Gaussian Elimination and Linear Regression

Reva Dhillon
AE21B108

August 31, 2022

1 Introduction

This report gives a general procedure to solve for a system of linear equations using Gaussian Elimination. It discusses the in-built functions $inv(A) * b$ and $A\backslash b$ to solve the same system. It also graphs:

- The L_∞ norm of error between the methods against the matrix size n.
- The wall clock time against the matrix size n for the three methods.

2 Gaussian Elimination

2.1 General procedure:

- Given a system of linear equations of the form $Ax = b$ with A of order $n \times n$, x of order $n \times 1$ and b of order $n \times 1$, first create the augmented matrix $[A|b]$.
- Now perform row transformation operations to reduce this augmented matrix to an upper triangular matrix.
- Use this reduced form of A and b to calculate x starting from x_n and going to x_1.

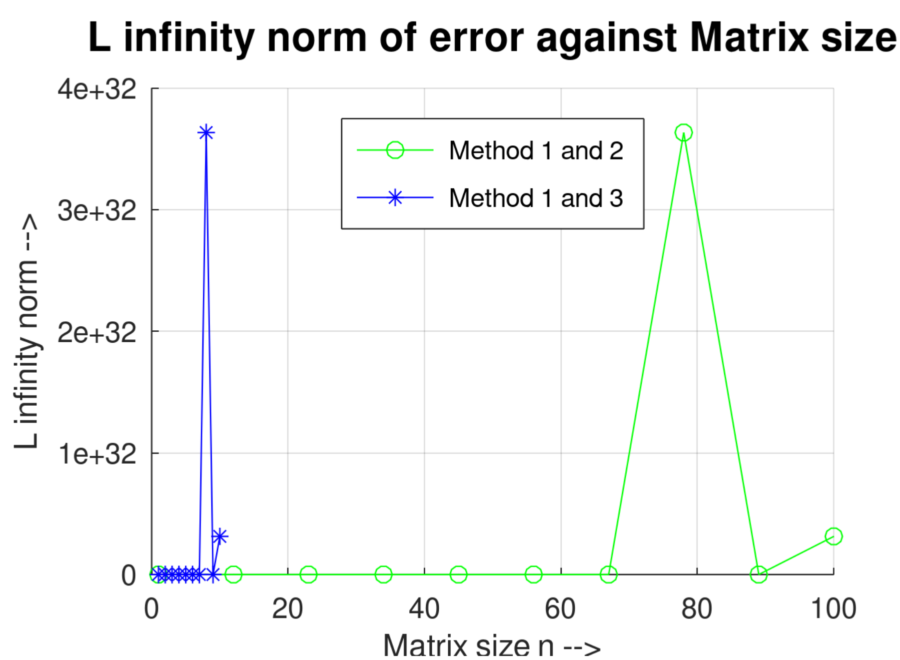

Figure 1: L_∞ norm of error against Matrix size n

2.2 L_∞ norm:

L_∞ norm gives the absolute value of that component of a vector that has the largest magnitude. Here L_∞ norm is that of error between methods 1 and 2, and methods 1 and 3 where:

- Method 1 uses the general algorithm described above the find the solutions to a system of linear equations.
- Method 2 uses the in-built function $inv(A) * b$ in octave.
- Method 3 uses the in-built function $A\backslash b$ in octave.

In figure 1, the L_∞ norm of error between methods 1 and 2, and methods 1 and 3 has been plotted.

2.3 Wall clock time and matrix size:

In figure 2, the time of operation taken by the methods 1, 2 and 3 mentioned above has been plotted against the matrix size n.
It can be concluded that the time taken by the in-built functions $inv(A) * b$ and $A\backslash b$ to solve the same system become orders of magnitude smaller as n increases.
The plot shows the result for random matrices upto an order of 100.

