MTH511: Statistical Simulation and Data Analysis

Extended Logit-Normal Regression

Team Pitchers

IIT Kanpur

January 1, 2023

Contents

Introduction

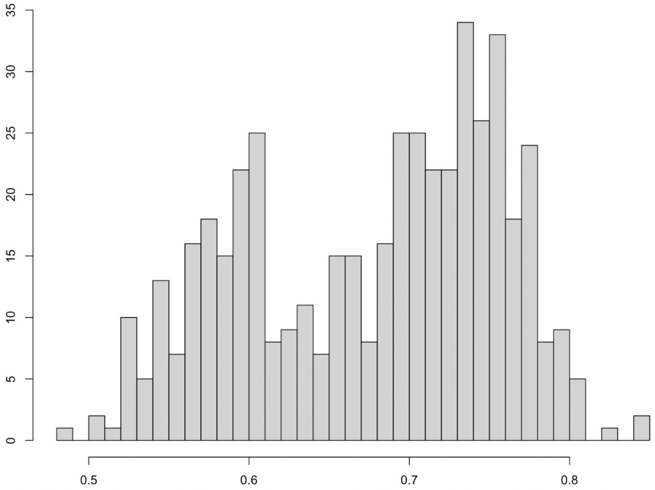
- In several real world applications, we are faced with situations where we want to be able to generate predictions about the frequency or magnitude of a phenomenom, but all we have is past data.
- In these scenarios, data fitting using a probability distributions can be a useful. For example, we commonly use the normal distribution to fit symmetric data centered around the mean.
- These distributions can have one peak, or they can have several peaks. The probability distribution that best fits them varies according to the shape of the distribution.
- The familiar normal distribution, or bell curve, has one peak, or unimodal. Similarly, we define distributions with two peaks as bimodal distributions.

Introduction

- Generally a mixture of distributions has to be used to model the bimodal proportional data.
- In this paper the author has proposed the *odd log-logistic logit-normal* (OLLLTN) distribution to model bimodal proportional data without the need of a mixture of distributions.
- The distribution is generated using the *odd log-logistic generator* (OLL-G) family on the *logit-normal* (LTN) distribution, as the name suggests
- The *logit-normal* (LTN) distribution is very useful to analyze data in the interval (0, 1). It is commonly used to generate other probability distributions
- The author argues that the OLLLTN distribution can be an interesting alternative to bimodal data in the (0, 1) interval instead of the classical beta and simplex models.

Introduction

Figure 1 displays the HDI histogram for 478 cities in the Brazilian states of Santa Catarina and Pernambuco. The formation of two sub-populations can be clearly noted, so these data have a bimodal shape.



What is OLLLTN distribution?

Lets start by defining LTN distribution.

Logit-Normal Distribution

If $W \sim Normal(logit(\mu)), \sigma^2)$, where $logit(\mu) = log(\mu/1-\mu)$ and $\mu \in (0,1)$, The LTN random variable is defined by $X = (1+e^{-W})^{-1}$.
The cdf of X (for $0 < x < 1$) is:

$$G_{\mu,\sigma}(x) = \psi(\frac{logit(x)-logit(\mu)}{\sigma}) = \frac{1}{2}[1 + erf(\frac{logit(x)-logit(\mu)}{\sqrt{2}\sigma})]$$

where $\psi(z)$ is the standard normal cdf and $erf(z) = 2\pi^{\frac{1}{2}} \int_0^z e^{\frac{-t^2}{2}}$ is the error function.

Now, from cdf we get the pdf of X to be

Probability density function of LTN

$$g_{\mu,\sigma}(x) = \frac{1}{x(1-x)\sqrt{2\pi\sigma^2}} e^{\frac{-[logit(x)-logit(\mu)]^2}{2\sigma^2}}$$

What is OLLLTN distribution?

The odd *log-logistic generator* (OLL-G) family has been widely employed in the last fifteen years. It follows by integrating the log-logistic density function

CDF of OLLTN Family

Based on this family, the OLLLTN cdf can be expressed as

$$F(y) = F(y;\mu,\sigma,\nu) = \int_0^{\frac{G_{\mu,\sigma(y)}}{1-G_{\mu,\sigma(y)}}} \frac{\nu x^{\nu-1}}{(1+x^{\nu})^2} dx = G_{\mu,\sigma}(y)^{\nu}$$

where $0 < \mu < 1$ is a position, $\sigma > 0$ is a scale, and $\nu > 0$ is a shape parameter.

