LMAT1271 - Project 2022

Statistical analysis of a company electricity consumption

Rousseau Mathieu, 67001800
Dieden Sam, 23731800

UCLouvain
Belgium
19/05/2022

1. Point estimation

Context

Our engineering team just landed a consulting contract with a company interested in the electricity consumption of its machines. In a first part, we would like to determine how electricity consumption is evenly distributed across the different machines of the same type. To this end, we use the Gini coefficient. In a nutshell, it is an index ranging from 0 to 1 measuring the inequality featured in a distribution. A value of 0 denotes that all our machines use the same amount of electricity while a value of 1 means that all the electricity is used by a single machine. We assume that all of the n machines operate independently and their daily electricity consumption (in MWh) can be modelled as a random variable X with the following probability density function (PDF),

$$f_{\theta_1,\theta_2}(x) = \begin{cases} \frac{\theta_1 \theta_2^{\theta_1}}{x^{\theta_1+1}}, & x \geq \theta_2 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

with $\theta_1 > 2$ and $\theta_2 > 0$. This is the PDF of the **Pareto distribution**.

(a) Derive the quantile function of X

We're looking to solve $P(X \leq x_t) = t$ for x_t.

First let's compute the cumulative distribution function (CDF) $P(X \leq x_t)$,

$$\begin{aligned} P(X \leq x_t) &= \int_{-\infty}^{x_t} f_{\theta_1,\theta_2}(x)dx \\ &= \int_{\theta_2}^{x_t} \theta_1 \theta_2^{\theta_1} x^{-(\theta_1+1)} dx \\ &= -\frac{\theta_1 \theta_2^{\theta_1}}{\theta_1} \left[x^{-\theta_1}\right]_{x=\theta_2}^{x=x_t} \\ &= -\theta_2^{\theta_1} \left(x_t^{-\theta_1} - \theta_2^{-\theta_1}\right) \\ &= 1 - \left(\frac{\theta_2}{x_t}\right)^{\theta_1} \end{aligned}$$

Let's solve $P(X \leq x_t) = t$ for x_t,

$$\begin{aligned} 1 - \left(\frac{\theta_2}{x_t}\right)^{\theta_1} = t &\iff (1-t)^{1/\theta_1} = \frac{\theta_2}{x_t} \\ &\iff x_t = \frac{\theta_2}{(1-t)^{1/\theta_1}} \end{aligned}$$

Therefore we have,

$$Q_{\theta_1,\theta_2}(t) = \frac{\theta_2}{(1-t)^{1/\theta_1}} \tag{2}$$

(b) Derive the Gini coefficient of X.

The Gini coefficient is defined as,

$$G_{\theta_1,\theta_2} = 2\int_0^1 \left(p - \frac{\int_0^p Q(t)dt}{E(X)}\right) dp \tag{3}$$

Let's first compute the expectation value of X,

$$\begin{aligned}
E(X) &= \int_{-\infty}^{+\infty} x \cdot f_{\theta_1,\theta_2}(x)dx \\
&= \int_{\theta_2}^{+\infty} x\frac{\theta_1\theta_2^{\theta_1}}{x^{\theta_1+1}}dx \\
&= \theta_1\theta_2^{\theta_1}\int_{\theta_2}^{+\infty} x^{-\theta_1}dx \\
&= -\frac{\theta_1\theta_2^{\theta_1}}{(\theta_1-1)}\left[x^{-(\theta_1-1)}\right]_{\theta_2}^{+\infty} \\
&= \begin{cases} -\frac{\theta_1\theta_2^{\theta_1}}{(\theta_1-1)}\left(-\frac{1}{\theta_2^{-(\theta_1-1)}}\right), & \theta_1 > 1 \\ +\infty, & \theta_1 \leq 1 \end{cases} \\
&= \begin{cases} \frac{\theta_1\theta_2}{(\theta_1-1)}, & \theta_1 > 1 \\ +\infty, & \theta_1 \leq 1 \end{cases}
\end{aligned}$$

So the Gini coefficient is defined for $\theta_1 > 1$,

$$G_{\theta_1,\theta_2} = 2\left(\int_0^1 pdp - \int_0^1 \frac{\int_0^p Q_{\theta_1,\theta_2}(t)dt}{E(X)}dp\right)$$

We compute each integral separately,

$$\int_0^1 pdp = \frac{1}{2}$$

Then,

$$\int_0^p Q_{\theta_1,\theta_2}(t)dt = \theta_2\int_0^p \frac{1}{(1-t)^{1/\theta_1}}$$

We use the change of variable $u = 1 - t \implies du = -dt$

The boundaries become,

$$\begin{cases} t = 0 & \implies u_1 \equiv 1 \\ t = p & \implies u_2 \equiv 1 - p \end{cases}$$

Then,

$$\begin{aligned}\int_0^p Q_{\theta_1,\theta_2}(t)dt &= -\theta_2 \int_{u_1}^{u_2} \frac{1}{(u)^{1/\theta_1}} du \\ &= -\theta_2 \left[\frac{(u)^{-(1/\theta_1 - 1)}}{-((1/\theta_1)-1)} \right]_{u_1}^{u_2} \\ &= \frac{\theta_2}{(1/\theta_1)-1} \left(\frac{1}{(1-p)^{1/\theta_1 - 1}} - \frac{1}{1^{1/\theta_1 - 1}} \right) \\ &= \frac{\theta_2}{(1/\theta_1)-1} \left(\frac{1}{(1-p)^{1/\theta_1 - 1}} - 1 \right)\end{aligned}$$

Therefore for $\theta_1 > 1$,

$$\begin{aligned}\frac{\int_0^p Q_{\theta_1,\theta_2}(t)dt}{E(X)} &= \frac{\frac{\theta_2}{(1/\theta_1)-1} \left(\frac{1}{(1-p)^{1/\theta_1 - 1}} - 1 \right)}{\frac{\theta_1 \theta_2}{(\theta_1 - 1)}} \\ &= \frac{\theta_2}{(1/\theta_1)-1} \left(\frac{1}{(1-p)^{1/\theta_1 - 1}} - 1 \right) \frac{(\theta_1 - 1)}{\theta_1 \theta_2} \\ &= \frac{\theta_1(1-(1/\theta_1))}{((1/\theta_1)-1)\theta_1} \left(\frac{1}{(1-p)^{1/\theta_1 - 1}} - 1 \right) \\ &= -\left(\frac{1}{(1-p)^{(1/\theta_1)-1}} - 1 \right) \\ &= 1 - \frac{1}{(1-p)^{(1/\theta_1)-1}}\end{aligned}$$

Then,

$$\int_0^1 \frac{\int_0^p Q_{\theta_1,\theta_2}(t)dt}{E(X)} dp = \underbrace{\int_0^1 1 dp}_{A} - \underbrace{\int_0^1 \frac{1}{(1-p)^{(1/\theta_1)-1}} dp}_{B}$$

Computing integral A and B.

$$A = \int_0^1 1 dp = 1$$

$$B = \int_0^1 \frac{1}{(1-p)^{(1/\theta_1)-1}} dp$$

We use the change of variable $u = 1 - p \implies du = -dp$.

The boundaries become,

$$\begin{cases} p = 0 & \implies u_1 \equiv 1 \\ p = 1 & \implies u_2 \equiv 0 \end{cases}$$

Then,

$$\begin{aligned}\int_0^1 \frac{1}{(1-p)^{(1/\theta_1)-1}}dp &= -\int_{u_1}^{u_2} \frac{1}{(u)^{(1/\theta_1)-1}}du \\ &= -\int_{u_1}^{u_2} u^{-((1/\theta_1)-1)}du \\ &= \frac{1}{((1/\theta_1)-1)-1}\left[(u)^{((1/\theta_1)-1-1)}\right]_1^0 \\ &= -\frac{1}{(1/\theta_1)-2} \\ &= \frac{1}{2-(1/\theta 1)}\end{aligned}$$

Eventually the Gini coefficient is (for $\theta_1 > 0$),

$$\begin{aligned}G_{\theta_1,\theta_2} &= 2\left(\frac{1}{2} - \frac{1}{2-(1/\theta 1)}\right) \\ &= 2\left(\frac{1}{2}\left[1 - \frac{1}{1-(1/2\theta_1)}\right]\right) \\ &= 1 - \frac{1}{1-(1/2\theta_1)} \\ &= \frac{1/2\theta_1}{1-(1/2\theta_1} \\ &= \frac{1}{2\theta_1\left(1-\frac{1}{2\theta_1}\right)} \\ &= \frac{1}{2\theta_1 - 1}\end{aligned}$$

(c) Derive the maximum likelihood estimator (MLE) of G_{θ_1,θ_2}. Call this estimator $\hat{G}_{\text{MLE}}$

Let's first compute the likelihood function $L(\theta_1, \theta_2)$,

$$\begin{aligned}L(\theta_1,\theta_2) &:= \Pi_{i=1}^n f_{\theta_1,\theta_2}(x) \\ &= \Pi_{i=1}^n \frac{\theta_1\theta_2^{\theta_1}}{x^{\theta_1+1}} \cdot I(X_i \geq \theta_2 > 0) \\ &= \theta_1^n\theta_2^{n\theta_1}\frac{1}{\Pi_{i=1}^n X_i^{\theta_1+1}} \cdot I(X_{(1)} \geq \theta_2 > 0)\end{aligned}$$

where $X_{(1)} \equiv \min(X_1, ..., X_n)$.

