BE de Statistiques:

Etude de la formule Breguet Indice de Sobol et modèle linéaire

RANEM Nadir

DIOP Abdoulaye

SHAUKAT Shawez Ali

MILLOGO Christopher

I - Introduction

Présentation du sujet :

On s'intéresse dans ce projet à l'étude de la formule de Breguet permettant de modéliser la consommation en fuel d'un avion donné pour un trajet. Cette formule nous est donnée par :

$$M_{fuel} = \left(M_{empty} + M_{pload}\right)\left(e^{\frac{SFC.g.Ra}{V.F}10^{-3}} - 1\right)$$

Cette formule dépend des paramètres suivants :

- $M_{empty} = 42600kg$, la masse à vide de l'avion
- $M_{pload} = 62500kg$, la masse maximale de l'avion
- $g = 9.8\ m/s^2$, la constante gravitationnelle
- $Ra = 3000km$, la distance parcourue

Et des variables aléatoires suivantes :

- V , la vitesse de croisière modélisée par une loi uniforme sur [226, 234]

- F "finesse" (modélise la qualité aérodynamique de l'avion), suit une loi Béta de paramètres (7, 2) sur [18.7, 19.05]. La densité de la loi Béta de paramètre (α,β) sur l'intervalle [a, b] est donnée pour $a \leq x \leq b$ par : $g(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}B(\alpha,\beta)}$ où B(a, b) est la fonction béta définie par $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}$

- SFC "consommation spécifique de carburant" (modélise la qualité du moteur), suit une loi exponentielle de paramètre 3.45 sur [17.23, +∞ [. La densité de la loi SFC est donnée par : $h(x) = 3.45e^{-3.45(x-17.23)}$

L'objectif de cette étude est donc de comprendre l'impact des différents paramètres qui régissent la formule de Breguet sur la consommation de fuel, en déterminant les paramètres qui impactent le plus la consommation de carburant d'un avion. Nous commencerons tout d'abord par modéliser les différentes variables aléatoires en déterminant leurs paramètres statistiques. Ensuite, nous estimerons le paramètre qui joue le plus sur la consommation à l'aide de deux méthodes différentes.

Répartition des tâches :

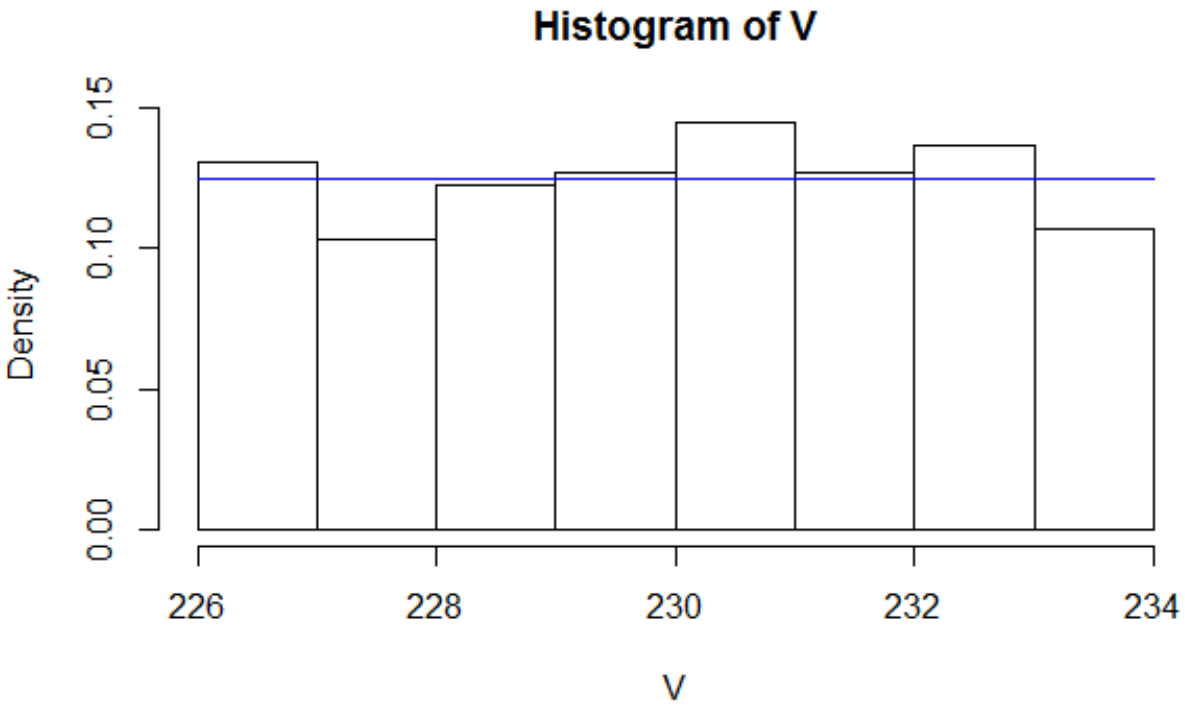
- Simulation du modèle comprenant la simulation des 3 variables aléatoires (V, F et SFC) ainsi que l'impact du bruit sur $Mfuel$.	- Nadir - Shawez - Abdoulaye - Christopher
- Estimation des indices de Sobol	- Christopher - Abdoulaye
- Test statistique de Wilcoxon pour comparer F et SFC	- Shawez - Nadir
- Recherche et estimation régression linéaire (modèle gaussien)	- Nadir - Shawez - Abdoulaye - Christopher

II - Réponses aux questions

II.1. Simulation du modèle

On cherche à déterminer l'impact du bruit sur M_{fuel}. Pour ce faire, on génère des échantillons de taille 1000 pour chaque variable *V*, *F* et *SFC*. Puis, pour modéliser ce bruitage, on ajoute à chacune des variables une quantité suivant une loi gaussienne centrée de variance σ^2.