Letting $\eta(y) = G_{\mu,\sigma}(y)$

PDF of OLLLTN Family

$$f(y) = f(y;\mu,\sigma,\nu) = \frac{\nu}{y(1-y)\sqrt{2\pi\sigma^2}} e^{\frac{-[logit(x)-logit(\mu)]^2}{2\sigma^2}}$$
$$\times[\eta(y)[1-\eta(y)]]^{\nu-1}[\eta(y)^{\nu}+[1-\eta(y)]]^{\nu-2}$$

It is evident from the above equations that *OLLLTN* reduces to *LTN* when $\nu = 1$.

What is OLLLTN distribution?

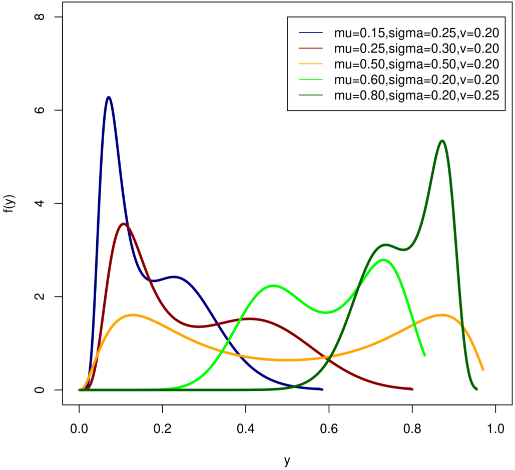


Figure: The pdf of OLLLTN distribution for various parameters

Plots

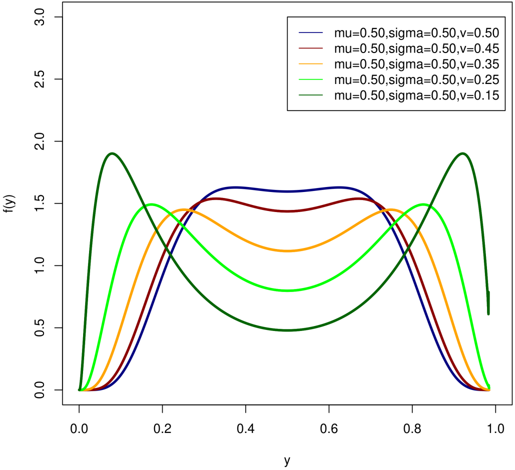


Figure: The pdf of OLLLTN distribution for various parameters

Properties of OLLLTN distribution

The quantile function Q_Y of the *OLLLTN* can be expressed in terms of the quantile function Q_{LTN} of the*LTN* distribution.

Quantile Functions for *LTN*

The quantile function of *LTN* distribution is given by,

$$Q_{LTN}(u) = G^{-1}_{\mu,\sigma}(u) = (1 + \exp(-v(u)))^{-1},$$

where $v(u) = logit(\mu) + \sqrt{2}\sigma^{-1}(2u - 1)$

Quantile function for *OLLLTN*

The quantile function of *OLLLTN* distribution is given by,

$$Q_Y(u) = Q_{LTN}\left(\frac{u^{\frac{1}{\nu}}}{u^{\frac{1}{\nu}} + (1-u)^{\frac{1}{\nu}}}\right)$$

