PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

MAY 2020: IN SEMESTER ASSESSMENT (ISA) B.TECH. IV SEMESTER

UE18MA251- LINEAR ALGEBRA

MINI PROJECT REPORT

ON

WEIGHTED LEAST SQUARES FOR DATA ANALYSIS

Submitted by

1.	SAIPRAKASH L SHETTY	PES2201800730
2.	NAVEEN SURESH	PES2201800508
3.	AKASH YADAV	PES2201800415

Branch & Section : COMPUTER SCIENCE ENGINEERING, SECTION - F

PROJECT EVALUATION

(For Official Use Only)

Sl.No.	Parameter	Max Marks	Marks Awarded
1	Background & Framing of the problem	4	
2	Approach and Solution	4	
3	References	4	
4	Clarity of the concepts & Creativity	4	
5	Choice of examples and understanding of the topic	4	
6	Presentation of the work	5	
	Total	25	

Name of the Course Instructor :

Signature of the Course Instructor :

Introduction

Linear algebra has wide range of applications in the field of mathematics. Every major organization or industry uses or generates data which is essential for them. Mostly, this data tends to be fed into a platform by a human or self-reported. Anytime a human is involved in a process, the possibility of error increases. There must be a lot of emphasis in producing an estimate for this error and modelling an estimated correct entry. There are various procedures to account for the bias of self-reporting and other such human errors. One such procedure is Least Squares Method for data analysis.

The method of least squares is a standard approach in regression analysis to approximate the solution of over determined systems by minimizing the sum of the squares of the residuals made in the results of every single equation. The most important application is in data fitting. But this project specifically focuses on weighted least squares which is a generalization of ordinary least squares and linear regression in which the errors covariance matrix is allowed to be different from an identity matrix.

In this project we are discussing how, 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.

Unlike linear and nonlinear least squares regression, weighted least squares regression is not associated with a particular type of function used to describe the relationship between the process variables. Instead, weighted least squares reflects the behaviour of the random errors in the model; and it can be used with functions that are either linear or nonlinear in the parameters. It works by incorporating extra nonnegative constants, or weights, associated with each data point, into the fitting criterion. The size of the weight indicates the precision of the information contained in the associated observation. Optimizing the weighted fitting criterion to find the parameter estimates allows the weights to determine the contribution of each observation to the final parameter estimates. It is important to note that the weight for each observation is given relative to the weights of the other observations; so different sets of absolute weights can have identical effects.

Each term in the weighted least squares criterion includes an additional weight that determines how much each observation in the data set influences the final parameter estimates and it can be used with functions that are either linear or nonlinear in the parameters.

In a weighted fit, less weight is given to the less precise measurements and more weight to more precise measurements when estimating the unknown parameters in the model. Using weights that are inversely proportional to the variance at each level of the explanatory variables yields the most precise parameter estimates possible. Weighting the sum of the squares of the differences may significantly improve the ability of the least square regression to fit the linear model to the data. Weighted least square is an efficient method that makes good use of small data set. It also shares the ability to provide different types of easily interpretable statistical intervals for estimation, prediction, calibration and optimization.

Literature Review:

Concept of Weighted Least Squares:

Weighted Least Squares is considered to be an extension of Ordinary Least Squares. As mentioned earlier, we would be focusing on Weighted Least Squares in this project. Weights are non-negative constants that are attached to data points. Another important concept that is associated with the weighted least squares is the concept of homoscedasticity and heteroscedasticity.

- **Homoscedasticity -** The Ordinary Least Squares assumes that there is a constant variance in the errors. This is homoscedasticity which essentially translates to same variance. It describes the situation in which the error term is same across all values of the independent variables.
- **Heteroscedasticity** - This arises when the size of the error term differs across all values of the independent variables. It essentially is the violation of homoscedasticity.

The Weighted Least Squares can be used when any of the following conditions hold true:

1) The data violates the assumption of homoscedasticity: The Weighted Least Squares can be used when the data does not have equal variances and there is no even scatter pattern that arises from the data
2) When the focus is restricted to certain specific areas such as low input value, when you want to highlight some important areas in your study, the weighted least squares can be used.
3) In a situation where the data points should not be treated equally.

