# On segregation rate of Brazilian districts and their relationship with student performance

Arkaprovo Das, Anuvab De, Arnav Kanaujia, Siddharto Das

April 2, 2025

Indian Statistical Institute, Bangalore

# Introduction

# Introduction

1. This is an attempt to replicate and extend analysis done on segregation rate of different districts in Brazil by Françozo et al.
2. The main goal is to study how segregation in Brazil has affected the quality of education of students with special needs.
3. We have decided to use the ENEM dataset for 2008-2023.
4. To do this, we have created a linear regression model over time.
5. We have also considered techniques to deal with heteroskedastic data.

# Introduction

**The index of dissimilarity (IoD),**

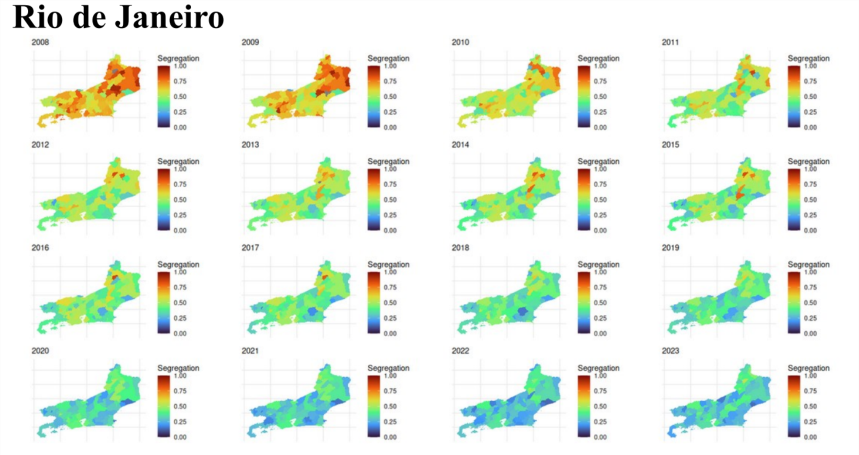
It is used to quantify the segregation between two populations in n spatial units, referred to as tracts. The IoD varies in the closed range [0,1] A sample of the IoD calculation is as follows:

$$D_g = \sum_{i=1}^{n} \frac{|X_i - Y_i|}{2}$$

With X and Y representing the proportion of the two populations being analysed . The value of $D_g$ varies between 0 and 1 and represents the proportion of a group (1 or 2) that would need to move in order to create a uniform distribution of the population.

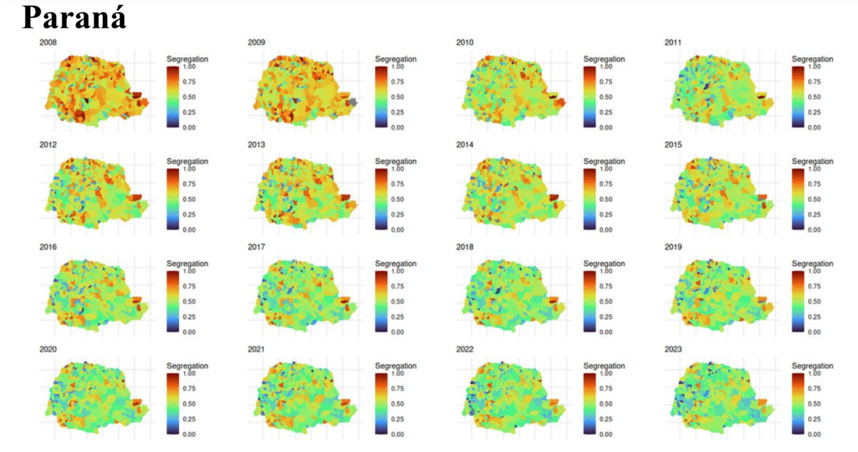
# Schematic Maps

# Maps



Rio de Janeiro
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Segregation
1.00
0.75
0.50
0.25
0.00

# Maps


Paraná
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Segregation
1.00
0.75
0.50
0.25
0.00

# Model

# Model

$$A_i = M(t) + \epsilon_i$$
$$B_i = M(t) - (C\nu(t) + D) + \epsilon_i$$

- $A_i$ is the marks obtained by the $i^t h$-student without special education needs categorized by district and year of data collection.
- M(t) is the score that a student would score if there is no irregularities present
- $\epsilon_i$ is any kind of irregularities caused because of factors irregularities present
- $B_i$ is the marks obtained by $i^{th}$ student with Special Education Needs categorized by district and year of data collection.