Properties of OLLLTN distribution

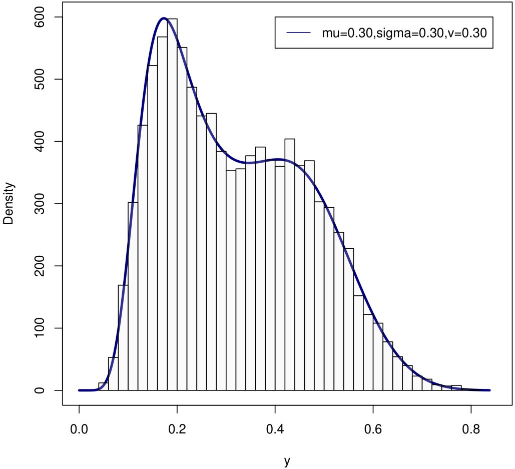


Figure: Histograms and plots of OLLLTN distribution

Properties of *OLLLTN* distribution

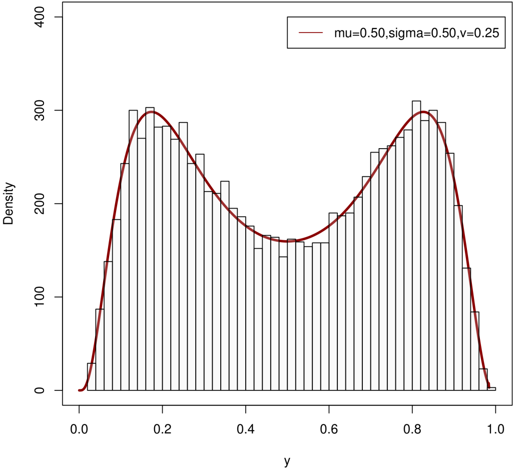


Figure: Histograms and plots of OLLLTN distribution

Properties of *OLLLTN* distribution

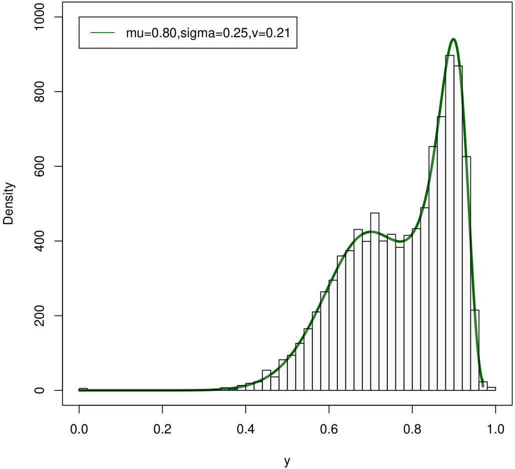


Figure: Histograms and plots of OLLLTN distribution

Moments, Skewness and Kurtosis

Moments of *OLLLTN*

The n^{th} ordinary moment of the *OLLLTN* model is given by

$$E(Y^n) = \int_0^1 Q_{LTN}\left(\frac{u^{\frac{1}{\nu}}}{u^{\frac{1}{\nu}} + (1-u)^{\frac{1}{\nu}}}\right)^n du$$

Skewness and Kurtosis

The *Bowley′s Skewness* and *Moors′ Kurtosis* for *OLLLTN* are less sensitive to outliers and are given by,

$$Skewness = \frac{Q_Y(1/4) + Q_Y(3/4) - 2Q_Y(1/2)}{Q_Y(3/4) - Q_Y(1/4)}$$

and,

$$Kurtosis = \frac{Q_Y(7/8) - Q_Y(5/8) + Q_Y(3/8) - Q_Y(1/8)}{Q_Y(6/8) - Q_Y(2/8)}$$

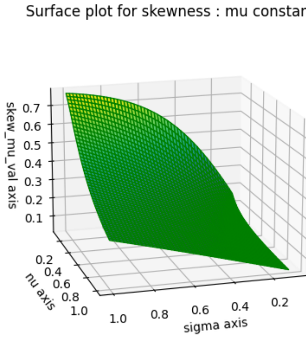
respectively.

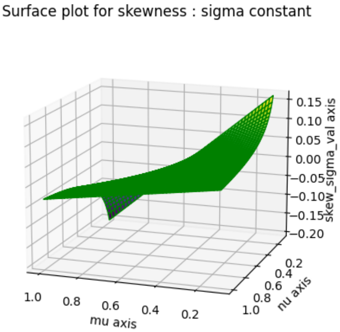
Moments of OLLLTN

μ	σ	ν	κ_1	κ_2	κ_3	κ_4
0.2	0.1	0.1	0.22532054	0.07068196	0.03576166	0.02706187
0.2	0.1	2	0.200141025	0.040131762	0.008062229	0.001622591
0.5	0.8	0.1	0.5000172	0.4329052	0.3993577	0.3758595
0.5	2	0.1	0.5	0.4781088	0.4671633	0.4596786
0.8	0.5	0.1	0.6579184	0.5436995	0.4856778	0.4484023
0.8	0.1	2	0.799859	0.6398497	0.5119104	0.4096002

Table: Table 1. Values for $\kappa_1, \kappa_2, \kappa_3 and \kappa_4$.

