Contents

List of Tables

List of Figures

1 Introduction and Background

Least squares approximation is a useful method for finding an estimated solution to a system of equations that would otherwise be unsolvable for the unknown parameters. The fact that these systems cannot be solved is typically a result of there being significantly more equations than there are unknowns in the system. One of the most popular applications of least squares approximation is in linear regression analysis, a method from statistics that seeks to find a way to estimate and quantify the relationship between 2 (or more) variables. Linear regressions have a host of applications, from Economics and Machine Learning (Wooldridge, 2019), to Geology (Lay, Lay & McDonald, 2016). This paper seeks to provide a brief outline of the Mathematics of least squares approximation, as well as a tutorial of its application in linear regression analysis. We begin by providing a simple explanation of least squares approximation and linear regression analysis below. Section 2 provides a short example of the application of least squares in *simple* linear regression (linear regressions with one dependent and one independent variables). Section 3 provides a more comprehensive example - using larger samples and numerical values - while Section 4 provides a complex example that builds on the previous examples by adding additional variables to the linear regression (known as *multiple* linear regression). Section 5 then provides a brief discussion and conclusion.

1.1 Least Squares Approximation

We begin by supposing that we have some system of the form $A\mathbf{x} = \mathbf{b} = \mathbf{y}$. Here, A and $\mathbf{y}$ are known, and our goal is to find an $\mathbf{x}$ which satisfies this equation. The matrix equation can be expressed as:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{1}$$

where A is an $m \times n$ matrix, $\mathbf{x}$ is an $n \times 1$ vector, and $\mathbf{y}$ is an $m \times 1$ vector. The system of equations that this results in can be expressed as:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n &= y_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n &= y_m \end{aligned} \tag{2}$$

We now suppose that this system of equations is inconsistent. This means that there are no set of values for $x_1, x_2, ..., x_n$ (i.e. there is no vector $\mathbf{x}$) that satisfies all m of these equations above. This might occur when, for example, 1 or more of the m equations is of the form $0 * x_1 + 0 * x_2 + ... + 0 * x_n = k$, where $k \neq 0$. Clearly, there are no values of $x_1, x_2, ..., x_n$ that, when multiplied by zero, produce a non-zero constant. When equations of the form $A\mathbf{x} = \mathbf{y}$ have no solution $\mathbf{x}$, we say that the vector $\mathbf{y}$ cannot be written as a linear combination of the columns of A. Therefore, we can conclude that $\mathbf{y}$ is not in the column space of A (i.e. $\mathbf{y} \notin Col(A)$).

But what if we wanted to approximate a solution to these equations? Well, this is precisely what least squares approximation allows us to do. We define $\hat{\mathbf{y}}$ as the projection of $\mathbf{y}$ onto the column space of A, which we formally write as: $\hat{\mathbf{y}} = proj_{Col(A)}(b)$. The vector $\hat{\mathbf{y}}$ will be the approximate solution we are looking for. Since we have defined $\hat{\mathbf{y}}$ to be in the column space of A, we know that the equation $A\mathbf{x} = \hat{\mathbf{y}}$ must be consistent - i.e. there is some $\mathbf{x} \in \mathbb{R}^m$ that satisfies the equation (Lay, Lay & McDonald, 2016). Therefore, we define $\hat{\mathbf{x}}$ as the particular vector(s) such that $A\hat{\mathbf{x}} = \hat{\mathbf{y}}$.

