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 (CDFa) $P(X \leq x_t)$,

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

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 becomes,

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

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,

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

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,

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

Using the following R code,

```r
source("src/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
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)
```

We get the following estimations,

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

(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 ?

```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", ]
```

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

# histogram of gini samples
par(mfrow = c(1, 2))
hist(gini_mle_sample, breaks = 50, main = "", xlab = "Gini MLE sample", col = "steelblue")
abline(v = gini_theoretical(theta_1), col = "green", lty = 2)
hist(gini_mme_sample, breaks = 50, main = "", xlab = "Gini MME sample", col = "red")
abline(v = gini_theoretical(theta_1), col = "green", lty = 2)
legend("topright", c("Gini MLE", "Gini MME", "True Gini"), fill = c("steelblue", "red", "green"))

# boxplot of gini samples
par(mfrow = c(1, 2))
boxplot(gini_mle_sample, col = "grey")
abline(h = gini_theoretical(theta_1), col = "brown", lty = 2)
boxplot(gini_mme_sample, col = "grey")
abline(h = gini_theoretical(theta_1), col = "brown", lty = 2)
```

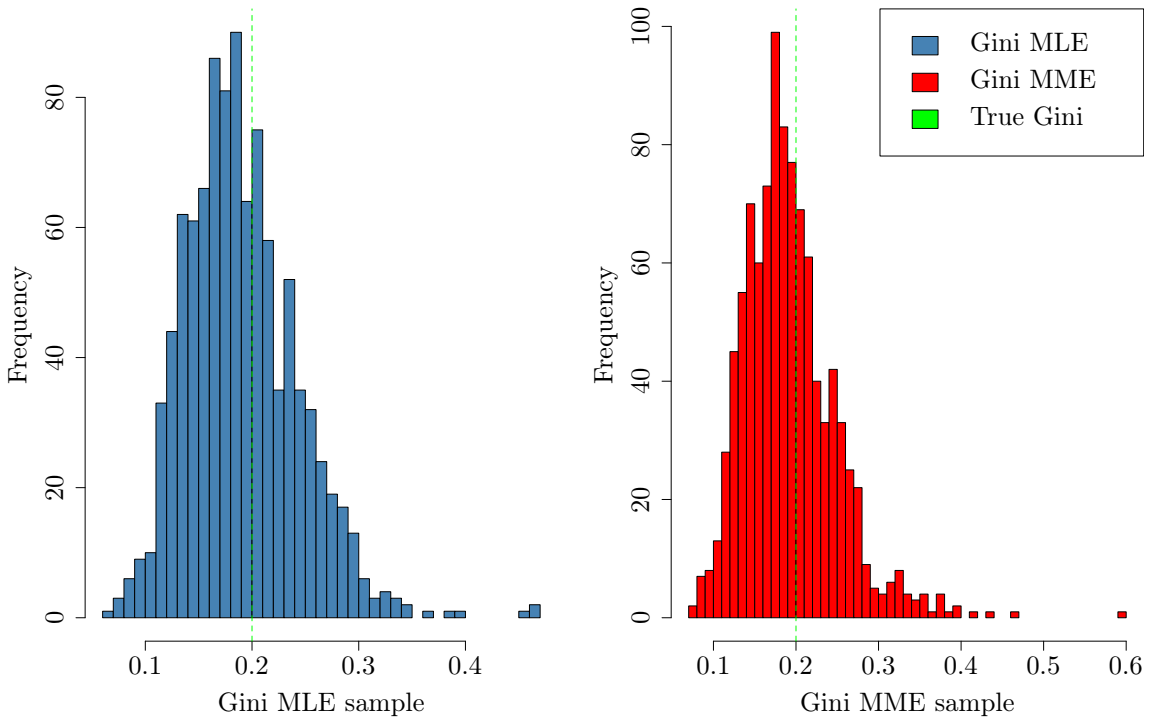


FIGURE 1 – histogramme of $N = 1000$ simulations of Gini coefficients estimated by the maximum likelihood method (*left*) and the moments method (*right*) base on a sample of size $n = 20$

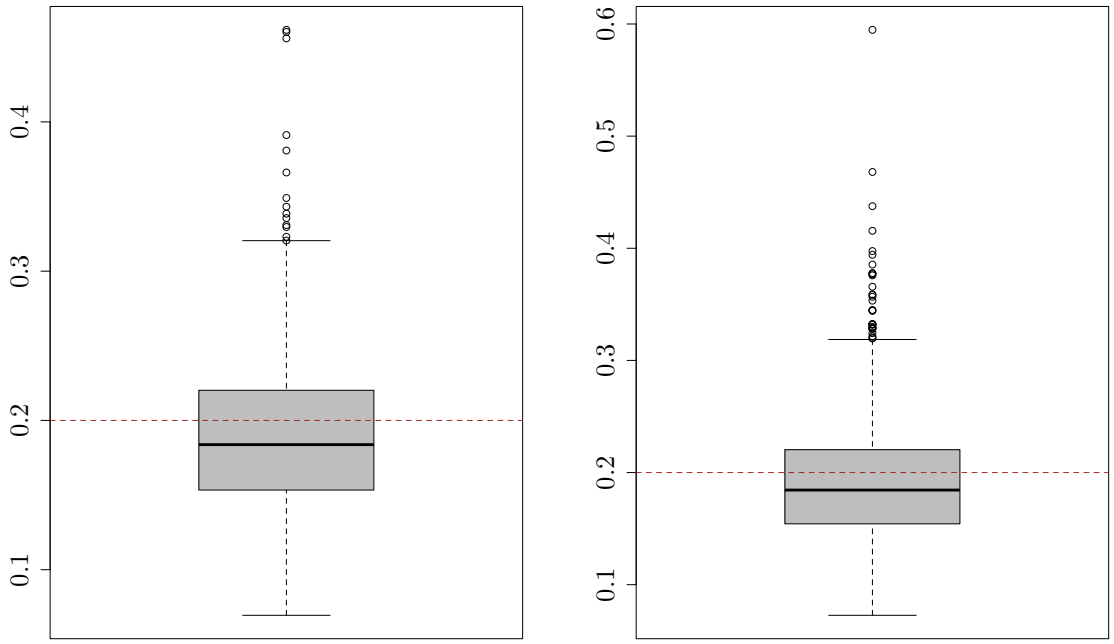


FIGURE 2 – boxplot of $N = 1000$ simulations of Gini coefficients estimated by the maximum likelihood method (*left*) and the moments method (*right*) base on a sample of size $n = 20$

As we can see the mean value of these Gini samples are pretty close to the theoretical values of the Gini coefficient. 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 outliers on the right.

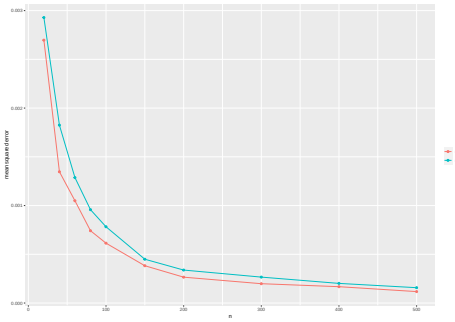