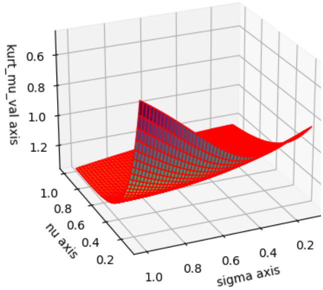
Skewness of OLLLTN


Surface plot for skewness : mu constar

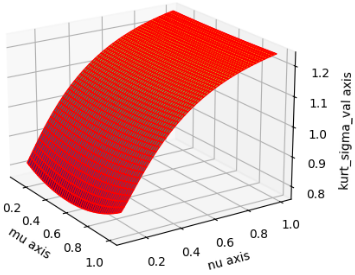

Surface plot for skewness : sigma constant

Kurtosis of OLLLTN

Surface plot for kurtosis : mu constant



Surface plot for kurtosis : sigma constant



The OLLLTN Regression

σ-link and μ-link for *OLLLTN* regression

In *OLLLTN* the parameters μ_i and σ_i of the responses Y_i are linked to the vector $\mathbf{x_i^T} = (x_{i1}, \cdots, x_{ip})$ of covariates as follows,

$$\mu_i = \frac{\exp(\mathbf{x_i^T}\beta_1)}{1+\exp(\mathbf{x_i^T}\beta_1)} \quad \text{and} \quad \sigma_i = \exp\left(\mathbf{x_i^T}\beta_2\right), \quad i = 1, \cdots, n,$$

where the vectors $\beta_1 = (\beta_{11}, \cdots, \beta_{1p})^T$ and $\beta_2 = (\beta_{21}, \ldots, \beta_{2p})^T$ are unknown.

log-likelihood for *OLLLTN* regression

The *log-likelihood* for $\theta = (\beta_1^{\mathbf{T}}, \beta_2^{\mathbf{T}}, \nu)^{\mathbf{T}}$ for *OLLLTN* regression given n independent observations is,

$$l(\theta) = n\log(\nu) + \sum_{i=1}^{n} \log\left[\frac{1}{y_i(1-y_i)}\right] - \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_i^2)$$

$$-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{logit(y_i) - logit(\mu_i)}{\sigma_i}\right]^2 + (\nu - 1)\sum_{i=1}^{n}\log(\eta(y_i)(1-\eta(y_i)))$$

$$-2\sum_{i=1}^{n}\log(\eta(y_i)^{\nu} + (1-\eta(y_i))^{\nu})$$

Monte Carlo Simulation Study

To prove that we can fit a *bidmodal* data using the *OLLLTN* distribution, the paper conduct the following simulation study:

- Set μ, σ, ν.
- Sample 5000 observations from the *OLLLTN* distribution by putting $u \sim U(0, 1)$ in the *Quantile* function defined above.
- Estimate the parameters for which the negative log likelihood is minimum.
- Repeat the above n times.
- Calculate the *Average Bias*, *Average Estimate* and *Average Mean Square Error*.
- The above experiment was conducted for $n = 100$, 500 and 1000 for three sets of parameters.
- Scenario 1: $\mu = 0.3, \sigma = 0.3, \nu = 0.3$.
- Scenario 2: $\mu = 0.5, \sigma = 0.5, \nu = 0.25$.
- Scenario 3: $\mu = 0.8, \sigma = 0.6, \nu = 0.5$.

Monte Carlo Simulation Study

Scenario 1				
n	Parameters	μ	σ	ν
	AE	0.2996319396	0.3179439145	0.3218377346
100	Bias	0.0003680604311	-0.01794391446	-0.02183773462
	MSE	0.004567852601	0.08863975505	0.09339966532
	AE	0.2999522011	0.3145330975	0.3169807203
500	Bias	4.78E-05	-0.01453309749	-0.01698072025
	MSE	0.004298254545	0.0845669961	0.0899673599
	AE	0.299942189	0.3141098351	0.3186003053
1000	Bias	5.78E-05	-0.01410983513	-0.01860030534
	MSE	0.004409591492	0.06866856481	0.07603868059

Table: Table 2. Simulation results from the OLLLTN distribution.

Monte Carlo Simulation Study

Remarks

- We performed the simulation using the *optim* function in *R*.
- The paper suggests using the *gamlss* package in *R*, but unfortunately we had some difficulty in using the package.
- The optimization was highly unstable for lower σ values using *optim*, and often failed to converge, possibly because of precision errors while evaluating *erf*() and *pnorm*().
- However, the general trend can still be observed using the simulation study. As *n* increases, the *average bias* and *MSE* tend to decrease and the *Average Estimates* tend to converge the true values.