We notice that $L(\theta_1, \theta_2)$ is not continuous along θ_2 and then not differentiable in θ_2. However, we observe that $L(\theta_1, \theta_2)$ increase with θ_2. Therefore, we have to take θ_2 the largest possible in order to maximize $L(\theta_1, \theta_2)$ respecting the condition $X_{(1)} \leq (\theta_2 > 0)$ otherwise we would have $L(\theta_1, \theta_2) = 0$,

$$\hat{\theta}_2 = X_{(1)}$$

For $\hat{\theta}_1$ we can compute the log-likelihood function $l(\theta_1, \theta_2)$,

$$\begin{aligned} l(\theta_1, \theta_2) &:= \ln(L(\theta_1, \theta_2)) \\ &= \ln(\theta_1^n) + \ln(\theta_2^{n\theta_1}) + \ln(1) - \ln(\pi_{i=1}^n X_i^{(\theta_1+1)}) \\ &= n\ln(\theta_1) + n\theta_1 \ln(\theta_2) - (\sum_{i=1}^n \ln(X_i^{(\theta_1+1)})) \\ &= n\ln(\theta_1) + n\theta_1 \ln(\theta_2) - \sum_{i=1}^n (\theta_1 + 1))\ln(X_i) \end{aligned}$$

We differentiate with respect to θ_1 in order to find the maximum,

$$\frac{\partial l(\theta_1, \theta_2)}{\partial \theta_1} = \frac{n}{\theta_1} + n\ln(\theta_2) - \sum_{i=1}^n \ln(X_i)$$

Then,

$$\begin{aligned} \frac{\partial l(\theta_1, \theta_2)}{\partial \theta_1} = 0 \iff \hat{\theta}_1 &= \frac{n}{\sum_{i=1}^n (\ln(X_i)) - n\ln(\hat{\theta}_2)} \\ &= \frac{n}{\sum_{i=1}^n (\ln(X_i) - \ln(X_{(1)}))} \\ &= \frac{n}{\sum_{i=1}^n \ln\left(\frac{X_i}{X_{(1)}}\right)} \end{aligned}$$

Now we can compute $\hat{G}_{\text{MLE}}$,

$$\begin{aligned} \hat{G}_{\text{MLE}} &:= G_{\hat{\theta}_1, \hat{\theta}_2} \\ &= \frac{1}{2\hat{\theta}_1 - 1} \\ &= \frac{1}{\left(\frac{2n}{\sum_{i=1}^n \ln\left(\frac{X_i}{X_{(1)}}\right)}\right) - 1} \end{aligned}$$

(d) Propose a method of moment estimator of G_{θ_1, θ_2}. Call this estimator $\hat{G}_{\text{MME}}$

We already have computed the expectation value of X,

$$E(X) = \begin{cases} \frac{\theta_1\theta_2}{(\theta_1 - 1)}, & \theta_1 > 1 \\ +\infty, & \theta_1 \leq 1 \end{cases}$$

We know that,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \equiv E(X)$$

Let's solve for θ_1,

$$\begin{aligned}\bar{X} = \frac{\hat{\theta}_1\hat{\theta}_2}{(\hat{\theta}_1 - 1)} &\iff \bar{X}\hat{\theta}_1 - \bar{X} = \hat{\theta}_1\hat{\theta}_2 \\ &\iff \hat{\theta}_1(\bar{X} - \hat{\theta}_2) = \bar{X} \\ &\iff \hat{\theta}_1 = \frac{\bar{X}}{(\bar{X} - \hat{\theta}_2)}\end{aligned}$$

In order to estimate $\hat{\theta}_2$ we know that the CDF is given by,

$$F_{\theta_1\theta_2}(x) = P(X \leq x) = 1 - \left(\frac{\theta_2}{x}\right)^{\theta_1}$$

Therefore,

$$\begin{aligned}P(X > x) &= 1 - P(X \leq x) \\ &= \left(\frac{\theta_2}{x}\right)^{\theta_1}\end{aligned}$$

The probability that all random variables $(X_1, \ldots, X_n)$ are greater than x is,

$$\begin{aligned}P((X_1, \ldots, X_n) > x) &= \Pi_{i=1}^{n} P(X > x) \\ &= \left(\frac{\theta_2}{x}\right)^{n\theta_1}\end{aligned}$$

Then, the probability that the minimum random variable $X_{(1)} \equiv \min(X_1, \ldots, X_n)$ is greater than x is also,

$$P(X_{(1)} > x) = \left(\frac{\theta_2}{x}\right)^{n\theta_1}$$

Therefore,

$$P(X_{(1)} \leq x) = 1 - \left(\frac{\theta_2}{x}\right)^{n\theta_1}$$

The corresponding probability density function is,

$$\begin{aligned}f_{\theta_1,\theta_2}(x) &= F'_{\theta_1,\theta_2}(x) \\ &= \frac{\mathrm{d}}{\mathrm{d}x}\left(1 - \left(\frac{\theta_2}{x}\right)^{n\theta_1}\right) \\ &= -\theta_2^{n\theta_1}\frac{\mathrm{d}}{\mathrm{d}x}\left(x^{-n\theta_1}\right) \\ &= n\theta_1\theta_2^{n\theta_1}x^{-(n\theta_1+1)} \\ &= \frac{n\theta_1\theta_2^{n\theta_1}}{x^{(n\theta_1+1)}}, \quad x \geq \theta_2\end{aligned}$$

The corresponding expectation value is,

$$\begin{aligned}
E(X) &= \int_{\theta_2}^{+\infty} x \cdot f_{\theta_1,\theta_2}(x)dx \\
&= \int_{\theta_2}^{+\infty} x \cdot \frac{n\theta_1\theta_2^{n\theta_1}}{x^{(n\theta_1+1)}}dx \\
&= n\theta_1\theta_2^{n\theta_1} \int_{\theta_2}^{+\infty} x^{(-n\theta_1)}dx \\
&= \frac{n\theta_1\theta_2^{n\theta_1}}{-(n\theta_1-1)} \left(-\frac{1}{\theta_2^{-(n\theta_1-1)}}\right) \\
&= \frac{n\theta_1\theta_2}{(n\theta_1-1)}
\end{aligned}$$

Setting expectation value $E(X)$ to be equal the minimum random variable $X_{(1)}$,

$$X_{(1)} = \frac{n\theta_1\theta_2}{(n\theta_1-1)} \iff \hat{\theta}_2 = X_{(1)}\frac{(n\hat{\theta}_1-1)}{n\hat{\theta}_1}$$

Therefore,

$$\begin{aligned}
\hat{\theta}_1 &= \frac{\bar{X}}{(\bar{X}-\hat{\theta}_2)} \\
&= \frac{\bar{X}}{\bar{X} - X_{(1)}\frac{(n\hat{\theta}_1-1)}{n\hat{\theta}_1}} \\
\iff \bar{X} &= \hat{\theta}_1\left(\bar{X} - X_{(1)}\frac{(n\hat{\theta}_1-1)}{n\hat{\theta}_1}\right) \\
&= \hat{\theta}_1\bar{X} - \hat{\theta}_1 X_{(1)}\frac{n\hat{\theta}_1}{n\hat{\theta}_1} + \hat{\theta}_1 X_{(1)}\frac{1}{n\hat{\theta}_1} \\
&= \hat{\theta}_1\left(\bar{X} - X_{(1)}\right) + \frac{X_{(1)}}{n} \\
\iff \hat{\theta}_1 &= \frac{\bar{X} - (X_{(1)}/n)}{(\bar{X} - X_{(1)})} \\
&= \frac{n\bar{X} - X_{(1)}}{n(\bar{X} - X_{(1)})}
\end{aligned}$$

Now we can compute $\hat{G}_{\text{MME}}$,

$$\begin{aligned}
\hat{G}_{\text{MME}} &:= G_{\hat{\theta}_1,\hat{\theta}_2} \\
&= \frac{1}{2\hat{\theta}_1 - 1} \\
&= \frac{1}{\left(\frac{2(n\bar{X}-X_{(1)})}{n(\bar{X}-X_{(1)})}\right) - 1}
\end{aligned}$$

(e) Set $\theta_1^0 = 3$ and $\theta_2^0 = 1$. Generate an i.i.d sample of size $n = 20$ from the density $f_{\theta_1^0,\theta_2^0}$. In order to achieve this, you can make use of the inverse transform sampling. Using this sample, compute $\hat{G}_{\text{MLE}}$ and $\hat{G}_{\text{MME}}$.

We have,

$$f_{\theta_1^0,\theta_2^0} = \begin{cases} \frac{3 \cdot 1^3}{x^{3+1}} = \frac{3}{x^4}, & x \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

We compute the CDF of X,

$$\begin{aligned} F_{\theta_1,\theta_2}(x) = \int_1^x \frac{3}{t^4} dt &= 3 \left[\frac{t^{-3}}{-3} \right]_1^x \\ &= -\left(\frac{1}{x^3} - \frac{1}{1^3} \right) \\ &= 1 - \frac{1}{x^3} \end{aligned}$$

The inverse is,

$$F^{-1}_{\theta_1,\theta_2}(y) = \frac{1}{(1-y)^{1/3}}$$

We get the following estimations,

$$\hat{G}_{\text{MLE}} = 0.2516462 \quad ; \quad \hat{G}_{\text{MME}} = 0.235356 \tag{4}$$

With these estimations, we can notice that the electricity consumption is not evenly distributed among all the computers.

(f) Repeat this data generating process $N = 1000$ times (with the same sample size $n = 20$ and the same (θ_1^0, θ_2^0)). Hence, you obtain a sample of size N of each estimator of G_{θ_1,θ_2}. Make a **histogram** and a **boxplot** of these two samples. What can you conclude ?

As we can see on the figure 1 and figure 2, the most presents values of these Gini coefficients for the two methods of estimation are pretty close to the exact value of the Gini coefficient. In fact, on the boxplot (figure 2), we see that the median value of these sample is a little bit lower than the exact of value of the Gini coefficient and in particular the median value of $\hat{G}_{\text{MME}}$ is almost identical to the exact value G. The samples distributions are close to a normal distribution even more true for the sample generated by the maximum likelihood method. For the two distributions, we observe some outliers on the right of the distribution.

