**MATHEMATICS AA HL EXPLORATION**

Exploring different methods to graph a Lorenz curve and calculate the Gini coefficient of India to accurately measure income inequality.

May 2023

**Table of Contents**

Introduction…………………………………………………………………………………… 3

Background Information…………………………………………………………………….. 4

Data……………………………………………………………………………………………. 6

Dividing the Lorenz Curve into Trapeziums to calculate the Gini Coefficient…………. 7

Using Quadratic Regression to plot the Lorenz Curve and find the Gini Coefficient…. 8

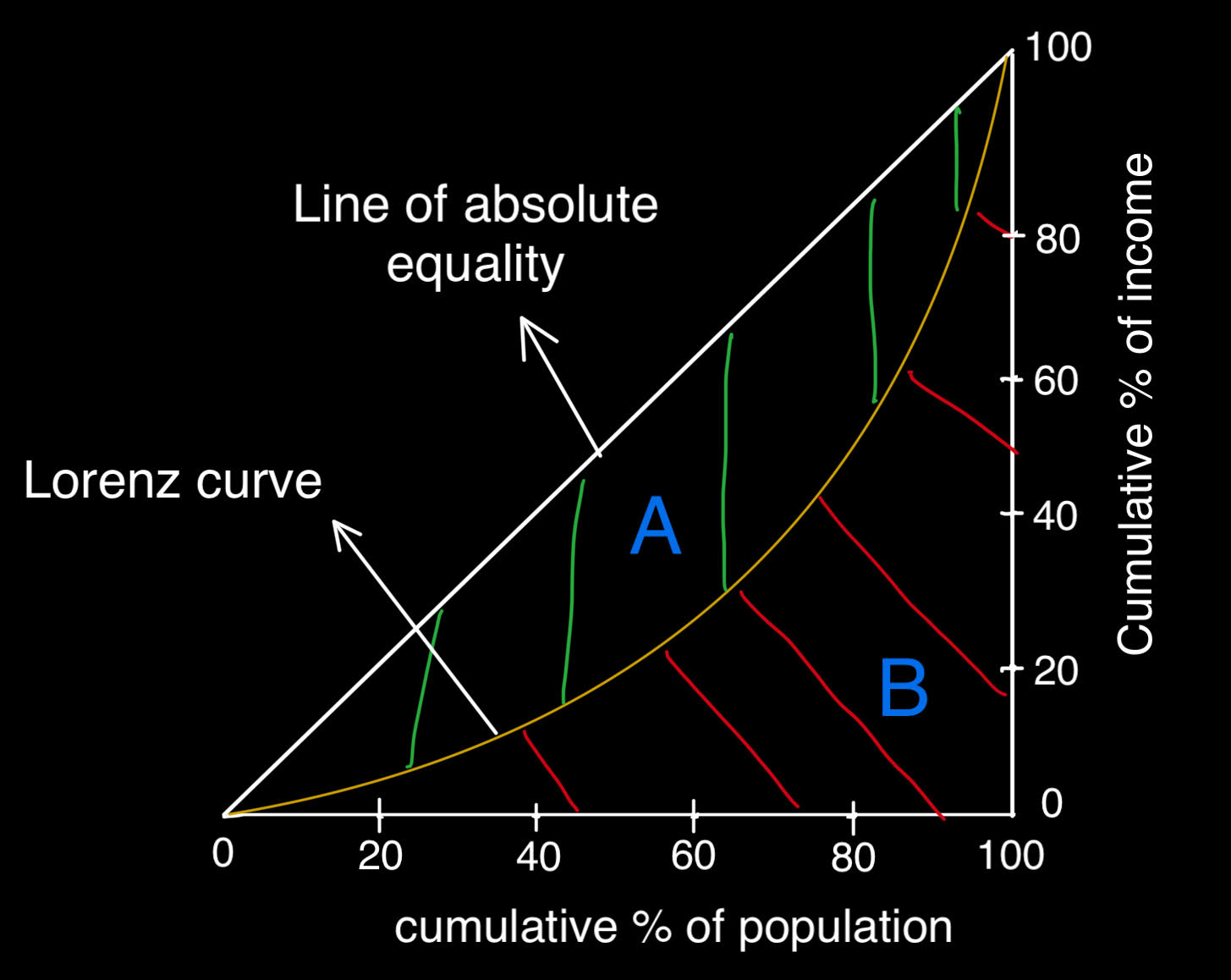
Using the Vandermonde matrix to plot the Lorenz Curve and find the Gini Coefficient 14