(g) Use the samples obtained in (f) to estimate the **bias**, the **variance** and the **mean squared error (MSE)** of both estimators What can you conclude ?

Redoing the $N = 1000$ simulations and computing the different quantities, we get

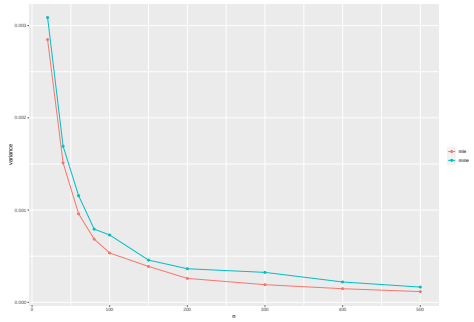
Estimator	Bias	Variance	Mean Squared Error
$\hat{G}_{\text{MLE}}$	-0.01004	0.00275	0.00284
$\hat{G}_{\text{MME}}$	-0.00769	0.00313	0.00319

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

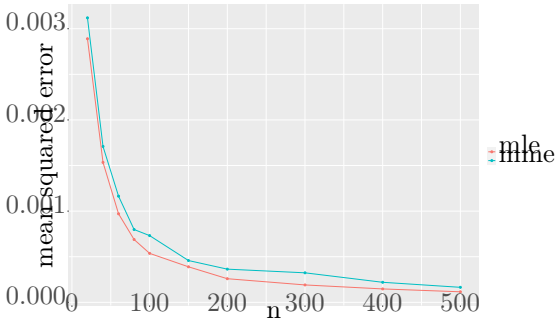
(h) Repeat the calculations in (f) for $n = 20, 40, 60, 80, 100, 150, 200, 300, 400, 500$. Compare the **biases**, the **variances** and the **mean squared errors** of both estimators graphically (make a separate plot for each quantity as a function of n). What can you conclude ? Which estimator is the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