Finally, we can conclude that the median value of the Gini sample estimated through the method of moment is closer to the exact value of the Gini coefficient than the median value estimated through the method of maximum likelihood. However, there seems to have more outliers in this last one compared to the one estimated through the maximum likelihood method.

(g) Use the samples obtained in (f) to estimate the **bias**, the **variance** and the **mean squared error (MSE)** of both estimators What can you conclude ?

Using the samples obtained in (f) and computing the different asked quantities, we get

Estimator	Bias	Variance	Mean Squared Error
$\hat{G}_{\text{MLE}}$	-0.00913594	0.002571429	0.002652323
$\hat{G}_{\text{MME}}$	-0.003834601	0.00286779	0.002879627

Table 1 – *Bias, variance and mean squared error of the estimators $\hat{G}_{MLE}$ and $\hat{G}_{MME}$ for $N = 1000$ simulations and a sample size of $n = 20$*

We notice that the bias of $\hat{G}_{\text{MLE}} < \hat{G}_{\text{MME}}$. $\hat{G}_{\text{MME}}$ being closer to zero confirms that this estimator is better in estimating the exact value of the Gini coefficient. A negative bias confirms that we tend to underestimate the value of that coefficient.

However, despite $\hat{G}_{\text{MLE}}$ being more biased, its variance and its mean squared error is lower than the variance of $\hat{G}_{\text{MME}}$. So the coefficient estimated by the method of maximum likelihood seems preferable.

Finally, obviously, the MSE for the two methods is not zero zero meaning that we do not predicts the exact value of the Gini coefficient with perfect accuracy.

(h) Repeat the calculations in (f) for $n = 20$, 40, 60, 80, 100, 150, 200, 300, 400, 500. Compare the **biases**, the **variances** and the **mean squared errors** of both estimators graphically (make a separate plot for each quantity as a function of n). What can you conclude ? Which estimator is the best ? Justify your answer.

Looking at figure 3, figure 4 and figure 5, the more we increase the sample size n, the less biased are the estimators. For a sample size of $n = 500$, we see that $\hat{G}_{\text{MME}}$ is **unbiased** whereas $\hat{G}_{\text{MLE}}$ is a tiny bit (positively) **biased** meaning that it is overpredicting a tiny bit the exact value of the Gini coefficient.

We notice that the variance and mean squared error of the two estimators are converging toward 0 but as we saw in (g), these two statistical quantities are still lower for $\hat{G}_{\text{MLE}}$ so that last one estimator is preferable.

(i) Create an histogram for $\sqrt{n}(\hat{G}_{\text{MLE}} - G_{\theta_1^0,\theta_2^0})$, for $n = 20$, $n = 100$ and $n = 500$. What can you conclude ?

If we check figure 6, as we increase the sample size n, the distribution of $\sqrt{n}(\hat{G}_{\text{MLE}} - G_{\theta_1^0,\theta_2^0})$ tends more and more toward a standard normal distribution $\mathcal{N}(0,1)$ meaning that $\hat{G}_{\text{MLE}}$ converges toward an exact estimation of the Gini coefficient as we increase the sample size.

2. Regression

The company wants to understand how electricity consumption is linked to productivity (i.e daily amount in 1000 euros that the company gains when the machine operates). We gather a dataset made of 40 independent observations for which we observe the following variables,

$$X \equiv \text{Electricity consumption in MWh} \quad ; \quad Y \equiv \text{productivity in thousands of euros per day} \tag{5}$$

(a) Is it reasonable to fit a linear regression model between **productivity** (Y) and **electricity consumption** (X) ? If no, what transformation of X and/or Y would you propose to retrieve a linear model ? Justify.

Hint : graphical representation may help visualize how the variables and the residuals behave.

For the rest of the exercise, we work we the transformed variables X^* and Y^*. Write down the obtained model.

Note : it may be that $Y = Y^*$ and/or $X = X^*$.

If we look at the **scatter plot** (7) of the productivity versus electricity consumption. We see clearly that the relationship between X and Y is not linear at all. As the electricity consumption goes up, the productivity variable is much more scattered. We can also notice an outlier at electricity consumption $\approx 48MWh$.

We can still try to fit a linear model using the **ordinary least square (OLS)** method that consists in minimizing the sum of the square of the error :

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{6}$$

Looking at figure 8, in the **Residuals vs Fitted** plot, we notice that only for the small values of $\hat{Y}$, the points are randomly distributed around $y = 0$. As the value of $\hat{Y}$ increase, it's less the case. Therefore, for high values of $\hat{Y}$, the residues are not normally distributed. Highlighting the non-linearity we saw on the previous scatter-plot. The red line shows that mean of the residuals $E(\epsilon) \neq 0$.

Checking the **QQplot** in the top right corner, we see that the extreme values are pulling away from the dashed line. This suggests that the explanatory variable X is heavy tailed.

Then, the **Scale-Location** plot indicates that the size of the residues gets smaller as the fitted values increase but especially the red line goes up and down suggesting again a poor fit.

Finally, the **Residuals vs Leverage** plot show us that there is a point that lies outside the Cook's distance. Therefore, this point is an influential observation that impact heavily our linear model.

Eventually, checking the summary of the linear model in R, we see the $R^2 = 0.2614$ so 26% of the variation of the **productivity** is explained by the **electricity consumption**. This suggest that the linear relation between the 2 variable is weak.

We can verify the observation made with the QQplot, by plotting an histogram of X (figure 9)

We observe a strong asymmetry in X. This variable is right skewed.

We conclude that we cannot use a simple linear model like $Y \sim X$,

$$Y = \beta_0 + \beta_1 X, \quad \beta_0, \beta_1 \in \mathbb{R}$$

We can try to transform the variable X. Having a right skewed explanatory variable suggests trying the following transformations,

- **Square Root transformation** : $X \to \sqrt{X}$
- **Log transformation** : $X \to \log_{10}(X)$
- **Reciprocal transformation** : $X \to 1/X$ (or higher order in X)

After trying the different models, we notice that the following transformation : $X \to 1/(X)^2 \equiv X^*$ seems the best at explaining the relationship between the **productivity** and the **electricity consumption** (figure 10, figure 11)

Indeed, looking at some plot to analyse this model (figure 12), we can see on the **Residuals vs Fitted** plot the the points are now randomly distributed around $y = 0$. Moreover, the mean of the residuals is tending to 0 for every fitted values.

On the **QQplot**, the low values are now following the dashed line. The high values are still pulling away from that line probably because of the outlier. So the model is still not adapted for the high values of X, removing the outlier or providing more data for X in order the fill the gap unto the outlier could be some solutions to improve the model and provide better predictions.

The points are randomly distributed around the red line on the **Scale-Location** plot and the red line is not going up and down anymore. Eventually, there is no points lying outside of the Cook's distance anymore.

Checking the summary of that model, the R-Squared is now,

$$R^2 = 0.4671 \tag{7}$$

Finally, our model is,

$$Y = 23 - \frac{203.822}{X^2} \tag{8}$$

(b) Mathematically derive the marginal impact of X on Y in your model. This is computed via the following formula,

$$\frac{\partial E(Y|X=x)}{\partial x} \tag{9}$$

Provide interpretation.

We have,

$$E(Y|X=x) = 23 - \frac{203.822}{x^2}$$

Therefore, the marginal impact of X on Y is,

$$\begin{aligned}\frac{\partial E(Y|X=x)}{\partial x} &= \frac{\partial}{\partial x}\left(23 - \frac{203.822}{x^2}\right) \\ &= 2 \cdot \frac{203.822}{x^3}\end{aligned}$$

The marginal impact is telling us how the response variable (**productivity**) changes when the explanatory variable (**electricity consumption**) changes of one unit at a given value of this last variable.

Therefore, the effect of increasing the electricity consumption by 1 MWh on the productivity in 1000 €/day is $2 \cdot 203.822 = 1607.644$, the effect of increasing the electricity consumption by 2 MWh on the productivity in 1000€/day is $2 \cdot \frac{203.822}{2^3} = 200.95$ and so on. The more we increase the electricity consumption, the less is the effect on the productivity.

(c) Is the linear effect significant ? Choose the adequate test for testing linear significance. Compute the p-value of this test. Based on the resulting p-value, what can we conclude ? Analyse the value of the linear effect.

We can use the t-test for β_0 et β_1 for testing linear significance.

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\beta_0}} \sim t_{n-2} \quad ; \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \sim t_{n-2}$$

where $\hat{\sigma}_{\beta_i}$ is the standard error of β_i.

The corresponding hypothesis testing are,

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

for $i = 0, 1$.

The p-value for this hypothesis test is given by $Pr(> |t|)$ and is provided by the summary of the model in r (figure 13)

Looking at that summary, we see that,

$$Pr(\beta_0 > |t|) < 2 \cdot 10^{-16}$$
$$Pr(\beta_1 > |t|) = 1.17 \cdot 10^{-6}$$

Therefore, β_0 and β_1 are significantly $\neq 0$ and then we reject the null hypothesis $H_0 : \beta_i = 0$ for $i = 0, 1$ for any choice of significance level α (and in particular for the standard level $\alpha = 5\%$). We can conclude there is a linear relationship between the **productivity** (Y) and the inverse square of the **electricity consumption** $(1/X^2)$.