Evaluation……………………………………………………………………………………… 17

Strengths and limitations of the exploration………………………………………………... 18

Conclusion…………………………………………………………………………………….. 19

**Introduction**

While attending an economics class, I came across a numerical index called "Gini coefficient". The Gini coefficient has a value between from 0 to 1 (it can also be measured as a percentage by multiplying 100) and is used to measure inequality in income in a given population.[[1]](#footnote-1) The Gini coefficient is calculated by plotting 2 curves, a "Line of absolute equality curve" and a "Lorenz Curve”. It is calculated by using the following formula (according to **Figure 1**) –

**Figure 1:** Lorenz Curve

(Area of A is the area shaded in green and Area of B is the area shaded in red) At the time, I was not familiar with integration and had no idea how you could calculate the area of regions A and B. I thought you would simply divide the areas of A and B into several trapeziums and triangles to calculate the area. I also was not sure how you could find a function that accurately models the Lorenz curve. I scanned through my economics textbook right after class but could not find any answers. Once I reached home that day, I decided to find how the Lorenz curve is plotted and found that there are several "interpolation" formulas that could be used to find a graph to accurately model a Lorenz curve with a polynomial function. I decided to use the Vandermonde matrix formula to find an accurate polynomial function to model the Lorenz curve. I will also be using quadratic regression to model the Lorenz curve as the curve looks very similar to a parabola when plotted (seen in **Figure 1**). I thought it would be interesting to see how the value of the Gini coefficient changes if I use a parabola. Additionally, I wanted to test how accuracy in the Gini coefficient changes when I simply divide the area into trapeziums and when I use integration. The reason why I decided to calculate the Gini coefficient specifically for India is because the country has accurate, freely available data from the World bank[[2]](#footnote-3). In addition, it is the country where I reside, and I would like to understand where India stands compared to other countries with regards to income inequality in recent years.

**Background Information**

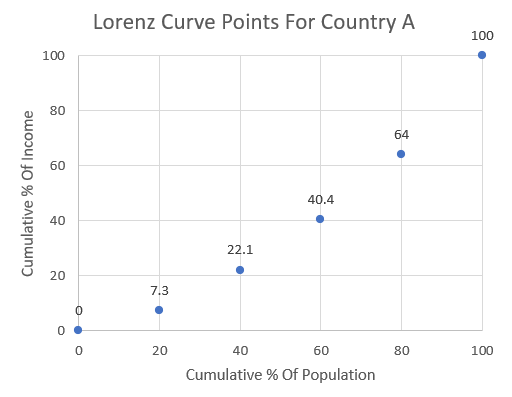
Gini Coefficient and Lorenz Curve

The Gini Coefficient is an index (ranging from 0-1 or 0-100 if taken as a percentage) used to measure the degree of economic inequality. To calculate the Gini Coefficient, we need to draw a Lorenz Curve. A Lorenz Curve is a curve that plots the cumulative percentage of population (x-axis) against the cumulative percentage of income (y-axis). Normally, its points are plotted by dividing the population into **quintiles** (20% of the population of a country), usually the poorest 20% until the richest 20% (a total of 5 quintiles). Each of the quintiles are plotted in chronological order according to the share of income they earn in an economy. For example, consider a hypothetical country, Country A, as seen in **Table 1**, the points of the Lorenz Curve would be plotted as seen in **Figure 2.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country A | **Poorest 20%** | **Second 20%** | **Third 20%** | **Fourth 20%** | **Richest 20%** |
| **Percentage of income earned** | 7.3% | 14.8% | 18.3% | 23.6% | 36.0% |
| **Table 1:** Country A, quintiles and percentage of income earned | | | | | |