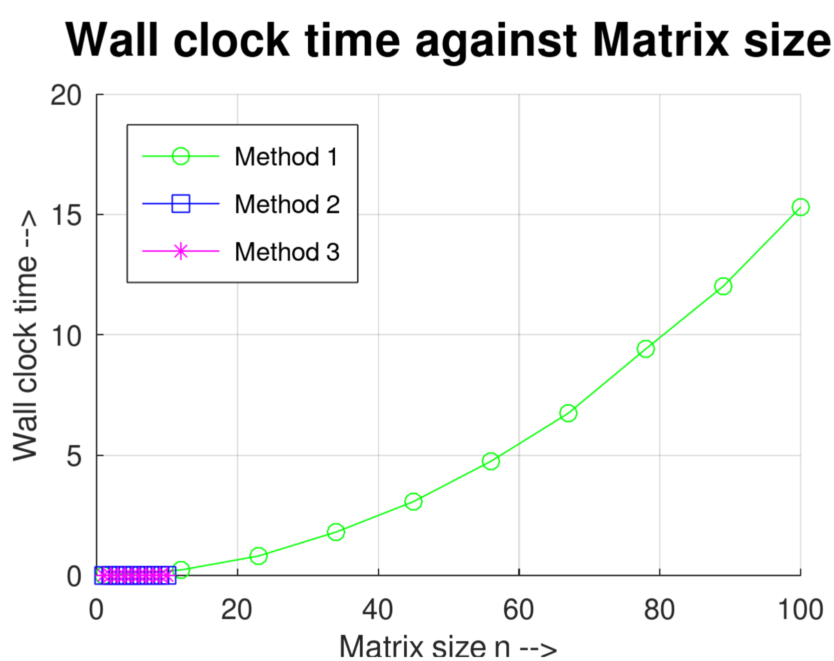

Figure 2: Wall clock time against Matrix size n

3 Regression Analysis

3.1 Regression analysis and Linear regression:

In statistical modelling, **regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.
Linear regression: It is the most common form of regression analysis in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.

3.2 Method for linear regression:

This subsection explains how to perform linear regression using matrices.

$$y = mx + c \qquad y_i = mx_i + c \tag{1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = m \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix} + \begin{bmatrix} c \\ c \\ . \\ . \\ . \\ c \end{bmatrix} \tag{2}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ . & \\ . & \\ . & \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} \tag{3}$$

$$y = Ak \tag{4}$$

$$A^T y = A^T A k \tag{5}$$

$$k = (A^T A)^{-1} A^T y \tag{6}$$

$$\begin{bmatrix} m \\ c \end{bmatrix} = (A^T A)^{-1} A^T y \tag{7}$$

The same method has been used to perform linear regression on the following three datasets.

3.3 Dataset 1:

x	y	x	y	x	y	x	y	x	y
1.0000	3.1000	11.0000	35.1000	21.0000	87.1000	31.0000	159.1000	41.0000	251.1000
2.0000	5.4000	12.0000	39.4000	22.0000	93.4000	32.0000	167.4000	42.0000	261.4000
3.0000	7.9000	13.0000	43.9000	23.0000	99.9000	33.0000	175.9000	43.0000	271.9000
4.0000	10.6000	14.0000	48.6000	24.0000	106.6000	34.0000	184.6000	44.0000	282.6000
5.0000	13.5000	15.0000	53.5000	25.0000	113.5000	35.0000	193.5000	45.0000	293.5000
6.0000	16.6000	16.0000	58.6000	26.0000	120.6000	36.0000	202.6000	46.0000	304.6000
7.0000	19.9000	17.0000	63.9000	27.0000	127.9000	37.0000	211.9000	47.0000	315.9000
8.0000	23.4000	18.0000	69.4000	28.0000	135.4000	38.0000	221.4000	48.0000	327.4000
9.0000	27.1000	19.0000	75.1000	29.0000	143.1000	39.0000	231.1000	49.0000	339.1000
10.0000	31.0000	20.0000	81.0000	30.0000	151.0000	40.0000	241.0000	50.0000	351.0000