# Model

- $C\nu(t) + D$ represents the disadvantage faced by people with disabilities due to Segregation, where $\nu(t)$ denotes segregation rate with respect to year t and, C and D are constants.
- D is the disadvantage that people with special educational needs have due to their disability.

**Linear Regression**

Our model: $f(\nu(t_i)) = C\nu(t_i) + D + \epsilon_i$.
Here, $f(\nu(t_i))$ represents the disadvantage due to segregation.

# Assumptions

# Assumptions

- We assume that the distributions of $\epsilon_i$ are independent normal with expectation zero.
- The main factor for the discrepancy in the marks of people with Special Educational Needs (SEN) is segregation rate.
- In a given district for a particular year the marks of the students with SEN come from one distribution and the marks of the students without SEN come from one distribution.
- We assume disadvantage due to segregation is uncorrelated between different years.

# Computations for the model

# Computations for the model

Let $\hat{f}(\nu(t_i))$ be an estimator of $C\nu(t_i) + D$ in the year $t_i$, and let it be normally distributed.
Let $\hat{\sigma}_i^2$ be a point estimate for the variance of the sampling distribution of $\hat{f}(\nu(t_i))$.

$$\text{Let } Y = \begin{bmatrix} \hat{f}(\nu(t_1)) \\ \hat{f}(\nu(t_2)) \\ \vdots \\ \hat{f}(\nu(t_k)) \end{bmatrix} \implies \mathbb{E}[Y] = \begin{bmatrix} \nu(t_1) & 1 \\ \nu(t_2) & 1 \\ \vdots & \vdots \\ \nu(t_k) & 1 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix}$$

## Computations for the model

### Notation

$$\mathbb{D}(Y) = \begin{bmatrix} {\sigma_1}^2 & & & \\ & {\sigma_2}^2 & & \\ & & \ddots & \\ & & & {\sigma_k}^2 \end{bmatrix} \approx \begin{bmatrix} \hat{\sigma}_1^2 & & & \\ & \hat{\sigma}_2^2 & & \\ & & \ddots & \\ & & & \hat{\sigma}_k^2 \end{bmatrix} = G$$

$$X = \begin{bmatrix} \nu(t_1)) & \nu(t_2)) & \dots & \nu(t_k) \\ 1 & 1 & \dots & 1 \end{bmatrix}^T \qquad \beta = \begin{bmatrix} C & D \end{bmatrix}^T$$

$$U = G^{-1/2}X$$

$$Z = G^{-1/2}Y$$

# Computations for the model

$$\mathbb{E}[Z] = G^{-1/2}\mathbb{E}[Y] = G^{-1/2}X\beta = U\beta \implies \mathbb{E}[Z] = U\beta$$

$$\mathbb{D}(Z) = G^{-1/2}\mathbb{D}(Y)(G^{-1/2})^T = G^{-1/2}G(G^{-1/2})^T = I$$
$$\implies \mathbb{D}(Z) = I$$

If $U^T U\hat{\beta} = U^T Z$ ,
then $||Z - U\beta_1||^2 \geq ||Z - U\hat{\beta}||^2 \quad \forall\beta_1$
So, let $\hat{\beta} = (U^T U)^{-1}U^T Z$
then, $\mathbb{E}[\hat{\beta}] = (U^T U)^{-1}U^T\mathbb{E}[Z] = (U^T U)^{-1}U^T U\beta = \beta$
and, $\mathbb{D}(\hat{\beta}) = (U^T U)^{-1}U^T\mathbb{D}(Z)U((U^T U)^{-1})^T$
$\qquad = (U^T U)^{-1}U^T U((U^T U)^T)^{-1} = (U^T U)^{-1}$

so that, $\hat{\beta}$ is a least square estimate for $\beta$

## Computations for the model

**Proof**

If $U^T U$ is invertible then let $\hat{\beta} = (U^T U)^{-1} U^T Z$

$(Z - U\beta)^T (Z - U\beta)$

$= [Z - U\hat{\beta} + U(\hat{\beta} - \beta)]^T [Z - U\hat{\beta} + U(\hat{\beta} - \beta)]$