Regression Simulation

To simulate regression, the following study was conducted:

- Set $\beta_{10} = 0.7, \beta_{11} = 0.1, \beta_{12} = 0.9, \beta_{20} = 0.3, \beta_{21} = 0.5, \beta_{22} = 0.2, \nu = 0.3$
- Sample $X_{i1} \sim \mathit{Normal}(0,1), X_{i2} \sim \mathit{Binomial}(1, 0.5)$ for $i = 1, \ldots, 5000$.
- $\mu_i = \frac{exp(\beta_{10}+\beta_{11}x_{i1}+\beta_{12}x_{i2}}{1+exp(\beta_{10}+\beta_{11}x_{i1}+\beta_{12}x_{i2})}$ and $\sigma_i = exp(\beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2})$.
- Sample $Y_i \sim \mathit{OLLLTN}(\mu_i, \sigma_i, \nu_i)$.
- Maximize the *MLE* to obtain estimates for β_i and calculate the *Average Estimates*, *Average Bias* and *Average Mean Square Error*.

Regression Simulation

n	Parameters	β_{10}	β_{11}	β_{12}
	AE	0.7463860069	0.1134486347	0.7699283599
10	Bias	-0.04638600685	-0.01344863468	0.1300716401
	MSE	0.07255056305	0.06103310921	0.1439022475
	AE	0.752577285	0.1206346289	0.7945489483
100	Bias	-5.26E-02	-0.02063462887	0.1054510517
	MSE	0.1147136143	0.05395788794	0.164917494
	AE	0.7247482891	0.1211346054	0.8024552818
200	Bias	-2.47E-02	-0.02113460545	0.0975447182
	MSE	0.09277026972	0.06050829266	0.1525868306

Table: Table 3. Simulation results from the OLLLTN distribution.

Regression Simulation

n	Params	β_{20}	β_{21}	β_{22}	ν
	AE	0.4916318	0.5010259	0.1925605	0.3979769
10	Bias	-0.1916318	-0.00102596	0.007439408	-0.09797696
	MSE	0.2014186	0.01747957	0.03146895	0.1036441
	AE	0.4697655	0.4976435	0.2025009	0.3886241
100	Bias	-0.1697655	0.002356415	-0.002500954	-0.08862416
	MSE	0.1835779	0.01863063	0.04470331	0.09539252
	AE	0.4785972	0.4992003	0.2027889	0.3932353
200	Bias	-0.1785972	0.0007996548	-0.002788988	-9.32E-02
	MSE	0.1921706	0.01967459	0.03761810	0.100755

Table: Table 4. Simulation results from the OLLLTN distribution.

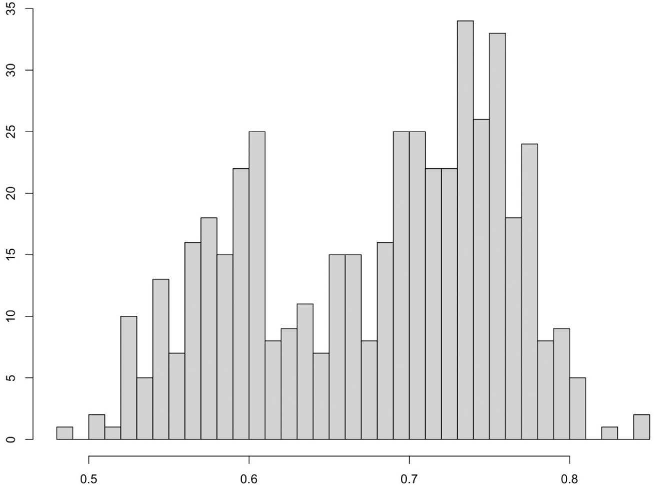
Regression Simulation

Remarks

- We used *optim* to maximize the log likelihood function and get estimates of β_i
- We set initial values of $\beta_{ij} = 0.5$ and $\nu = 1$ and conducted regular *logit-normal log likelihood* maximization to obtain β_1 and β_2.
- We then used these as starting values for the *OLLLTN loglikelihood* maximization. Thus our starting values were $(\beta_1, \beta_2, 1)$.
- The regression was highly unstable and it failed to converge in many cases. Thus, while doing the above studies, we only considered cases where the function actually converged, which was roughly about 40% of the cases
- However, we can still observe the general trend that the β_{ij} converge to their true values, although the *bias* is high in this case. This is due to precision errors and rounding off leading to unstable convergence.
- The results would have probably been more accurate had we been able to use the gamlss package

Application to Brazil HDI Data

- The Brazilian HDI data is a collection of the HDI indexes of 478 municipalities in the Brazilian states of Pernambuco and Santa Catarina
- The data follows a bimodal shape roughly.