m	c
7.1000	-43.2000

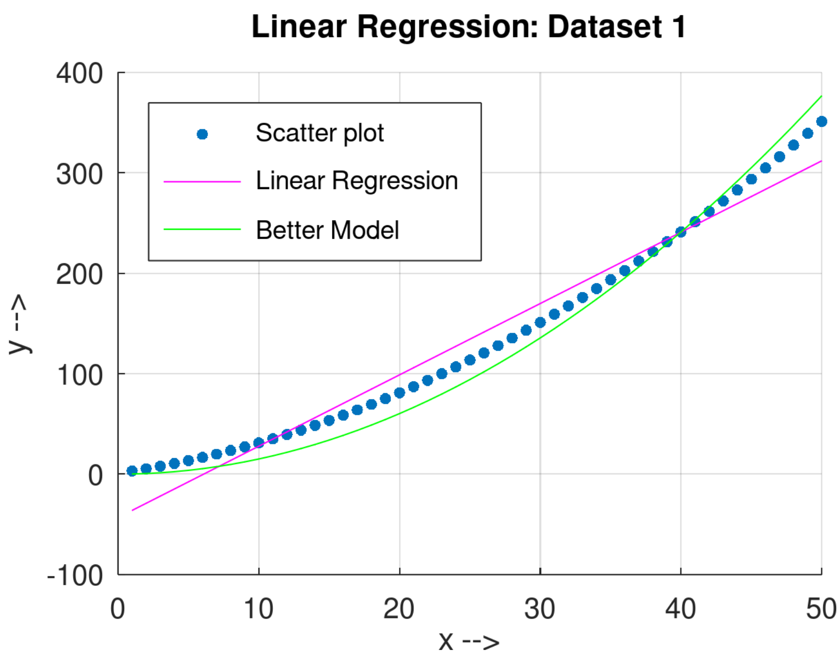

Figure 3: Linear regression on dataset 1.

Figure 3 shows the scatter plot for the dataset given above and also displays the best-fit line for it.
However, for this particular set of datapoints, a parabolic curve is a better model. This parabolic model has been shown in green in figure 3 with (40.0000, 241.0000) being a point on the curve. This is apparent as all the datapoints are closer to the curve than they are to the line.

3.4 Dataset 2:

x	y	x	y	x	y	x	y	x	y
1.0000	3.1000	21.0000	43.4583	41.0000	83.6403	61.0000	123.7810	81.0000	163.9000
2.0000	5.1414	22.0000	45.4690	42.0000	85.6481	62.0000	125.7874	82.0000	165.9055
3.0000	7.1732	23.0000	47.4796	43.0000	87.6557	63.0000	127.7937	83.0000	167.9110
4.0000	9.2000	24.0000	49.4899	44.0000	89.6633	64.0000	129.8000	84.0000	169.9165
5.0000	11.2236	25.0000	51.5000	45.0000	91.6708	65.0000	131.8062	85.0000	171.9220
6.0000	13.2449	26.0000	53.5099	46.0000	93.6782	66.0000	133.8124	86.0000	173.9274
7.0000	15.2646	27.0000	55.5196	47.0000	95.6856	67.0000	135.8185	87.0000	175.9327
8.0000	17.2828	28.0000	57.5292	48.0000	97.6928	68.0000	137.8246	88.0000	177.9381
9.0000	19.3000	29.0000	59.5385	49.0000	99.7000	69.0000	139.8307	89.0000	179.9434
10.0000	21.3162	30.0000	61.5477	50.0000	101.7071	70.0000	141.8367	90.0000	181.9487
11.0000	23.3317	31.0000	63.5568	51.0000	103.7141	71.0000	143.8426	91.0000	183.9539
12.0000	25.3464	32.0000	65.5657	52.0000	105.7211	72.0000	145.8485	92.0000	185.9592
13.0000	27.3606	33.0000	67.5745	53.0000	107.7280	73.0000	147.8544	93.0000	187.9644
14.0000	29.3742	34.0000	69.5831	54.0000	109.7348	74.0000	149.8602	94.0000	189.9695
15.0000	31.3873	35.0000	71.5916	55.0000	111.7416	75.0000	151.8660	95.0000	191.9747
16.0000	33.4000	36.0000	73.6000	56.0000	113.7483	76.0000	153.8718	96.0000	193.9798
17.0000	35.4123	37.0000	75.6083	57.0000	115.7550	77.0000	155.8775	97.0000	195.9849
18.0000	37.4243	38.0000	77.6164	58.0000	117.7616	78.0000	157.8832	98.0000	197.9899
19.0000	39.4359	39.0000	79.6245	59.0000	119.7681	79.0000	159.8888	99.0000	199.9950
20.0000	41.4472	40.0000	81.6325	60.0000	121.7746	80.0000	161.8944	100.0000	202.0000