Advantages and Disadvantages of Weighted Least Squares

Weighted Least squares has certain striking advantages which include:

- It is well suited for extracting maximum information from small data sets.
- It is the only method that can be used for data points for varying quality.

Some of the Disadvantages include:

- Accuracy of weights is expected in this method. Weight estimation can have unpredictable results while dealing with small samples.
- Sensitivity to outliers is a problem.

Identifying the Weights

The principal difficulty is to determine the value for the weights. We need to use the Weighted Least Squares when there is a non-constant variance. Most commonly, the pattern of non-constant variance is that either the standard deviation or the variance of the residuals is linearly related to the mean (the fits).

The absolute residuals essentially are estimates of standard deviation. So we might plot absolute residuals versus fits. If this looks linear, we could fit a regression line (response = absolute residuals, predictor = fits) to the pattern. The predicted values from this regression could be viewed as smoothed estimates of the standard deviations of the points. So, our weights in a weighted least squares regression would be, $w_i = \frac{1}{(\hat{S}_i)^2}$. Note that the "predicted" standard deviations would have to be squared in the weight function.

The squared residuals essentially are estimates of variances. We might plot squared residuals versus fits. If this looks linear, we could fit a regression line (response = squared residuals, predictor = fits) to determine smoothed estimates of the variance. Denote these estimates as $\hat{V}_i^{\,2}$. Then, in a weighted regression, use weights $w_i = \frac{1}{\hat{V}_i^{\,2}}$.

REPORT

Derivation of Weighted Least Square (WLS):

In this method, the deviation between the observed and expected values of y_i is multiplied by a weight w_i where w_i is chosen to be inversely proportional to the variance of y_i.

For simple linear regression model $y_i = \alpha + \beta x_i + e_i$

The *Weighted Least Squares function* is,

$$\Sigma\, w_i\, e_i^2 = \Sigma\, w_i(y_i - \alpha - \beta\, x_i)^2$$

For easy computation, let L be represented by the sum of square of the residuals, so that $L = \Sigma\, w_i e_i^2$

$$L = \Sigma\, w_i(y_i - \alpha - \beta\, x_i)^2$$

We want to minimize L with respect to α and β. The least square estimate (a and b) of α and β is obtained by differentiating L and equate the derivative to zero (0).

Thus $\quad \frac{dL}{d\alpha} = -2\sum w_i(y_i - a - bx_i) = 0$

Or $\quad -2\sum w_i(y_i - a - bx_i) = 0$

$\sum w_i(y_i - a - bx_i) = 0$

$\sum w_i y_i - \sum a w_i - \sum b w_i x_i = 0$

$\sum w_i y_i - a\sum w_i - b\sum w_i x_i = 0$

$a\sum w_i = \sum w_i y_i - b\sum w_i x_i$

$\boldsymbol{a = \frac{\sum w_i y_i - b\sum w_i x_i}{\sum w_i}}$ (i)

Also $\quad \frac{dL}{d\beta} = -2\sum w_i(y_i - a - bx_i)x_i = 0$

Or $\quad -2\sum w_i(x_i y_i - ax_i - bx_i^2) = 0$

$\sum w_i(x_i y_i - ax_i - bx_i^2) = 0$

$\sum w_i x_i y_i - \sum a w_i x_i - \sum b w_i x_i^2 = 0$

$\sum w_i x_i y_i - a\sum w_i x_i - b\sum w_i x_i^2 = 0$

$b\sum w_i x_i^2 = \sum w_i x_i y_i - a\sum w_i x_i$

$b = \frac{\sum w_i x_i y_i - a\sum w_i x_i}{\sum w_i x_i^2}$ (ii)