A. Plots

A.1. Histogram of Gini sample

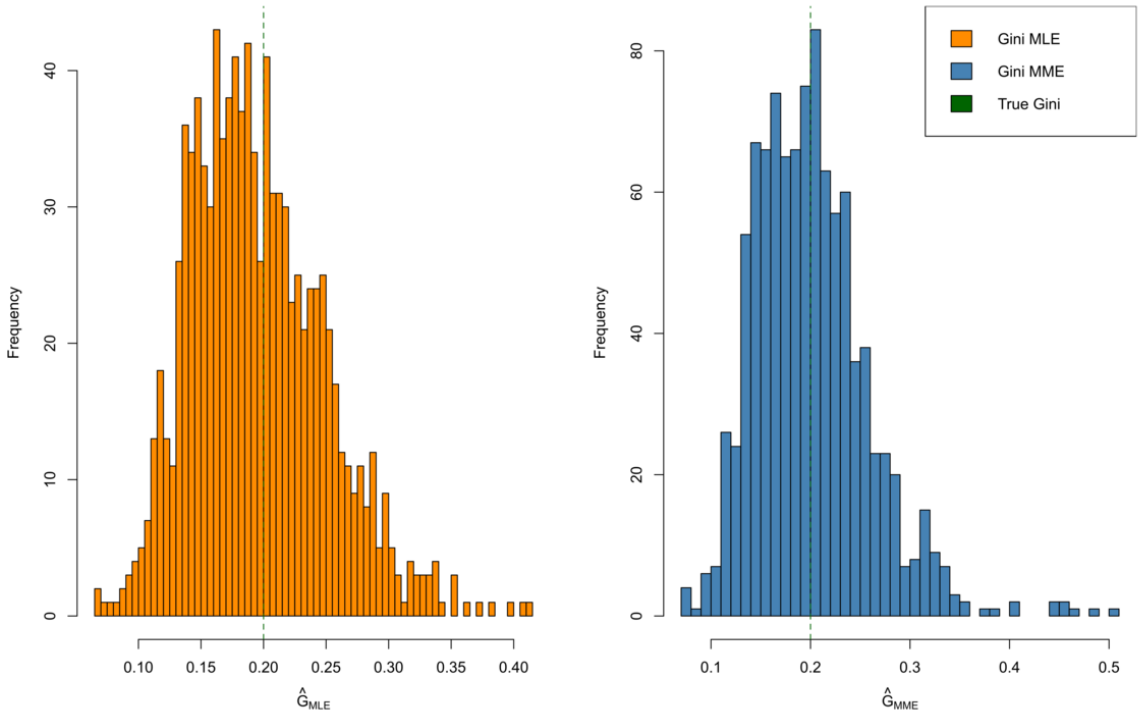

Figure 1 – *histogram of $N = 1000$ simulations of Gini coefficients estimated by the maximum likelihood method (left) and the moments method (right) base on a sample of size $n = 20$. In green we have the exact value of the Gini coefficient*

A.2. Boxplot of Gini sample

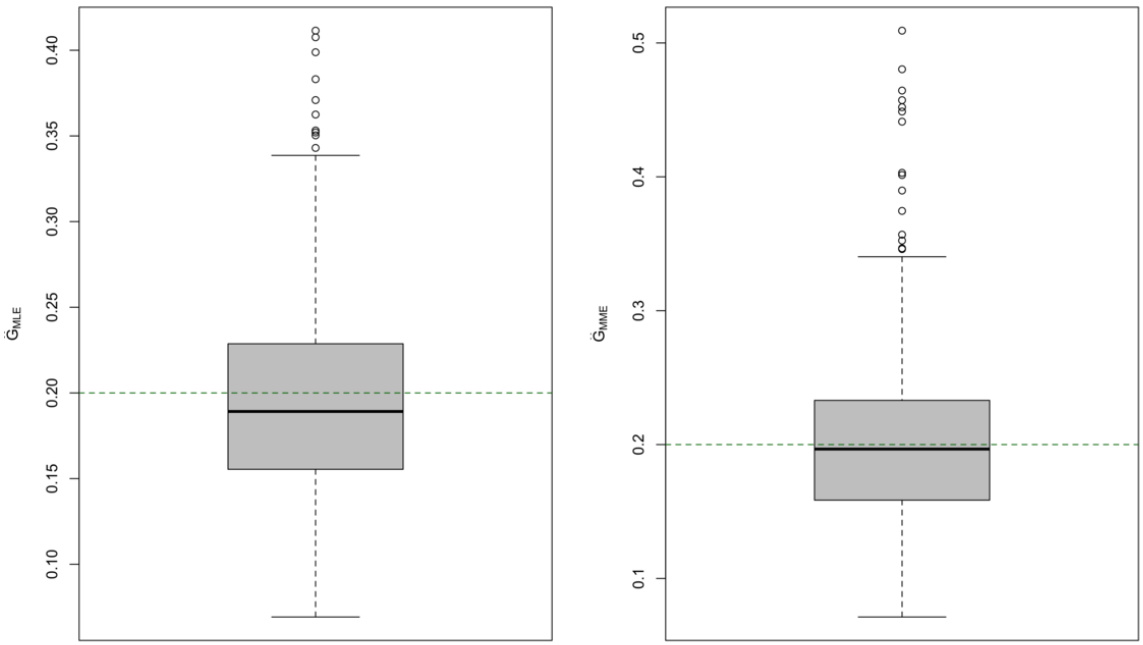

Figure 2 – *boxplot of $N = 1000$ simulations of Gini coefficients estimated by the maximum likelihood method (left) and the moments method (right) base on a sample of size $n = 20$. In green we have the exact value of the Gini coefficient*

A.3. Gini estimators : bias plot

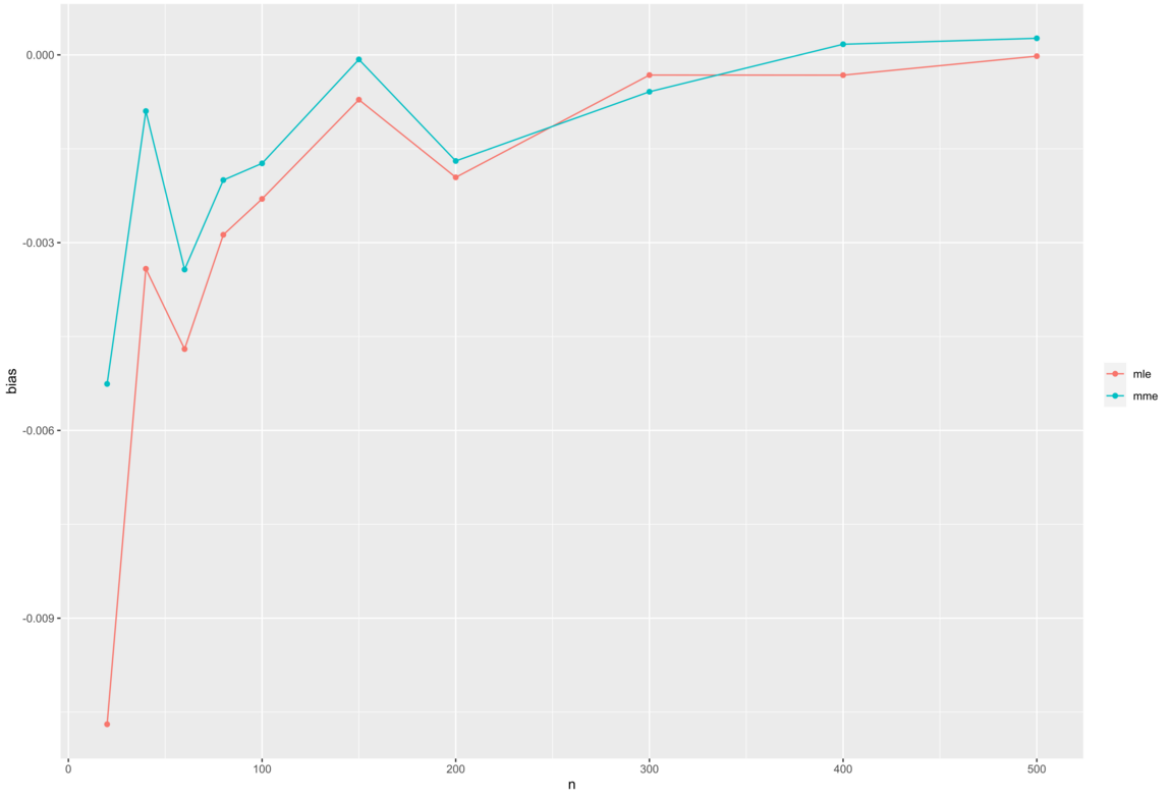

Figure 3 – *Plot of the bias of the Gini estimators (obtained by maximum likelihood in* ***red*** *and by method of moment in* ***turquoise****) as a function of the sample size n.*