m	c
2.0079	1.2719

Figure 4 shows the scatter plot for the dataset given above and also displays the best-fit line for it.
For this particular set of datapoints, the linear model shown is the best model.

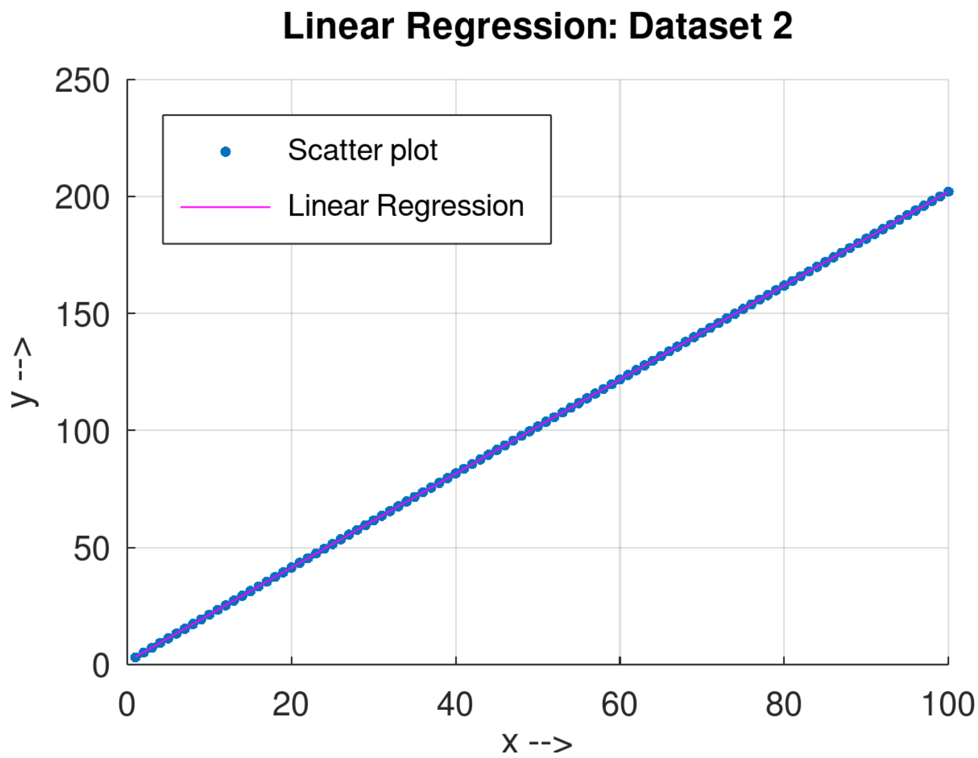

Figure 4: Linear regression on dataset 2.