Applications to Brazil HDI Data

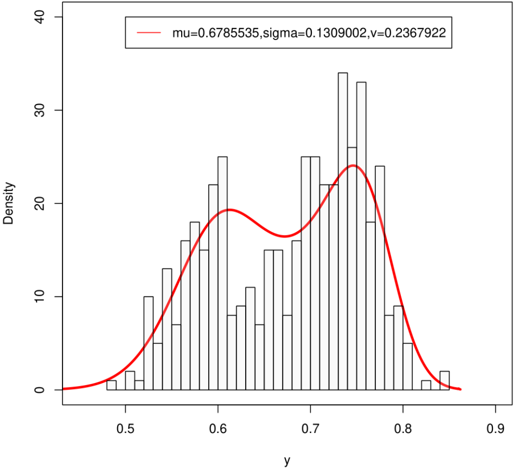


Figure: OLLLTN fitted to HDI data

Application to Brazil HDI Data

We fit four distributions to the Brazilian HDI data and measured some goodness-of-fit statistics. The results are tabulated below:

Distributions	μ	σ	ν	*GD*	*AIC*
OLLLTN	0.6785535	0.1309002	0.2367922	-1167.249	-1161.249
SIMPLEX	0.6790461	0.7926546	-	-1104.26	-1100.26
LTN	0.6842778	0.3652149	-	-1098.859	-1094.859
BE	0.6790925	0.1663487	-	-1092.7	-1088.7

Table: Table 4. Results from four fitted distributions to the *HDI* data

To estimate parameters for the data, we used the gamlss package to define an OLLLTN family from scratch. Through gamlss, we were able to get parameters for different distributions and their goodness of fit measures more easily than by using optim

Application to Brazil HDI Data

Remarks

- We fit four different distributions on the Brazilian HDI data, which is approximately bimodal in shape
- The *Global Deviance,* which is equal to two times the negative log likelihood, is the lowest for OLLLTN which indicates that the likelihood function is the highest for it among the other distributions
- The *Akaike Information Criterion (AIC)* is also a measure of the goodness of a fitting model. It takes into account the negative log likelihood as well as the number of independent parameters used by the model. OLLLTN outperforms the others in this case as well
- The graphs also suggest that only OLLLTN distribution is able to account for the bimodality present in the data distribution. The other distributions are not able to account for the bimodality.
- The values in the table are in accordance to the values presented in the paper

Concluding Remarks

- The author presented the OLLLTN distribution as an alternative for modelling data in the interval (0,1)
- First, we established the theoretical properties of the distribution such as pdf,cdf,moments,etc.
- We plotted the pdf of OLLLTN for various parameters and showed that it can be used to account for all kinds of bimodal data in the interval (0,1)
- Based on the Logit Normal Regression model, we defined the OLLLTN regression theoretically
- We conducted Monte Carlo Simulation studies to show that the parameters of a fitting-OLLLTN distribution can be correctly estimated for a given data
- We then simulated OLLLTN regression to show that we can correctly estimate the coeffecients of covariates in a regression setting for bimodal data
- Finally, we established the usefulness of the OLLLTN distribution by showing its application to real-world data, and that it even performs better than other distributions in some cases.

Concluding Remarks

- We were successfully able to reproduce the trends of most of the simulation studies, graphs and tables in the paper
- The paper went a step furthur and performed OLLLTN regression on HDI data using educational and income measures as covariates
- However, we were unable to reproduce the regression on the HDI data as the authors had not posted the data of what they used as the covariates and we were unable to find the exact information anywhere on the internet
- We attempted to perform the regression using other similar data, but the results were not reproduced
- The main difficulty was that the regression was extremely unstable using optim, and often did not converge, especially when the coeffecients were greater than one.

End of Presentation

Thank You!