(a)



(b)



(c)

(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 ?

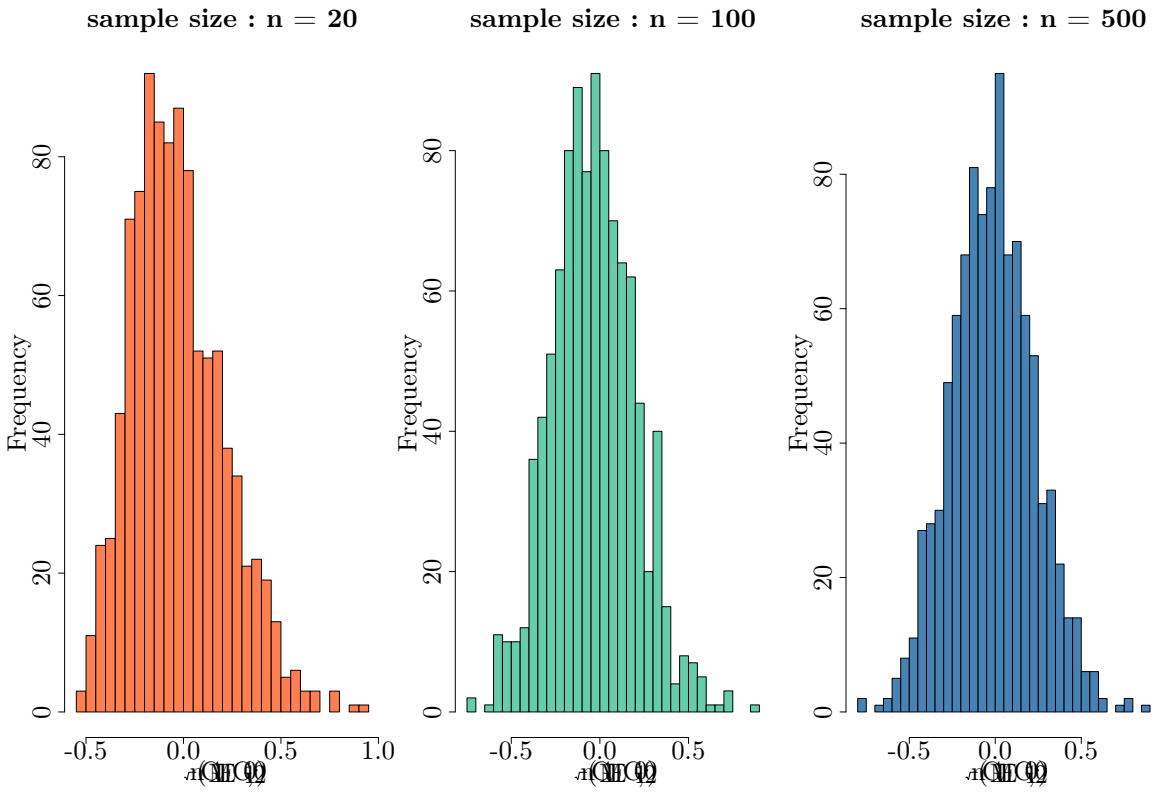


FIGURE 4 – histogram for $\sqrt{n}(\hat{G}_{\text{MLE}} - G_{\theta_1^0,\theta_2^0})$, for sample size $n = 20$, $n = 100$ and $n = 500$

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 normal distribution.

---

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 reasonnable 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** 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 $\approx 48MWh$.

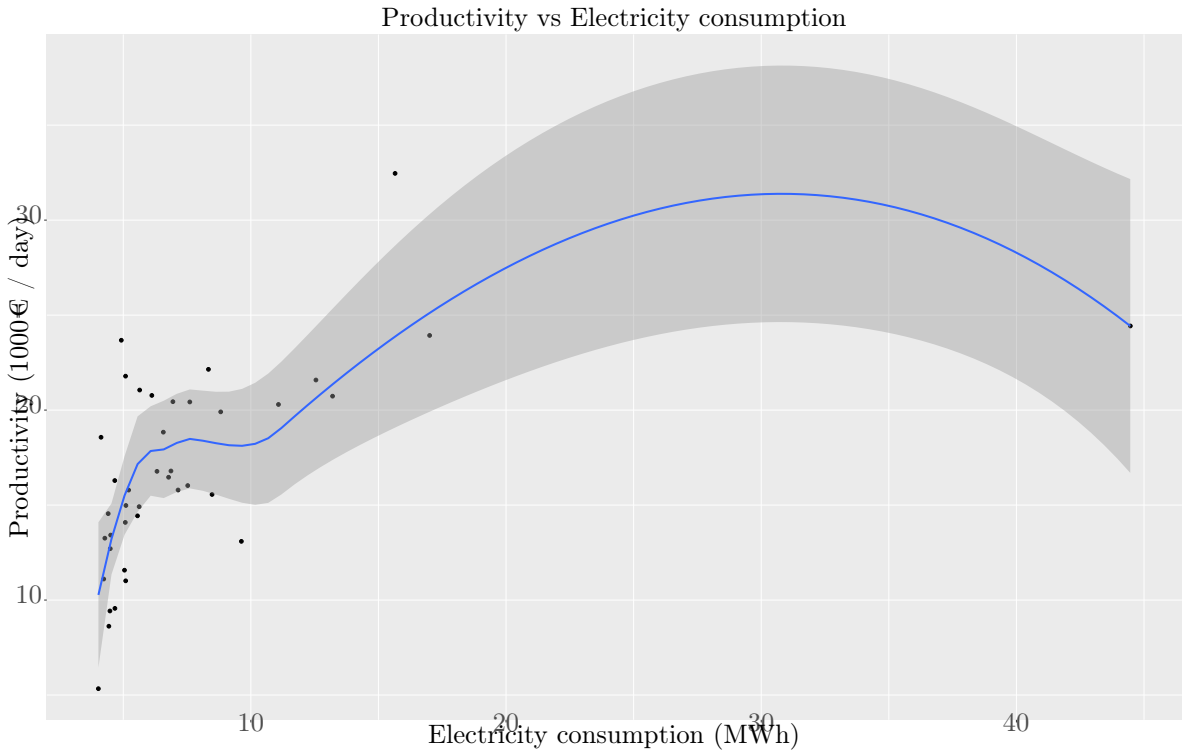


FIGURE 5 – scatter plot of productivity versus electricity consumption

If we 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 :