$= (Z - U\hat{\beta})^T (Z - U\hat{\beta}) + (\hat{\beta} - \beta)^T U^T U (\hat{\beta} - \beta)$

$\geq (Z - U\hat{\beta})^T (Z - U\hat{\beta})$

This shows the minimum of $(Z - U\beta)^T (Z - U\beta)$ is $(Z - U\hat{\beta})^T (Z - U\hat{\beta})$ and is attained at $\beta = \hat{\beta}$

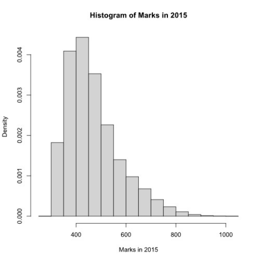
# Sources of Data

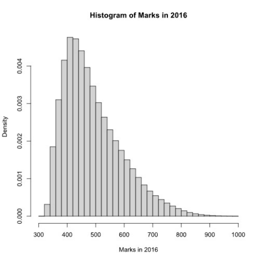
# Sources of Data

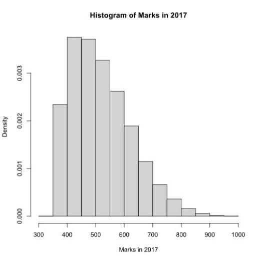
The dataset provided shows the segregation rate between students with and without disabilities in Brazilian cities between 2008 and 2023. Data are published on mendeley.com by Rafael Françozo source: https://data.mendeley.com/datasets/hxx5gfkhms/2

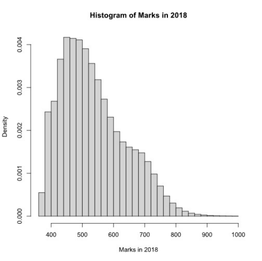
The data of marks obtained by students with and without disabilities are extracted from the ENEM data provided by the Brazilian government. The Exame Nacional do Ensino Médio (ENEM), or National High School Exam, is a standardized Brazilian national exam that assesses high school students in Brazil.

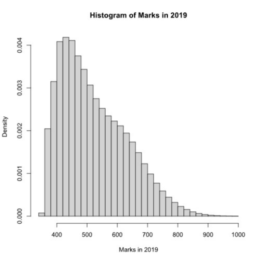
# Sources of data



# Computation with Data

# Maximum Likelihood Estimate (MLE)

Maximum Likelihood Estimation (MLE) is a method used to estimate the parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ of a probability distribution by maximizing the likelihood function. Given an independent and identically distributed (i.i.d.) sample $X = \{x_1, x_2, \ldots, x_n\}$, the likelihood function is:

$$L(\theta) = \prod_{i=1}^{n} P(x_i|\theta_1, \theta_2, \ldots, \theta_k)$$

We normally take the log of the maximum likelihood function to ease computations:

$$\ell(\theta) = \sum_{i=1}^{n} \log P(x_i|\theta_1, \theta_2, \ldots, \theta_k)$$

# Maximum Likelihood Estimate (MLE)

Now we can use any numerical method to maximize this log likelihood function. Usually we take partial derivatives and setting them to zero. In our data we have a truncated normal so we use a MLE to estimate $\mu$ and $\sigma$.

The distribution from which the marks of the students with SEN and those without SEN coming from a particular year and district is assumed to be normal.
However, in this case, we have a truncated sample.
Thus, the mean of the true normal is estimated using the MLE $\hat{\mu}_M$.