On substituting (i) into (ii) we have

$$b = \frac{\sum w_i x_i y_i \; - \left[\frac{\sum w_i y_i - b\sum w_i x_i}{\sum w_i}\right]\sum w_i x_i}{\sum w_i x_i^2}$$

$$b = \frac{\sum w_i x_i y_i - \left[\frac{\sum w_i x_i \sum w_i y_i - b\left(\sum w_i x_i\right)^2}{\sum w_i}\right]}{\sum w_i x_i^2}$$

$$b = \frac{\frac{\sum w_i \sum w_i x_i y_i \; - \left[\sum w_i x_i \sum w_i y_i - b\left(\sum w_i x_i\right)^2\right]}{\sum w_i}}{\sum w_i x_i^2}$$

$$b = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i \; + \; b\left(\sum w_i x_i\right)^2}{\sum w_i \sum w_i x_i^2}$$

$$b \sum w_i \sum w_i x_i^2 \; = \; \sum w_i \sum w_i x_i y_i \; - \; \sum w_i x_i \sum w_i y_i + b\left(\sum w_i x_i\right)^2$$

$$b \sum w_i \sum w_i x_i^2 \; - \; b\left(\sum w_i x_i\right)^2 \; = \; \sum w_i \sum w_i x_i y_i \; - \; \sum w_i x_i \sum w_i y_i$$

$$b\left[\sum w_i \sum w_i x_i^2 \; - \; \left(\sum w_i x_i\right)^2\right] \; = \; \sum w_i \sum w_i x_i y_i \; - \; \sum w_i x_i \sum w_i y_i$$

$$\boldsymbol{b} \; = \; \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2}$$

Summary of derivation:

For simple linear regression model (Weighted Least Square):

y = a +bx

$$\boxed{b \; = \; \frac{\sum w_i \sum w_i x_i y_i \; - \; \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 \; - \; \left(\sum w_i x_i\right)^2}}$$

$$a = \bar{y} - b\bar{x}$$

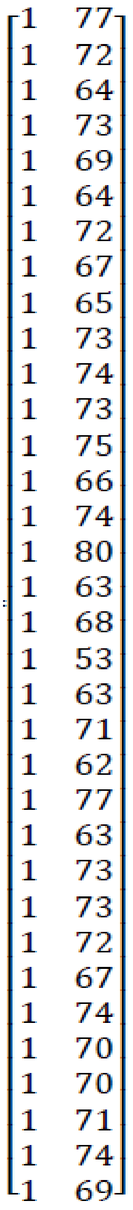
$$\boxed{a = \frac{\sum w_i y_i \; - \; b\sum w_i x_i}{\sum w_i}}$$

Problem:

Given below is height and weight of students in a school. 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 to calculate an expected weight.

Matrix A

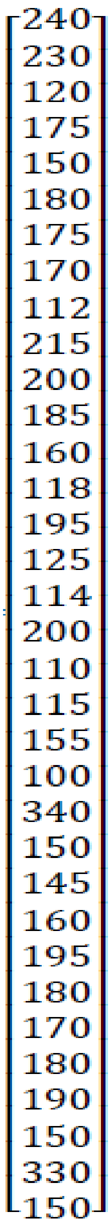
Number of students Height(inches)



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

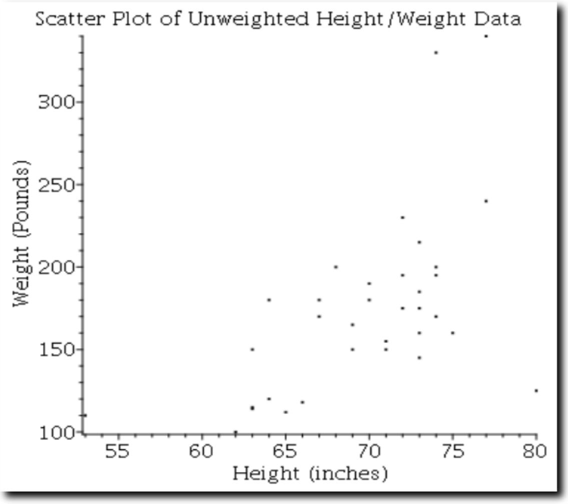
Matrix B

Weight(pounds)



$$\begin{bmatrix} 240 \\ 230 \\ 120 \\ 175 \\ 150 \\ 180 \\ 175 \\ 170 \\ 112 \\ 215 \\ 200 \\ 185 \\ 160 \\ 118 \\ 195 \\ 125 \\ 114 \\ 200 \\ 110 \\ 115 \\ 155 \\ 100 \\ 340 \\ 150 \\ 145 \\ 160 \\ 195 \\ 180 \\ 170 \\ 180 \\ 190 \\ 150 \\ 330 \\ 150 \end{bmatrix}$$

Below is a scatterplot with each point corresponding to a point within the data set.