A.4. Gini estimators : variance plot

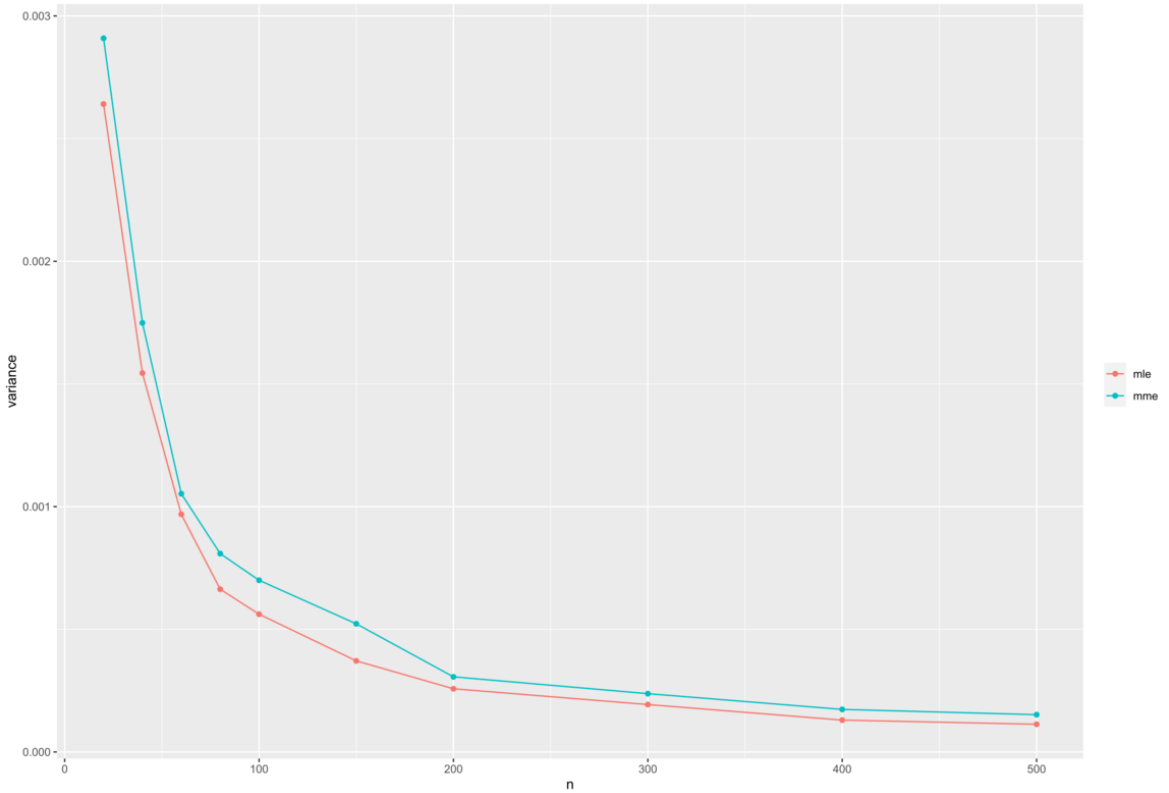

Figure 4 – *Plot of the variance of the gini estimators (obtained by maximum likelihood in* ***red*** *and by method of moment in* ***turquoise****) as a function of the sample size n.*

A.5. Gini estimators : mean squared error plot

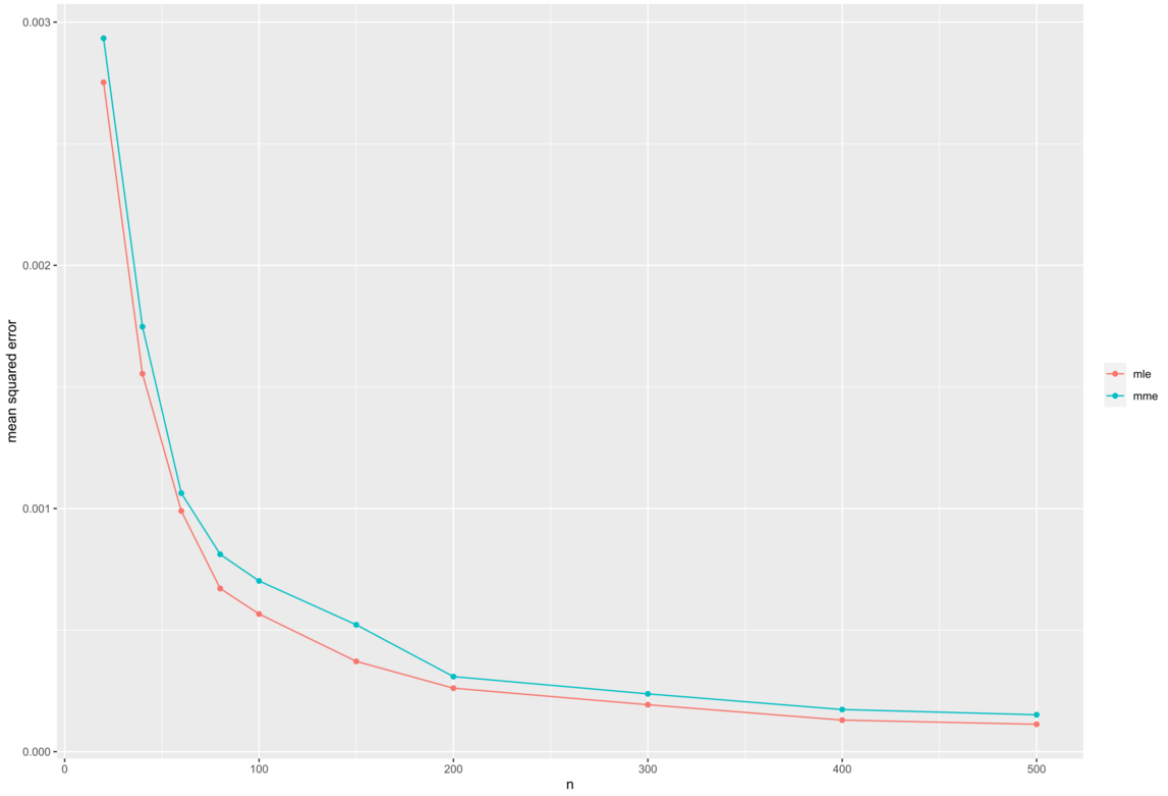

Figure 5 – *Plot of the mean squared error (MSE) of the gini estimators (obtained by maximum likelihood in* ***red*** *and by method of moment in* ***turquoise****) as a function of the sample size n.*

A.6. Histograms of the formula (question (i))

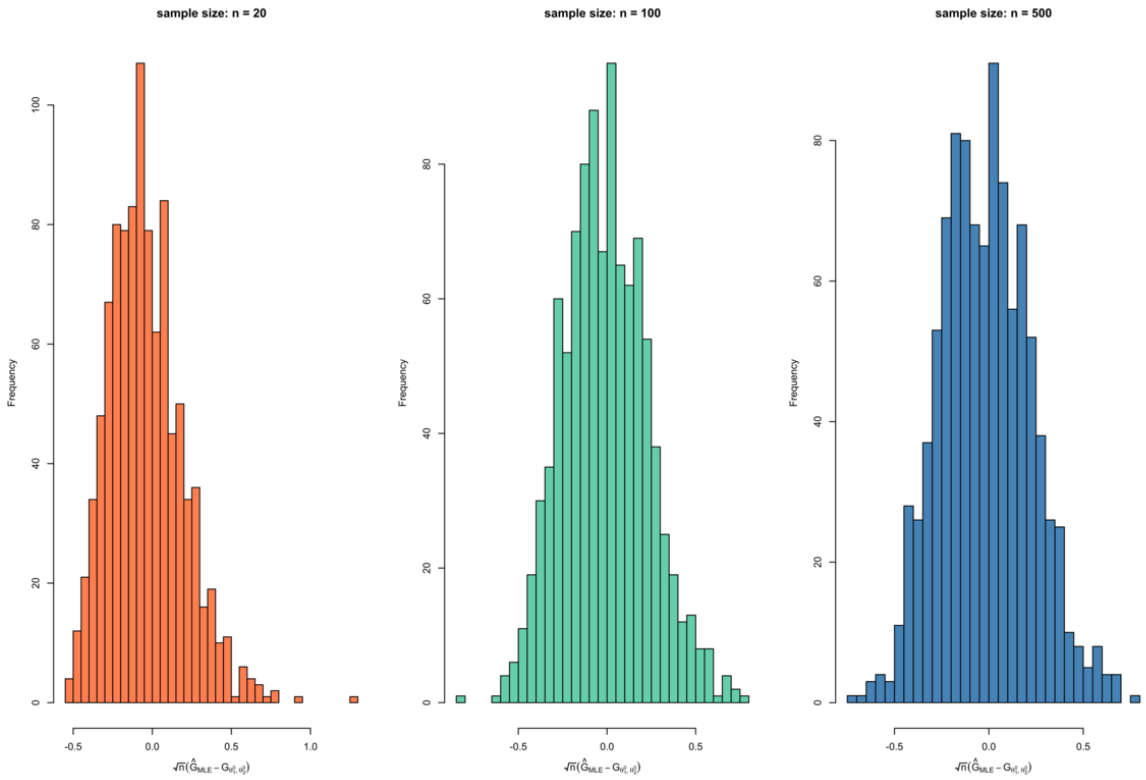

Figure 6 – *histogram for* $\sqrt{n}(\hat{G}_{MLE} - G_{\theta_1^0,\theta_2^0})$*, for sample size* $n = 20$*,* $n = 100$ *and* $n = 500$