# Maximum Likelihood Estimate (MLE)

**Likelihood Function:**

$$L(\mu, \sigma) = \prod_{i=1}^{N} \frac{\varphi_{\mu,\sigma}(x_i)}{1 - \Phi_{\mu,\sigma}(a)}$$

**Log Likelihood Function:**

$$\ell(\mu, \sigma) = \sum_{i=1}^{N} log(\varphi_{\mu,\sigma}(x_i)) - Nlog(1 - \Phi_{\mu,\sigma}(a))$$

We have the MLE's $\hat{\mu}_M$,$\hat{\sigma}_M$ such that L($\mu$,$\sigma$) (or equivalently $\ell$) is maximized at $\mu = \hat{\mu}$, $\sigma = \hat{\sigma}$

# Maximum Likelihood Estimate (MLE)

```r
# data in reduced
M <- mean(reduced$NU_NOTA_MT)
S <- sd(reduced$NU_NOTA_MT)
ui <- c(0, 1)
dim(ui) <- c(1, 2)
ci <- c(0)
llike = function (dat, k) {
  function (ms) {
    m <- ms[1]
    s <- ms[2]
    -sum(log(dnorm(dat, m, s)))+length(dat)*log(1-pnorm(k, m, s))
  }
}
abled <- by(reduced, reduced$NO_MUNICIPIO_PROVA, function (dat) {
  constrOptim(c(M, S), llike(dat[dat$D_UNION == 0, 'NU_NOTA_MT'], a),
              NULL, ui, ci)
})
abled <- sapply(abled, `[[`, 'par')
disabled <- by(reduced, reduced$NO_MUNICIPIO_PROVA, function (dat) {
  constrOptim(c(M, S), llike(dat[dat$D_UNION == 1, 'NU_NOTA_MT'], a),
              NULL, ui, ci)
})
disabled <- sapply(disabled, `[[`, 'par')
tot <- as.data.frame(cbind(t(abled), t(disabled)))
```

**Figure 1:** Code for MLE

# Maximum Likelihood Estimate (MLE)

To estimate the variance of $\hat{\sigma}_M$ we see the fisher information matrix. Here we express the observed Fisher information matrix of unknown variables $\mu$ and $\sigma$ in the following form:

$$I_{obs}(\mu, \sigma) = -\begin{bmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \mu} & \frac{\partial^2 \ell}{\partial \sigma^2} \end{bmatrix}_{(\mu,\sigma)=\hat{\mu}_M,\hat{\sigma}_M}$$

The asymptotic variance-covariance matrix is derived from

$$I_{obs}^{-1}(\mu, \sigma) = \begin{bmatrix} var(\hat{\mu}_M) & cov(\hat{\mu}_M, \hat{\sigma}_M) \\ cov(\hat{\sigma}_M, \hat{\mu}_M) & var(\hat{\sigma}_M) \end{bmatrix}$$