Where number of students is w_i , Height is x_i and weight to be predicted is considered as y_i. We substitute the previous entries in the following equation to obtain the equation of line in the form **y = a + bx.** Where,

$$b = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2}$$

$$a = \frac{\sum w_i y_i - b \sum w_i x_i}{\sum w_i}$$

<u>(i) Results and Discussions:</u>

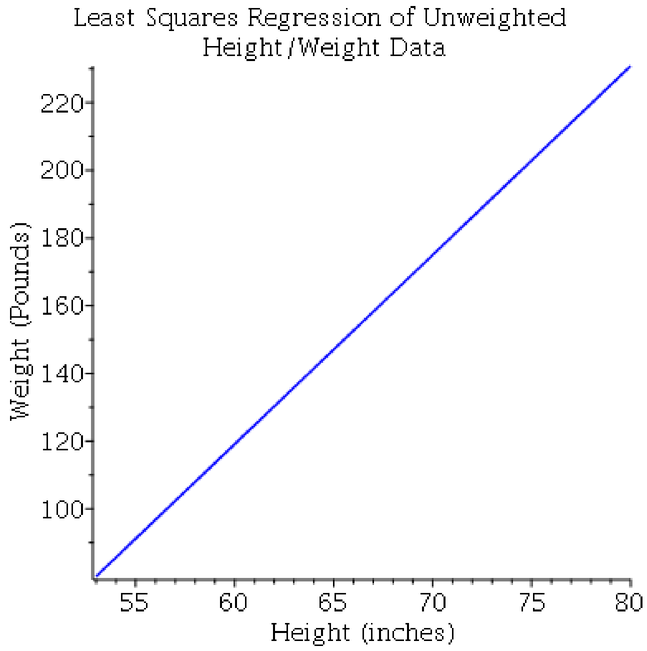
By computing we obtain:

$$a = -\frac{7200118}{33341}$$

$$b = \frac{186201}{33341}$$

Therefore, the equation of the line:

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

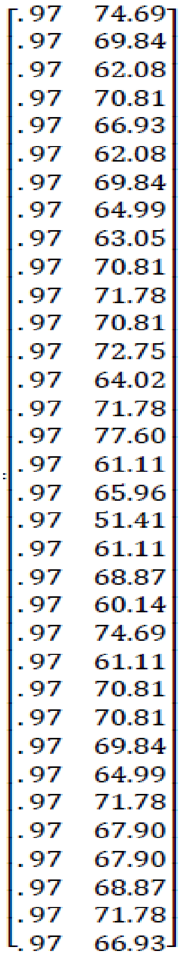


We can use the data to make predictions about the expected weight of a person based on their height. For example, an estimate of the expected weight of a person who is 5'10" (70"), can be calculated as follows:

$$\boldsymbol{y = -720011833341 + 18620133341(70) = 174.98}$$

(Note): While this line will approximate the data as given, we must consider bias as a result of certain sampling types. 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, i.e., Multiply $W=\begin{bmatrix} .97 & 0 \\ 0 & .97 \end{bmatrix}$ with Matrix A.