Since we must plot cumulative percentages, we can calculate the points by adding the percentage of income share of the current quintile along with the value of the cumulative percent for the previous quintile (calculating the cumulative percentage). For example -

|  |  |  |
| --- | --- | --- |
| Cumulative percent of population | Percent of Income Share | Cumulative Percent of Income Share |
| 20% | 7.3% | 7.3% |
| 40% | 14.8% | 22.1% |
| 60% | 18.3% | 40.4% |
| 80% | 23.6% | 64.0% |
| 100% | 36.0% | 100.0% |
| **Table 2**: Country A cumulative percent of income share | | |



**Figure 2:** Lorenz Curve Points For Country A (Excel)

(Note when observing **Figure 2** that all Lorenz Curves contain the points (0,0) and (100,100), this stems from the fact that 0% of a population must earn 0% of a country’s total income and 100% of a population must earn 100% of a country’s income).

By using excel, we can plot the points as seen in **Figure 2**. An equation for a best fit curve can be calculated using a regression equation and for an exact equation, a polynomial interpolation formula can be used.

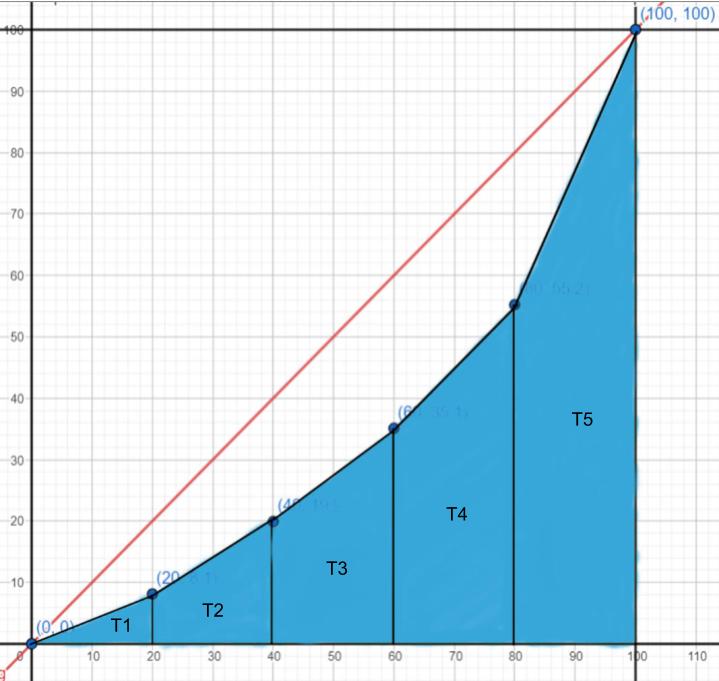
Another curve needed to calculate the Gini coefficient is the line of absolute equality, but this curve is simple to plot as it just consists of a 45° line from point (0,0) to (100,100) [[3]](#footnote-5) (Seen in **Figure 1**). This can be plotted by making a curve where y = x as seen. All y-values equal all x-values.

**Data**

Note that data was taken for the year 2019 as it is the latest year with accurate data from a reliable source. Additionally, the Gini coefficient value was included to be able to test how accurate my models are.

|  |  |  |
| --- | --- | --- |
| Cumulative percent of population | Percent of Income Share | Cumulative Percent of Income Share |
| 20% | 8.1% | 8.1% |
| 40% | 11.8% | 19.9% |
| 60% | 15.2% | 35.1% |
| 80% | 20.1% | 55.2% |
| 100% | 44.8% | 100% |
| Gini Coefficient – 35.7[[4]](#footnote-6) | | |
| **Table 3:** Data on Quintile wise earnings for India | | |