# Maximum Likelihood estimate

```r
erf <- function(x) {
  2*pnorm(x*sqrt(2))-1
}

phim <- function (x, m, s) {
  ((x-m)*exp(-(x-m)^2/(2*s*s)))/(sqrt(2*pi)*s^3)
}

phimm <- function (x, m, s) {
  ((((x-m)^2/s^2)-1)*exp(-((x-m)^2)/(2*s^2)))/(sqrt(2*pi)*s^3)
}

phis <- function (x, m, s) {
  ((((x-m)^2)*exp(-(x-m)^2/(2*s^2)))/(sqrt(2*pi)*s^4))-(exp(-(x-m)^2/(2*s^2))/(sqrt(2*pi)*s^2))
}

phiss <- function (x, m, s) {
  sqrt(2)*(1-(3-(((m-x)^2/s^2))*(m-x)^2/(2*s^2))-((m-x)^2/s^2))*exp(-(m-x)^2/(2*s^2))/(sqrt(pi)*s^3)
}

phims <- function (x, m, s) {
  sqrt(2)*(3-((m-x)^2/s^2))*(m-x)*exp(-(m-x)^2/(2*s)^2)/(2*sqrt(pi)*s^4)
}

E<-function (a, m, s) {exp(-((a-m)**2)/(2*s**2))}
PHIm<-function (a, m, s) {-E(a,m,s)/(sqrt(2*pi)*s)}
PHIs<-function (a, m, s) {-((a-m)*E(a,m,s))/(sqrt(2*pi)*s**2)}
PHImm<-function (a, m, s) {-((a-m)*E(a,m,s))/(sqrt(2*pi)*s**3)}
PHIss<-function (a, m, s) {((a-m)*E(a,m,s)*(1-((a-m)/(sqrt(2)*s))**2))/(sqrt(pi/2)*s**3)}
PHIsm<-function (a, m, s) {(E(a,m,s)*(1-((a-m)/s)**2))/(sqrt(2*pi)*s**2)}
PHIms<-PHIsm

d2ldm2 <- function(x, a, m, s) {
  n <- length(x)
  sum(phimm(x, m, s)/dnorm(x, m, s))-sum(phim(x, m, s)^2/dnorm(x, m, s)^2) +
    n*PHImm(a, m, s)/(1-pnorm(a, m, s)) +
    n*(PHIm(a, m, s)^2/(1-pnorm(a, m, s))^2)
}

d2lds2 <- function(x, a, m, s) {
  n <- length(x)
  sum(phiss(x, m, s)/dnorm(x, m, s))-sum(phis(x, m, s)^2/dnorm(x, m, s)^2) +
    n*PHIss(a, m, s)/(1-pnorm(a, m, s)) + n*(PHIs(a, m, s)^2/(1-pnorm(a, m, s))^2)+(n/s^2)
}

d2ldmds <- function(x, a, m, s) {
  n <- length(x)
  sum(phims(x, m, s)/dnorm(x, m, s))-sum(phim(x, m, s)*phis(x, m, s)/dnorm(x, m, s)^2) +
    n*(PHIms(a, m, s)/(1-pnorm(a, m, s)))+n*(PHIm(a, m, s)*PHIs(a, m, s)/(1-pnorm(a, m, s))^2)
}

d2ldsdm <- function(x, a, m, s) {
  n <- length(x)
  sum(phims(x, m, s)/dnorm(x, m, s))-sum(phim(x, m, s)*phis(x, m, s)/dnorm(x, m, s)^2) +
    n*(PHIsm(a, m, s)/(1-pnorm(a, m, s)))+n*(PHIm(a, m, s)*PHIs(a, m, s)/(1-pnorm(a, m, s))^2)
}

esigma <- function(x, a, m, s) {
  m11 <- d2ldm2(x, a, m, s)
  m12 <- d2ldmds(x, a, m, s)
  m21 <- d2ldsdm(x, a, m, s)
  m22 <- d2lds2(x, a, m, s)
  (d2lds2(x, a, m, s)/(d2ldm2(x, a, m, s)*d2lds2(x, a, m, s)-d2ldmds(x, a, m, s)*d2ldsdm(x, a, m, s)))
}
procdat <- function(dat) {
  m <- dat
  m <- m[!is.na(m$NU_NOTA_MT),]
  m <- m[m$NU_NOTA_MT != 0,]
  m
}
```

# Maximum Likelihood estimate

```r
cities <- c("São Paulo","Rio de Janeiro","Brasília","Fortaleza","Salvador","Belo Horizonte","Manaus","Curitiba","Recife","Goiânia","Belém","Porto Alegre","Guarulhos","Campinas",
            "São Luís","Maceió","São Gonçalo","Campo Grande","Teresina","João Pessoa","Duque de Caxias","Nova Iguaçu","São Bernardo do Campo","Natal","Santo André")

# with data in reduced, info in qest, a set.

abled <- sapply(cities, function (city) {
  m <- qest[qest[1] == city, 2]
  s <- qest[qest[1] == city, 3]
  x <- reduced[reduced$D_UNION == 0 &
                 reduced$NO_MUNICIPIO_PROVA == city,
               "NU_NOTA_MT"]
  esigma(x, a, m, s)
})

disabled <- sapply(cities, function (city) {
  m <- qest[qest[1] == city, 4]
  s <- qest[qest[1] == city, 5]
  x <- reduced[reduced$D_UNION == 1 &
                 reduced$NO_MUNICIPIO_PROVA == city,
               "NU_NOTA_MT"]
  esigma(x, a, m, s)
})
disabled
```