A.7. Scatter plot of productivity vs electricity consumption

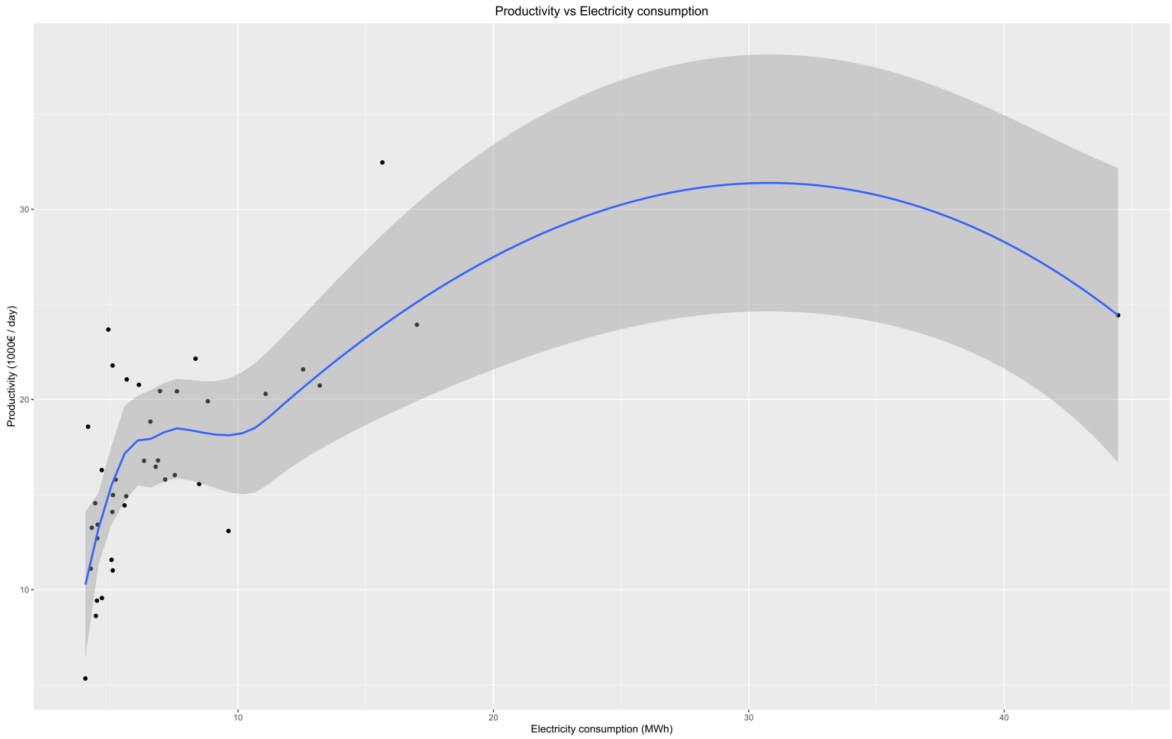

Figure 7 – *scatter plot of productivity versus electricity consumption*

A.8. Linear model : analysis

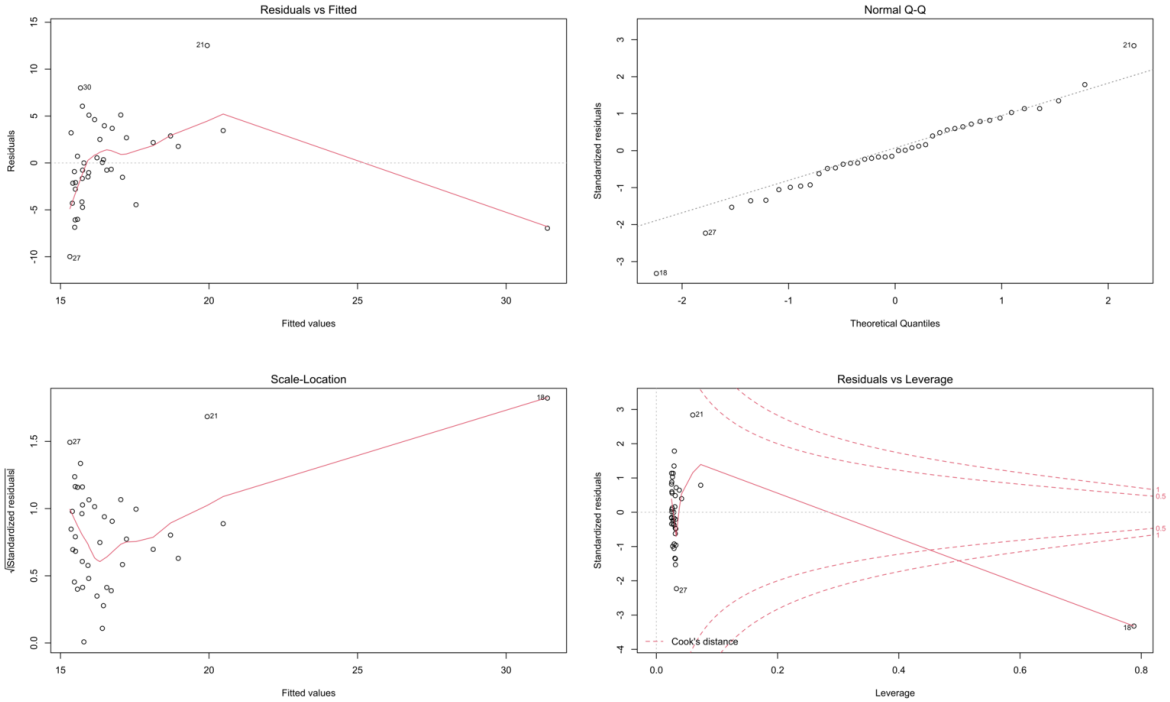

Figure 8 – *Analysis of the linear model* $Y \sim X$

A.9. Linear model : histogram of explanatory variable

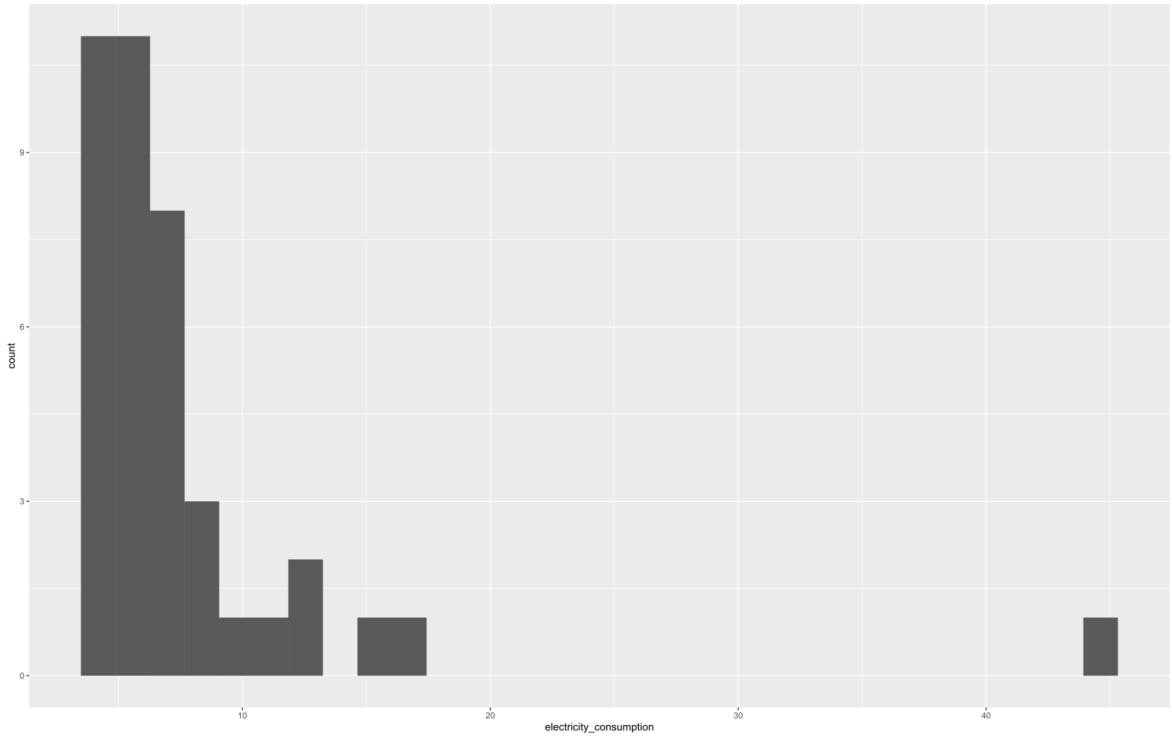

Figure 9 – *Histogram of the explanatory variable X (electricity consumption)*

A.10. Reciprocal model : scatter plot

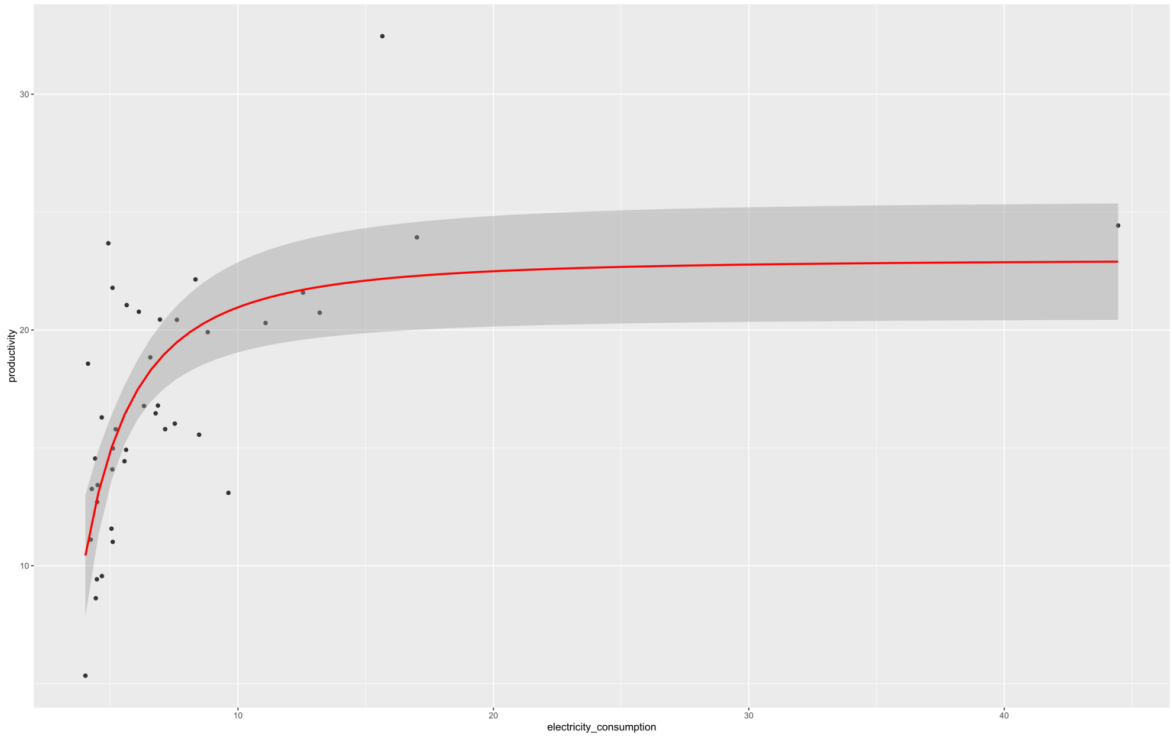

Figure 10 – *Scatter plot of Y versus X. In red we have the following regression line : $Y_i = \beta_0 + \beta_1(1/(X_i)^2)$*