**Dividing the Lorenz Curve into Trapeziums to calculate the Gini Coefficient**



In this method, we don’t need to find a function for the Lorenz curve, all we must do is form a trapezium by joining each point with the point after and draw a vertical line touching the x-axis from each point. If we repeat this process for every point, we see that we have 5 trapeziums enclosed within the points. To calculate the area of B (according to **figure 1**) all we must do is calculate the area of each trapezium and add them (Note that T1 in **Figure 3** is a triangle since it begins at the point (0,0)). To calculate the area of A + B (**figure 1**) we can simply use the area of the triangle formula –

**Figure 3:** Gini coefficient calculation by using trapeziums (GeoGebra)

Then to calculate the Gini coefficient, we can use **formula (0)**

Comparing the value to the expected value

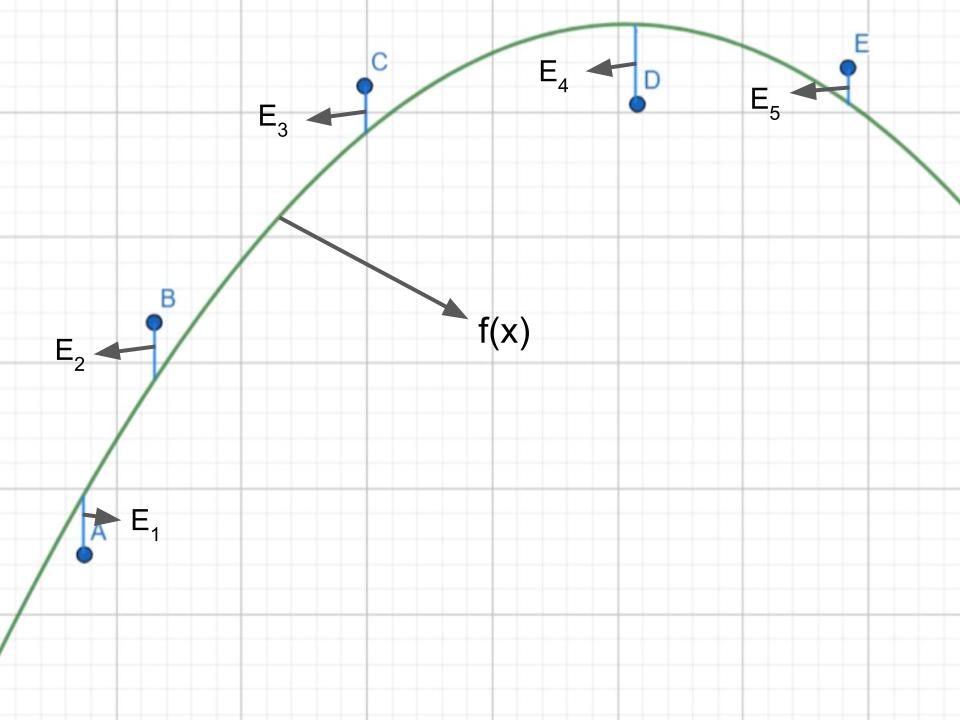
The expected value according to the world bank is 35.7, but when brought down to a scale of 0-1, the expected value is 0.357. To find the error in the value of the Gini coefficient, we can find the percentage error in the value of the Gini coefficient using the following formula -

Substituting the values into this formula –

As we can see the error is quite high. This is likely because using the trapezium method functions the Lorenz curve by dividing it into multiple (5) piecewise functions, all being linear, but in the real world the Lorenz curve is a curved line as seen in **Figure 1.** This could lead to the area of B according to **Figure 1** being much lesser than it should, hence reducing the value of the Gini coefficient.

**Using Quadratic Regression to plot the Lorenz Curve and find the Gini Coefficient**

The method I will be using to do Quadratic Regression is the least squares method.[[5]](#footnote-7) The least squares method tries to find a Quadratic for a given set of points by minimizing the value of the squares of the error. For example, let us say we have a set of points A, B, C, D, E and a quadratic function f(x) models these points (**Figure 4**).