# Maximum Likelihood estimate

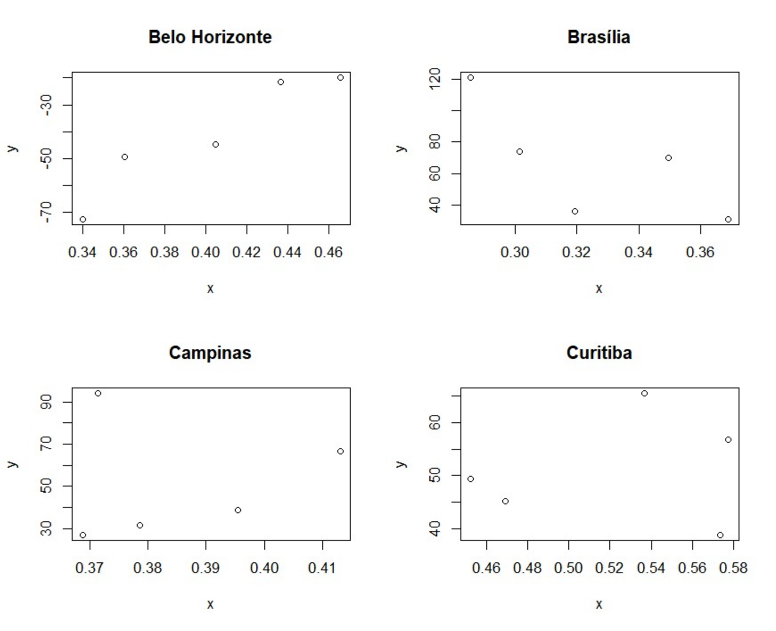
```
abled
          São Paulo      Rio de Janeiro            Brasília           Fortaleza
        1.808484269        -0.323424832         0.256087908         0.332452311
           Salvador      Belo Horizonte              Manaus            Curitiba
        0.050658072        -0.352743172         0.002733032        -0.583084533
             Recife             Goiânia               Belém        Porto Alegre
       -2.122562774        -1.638324662         0.013168394        -0.940740617
          Guarulhos            Campinas           São Luís              Maceió
        0.078256219        -1.541138550         0.010319611         0.081619226
         São Gonçalo       Campo Grande            Teresina         João Pessoa
        0.149159493         0.152261740         0.129356895         0.490435288
     Duque de Caxias         Nova Iguaçu São Bernardo do Campo             Natal
        0.036245633         0.042732088         3.235402599        -6.208057122
        Santo André
       -3.127630850
```