Since $\hat{\mathbf{y}} = A\hat{\mathbf{x}}$ is in $Col(A)$ we have that $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - A\hat{\mathbf{x}}$ is orthogonal to $Col(A)$ - meaning that every column vector in A is orthogonal to $\mathbf{y} - A\hat{\mathbf{x}}$ (van den Doel, 2022). Since $\mathbf{y} - A\hat{\mathbf{x}}$ is orthogonal to $Col(A)$, it must be the case that it is orthogonal to every column vector of A. If we let $\mathbf{a}_j$ be the j-th column vector of A, then we have that the dot product of every $\mathbf{a}_j$ with $\mathbf{y} - A\hat{\mathbf{x}}$ must evaluate to 0. In other words: $\mathbf{a}_j \cdot (\mathbf{y} - A\hat{\mathbf{x}}) = 0$. This can be written as $\mathbf{a}_j^T(\mathbf{y} - A\hat{\mathbf{x}}) = 0$, where $\mathbf{a}_j^T$ is the j-th row of A^T. Applying this to every row of A^T provides the result:

$$A^T(\mathbf{y} - A\hat{\mathbf{x}}) = 0 \tag{3}$$

Simplifying this expression, we have that:

$$A^T\mathbf{y} = A^T A\hat{\mathbf{x}} \tag{4}$$

These are known as the normal equations. Since A and $\mathbf{y}$ are known, we can use the normal equations to find $\hat{\mathbf{x}}$. This is then used to find the least squares approximation to the system of equations $\hat{\mathbf{y}} = A\hat{\mathbf{x}}$. It is useful to note that the reason for name "least squares approximation" is that our approximation for $\hat{\mathbf{y}}$ is the approximation that minimizes the norm of $\mathbf{y} - \hat{\mathbf{y}}$ denoted by: $||\mathbf{y} - \hat{\mathbf{y}}||= \sqrt{(\mathbf{y} - \hat{\mathbf{y}})_1^2 + (\mathbf{y} - \hat{\mathbf{y}})_2^2 + ... + (\mathbf{y} - \hat{\mathbf{y}})_m^2}$, where $(\mathbf{y} - \hat{\mathbf{y}})_i$ is the i-th entry of the vector $(\mathbf{y} - \hat{\mathbf{y}})$ for $1 \leq i \leq m$. This can also be written as: $||\mathbf{y} - \hat{\mathbf{y}}||^2= (\mathbf{y} - \hat{\mathbf{y}})_1^2 + (\mathbf{y} - \hat{\mathbf{y}})_2^2 + ... + (\mathbf{y} - \hat{\mathbf{y}})_m^2$. This indicates that the least squares approximation $\hat{\mathbf{y}}$ is the one that *minimises* the sum of *squares* in the expression for the norm of $\mathbf{y} - \hat{\mathbf{y}}$ - thereby minimizing the norm of itself.

1.2 Linear Regression

Here we will only consider the case of simple linear regression, before extending the discussion later on. We begin by assuming that we have a set of N pairs of values for 2 variables (or quantities): an independent x variable, and a dependent y variable. We want to investigate if there is some relationship between these 2 variables. In particular, for a given pair of x and y values, we want to see if the value of y is related to the value of its x in the pair. We assume this relationship can be modelled by the simple linear equation $y_i = \beta_0 + \beta_1 x_i$, where $1 \leq i \leq N$, β_0 is some fixed constant, and β_1 is some fixed slope coefficient measuring the relationship between x and y. This notation is commonly used in Econometrics (Wooldridge, 2019). We can write these equations as:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 \\ y_2 &= \beta_0 + \beta_1 x_2 \\ &\vdots \\ y_N &= \beta_0 + \beta_1 x_N \end{aligned} \tag{5}$$

Which can also be written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{6}$$

where we would be solving for the unknowns β_0 and β_1. Upon solving these, we could plot the *regression line* $f(x) = y = \beta_0 + \beta_1 x$, which models the estimated linear relationship between the 2 variables x and y. Unfortunately, because N is typically greater than the number of unknowns (2 in this case), this will be an over-determined system - i.e. it will have more equations than unknowns. Over-determined systems are often inconsistent, and therefore it is often not possible to solve for a β_0 and β_1. As a result, least squares approximation can be used to find an estimated regression line. If we let:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} ; A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} ; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{7}$$

