & 74 \\ 1 & 73 \\ 1 & 75 \\ 1 & 66 \\ 1 & 74 \\ 1 & 80 \\ 1 & 63 \\ 1 & 68 \\ 1 & 53 \\ 1 & 63 \\ 1 & 71 \\ 1 & 62 \\ 1 & 77 \\ 1 & 63 \\ 1 & 73 \\ 1 & 73 \\ 1 & 72 \\ 1 & 67 \\ 1 & 74 \\ 1 & 70 \\ 1 & 70 \\ 1 & 71 \\ 1 & 74 \\ 1 & 69 \end{bmatrix} \qquad y=\begin{bmatrix} 240 \\ 230 \\ 120 \\ 175 \\ 150 \\ 180 \\ 175 \\ 170 \\ 112 \\ 215 \\ 200 \\ 185 \\ 160 \\ 118 \\ 195 \\ 125 \\ 114 \\ 200 \\ 110 \\ 115 \\ 155 \\ 100 \\ 340 \\ 150 \\ 145 \\ 160 \\ 195 \\ 180 \\ 170 \\ 180 \\ 190 \\ 150 \\ 330 \\ 150 \end{bmatrix}$$

In order to calculate for the unknown vector ß, we obtain the normal equations by applying the transpose of *X* to both sides of the equation and multiplying the matrices as follows.

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 77 & 72 & 64 & 73 & 69 & 64 & 72 & 67 & 65 & 73 & 74 & 73 & 75 & 66 & 74 & 80 & 63 & 68 & 53 & 63 & 71 & 62 & 77 & 63 & 73 & 73 & 72 & 67 & 74 & 70 & 70 & 71 & 74 & 69 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 34 & 2371 \\ 2371 & 166323 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 5899 \\ 416845 \end{bmatrix}$$

$$\begin{bmatrix} 34 & 2371 \\ 2371 & 166323 \end{bmatrix} \text{ß} = \begin{bmatrix} 5899 \\ 416845 \end{bmatrix}$$

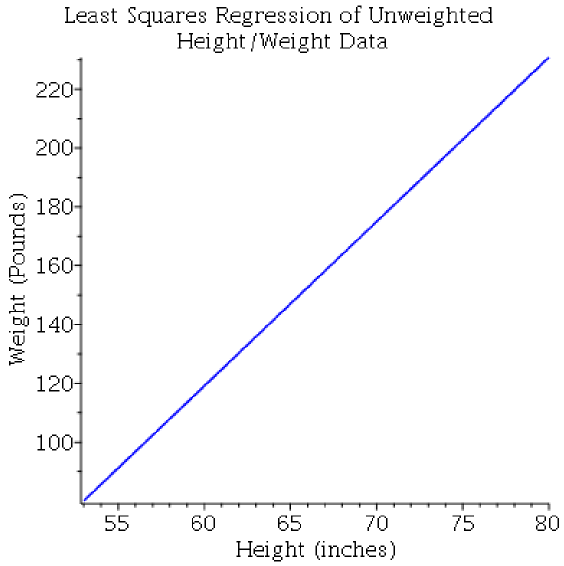
The resulting equation leaves the vector ß to be solved for by multiplying both sides by the inverse of $X^T X$.

$$(X^T X)^{-1} = \begin{bmatrix} \frac{166323}{33341} & \frac{-2371}{33341} \\ \frac{-2371}{33341} & \frac{34}{33341} \end{bmatrix}$$

$$\text{ß} = \begin{bmatrix} \text{ß}0 \\ \text{ß}1 \end{bmatrix} = \begin{bmatrix} \frac{166323}{33341} & \frac{-2371}{33341} \\ \frac{-2371}{33341} & \frac{34}{33341} \end{bmatrix} \begin{bmatrix} 5899 \\ 416845 \end{bmatrix} = \begin{bmatrix} \frac{-7200118}{33341} \\ \frac{186201}{33341} \end{bmatrix}$$

We substitute the entries in ß into the equation for a least squares line, y=ß$_0$+ß$_1$x to obtain the following equation for the least squares regression line that approximates the given height and weight data.

$$\boldsymbol{y = -\frac{7200118}{33341} + \frac{186201}{33341}x}$$



We can use the data to make predictions about the expected weight of a person based on their height. For example, an estimate of the expected weight of a person who is 5'10" (70"), can be calculated as follows:

$$y = -\frac{7200118}{33341} + \frac{186201}{33341}(70) = 174.98$$

While this line will approximate the data as given, we must consider bias as a result of certain sampling types. If we assume that people are prone to randomly underestimate their weight by 2-4%, we can calculate a regression line equation that considers this underestimate. If the amount of underestimation is truly random, the expected average value of underestimation is around 3%. We can calculate a weight matrix as the identity matrix multiplied by .97. We then apply that weight matrix to the original equation and proceed as before calculating a new matrix ß which we will call ß* (for differentiation purposes). Our new equation becomes *WXß*=y*.