# Maximum Likelihood estimate

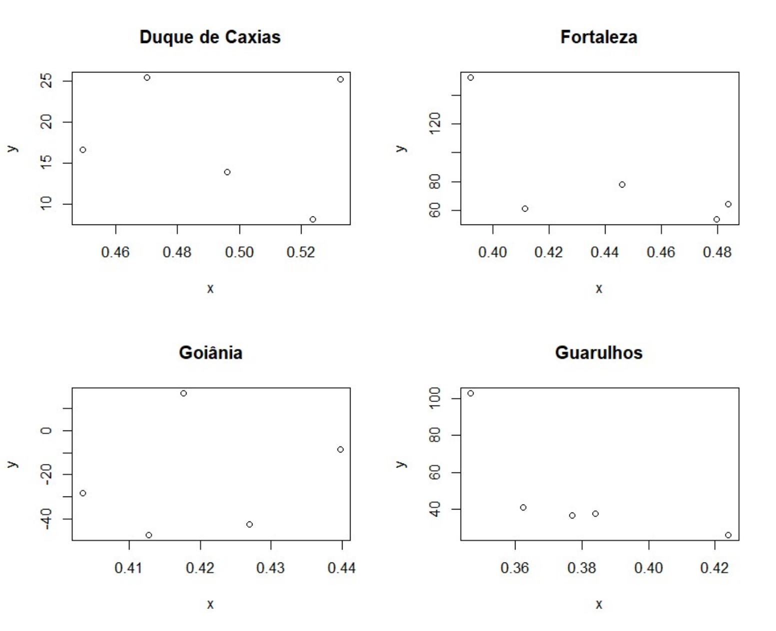
```
disabled
       São Paulo      Rio de Janeiro               Brasília            Fortaleza
    3.529495e+00       -4.614621e+01          -4.211090e+02         4.997112e+01
        Salvador      Belo Horizonte                 Manaus             Curitiba
    5.533500e+01       -6.935462e+01           5.186213e-01        -1.068980e+04
          Recife             Goiânia                  Belém         Porto Alegre
   -9.947888e+01       -1.501326e+02           6.356508e-02        -2.648524e+02
       Guarulhos            Campinas              São Luís               Maceió
    2.382943e-01        1.468275e+02           8.634219e-02         2.216615e+03
      São Gonçalo       Campo Grande               Teresina          João Pessoa
    3.004449e+01        2.104113e+01           1.855319e+00         3.690409e+02
   Duque de Caxias        Nova Iguaçu São Bernardo do Campo                Natal
    1.016681e+00        4.828897e+01          -8.786814e+02         3.884905e+01
     Santo André
   -3.477164e+02
```