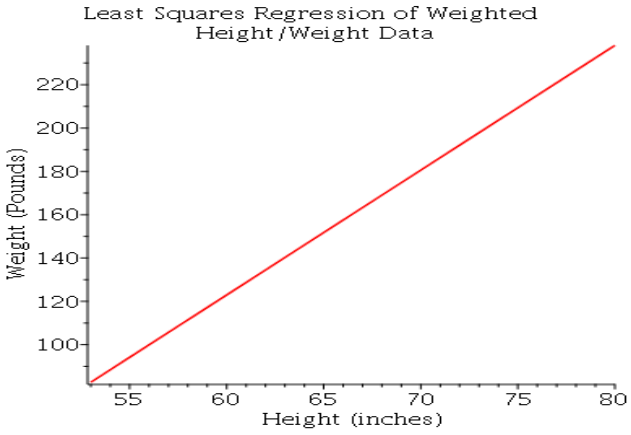
$$WA$$



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

(ii) Results and Discussions:

The new equation of line after substituting new values of **a** and **b**, **-223.63** and **5.76** respectively is **y = -223.63 + 5.76x**.

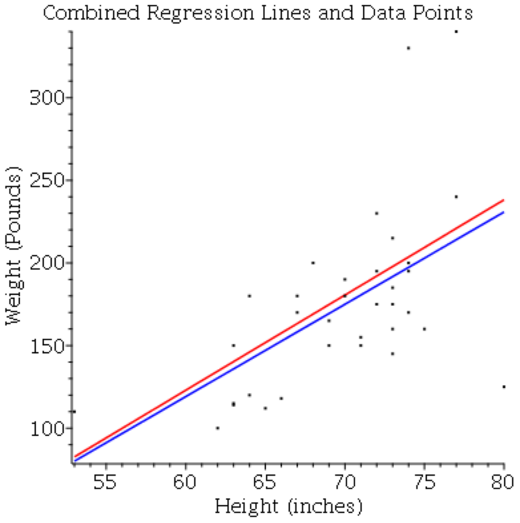


Since we are assuming each person in the data set actually weighs more than the data set states by an average of 3%, we would expect that when we calculate the expected weight for a person of any height, it would be approximately 3% greater than the weight calculated using the previous model. We can 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***

180.57/174.98 ~ 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.



Conclusion:

(i) The *Weighted Least Squares function* is

$$\Sigma\ w_i\ e_i^2 = \Sigma\ w_i(y_i\ - \alpha - \beta\ x_i\ )^2$$

(ii)

$$b\ =\ \frac{\sum w_i \sum w_i x_i y_i\ -\ \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2\ -\ (\sum w_i x_i)^2}$$

(iii)

$$a = \frac{\sum w_i y_i\ -\ b \sum w_i x_i}{\sum w_i}$$

(iv) For the given problem the equation of the line is $\mathbf{y} = -\frac{\mathbf{7200118}}{\mathbf{33341}} + \frac{\mathbf{186201}}{\mathbf{33341}}\,\mathbf{x}$

Uses of Weighted least square functions in present day scenario in mathematical and non-mathematical fields:

As health professionals are continually trying to get an accurate picture of our nation's health as a whole. 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 modelled 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.

An alternative approach/Future Enhancements

Weighted Least Squares can only be used in rare cases when we accurately know the values of weight estimates for each data point. If heteroskedasticity is a problem then it is better to use the ordinary least squares method by using a difference variance estimator, as suggested in the reference from White, Halbert (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". Econometrica. 48 (4): 817–838. doi:10.2307/1912934.

Bibliography

1. Shalizi, C. (20150. Lecture 24–25: Weighted and Generalized Least Squares. Retrieved February 20, 2018 from: http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/24/lecture-24–25.pdf

2. James R. Knaub (2012). Properties of Weighted Least Squares Regression for Cutoff Sampling in Establishment Surveys. Energy Information Administration.

3. Carroll, R.J., and Ruppert, D. (1988): Transformation and Weighting in Regression, Chapman &Hall.

4. Plackett, R. L. (1950). Some Theorems in Least Squares. Biometrika, 1/2, 149-157.

5. Least Squares Data Fitting with Applications by Hansen, Per Christian, Pereyra, Víctor, Scherer, Godela Chapter 2 & 3.

6. Introduction to Linear Regression Analysis 5th Edition by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining

7. Linear Algebra and its applications 5th Edition by David C. Lay