**Figure 4:** Error in quadratic function modelling (GeoGebra)

The difference between the actual points and the points predicted by the model f(x) is the error (represented by E1, E2, E3, E4 and E5). The goal of the least squares method is to find a quadratic with the least value for –

That is, to minimize the square of the errors between the actual value and the predicted value by the model.

Let us say we have a best fit quadratic that can be modelled by the expression - . Let a point on the Lorenz curve be represented by (xi,yi). Then the predicted y-value or will be - . Then the error (squared) for this one point can be written as –

If we want to find the sum of the error (squared), given all the points can be represented by (xi,yi) and given the total number of points is represented by “n”, we can use the following formula –

Now, we can use optimization to find the minimum error. We can partially derivate the total error equation with respect to a, b and c[[6]](#footnote-8) and equate each of the expressions to 0.

Then we will get 3 simultaneous equations which we can solve to get the values of a, b and c. (Note, when we calculate the derivative of a summation, we can bring the summation outside the derivative)

Now, finding the derivative of **equation (1)** with respect to a, using power rule and chain rule –

Since we’re taking the partial derivative with respect to , and are taken to be constants.

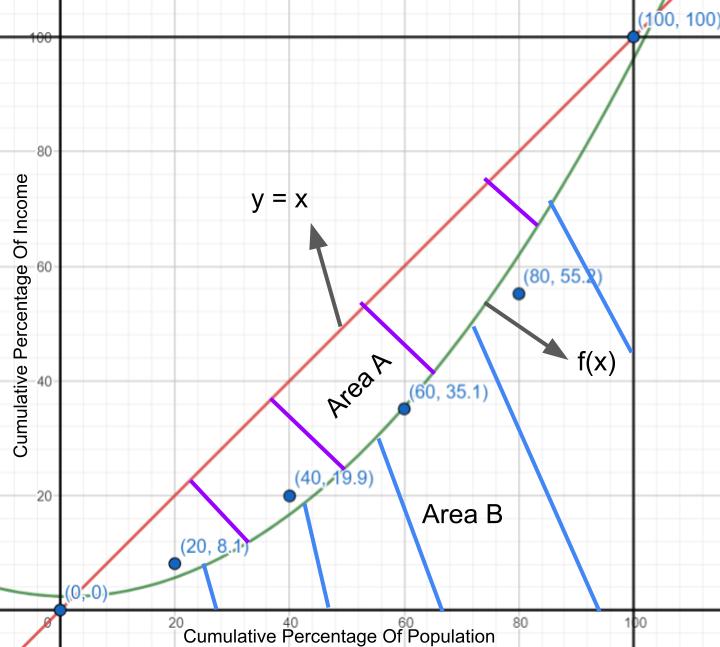
Now this must be equated to 0, so we get -

Now repeating a similar process with respect to b and c, we get the following equations.

Now we can calculate the value for and equate it to 0 to find the first equation.

Taking data points from **table 3** and substituting them into **equation (2)**. (Note that the first point of the lorenz curve is taken as (0,0))

This is the first equation of the simultaneous equations. If we perform the same calculations for equation (3) and (4), we get the following simultaneous equations –

Now solving the simultaneous equations (5), (6) and (7), we get

This means that the Lorenz curve f(x) can be modelled by the equation (**figure 5**)–

As we can see in **figure 5**, the graph is quite accurate, on further calculation we find that the R-value is 0.994 which means we have an extremely good model for the initial data points.