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

By using the following code in R,

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

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

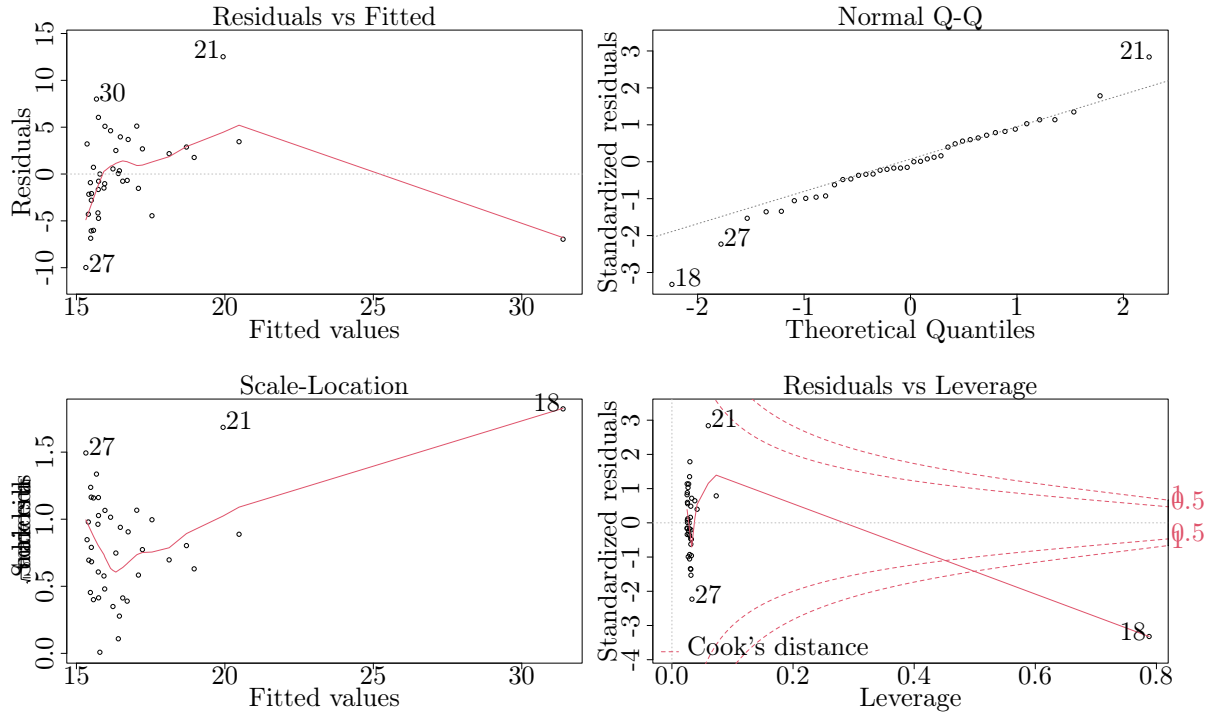


FIGURE 6 – Analysis of the linear model Y X

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.

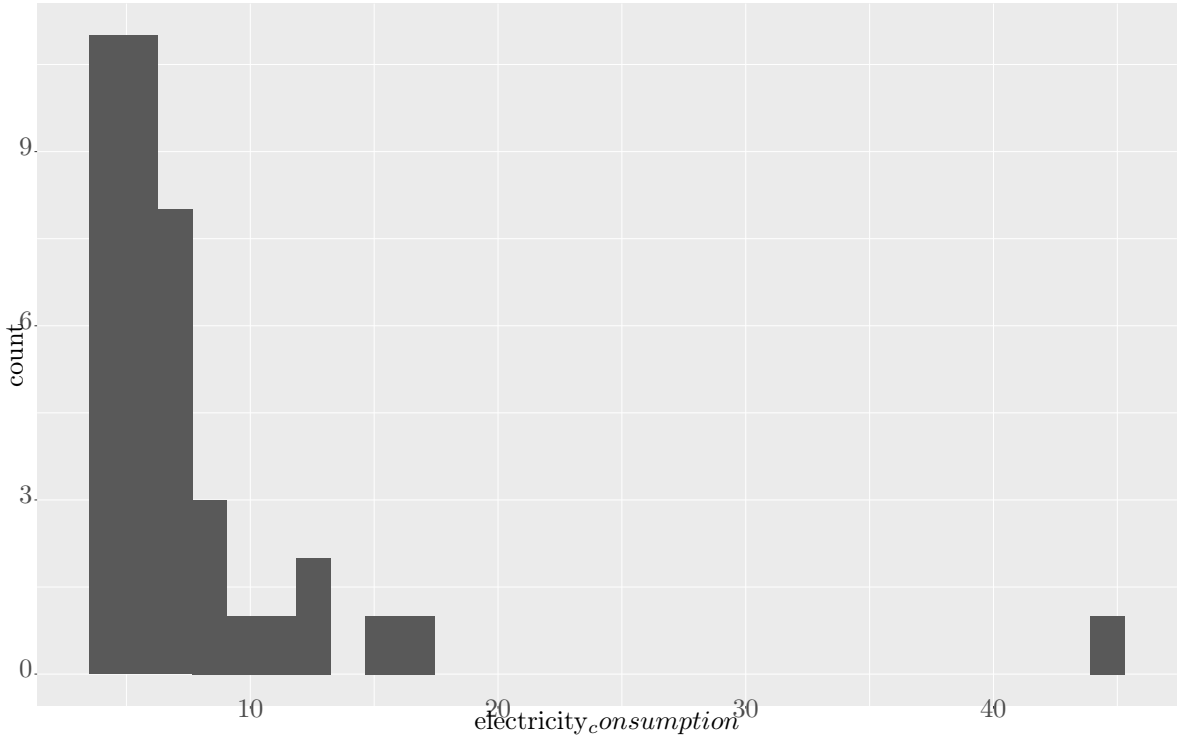


FIGURE 7 – Histogram of the explanatory variable X (electricity consumption)

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

We conclude that we cannot use a simple linear model like $Y_i \sim X_i$,

$$Y_i = \beta_0 + \beta_1 X_i, \quad \beta_0, \beta_1 \in \mathbb{R}$$

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

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