A.11. Reciprocal model : linear scatter plot

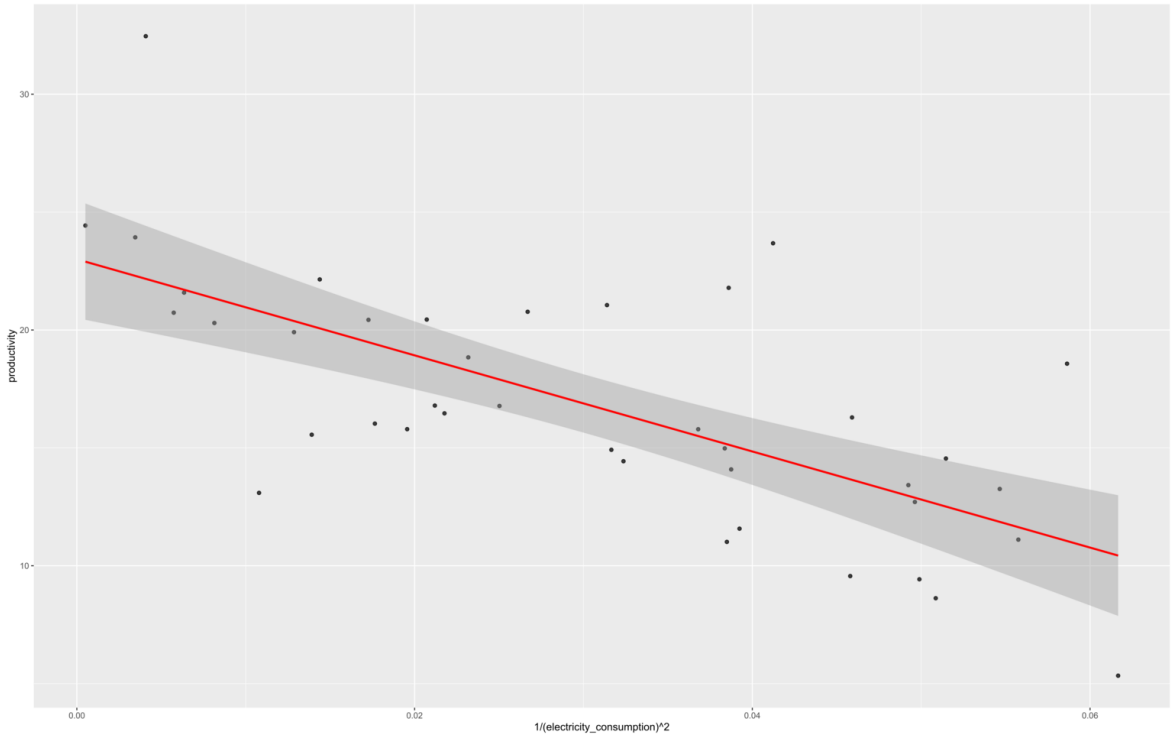

Figure 11 – *Scatter plot of* Y *versus* $1/X^2$*. In red we have the following regression line :* $Y_i = \beta_0 + \beta_1(1/(X_i)^2)$

A.12. Reciprocal model : analysis

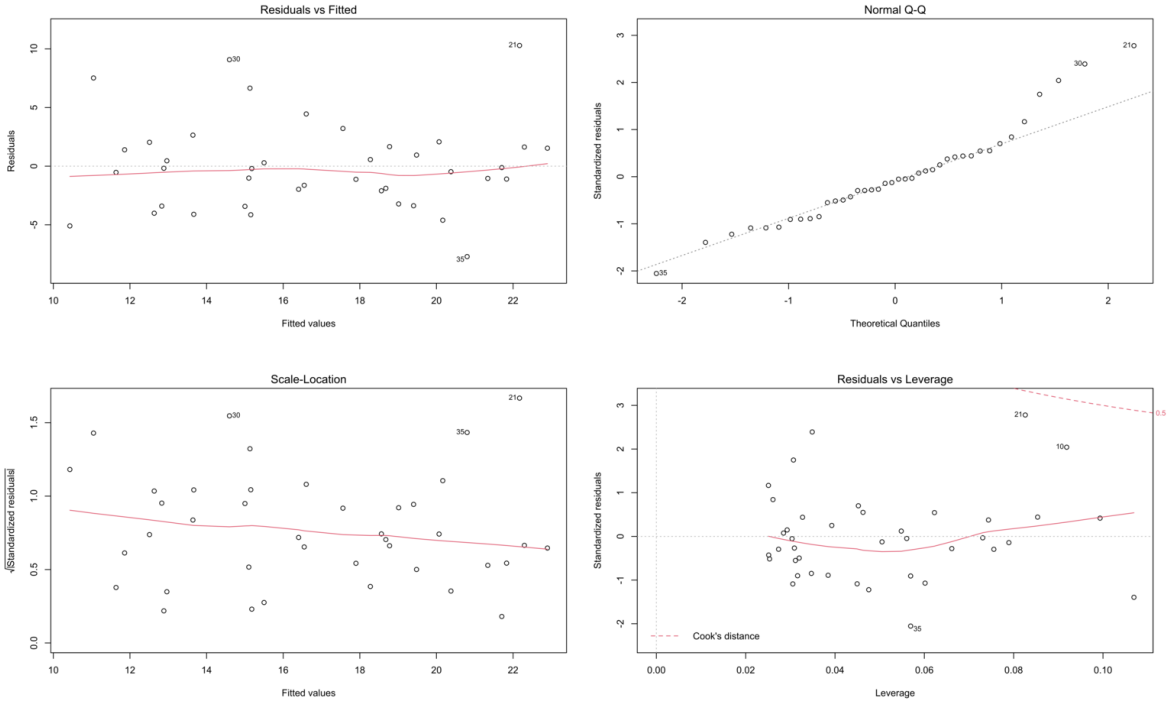

Figure 12 – *Analysis of the model : $Y \sim 1/X^2$*

A.13. Reciprocal model : summary

```
> summary(reciprocal_lm)

Call:
lm(formula = productivity ~ I(1/electricity_consumption^2), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7098 -2.3789 -0.3365  1.6413 10.2914

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     23.000      1.234  18.644  < 2e-16 ***
I(1/electricity_consumption^2) -203.822     35.320  -5.771 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.865 on 38 degrees of freedom
Multiple R-squared:  0.4671,    Adjusted R-squared:  0.453
F-statistic:  33.3 on 1 and 38 DF,  p-value: 1.175e-06
```

Figure 13 – *Summary of the model : $Y \sim 1/X^2$*

B. R code

B.1. utils.r

```
theta_1 <- 3
theta_2 <- 1

# cumulative density function
cdf <- function(x) {
  (-1 / x^3)
}

# inverse of cumulative density function
inv_cdf <- function(y) {
  (1 / ((1 - y)^(1 / 3)))
}

# generate random variables vector from the inverse cdf
inverse_transform_sampling <- function(n, inv_cdf) {
  # generate randoms numbers from the uniform distribution U(0,1)
  data_unif <- runif(n)
  rv_vector <- inv_cdf(y = data_unif)
}

# maximum likelihood method for gini coefficient estimator
gini_mle <- function(rv_vector, n) {
  return(1 / (( (2 * n) / (sum(log(rv_vector / min(rv_vector)))) ) - 1))
}

# method of moment for gini coefficient estimator
gini_mme <- function(rv_vector, n) {
  return(1 / (( (2 * ((n * mean(rv_vector)) - min(rv_vector))) / (n * (mean(rv_vector) - min(rv_vecto
}

gini_theoretical <- function(theta_1) {
  return(1 / ((2 * theta_1) - 1))
}

bias <- function(sample, theoretical) {
  mean(sample) - theoretical
}

mse <- function(sample, theoretical) {
  mean((sample - theoretical)^2)
}

# x: simulation of sample size n
compute_statistical_quantities <- function(x, n) {
  mean <- mean(x)

  gini_mle_estimator <- gini_mle(rv_vector = x, n = n)
  gini_mme_estimator <- gini_mme(rv_vector = x, n = n)

  c(
```

```
    mean,
    gini_mle_estimator,
    gini_mme_estimator
  )
}

# N: simulation size (i.e. number of samples)
# n: sample size
# f: function to generate random variables
# ... any other parameters given to f
sim <- function(N = 1000, n = 20, f, ...) {
  # compute a matrix of random variables based on the distribution f
  # each column correspond to one simulation
  x <- matrix(f(N * n, ...), nrow = n)

  # for each column (i.e. each simulation of sample size n)
  # we compute statistical quantities (mean, gini estimators,...)
  # the function "FUN" is called for each column
  stats <- apply(
    X = x,
    MARGIN = 2,
    FUN = compute_statistical_quantities,
    n = n
  )

  rownames(stats) <- c("mean-sample", "gini-mle-sample", "gini-mme-sample")

  return(stats)
}
```