**Figure 5:** Quadratic regression for the Lorenz Curve (GeoGebra)

Now that we have the function for the graph, we can calculate the area of B by using definite integrals and to calculate the area of A (according to **Figure 1**) we can calculate the area of the triangle bounded by the x-axis and y = x until the x value of 100 (or the y value of 100 since y=x). Additionally, we can substitute the value of Area A + Area B from **value (0.1)**

Now substituting these values into **formula (0)**, we get the following value for the Gini Coefficient

Comparing the value to the expected value

To see how accurate quadratics are at calculating the Gini coefficient, we can find the magnitude of the percentage error according to **formula (0.2)**

This means that there is an error of in the value calculated through a quadratic compared to the actual value. This is much better than dividing the Lorenz curve into trapeziums as it results in a curved line. Though the R value of the model is extremely high, it’s possible error still exists because the quadratic does not pass through all the points plotted. Additionally, the curve has a y-intercept higher than 0 (though Lorenz curve theoretically starts from point (0,0)) and has a y-value lesser than 100 when x = 100 (contrary to the theoretical Lorenz curve). This could have affected the area covered by region B, hence affecting the final Gini coefficient value. The advantage of this method is that is much quicker than using polynomial regression but still is a good approximation for the actual Lorenz curve.

**Using the Vandermonde matrix to plot the Lorenz Curve and find the Gini Coefficient**

Though the quadratic function was an extremely good estimate for the Lorenz curve (with a R value of 0.994), it is possible to get an even better function for the Lorenz curve by interpolating the data using a polynomial function. The polynomial interpolation method I will be using is the Vandermonde matrix. [[7]](#footnote-9)

Given a function that models the points (xi,yi) on the Lorenz curve, we know that the the total number of points on the Lorenz curve is 6 (including (0,0)). When using the Vandermonde matrix to interpolate the function with a polynomial, we use a polynomial with 1 degree lesser than the total number of nodes. So, the polynomial degree to interpolate the set of points is 5 (representing this with n).

Given a polynomial of n points where “a” represents the set of coefficients, the polynomial can be represented in the following manner.

Since we have only 6 points, can be represented as a polynomial of degree 5.

Suppose i = 0 and the point is (xo,yo), additionally, since P(xo) must equal yo, we get the following expression

If this same pattern continues till i = 5, that is, point (x1,y1) till (x5,y5) we get the following polynomials.

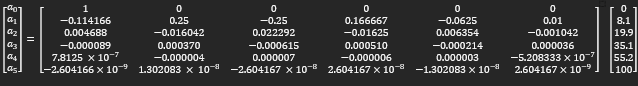
Until

This set of polynomials can be represented in matrix form to find the value of the coefficients of each term. Since there are 6 unknown constants and we are using a polynomial of degree 5, the matrix will also consist of 6 rows.

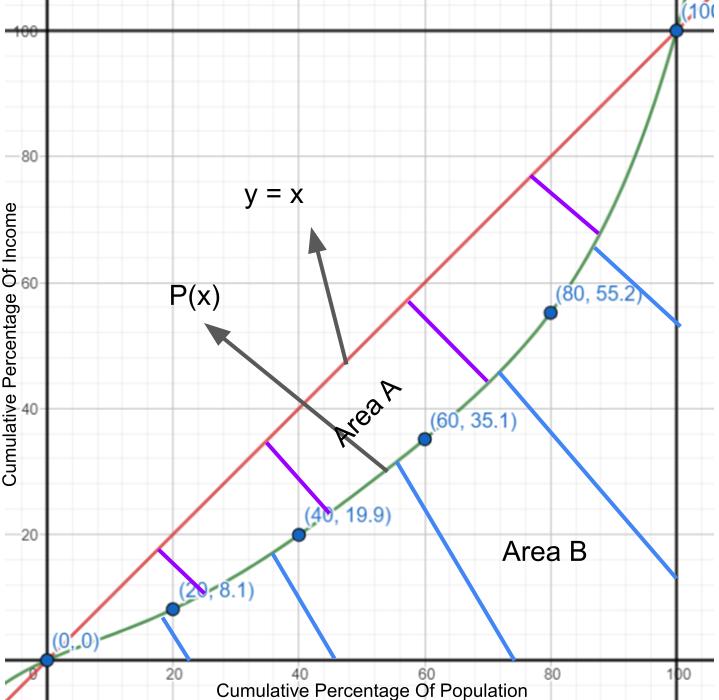
Over here, all the values are known except the values of the constants, a0, a1, … a4, a5.