# Verification

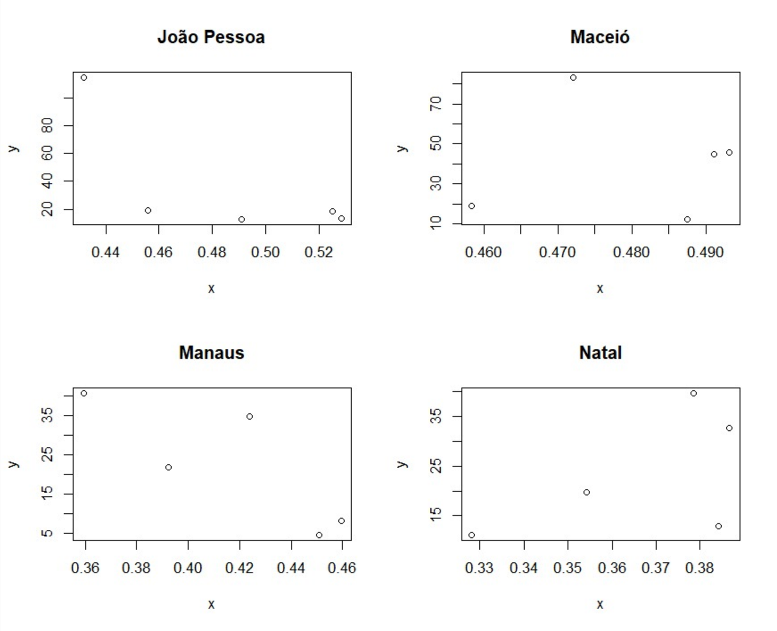
# Verification



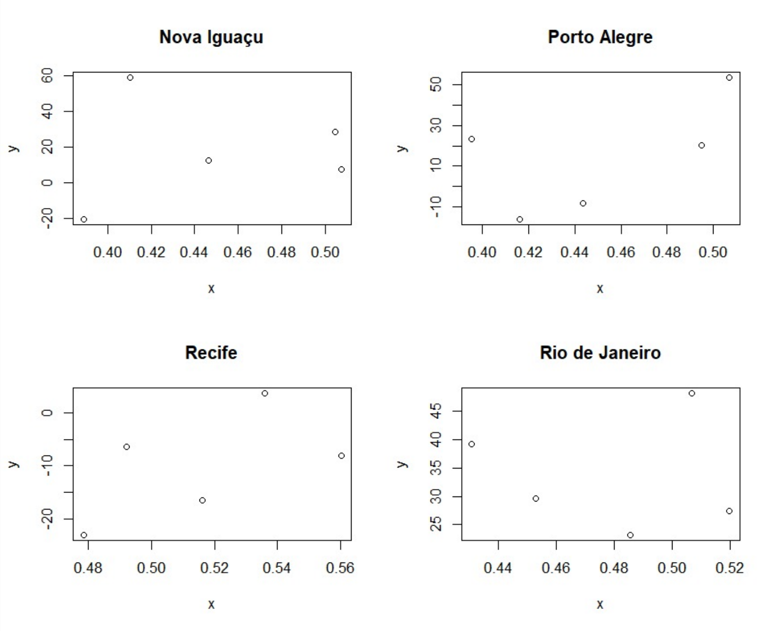
# Verification



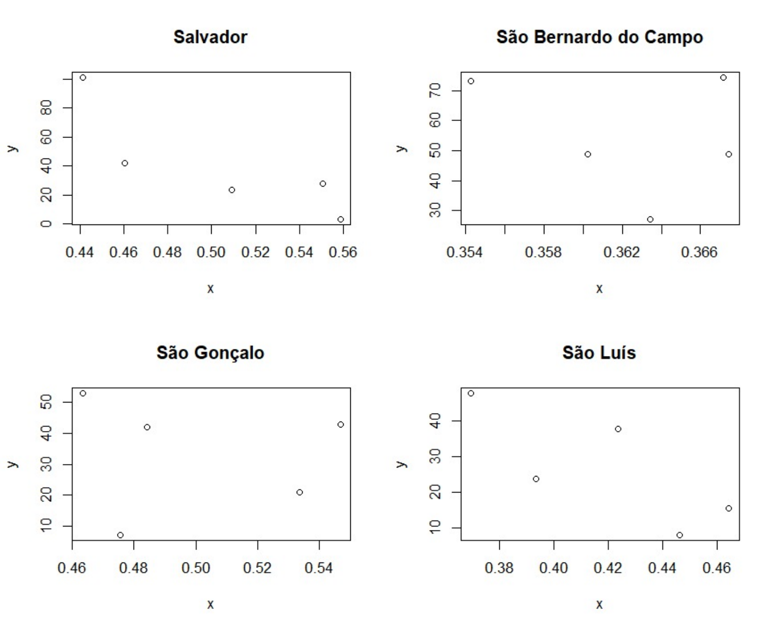
# Verification



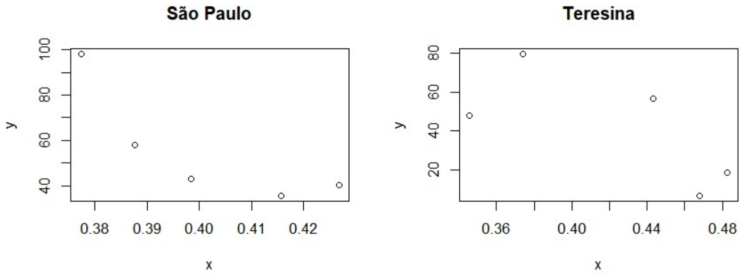
# Verification



## Verification



# Verification



# Conclusion

# Conclusion

There does not appear to be a general relationship between the Segregation Rate and the disadvantage.

# Bibliography

1. C. R. Rao, Linear Statistical Inference and Its Applications, Wiley, 2002, Chapter 4, p. 220
2. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) Brazil, `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem`
3. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) Brazil, `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar`

4. Rafael Françozo, et al., https://data.mendeley.com/datasets/hxx5gfkhms/2, 2024
5. Zeng, X.; Gui, W. Statistical Inference of Truncated Normal Distribution Based on the Generalized Progressive Hybrid Censoring. Entropy 2021, 23, 186. https://doi.org/10.3390/e23020186, 2021
6. Peter J. Bickel and Kjell A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Prentice Hall, Chapter 2, p. 114, 'Maximum Likelihood Estimate'

# Acknowledgement

We would like to express our gratitude to our professor Rituparna Sen who gave us this wonderful project and an opportunity to learn many new things.