$$W=\begin{bmatrix} .97 & 0 \\ 0 & .97 \end{bmatrix}$$

$$WX=\begin{bmatrix} .97 & 74.69 \\ .97 & 69.84 \\ .97 & 62.08 \\ .97 & 70.81 \\ .97 & 66.93 \\ .97 & 62.08 \\ .97 & 69.84 \\ .97 & 64.99 \\ .97 & 63.05 \\ .97 & 70.81 \\ .97 & 71.78 \\ .97 & 70.81 \\ .97 & 72.75 \\ .97 & 64.02 \\ .97 & 71.78 \\ .97 & 77.60 \\ .97 & 61.11 \\ .97 & 65.96 \\ .97 & 51.41 \\ .97 & 61.11 \\ .97 & 68.87 \\ .97 & 60.14 \\ .97 & 74.69 \\ .97 & 61.11 \\ .97 & 70.81 \\ .97 & 70.81 \\ .97 & 69.84 \\ .97 & 64.99 \\ .97 & 71.78 \\ .97 & 67.90 \\ .97 & 67.90 \\ .97 & 68.87 \\ .97 & 71.78 \\ .97 & 66.93 \end{bmatrix}$$

$$(WX)^T WX = \begin{bmatrix} 31.99 & 2230.87 \\ 2230.87 & 156100 \end{bmatrix}$$

$$(WX)^T Y = \begin{bmatrix} 5722.03 \\ 404100 \end{bmatrix}$$

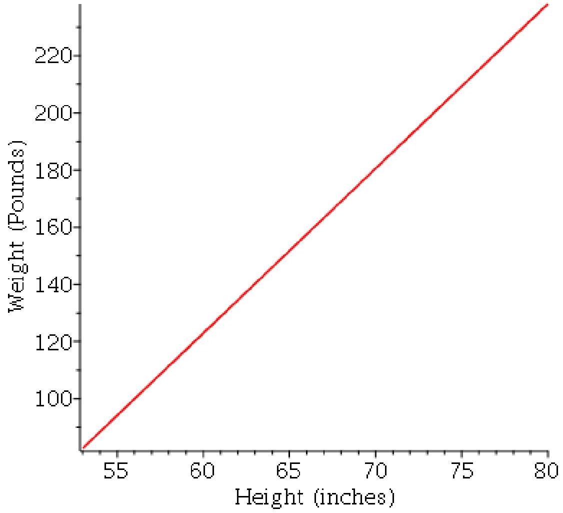
$$((WX)^T WX)^{-1} = \begin{bmatrix} 5.30 & -.08 \\ -.08 & 0 \end{bmatrix}$$

$$ß* = \begin{bmatrix} ß_0^* \\ ß_1^* \end{bmatrix} = \begin{bmatrix} 5.30 & -.08 \\ -.08 & 0 \end{bmatrix} \begin{bmatrix} 5722.03 \\ 404100 \end{bmatrix} = \begin{bmatrix} -222.63 \\ 5.76 \end{bmatrix}$$

The result is the following equation for the least squares regression line using the weighted approach.

$$y = -222.63 + 5.76x$$

Least Squares Regression of Weighted Height/Weight Data

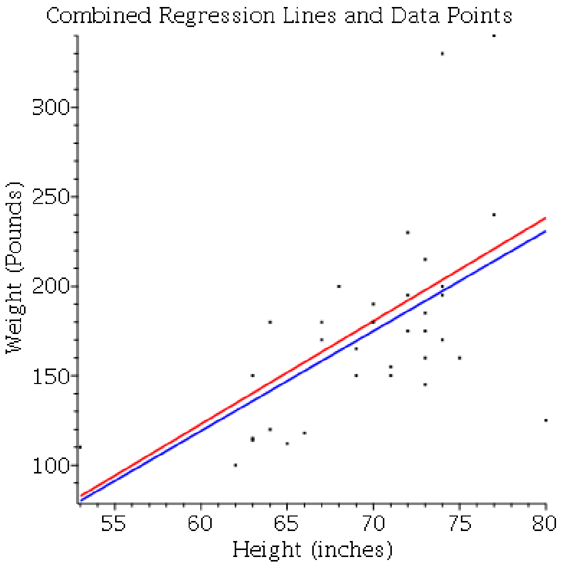


Since we are assuming each person in the data set actually weighs more than the data set states by an average of 3%, we would expect that when we calculate the expected weight for a person of any height, it would be approximately 3% greater than the weight calculated using the previous model. We can check this by calculating the expected weight of a person who is 5’10” (70”) and comparing it to the expected weight calculated using the previous model.

$$y = -222.63 + 5.76\,(70) = 180.57$$

$$\frac{180.57}{174.98} \sim 1.03$$

The intuition was correct. Below is a visual representation of the original data points and the two calculated least squares lines representing model 1 and model 2.



Calculating models as such is relevant for statistical analysis, in any field imaginable. As health professionals are continually trying to get an accurate picture of our nation’s health as a whole, they must analyze data from samples of American citizens. Accounting for bias using weighted least squares methods can help them to get the most accurate prediction of the measurements of people in the country. For example, self-reported BMI, daily water intake, height, weight, daily exercise time, daily computer time - this is all self-reported data that can be modeled accurately to account for bias. In addition, creating weighted regression lines can help professionals like doctors compare their patients’ real weight to their expected weight in order to make decisions about their health and well-being. In conclusion, using the power of the weighted least squares approach with linear algebra can help to produce more accurate statistical information that allows for those who are performing analysis of data to make more informed inferences.

References

Lay, D. C., Lay, S. R., & McDonald, J. (n.d.). Linear algebra and its applications (5th ed.).

Least Squares Fitting. (n.d.). http://uspas.fnal.gov/materials/05UCB/4_LSQ.pdf