3.5 Dataset 3:

x	y	x	y	x	y	x	y	x	y
1.0000	3.0038	41.0000	83.0004	81.0000	163.0040	121.0000	243.0048	161.0000	323.0065
2.0000	5.0047	42.0000	85.0098	82.0000	165.0044	122.0000	245.0001	162.0000	325.0057
3.0000	7.0082	43.0000	87.0046	83.0000	167.0015	123.0000	247.0044	163.0000	327.0081
4.0000	9.0062	44.0000	89.0074	84.0000	169.0065	124.0000	249.0041	164.0000	329.0082
5.0000	11.0029	45.0000	91.0060	85.0000	171.0083	125.0000	251.0028	165.0000	331.0039
6.0000	13.0013	46.0000	93.0022	86.0000	173.0031	126.0000	253.0021	166.0000	333.0063
7.0000	15.0031	47.0000	95.0001	87.0000	175.0090	127.0000	255.0060	167.0000	335.0041
8.0000	17.0027	48.0000	97.0050	88.0000	177.0081	128.0000	257.0040	168.0000	337.0086
9.0000	19.0085	49.0000	99.0082	89.0000	179.0025	129.0000	259.0089	169.0000	339.0015
10.0000	21.0052	50.0000	101.0022	90.0000	181.0094	130.0000	261.0067	170.0000	341.0046
11.0000	23.0091	51.0000	103.0032	91.0000	183.0035	131.0000	263.0025	171.0000	343.0025
12.0000	25.0050	52.0000	105.0086	92.0000	185.0041	132.0000	265.0096	172.0000	345.0024
13.0000	27.0030	53.0000	107.0092	93.0000	187.0026	133.0000	267.0003	173.0000	347.0039
14.0000	29.0008	54.0000	109.0067	94.0000	189.0088	134.0000	269.0030	174.0000	349.0034
15.0000	31.0068	55.0000	111.0065	95.0000	191.0070	135.0000	271.0053	175.0000	351.0072
16.0000	33.0049	56.0000	113.0058	96.0000	193.0057	136.0000	273.0016	176.0000	353.0039
17.0000	35.0026	57.0000	115.0032	97.0000	195.0002	137.0000	275.0046	177.0000	355.0081
18.0000	37.0025	58.0000	117.0064	98.0000	197.0013	138.0000	277.0068	178.0000	357.0029
19.0000	39.0037	59.0000	119.0040	99.0000	199.0001	139.0000	279.0016	179.0000	359.0004
20.0000	41.0000	60.0000	121.0071	100.0000	201.0070	140.0000	281.0089	180.0000	361.0098
21.0000	43.0100	61.0000	123.0081	101.0000	203.0074	141.0000	283.0026	181.0000	363.0089
22.0000	45.0085	62.0000	125.0033	102.0000	205.0021	142.0000	285.0091	182.0000	365.0025
23.0000	47.0055	63.0000	127.0087	103.0000	207.0031	143.0000	287.0001	183.0000	367.0039
24.0000	49.0037	64.0000	129.0074	104.0000	209.0043	144.0000	289.0023	184.0000	369.0059
25.0000	51.0053	65.0000	131.0038	105.0000	211.0032	145.0000	291.0071	185.0000	371.0022
26.0000	53.0083	66.0000	133.0064	106.0000	213.0097	146.0000	293.0096	186.0000	373.0001
27.0000	55.0093	67.0000	135.0055	107.0000	215.0070	147.0000	295.0053	187.0000	375.0072
28.0000	57.0030	68.0000	137.0025	108.0000	217.0054	148.0000	297.0040	188.0000	377.0062
29.0000	59.0067	69.0000	139.0034	109.0000	219.0084	149.0000	299.0006	189.0000	379.0093
30.0000	61.0045	70.0000	141.0019	110.0000	221.0017	150.0000	301.0046	190.0000	381.0081
31.0000	63.0019	71.0000	143.0066	111.0000	223.0001	151.0000	303.0063	191.0000	383.0057
32.0000	65.0008	72.0000	145.0072	112.0000	225.0084	152.0000	305.0080	192.0000	385.0088
33.0000	67.0057	73.0000	147.0088	113.0000	227.0076	153.0000	307.0037	193.0000	387.0011
34.0000	69.0024	74.0000	149.0073	114.0000	229.0096	154.0000	309.0066	194.0000	389.0024
35.0000	71.0041	75.0000	151.0077	115.0000	231.0058	155.0000	311.0057	195.0000	391.0004
36.0000	73.0001	76.0000	153.0085	116.0000	233.0068	156.0000	313.0085	196.0000	393.0032
37.0000	75.0081	77.0000	155.0067	117.0000	235.0029	157.0000	315.0032	197.0000	395.0066
38.0000	77.0046	78.0000	157.0050	118.0000	237.0032	158.0000	317.0077	198.0000	397.0045
39.0000	79.0088	79.0000	159.0083	119.0000	239.0001	159.0000	319.0069	199.0000	399.0047
40.0000	81.0011	80.0000	161.0013	120.0000	241.0057	160.0000	321.0079	200.0000	401.0044