Now substituting the appropriate values into the matrix, we get the following resultant matrices.

Now to find the constant values, we can take the inverse of the first matrix on both sides.



Using matrix multiplication, we get the following answers for each of the coefficients.

This means the Lorenz curve can be plotted by the polynomial (according to **Formula (8)**) –

(see **figure 6**).

This function is an excellent approximation for the points as it has an R value of 1 and passes through every single point exactly.

In **formula (0)**, we already have one of the variables, Area A + Area B (from **value (0.1)**). Area of B must be calculated by taking a definite integral from x = 0 to x = 100.

**Figure 6:** Plotting the Lorenz Curve using the Vandermonde matrix (GeoGebra)

This is equal to the following –

Now according to formula (1) the gini coefficient is equal to –

Comparing the value to the expected value

Though the polynomial function approximation of the Lorenz curve passes through all the appropriate points, we can observe that it still slightly deviates from the expected value. Substituting the calculated value and the expected value into equation (0.2), we get the percentage error as follows –

We see that the error is lower than the quadratic model. This is likely because the polynomial function passes through all the quintile points of the Lorenz curve. It is possible that error persists because rather than quintiles, the world bank could have taken ranges of 10%, leading to more points being plotted on the Lorenz curve and hence a more accurate Lorenz curve. Though the method gives rise to slight errors, its advantage is that the Lorenz curve plotted passes through every single point, hence providing an exact approximation for the Lorenz curve, even if not for the Gini coefficient.

**Evaluation**

Which model is the most effective?

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | Vandermonde Matrix | Quadratic Regression | Triangles and Trapeziums |
| **Error** | 2.88% | 5.93% | 9.17% |
| **Value** | 0.347 | 0.337 | 0.327 |
| **Table 4:** Percentage error and Gini Coefficient value for all methods | | | |

As we can see all 3 methods give values that are very similar, but the method that is closest to the value calculated by the world bank is the Vandermonde matrix. This is expected as it is the only model that is curved and passes through all points of the Lorenz curve (that is, it is the most similar to an actual Lorenz curve). Since the Vandermonde matrix most effectively models the Lorenz curve, it is also able to calculate the Gini coefficient most effectively, as the value of the coefficient largely depends on the equation of the curve. This means that if we want to calculate the value of the Gini coefficient, it is best to use polynomial interpolation.

**Strengths and limitations of the exploration -**

Strengths of the exploration **–**

* Considers multiple methods to calculate the Gini coefficient to calculate the most accurate method.
* Considers percentage error to exactly see how much the calculated value deviates from the predicted value.
* Considers reasons as to why the calculated value might deviate from the predicted value.

Weaknesses of the exploration **–**

* Does not consider how the World Bank calculates their value, this could help find a more accurate value for the Gini coefficient.
* Does not consider multiple countries.

**Conclusion**

The aim of the exploration was to calculate the Gini coefficient to understand where India stands in terms of income inequality. To carry out my exploration, I expanded my knowledge on differentiation and matrices by learning partial differentiation, matrix multiplication and determinant calculation. Based on the points of the Lorenz curve (using data from the world bank), I investigated multiple methods to calculate the Gini coefficient. This included understanding why a method was more accurate than another and how to accurately make calculations and derivations from each method.

Additionally, I learnt how to use visualization tools such as GeoGebra and Excel to help carry out my exploration more effectively and communicate my ideas through graphs throughout the exploration.

Future Scope

If possible, it would be interesting to understand exactly how world bank calculates their Gini Coefficient value and compare that method to other interpolation methods such as Lagrange’s interpolation formula. Additionally, instead of using quintiles, we could divide the population into sets of 10%, hence find a better approximation for the Gini Coefficient value.