Then we have a system of equations of the form $\mathbf{y} = A\mathbf{x} \rightarrow \mathbf{y} = A\beta$. As a result, we can use the normal equations with modified notation: $A^T\mathbf{y} = A^T A\hat{\beta}$ to find a least squares approximation for $\hat{\beta}$. Here, we are mainly concerned with $\hat{\beta} = [\hat{\beta}_0 \ \hat{\beta}_1]^T$, since we can use these to plot our approximation of the regression line. We define this approximate regression line as:

$$\hat{f(\mathbf{x})} = \hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x} \tag{8}$$

which is just $\hat{\mathbf{y}} = A\hat{\beta}$. Here we can define the norm (which is also a measure of the error - or deviation - in the regression line) in another way. Recall that the norm of $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - A\hat{\beta}$ is given by $||\mathbf{y} - A\hat{\beta}|| = \sqrt{(\mathbf{y}_1 - (A\hat{\beta})_1)^2 + (\mathbf{y}_2 - (A\hat{\beta})_2)^2 + ... + (\mathbf{y}_N - (A\hat{\beta})_N)^2}$, where $\mathbf{y}_i$ and $(A\hat{\beta})_i$ denote the i-th entry of the vectors $\mathbf{y}$ and $(A\hat{\beta})$ respectively for $1 \leq i \leq N$. We can also define this as: the least squares error $=$ $||\mathbf{y} - A\hat{\beta}|| = \sqrt{(\mathbf{y}_1 - \hat{f(x_1)})^2 + (\mathbf{y}_2 - \hat{f(x_2)})^2 + ... + (\mathbf{y}_N - \hat{f(x_N)})^2}$, where $\hat{f(x_i)}$ is the estimated $\hat{y_i}$ value for a given x_i, the i-th entry of the vector $\mathbf{x}$ $(1 \leq i \leq N)$.

2 Simple Example: Small Sample of Integer Values

We begin the tutorial application with a simple scenario. We suppose that we have $N = 5$ pairs of observations for 2 variables x and y given by:

Example 1 Data	
x	y
5	21
11	43
7	32
9	36
12	45

Table 1: Sample of integers for simple example with N=5.

This gives rise to the following over-determined system of linear equations:

$$\begin{aligned} \beta_0 + \beta_1 * 5 &= 21 \\ \beta_0 + \beta_1 * 11 &= 43 \\ \beta_0 + \beta_1 * 7 &= 32 \\ \beta_0 + \beta_1 * 9 &= 36 \\ \beta_0 + \beta_1 * 12 &= 45 \end{aligned} \tag{9}$$

Which can be rewritten as the matrix equation $A\beta = \mathbf{y}$, where:

$$A = \begin{bmatrix} 1 & 5 \\ 1 & 11 \\ 1 & 7 \\ 1 & 9 \\ 1 & 12 \end{bmatrix} ; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} ; \mathbf{y} = \begin{bmatrix} 21 \\ 43 \\ 32 \\ 36 \\ 45 \end{bmatrix} \tag{10}$$

We can then set up the augmented matrix $[A \mid \mathbf{y}]$ and attempt to solve for β_0 and β_1:

$$\begin{bmatrix} 1 & 5 & 21 \\ 1 & 11 & 43 \\ 1 & 7 & 32 \\ 1 & 9 & 36 \\ 1 & 12 & 45 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{11}$$