m	c
2.0000	1.0050

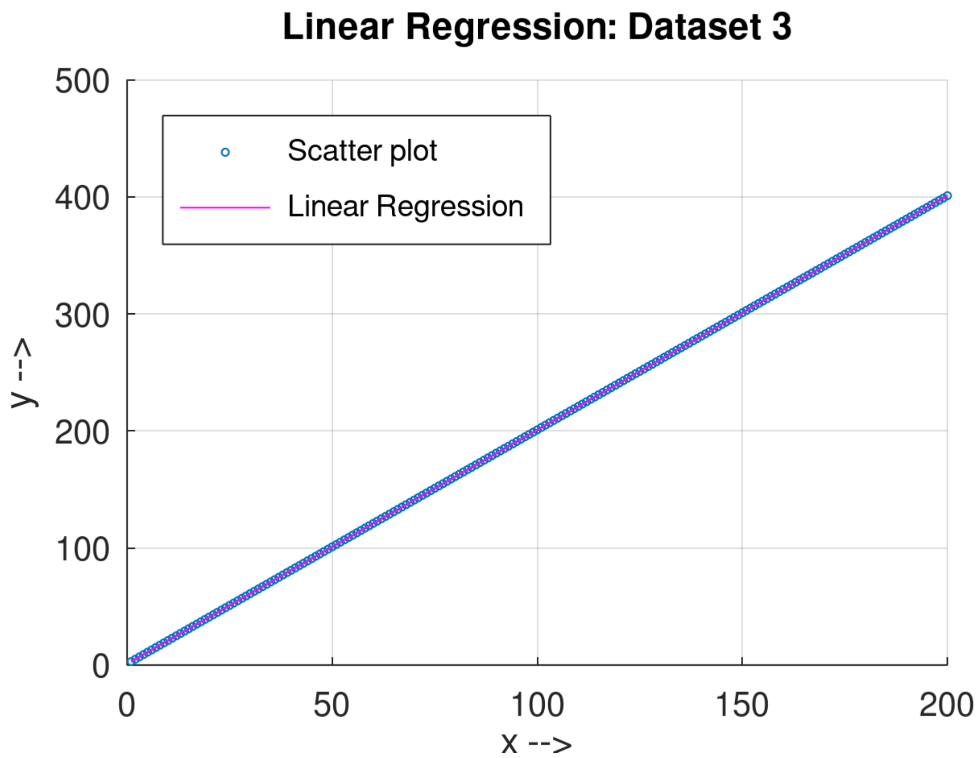

Figure 5: Linear regression on dataset 3.

Figure 5 shows the scatter plot for the dataset given above and also displays the best-fit line for it.
For this particular set of datapoints, the linear model shown is the best model.

4 Conclusion

Linear Regression is a great tool to analyze the relationships among the variables.
It is simple to implement and easier to interpret the output coefficients.
When you know that the independent and dependent variable have a linear relationship, this algorithm is the best to use because it has less complexity as compared to other algorithms.