**Bibliography**

1. Adam, Hayes. “Gini Index Explained and Gini Co-Efficients around the World.” Investopedia. Investopedia, Accessed January 13, 2023. https://www.investopedia.com/terms/g/gini-index.asp.

Ellie Tragakes, "Economics of Inequality and Poverty", Economics for the IB Diploma, 354, 355

1. Nick Higham. Accessed January 26, 2023. <https://nhigham.com/2021/06/15/what-is-a-vandermonde-matrix/>
2. Paul. “Partial Derivatives.” Paul's Notes. Accessed January 30, 2023. https://tutorial.math.lamar.edu/classes/calciii/partialderivatives.aspx.
3. “Gini Coefficient Explained: How to Calculate the Gini Ratio.” MasterClass. Accessed January 29, 2023. https://www.masterclass.com/articles/gini-coefficient.
4. Gini Index - India. World Bank, Accessed January 22, 2023. https://data.worldbank.org/indicator/SI.POV.GINI?locations=IN&most\_recent\_value\_desc=false .
5. Lowest 20%, India Income Distribution. World Bank, Accessed January 22, 2023. https://data.worldbank.org/indicator/SI.DST.FRST.20?locations=IN&most\_recent\_value\_desc=false
6. Second 20%, India Income Distribution. World Bank, Accessed January 22, 2023. https://data.worldbank.org/indicator/SI.DST.02ND.20?locations=IN&most\_recent\_value\_desc=false
7. Third 20%, India Income Distribution. World Bank, Accessed January 22, 2023. https://data.worldbank.org/indicator/SI.DST.03RD.20?locations=IN&most\_recent\_value\_desc=false
8. Fourth 20%, India Income Distribution. World Bank, Accessed January 22, 2023. https://data.worldbank.org/indicator/SI.DST.04TH.20?locations=IN&most\_recent\_value\_desc=false
9. Richest 20%, India Income Distribution. World Bank, Accessed January 22, 2023. https://data.worldbank.org/indicator/SI.DST.05TH.20?locations=IN&most\_recent\_value\_desc=false
10. "Least-Squares Regression: Definition, Equations & Examples." Study.com. August 1, 2016. Accessed, January 27 2023, https://study.com/academy/lesson/least-squares-regression-definition-equations-examples.html.

1. Hayes, Adam. “Gini Index Explained and Gini Co-Efficients around the World.” Investopedia. Investopedia, Accessed January 13, 2023. https://www.investopedia.com/terms/g/gini-index.asp. [↑](#footnote-ref-1)
2. Gini Index - India. World Bank, Accessed January 22, 2023. https://data.worldbank.org/indicator/SI.POV.GINI?locations=IN&most\_recent\_value\_desc=false . [↑](#footnote-ref-3)
3. Hayes, Adam. “Gini Index Explained and Gini Co-Efficients around the World.” Investopedia. Investopedia, Accessed January 13, 2023. https://www.investopedia.com/terms/g/gini-index.asp. [↑](#footnote-ref-5)
4. World Bank. (2019). Gini Index - India. Accessed January 22, 2023, from https://data.worldbank.org/indicator/SI.POV.GINI?locations=IN&most\_recent\_value\_desc=false [↑](#footnote-ref-6)
5. "Least-Squares Regression: Definition, Equations & Examples." Study.com. August 1, 2016. Accessed, January 27 2023, https://study.com/academy/lesson/least-squares-regression-definition-equations-examples.html. [↑](#footnote-ref-7)
6. Paul. “Partial Derivatives.” Paul's Notes. Accessed January 30, 2023. https://tutorial.math.lamar.edu/classes/calciii/partialderivatives.aspx. [↑](#footnote-ref-8)
7. Higham, Nick. Accessed January 26, 2023. https://nhigham.com/2021/06/15/what-is-a-vandermonde-matrix/. [↑](#footnote-ref-9)