Where the 3rd row indicates that the system is indeed inconsistent, as we have $\beta_0 * 0 + \beta_1 * 0 = 1$. Therefore, we cannot solve for β_0 and β_1 directly. However, we can use least squares approximation to solve for $\hat{\beta} = [\hat{\beta_0}\ \hat{\beta_1}]^T$ as well as $\hat{y}$. Since we are required to solve the normal equations $A^T\mathbf{y} = A^T A\hat{\beta}$ for $\hat{\beta}$, we know that we need: $\hat{\beta} = (A^T A)^{-1}A^T\mathbf{y}$. We can solve the expression $(A^T A)^{-1}A^T\mathbf{y}$ using our known A and $\mathbf{y}$ in *Mathematica*, but an outline of the steps are performed below:

$$A^T A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 5 & 11 & 7 & 9 & 12 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 1 & 11 \\ 1 & 7 \\ 1 & 9 \\ 1 & 12 \end{bmatrix} = \begin{bmatrix} 5 & 44 \\ 44 & 420 \end{bmatrix} \tag{12}$$

$$(A^T A)^{-1} = \frac{1}{5 * 420 - 44 * 44} \begin{bmatrix} 420 & -44 \\ -44 & 5 \end{bmatrix} = \begin{bmatrix} \frac{105}{41} & -\frac{11}{41} \\ -\frac{11}{41} & \frac{5}{164} \end{bmatrix} \tag{13}$$

$$(A^T A)^{-1}A^T\mathbf{y} = \begin{bmatrix} \frac{105}{41} & -\frac{11}{41} \\ -\frac{11}{41} & \frac{5}{164} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 5 & 11 & 7 & 9 & 12 \end{bmatrix} \begin{bmatrix} 21 \\ 43 \\ 32 \\ 36 \\ 45 \end{bmatrix} \tag{14}$$

$$\hat{\beta} = (A^T A)^{-1}A^T\mathbf{y} = \begin{bmatrix} \frac{259}{41} \\ \frac{271}{82} \end{bmatrix} \approx \begin{bmatrix} 6.32 \\ 3.30 \end{bmatrix} \tag{15}$$

Therefore, the estimated regression line $f(\hat{\mathbf{x}}) = \hat{\mathbf{y}} = \beta_0 + \beta_1\mathbf{x}$ is given as:

$$f(\hat{\mathbf{x}}) = \hat{\mathbf{y}} \approx 6.32 + 3.30\mathbf{x} \tag{16}$$

We can plot the estimated regression line in 2D space, along with the data points below in Figure 1:

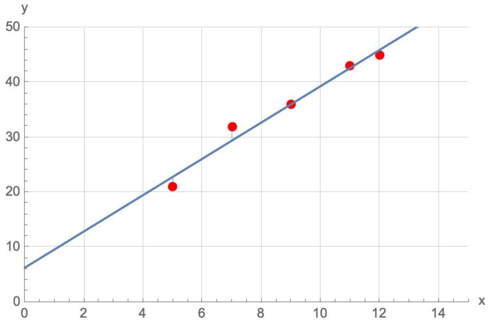

Figure 1: Plot of estimated least squares regression line for simple example where N=5.

As we can see in Figure 1 above, the line seems to fit the data quite well. However, we were lucky in this case, because the data in the small sample were not very spread out (i.e the sample contained no outliers). This is not always the case, and when working with samples as small as 5, it only takes one large outlier to drastically change the $\hat{\beta}_0$ and $\hat{\beta}_1$ parameters of the regression line - which would cause the straight line to take very different form.

Lastly, we can also calculate the norm and least squares error for this example. Since we have defined the regression equation, we can calculate the 5 individual $\hat{y}$ values:

Example 1 Calculations for $\hat{y}$		
$\widehat{f(x)}$		$\hat{y}$
$\widehat{f(5)}$	$\frac{259}{41} + \frac{271}{82}(5)$	$\frac{1873}{82}$
$\widehat{f(11)}$	$\frac{259}{41} + \frac{271}{82}(11)$	$\frac{3499}{82}$
$\widehat{f(7)}$	$\frac{259}{41} + \frac{271}{82}(7)$	$\frac{2415}{82}$
$\widehat{f(9)}$	$\frac{259}{41} + \frac{271}{82}(9)$	$\frac{2957}{82}$
$\widehat{f(12)}$	$\frac{259}{41} + \frac{271}{82}(12)$	$\frac{1885}{41}$

Table 2: Calculations of $\widehat{f(x_i)} = \hat{y}_i$ for simple example where N=5.