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

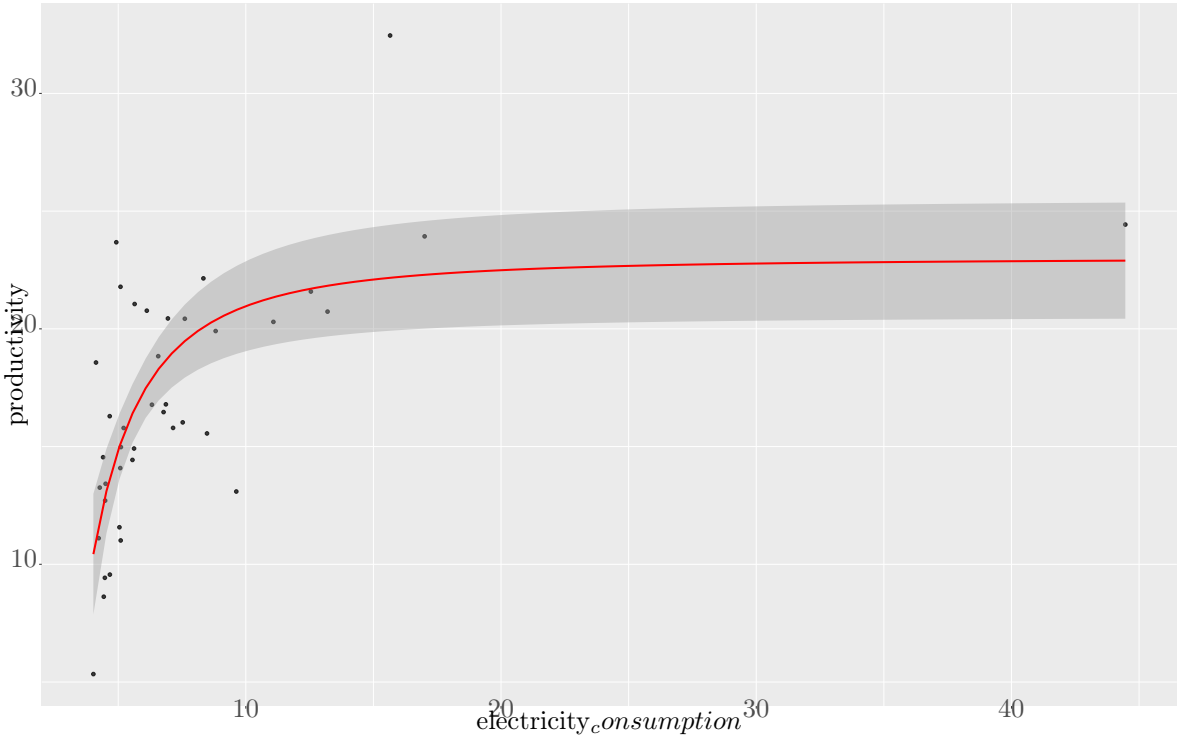


FIGURE 8 – Scatter plot of Y versus X. In red we have the following regression line : $Y_i = \beta_0 + \beta_1(1/(X_i)^2)$

Indeed,

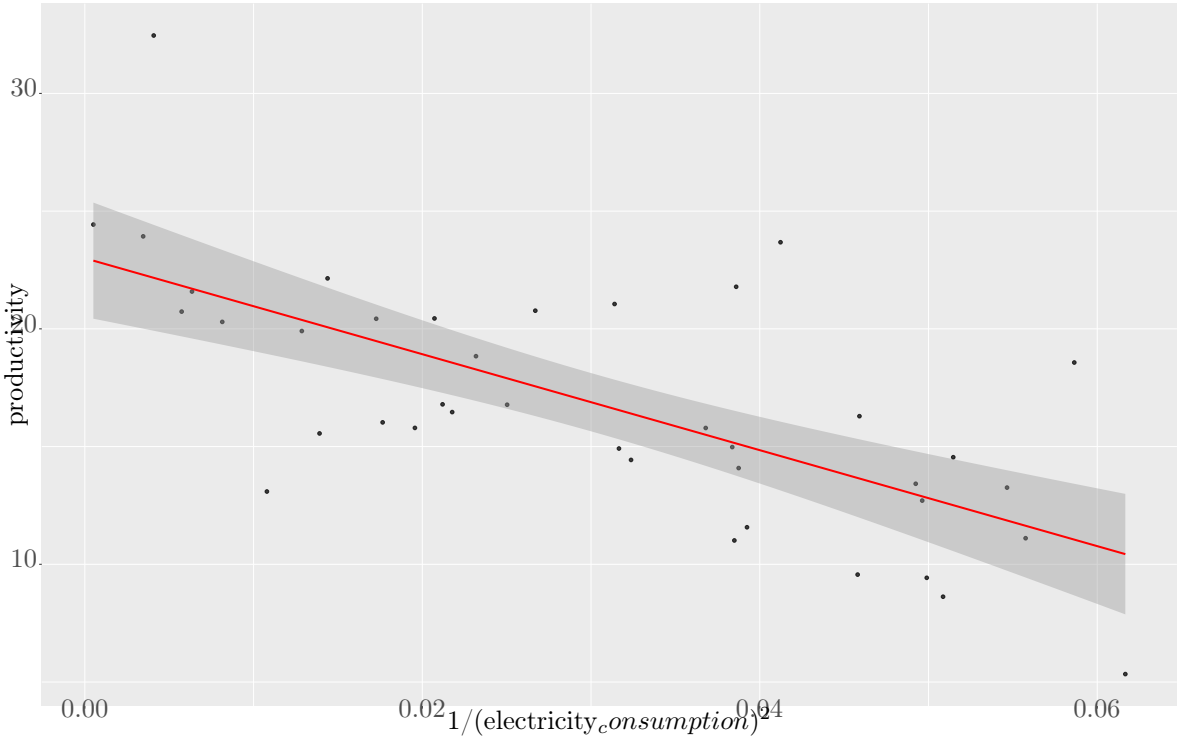


FIGURE 9 – Scatter plot of Y versus $1/X^2$. In red we have the following regression line : $Y_i = \beta_0 + \beta_1(1/(X_i)^2)$

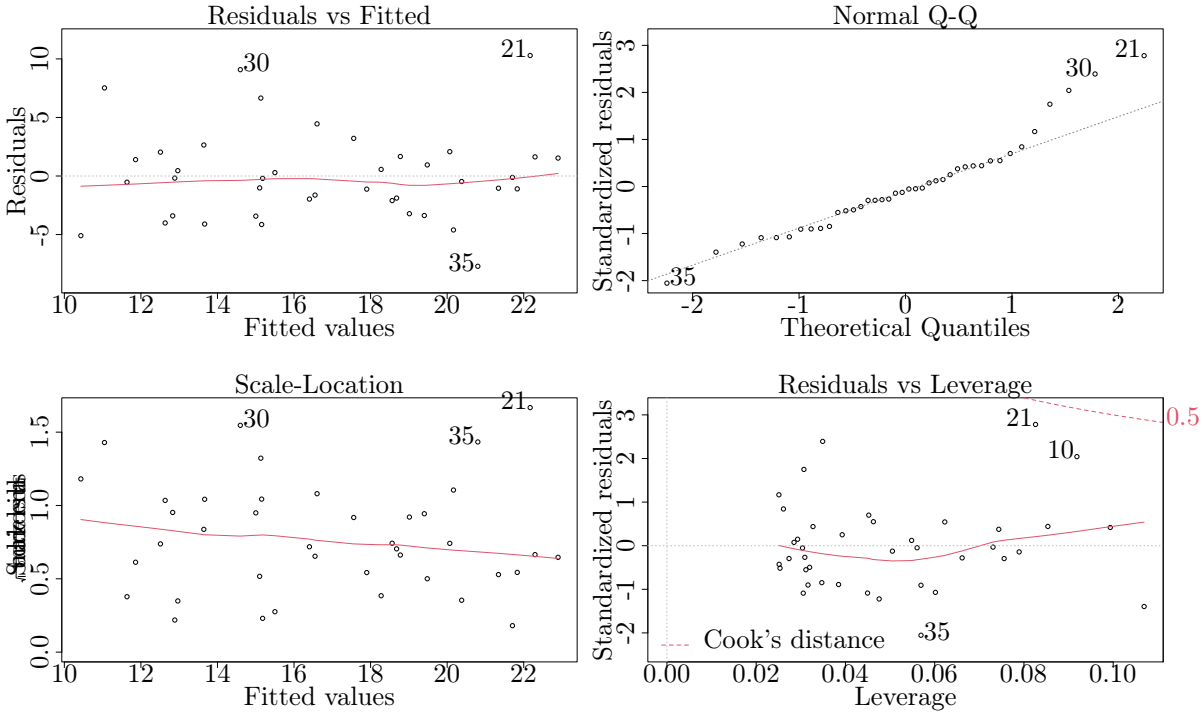


FIGURE 10 – <caption>

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,

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

Eventually, 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 too.

(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.

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

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 test is given by $Pr(> |t|)$ and is provided by the summary of the model in r (see table ...).

Looking at that summary, we see that,

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

$$Pr(\beta_1 > |t|) = 9.2 \cdot 10^{-6}$$

Therefore, β_0 and β_1 are significatively $\neq 0$ therefore we reject the null hypothesis $H_0 : \beta_i = 0$ for $i = 0, 1$ and there is a linear law between the **productivity** and the inverse square of the **electricity consumption**.

A Code R : fichier utils.r

```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_vector)
}

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)

  print(gini_mle_estimator)
  print("next simulation")
```

```
    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)
  }
```