B.2. code for part. 1

```
library(tidyverse)
library(reshape2)
library(latex2exp)

setwd("/Users/mathieu/Lab/company-electricity-statistical-analysis/src/")
source("utils.r")

# set a seed for reproductability
set.seed(42)

# Generate sample of size n = 20 by using inverse transform sampling
rv_vector <- inverse_transform_sampling(n = 20, inv_cdf = inv_cdf)

# plot an histogram of the random variable vector
par(mfrow = c(1,1))
hist(rv_vector, breaks = 50, freq = FALSE, xlab = "X", main = "random sample")

# compute Gini coefficients
gini_mle(rv_vector = rv_vector, n = 20)
gini_mme(rv_vector = rv_vector, n = 20)
```

```r
# Generate N = 1000 times the sample
x <- sim(N = 1000, n = 20, f = inverse_transform_sampling, inv_cdf)

gini_mle_sample <- x["gini-mle-sample", ]
gini_mme_sample <- x["gini-mme-sample", ]

# histogram of gini samples
par(mfrow = c(1, 2))
hist(gini_mle_sample, breaks = 50, main = "", xlab = TeX(r"($\hat{G}_{\textrm{MLE}}$)"), col = "darko
abline(v = gini_theoretical(theta_1), col = "darkgreen", lty = 2)
hist(gini_mme_sample, breaks = 50, main = "", xlab = TeX(r"($\hat{G}_{\textrm{MME}}$)"), col = "steel
abline(v = gini_theoretical(theta_1), col = "darkgreen", lty = 2)
legend("topright", c("Gini MLE", "Gini MME", "True Gini"), fill = c("darkorange", "steelblue", "darkg

# boxplot of gini samples
par(mfrow = c(1, 2))
boxplot(gini_mle_sample, col = "grey", ylab = TeX(r"($\hat{G}_{\textrm{MLE}}$)"))
abline(h = gini_theoretical(theta_1), col = "darkgreen", lty = 2)
boxplot(gini_mme_sample, col = "grey", ylab = TeX(r"($\hat{G}_{\textrm{MME}}$)"))
abline(h = gini_theoretical(theta_1), col = "darkgreen", lty = 2)

bias_mle <- bias(sample = gini_mle_sample, theoretical = gini_theoretical(theta_1))
bias_mme <- bias(sample = gini_mme_sample, theoretical = gini_theoretical(theta_1))

variance_mle <- var(gini_mle_sample)
variance_mme <- var(gini_mme_sample)

mse_mle <- mse(sample = gini_mle_sample, theoretical = gini_theoretical(theta_1))
mse_mme <- mse(sample = gini_mme_sample, theoretical = gini_theoretical(theta_1))

print(bias_mle)
print(variance_mle)
print(mse_mle)

print(bias_mme)
print(variance_mme)
print(mse_mme)

sample_sizes <- c(20, 40, 60, 80, 100, 150, 200, 300, 400, 500)

statistical_quantities <- c("gini-bias-mle", "gini-bias-mme", "gini-variance-mle", "gini-variance-mme

matrix <- matrix(NA, nrow = length(statistical_quantities), ncol = length(sample_sizes))
rownames(matrix) <- statistical_quantities
colnames(matrix) <- sample_sizes

# create the different samples, one for each sample size n
for (n in sample_sizes) {
  x <- sim(N = 1000, n = n, f = inverse_transform_sampling, inv_cdf)

  gini_mle_sample <- x["gini-mle-sample", ]
  gini_mme_sample <- x["gini-mme-sample", ]
```

```
  bias_mle <- bias(sample = gini_mle_sample, theoretical = gini_theoretical(theta_1))
  bias_mme <- bias(sample = gini_mme_sample, theoretical = gini_theoretical(theta_1))

  variance_mle <- var(gini_mle_sample)
  variance_mme <- var(gini_mme_sample)

  mse_mle <- mse(sample = gini_mle_sample, theoretical = gini_theoretical(theta_1))
  mse_mme <- mse(sample = gini_mme_sample, theoretical = gini_theoretical(theta_1))

  matrix[, as.character(n)] <- c(bias_mle, bias_mme, variance_mle, variance_mme, mse_mle, mse_mme)
}

# create dataframes
bias_df <- data.frame(n = sample_sizes, mle = matrix["gini-bias-mle", ], mme = matrix["gini-bias-mme"
var_df <- data.frame(n = sample_sizes, mle = matrix["gini-variance-mle", ], mme = matrix["gini-varian
mse_df <- data.frame(n = sample_sizes, mle = matrix["gini-mse-mle", ], mme = matrix["gini-mse-mme", ]

# converting to long format for easier plotting
bias_df <- melt(bias_df, id = "n")
var_df <- melt(var_df, id = "n")
mse_df <- melt(mse_df, id = "n")

par(mfrow = c(1,1))

# we compare the gini estimators through a plot as a function of n
ggplot(bias_df, aes(x = n, y = value, color = variable)) +
  geom_point() +
  geom_line() +
  labs(x = "n", y = "bias") +
  scale_fill_hue(labels = c("Gini MLE bias", "Gini MME bias")) +
  theme(legend.title = element_blank())

ggplot(var_df, aes(x = n, y = value, color = variable)) +
  geom_point() +
  geom_line() +
  labs(x = "n", y = "variance") +
  scale_fill_hue(labels = c("Gini MLE: variance", "Gini MME: variance")) +
  theme(legend.title = element_blank())

ggplot(mse_df, aes(x = n, y = value, color = variable)) +
  geom_point() +
  geom_line() +
  labs(x = "n", y = "mean squared error") +
  scale_fill_hue(labels = c("Gini MLE: mse", "Gini MME: mse")) +
  theme(legend.title = element_blank())

sim_n20 <- sim(N = 1000, n = 20, f = inverse_transform_sampling, inv_cdf)
result_n20 <- sqrt(20) * (sim_n20["gini-mle-sample", ] - gini_theoretical(theta_1))

sim_n100 <- sim(N = 1000, n = 100, f = inverse_transform_sampling, inv_cdf)
result_n100 <- sqrt(100) * (sim_n100["gini-mle-sample", ] - gini_theoretical(theta_1))

sim_n500 <- sim(N = 1000, n = 500, f = inverse_transform_sampling, inv_cdf)
```

```
result_n500 <- sqrt(500) * (sim_n500["gini-mle-sample", ] - gini_theoretical(theta_1))

par(mfrow = c(1, 3))
hist(result_n20, breaks = 50, main = "sample size: n = 20", xlab = TeX(r"($\sqrt{n}(\hat{G}_{\textrm
hist(result_n100, breaks = 50, main = "sample size: n = 100", xlab = TeX(r"($\sqrt{n}(\hat{G}_{\textr
hist(result_n500, breaks = 50, main = "sample size: n = 500", xlab = TeX(r"($\sqrt{n}(\hat{G}_{\textr
```

B.3. code for part. 2

```
library(tidyverse)

setwd("/Users/mathieu/Lab/company-electricity-statistical-analysis/src/")
df <- read.csv2(file = "electricity_consumption_dataset.txt", sep = ";", dec = ".")
head(df)

df <- rename(df, electricity_consumption = X, productivity = Y)

ggplot(df, aes(x = electricity_consumption, y = productivity)) +
  geom_jitter() +
  geom_smooth() +
  ggtitle("Productivity vs Electricity consumption") +
  xlab("Electricity consumption (MWh)") +
  ylab("Productivity (1000 euros / day)") +
  theme(plot.title = element_text(hjust = 0.5))

simple_lm <- lm(productivity ~ electricity_consumption, data = df)

par(mfrow = c(2,2))
plot(simple_lm)

summary(simple_lm)

ggplot(df, aes(electricity_consumption)) +
  geom_histogram()

ggplot(df, aes(productivity)) +
  geom_histogram()

# square root transformation
square_root_lm <- lm(productivity ~ sqrt(electricity_consumption), data = df)
summary(square_root_lm)

par(mfrow = c(2,2))
plot(square_root_lm)

ggplot(df, aes(x = electricity_consumption, y = productivity)) +
  geom_point(alpha = 0.7) +
  geom_point(aes(x = electricity_consumption, y = square_root_lm$fitted.values), color = "red", alpha
  geom_segment(aes(xend = electricity_consumption, yend = square_root_lm$fitted.values), color = "red

ggplot(df, aes(x = electricity_consumption, y = productivity)) +
  geom_point(alpha = 0.75) +
  geom_smooth(method = "lm", formula = y ~ sqrt(x), color = "red")
```

```r
# log transformation
log_lm <- lm(productivity ~ log(electricity_consumption), data = df)
summary(log_lm)

par(mfrow = c(2,2))
plot(log_lm)

ggplot(df, aes(x = electricity_consumption, y = productivity)) +
  geom_point(alpha = 0.7) +
  geom_point(aes(x = electricity_consumption, y = log_lm$fitted.values), color = "red", alpha = 0.5)
  geom_segment(aes(xend = electricity_consumption, yend = log_lm$fitted.values), color = "red", alpha

ggplot(df, aes(x = electricity_consumption, y = productivity)) +
  geom_point(alpha = 0.75) +
  geom_smooth(method = "lm", formula = y ~ log(x), color = "red")

ggplot(df, aes(x = log(electricity_consumption), y = productivity)) +
  geom_point(alpha = 0.75) +
  geom_smooth(method = "lm", color = "red")

ggplot(df, aes(x = log(electricity_consumption))) +
  geom_histogram()

# reciprocal transformation
reciprocal_lm <- lm(productivity ~ I(1 / electricity_consumption^2), data = df)
summary(reciprocal_lm)

par(mfrow = c(2,2))
plot(reciprocal_lm)

ggplot(df, aes(x = electricity_consumption, y = productivity)) +
  geom_point(alpha = 0.7) +
  geom_point(aes(x = electricity_consumption, y = reciprocal_lm$fitted.values), color = "red", alpha
  geom_segment(aes(xend = electricity_consumption, yend = reciprocal_lm$fitted.values), color = "red'

ggplot(df, aes(x = electricity_consumption, y = productivity)) +
  geom_point(alpha = 0.75) +
  geom_smooth(method = "lm", formula = y ~ I(1/x^2), color = "red")

ggplot(df, aes(x = 1 / (electricity_consumption)^2, y = productivity)) +
  geom_point(alpha = 0.75) +
  geom_smooth(method = "lm", color = "red")

ggplot(df, aes(x = 1 / electricity_consumption)) +
  geom_histogram()
```