Therefore, the norm and error would be given by $||\mathbf{y} - \hat{\mathbf{y}}||$:

$$= \sqrt{(21 - \frac{1873}{82})^2 + (43 - \frac{3499}{82})^2 + (32 - \frac{2415}{82})^2 + (36 - \frac{2957}{82})^2 + (45 - \frac{1885}{41})^2}$$

$$= \sqrt{\frac{449}{41}} \approx 3.3093$$

3 Moderate Example: Large Sample of Numerical Values

In this example, we will work with a bigger sample while also making use of non-integer real numbers. We begin by generating a random sample of 100 values for the x variable over the interval: $0 \leq x < 100$. We then generate a non-random sample of 100 y values that has the following relationship with with the x-variable:

$$y_i = -6 + z + 4x_i \tag{17}$$

where $1 \leq i \leq 100$ and z is a random number generated over the interval $-100 \leq z < 100$. The augmented matrix $[1\ \mathbf{x} \mid \mathbf{y}]$ is formed, where $\mathbf{x}$ and $\mathbf{y}$ are vectors containing the the data for the x and y variables respectively. This is then row reduced using *Mathematica*, where we confirm that the system is indeed inconsistent. This process is omitted to save space - as the Augmented matrix has dimensions 100×3. While a complete table of the data is not displayed for the same reasons, we can show a plot of the data points in 2D space in Figure 2 below:

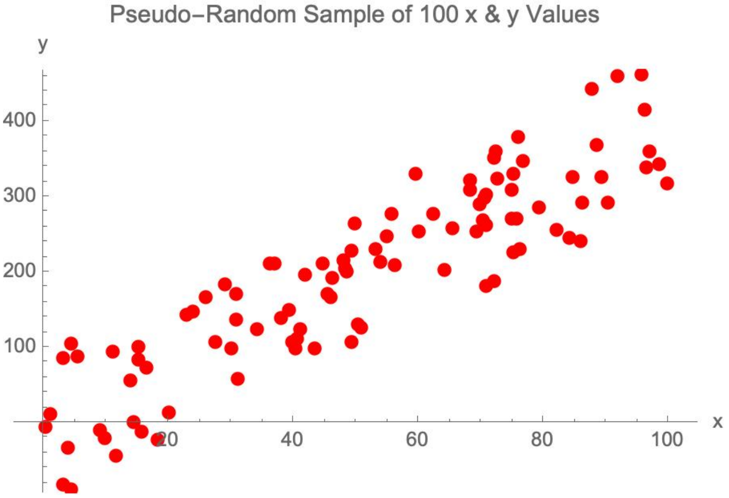

Figure 2: Plot of data for moderate example using numerical numbers where N=100

Knowing that the system is inconsistent, we can proceed to use least squares approximation to find the estimated regression line. We can begin by forming the 100×2 matrix $A = [1 \ x]$. Then, we can solve the normal equations for $\hat{\beta} = (A^TA)^{-1}A^T\mathbf{y}$, as we did in the simple example. We complete this process in *Mathematica* and find the vector $\hat{\beta}$ to be:

$$\hat{\beta} = (A^TA)^{-1}A^T\mathbf{y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \approx \begin{bmatrix} -4.42 \\ 3.97 \end{bmatrix} \tag{18}$$

This suggests that the regression equation is given by:

$$f(\hat{\mathbf{x}}) = \hat{\mathbf{y}} \approx -4.42 + 3.97\mathbf{x} \tag{19}$$

Notice that the x coefficient in the equation above is almost exactly equal to the x coefficient given when defining our sample values for y: $y_i = -6 + z + 4x_i$. This suggests that the least squares approximation has very accurately predicted part of the relationship between the x and y variable here. We can also visualise this solution in 2D space in Figure 3 below:

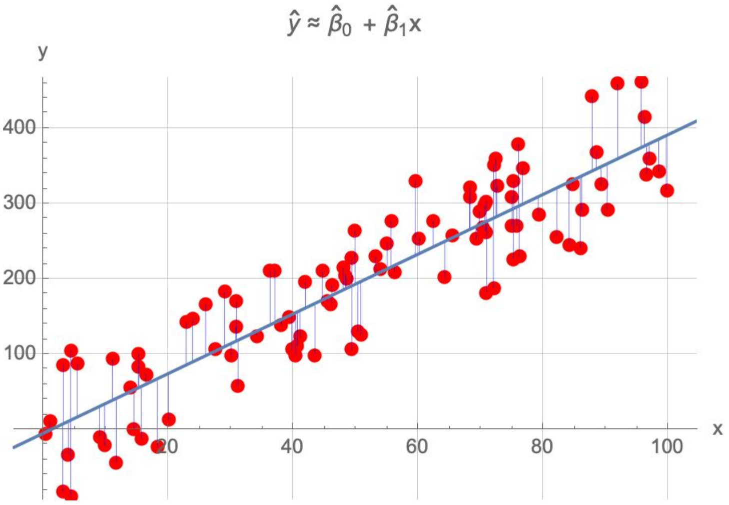

Figure 3: Plot of estimated least squares regression line for moderate example where N=100.

We can also calculate the norm and least squares error. Recalling that the norm and error are given by:

$$||\mathbf{y} - A\hat{\beta}|| = \sqrt{(\mathbf{y}_1 - (A\hat{\beta})_1)^2 + (\mathbf{y}_2 - (A\hat{\beta})_2)^2 + ... + (\mathbf{y}_{100} - (A\hat{\beta})_{100})^2}$$

$$= \sqrt{(\mathbf{y}_1 - f(\hat{x_1}))^2 + (\mathbf{y}_2 - f(\hat{x_2}))^2 + ... + (\mathbf{y}_{100} - f(\hat{x_{100}}))^2}$$

We can calculate this in *Mathematica* as 542.39. We can also interpret this geometrically. Suppose we take the distances between between the y_i (red dots) and their corresponding $\hat{y}_i$ (the point on the regression line for the same x_i value, which is given by $f(\hat{x_i})$). If we then square each of these values, sum them, and then take the square root of this sum, we will get a result of 542.39 for this semi-random sample.

4 Extra, Complex Example: 2 Independent Variables

The entire discussion above has covered linear regression with respect to 2 variables - one dependent and one independent. However, there is no reason that the analysis could not be extended to scenarios where there are 2 (or indeed more) independent variables. We provide a simple introduction to such a scenario here.

Suppose that we have 100 observations of a dependent variables y, as well as 2 independent variables x_1 and x_2. It is simple to extend the linear regression analysis to this case, as we simply suppose the the dependent variable is related to the independent variables in the following way: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$, where where the index i for y_i, $x_{1,i}$ and $x_{2,i}$ runs over $1 < i < 100$. We then have that the matrix equation $\mathbf{y} = A\beta$ simply has entries with different dimensions:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{100} \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \\ 1 & x_{1,100} & x_{2,100} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \tag{20}$$

Given the general discussion in Section 1, we know that a solution for this matrix equation can be estimated using Least Squares. Therefore, we proceed to construct an example. As we did previously, we randomly generate 100 values for the x_1 variable over the interval $0 \leq x_1 < 100$. In addition, we randomly generate 100 values for the x_2 variable over the interval $-100 \leq x_2 < 100$. With the values for our 2 independent variables generated, we proceed to construct a non-random sample of 100 values for the dependent variable y, where y is arbitrarily related to the independent variables in the following way:

$$y_i = 7 - z + 5x_{1,i} - 3x_{2,i} \tag{21}$$

where the index i for y_i, $x_{1,i}$ and $x_{2,i}$ runs over $1 \leq i \leq 100$, and z is a random number generated over the interval $-100 \leq z < 100$. We can proceed to construct the augmented matrix $[1\ \mathbf{x_1}\ \mathbf{x_2} \mid \mathbf{y}]$, where $\mathbf{x_1}$, $\mathbf{x_2}$ and $\mathbf{y}$ are vectors containing the the data for the variables x_1, x_2 and y respectively. Row reducing this in *Mathematica*, it is then confirmed that the system is inconsistent. We omit this process again to save space - since the augmented matrix has dimensions 100×4. However, we do give a visual plot of the data in Figure 4 below. Note that because we are dealing with 2 independent variables - and thus we are dealing with 3 variables in total - this will need to be plotted in 3D space:

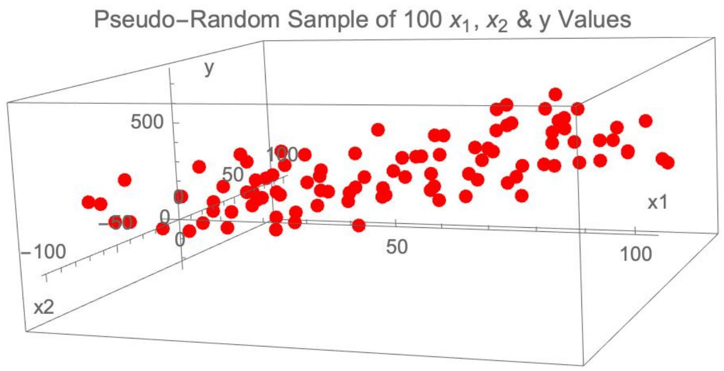

Figure 4: Plot of data for complex example using numerical numbers, 2 independent variables and with N=100.

We can now proceed to find an approximate solution to this system using least squares. We begin by forming the matrix A, with dimensions 100×3, given by: $A = [1\ \mathbf{x_1}\ \mathbf{x_2}]$. Using this to solve the normal equations for $\hat{\beta} = (A^T A)^{-1} A^T \mathbf{y}$ in *Mathematica*, we find the vector $\hat{\beta}$ to be:

$$\hat{\beta} = (A^T A)^{-1} A^T \mathbf{y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \approx \begin{bmatrix} 9.91 \\ 5.02 \\ -3.03 \end{bmatrix} \tag{22}$$

The above result indicates that the estimated regression equation is given by:

$$\widehat{f(\mathbf{x_1}, \mathbf{x_2})} = \hat{\mathbf{y}} \approx 9.91 + 5.02\mathbf{x_1} - 3.03\mathbf{x_2} \tag{23}$$

Again, we can note that that the x_1 and x_2 coefficients in the equation above are almost identical to the x_1 and x_2 coefficients given when defining the sample values for the dependent variable: $y_i = 7 - z + 5x_{1,i} - 3x_{2,i}$. Therefore, we again have evidence that least squares approximation can accurately predict the part of the relationship between the independent and dependent variables - even in the case where there are multiple independent variables. Furthermore, we can provide a visualisation of the estimated solution in 3D space in Figure 5 below:

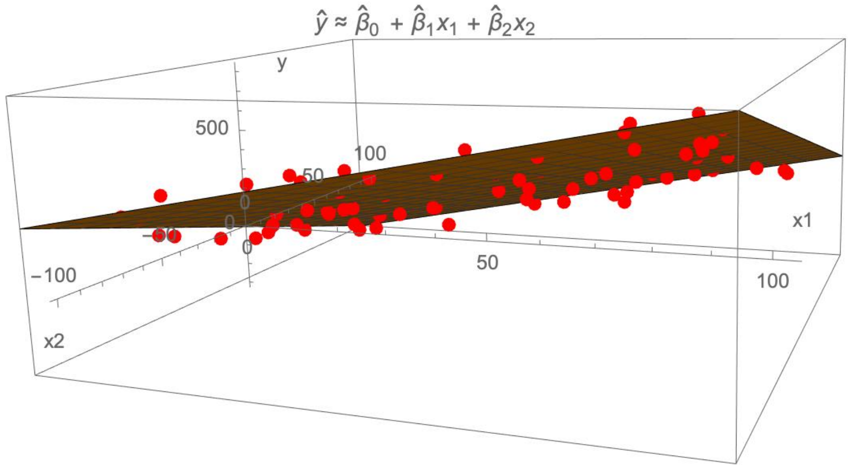

Figure 5: Plot of estimated least squares regression *plane* for complex example with 2 independent variables and N=100.

It should be noted that - because the points need to be plotted in 3D - the least squares approximation of the solution ($\hat{\mathbf{y}}$) is given by a *plane*, rather than the line we are now accustomed to in the case of a single independent variable. Furthermore, we can calculate the norm and least squares error again. The equations in this case will simply be given as:

$$||\mathbf{y} - A\hat{\beta}|| = \sqrt{(\mathbf{y}_1 - (A\hat{\beta})_1)^2 + (\mathbf{y}_2 - (A\hat{\beta})_2)^2 + ... + (\mathbf{y}_{100} - (A\hat{\beta})_{100})^2}$$

$$= \sqrt{(\mathbf{y}_1 - f(x_{1,\hat{1}}, x_{2,1}))^2 + (\mathbf{y}_2 - f(x_{1,\hat{2}}, x_{2,2}))^2 + ... + (\mathbf{y}_{100} - f(x_{1,\widehat{100}}, x_{2,100}))^2}$$

Where the only thing that has changed is that $\hat{f}$ is now a function of 2 variables: x_1 and x_2. We calculate this as 577.539 using *Mathematica*. Lastly, can also proceed to interpret this geometrically again. This value is simply the square root of the sum of the squared distances between between the y_i (red dots) and their corresponding $\hat{y}_i$ (the point on the regression *plane* for the same $x_{1,i}$ and $x_{2,i}$ values, which is given by $f(x_{1,\hat{i}}, x_{2,i})$).

5 Discussion and Conclusion

The above examples should serve as a good introduction to the role of least squares approximation in linear regression analysis. However, there are some notable drawbacks to the simple overview provided above.

In the simple example, a small sample was chosen deliberately, where there was a very clear correlation between the x and y values. However, not only is this not often the case, but it fails to point out a significant drawback of linear regression when using small sample sizes: that the approximated solution can be heavily affected by outlier values. These are values where there is a clear and significant break in the general correlation observed between the variables in the data set. The fact that least squares approximations for linear regression equations tend to be easily skewed by outliers in small samples would be an interesting topic for future research papers.

Similarly, this paper only provided a short introduction to multiple linear regression, using one additional independent variable. However, in practice, it is often desirable to use more than 2 independent, explanatory variables (Wooldridge, 2019). This short paper did not touch on how to solve these larger systems - although the underlying principles remain mostly unchanged. Furthermore, when there are more than 2 independent variables, the results becomes much harder to represent graphically. Therefore, another interesting topic for future research papers would be to discuss the above process for > 2 independent variables, as well as how these results (or a simplified version of these results) could be represented graphically.

Nevertheless, this short paper has provided a good overview of linear regression analysis using least squares approximation. All of the key aspects of the topic have been covered in varying degrees of detail, including: the definition of and reasoning behind least squares approximation, how least squares approximation forms the foundations of simple and multiple linear regression, as well as how to apply least squares approximation to 3 different examples of (simple or multiple) linear regression.

References

Lay, D., Lay, S., & McDonald, J. (2016). Linear Algebra and Its Applications, Global Edition (5th ed.). Pearson Education.

van den Doel, L. (2022). Least Squares (Ch. 6). Lecture, University College Roosevelt.

Wooldridge, J. (2019). Introductory Econometrics: A Modern Approach (7th ed.). Cengage.