1. a) *V* suit une loi uniforme sur [226,234]. Le logiciel R permet directement de générer un échantillon de taille N=1000 suivant la loi de *V*. On trouve alors, de manière empirique : $\begin{cases} \bar{V} = 230.005 \\ \overline{V^2} - \bar{V}^2 = 5.079 \end{cases}$.



On remarque que la répartition des valeurs est à peu près uniforme entre 226 et 234. La moyenne empirique est très proche de l'espérance ($= \frac{b+a}{2} = 230$). De même, la variance réelle vaut $\frac{(b-a)^2}{12} = 5.333$ soit un écart relatif de 4,8% avec la variance empirique.

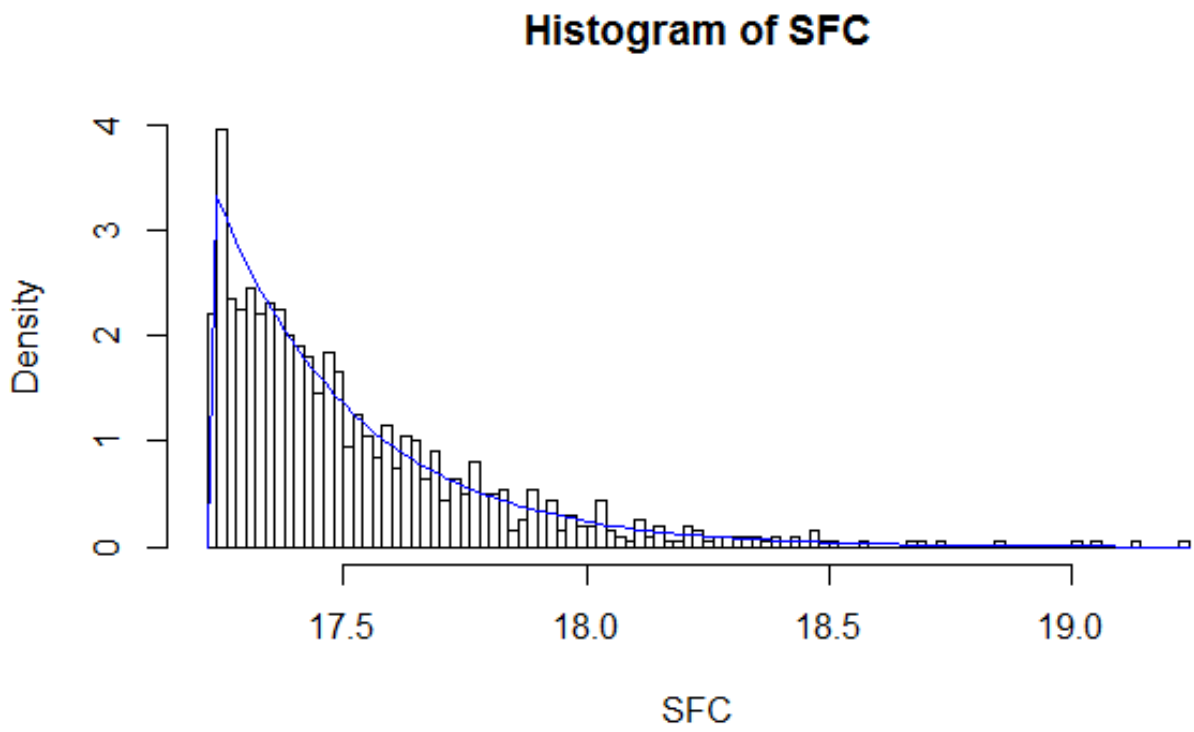
b) SFC suit une loi de densité $h(x) = 3.45e^{-3.45(x-17,23)}\mathbb{1}_{[17.23,+\infty[}$. Or sur le logiciel R les lois exponentielles sont définies uniquement sur [0,+∞[.

On pose *Y = SFC – 17.23* ∈ $[0,+\infty[$
$P(Y \leq y) = P(SFC \leq y + 17.23)$
$P(Y \leq y) = H(y + 17,23)$, H la fonction de répartition associée à SFC.

On en déduit la densité de Y :
$f_Y(y) = h(y + 17.23)\mathbb{1}_{[0,+\infty[}$
$f_Y(y) = 3.45e^{-3.45y}\mathbb{1}_{[0,+\infty[}$

D'où $Y \sim Exp(3.45)$
On génère alors un échantillon pour Y et on en déduit $SFC = Y + 17.23$.

On trouve alors, de manière empirique : $\begin{cases} \overline{SFC} = 17.520 \\ \overline{SFC^2} - \overline{SFC}^2 = 0.0825 \end{cases}$.



On remarque que la moyenne empirique est très proche de l'espérance ($= \frac{1}{3.45} + 17.23 = 17.52$). De même, la variance réelle vaut $\frac{1}{3.45^2} = 0.0840$ soit un écart relatif de 1,8% avec la variance empirique.

c) F suit une loi *Béta*(α, β) sur [a, b] avec $\alpha = 7, \beta = 2, a = 18.7$ et $b = 19.05$. La densité g vaut : $g(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}B(\alpha,\beta)}$.
Or, les lois Béta sur R sont définies uniquement sur [0,1]. On cherche donc à exprimer F en fonction d'une variable aléatoire Z de loi Béta sur [0,1].

On pose $Z = \frac{F-a}{b-a} \in [0,1]$:
$P(Z \leq z) = P(F \leq (b-a)z + a)$
$P(Z \leq z) = G\big((b-a)z + a\big)$, G fonction de répartition associé à F.

Pour tout z dans [0,1] :

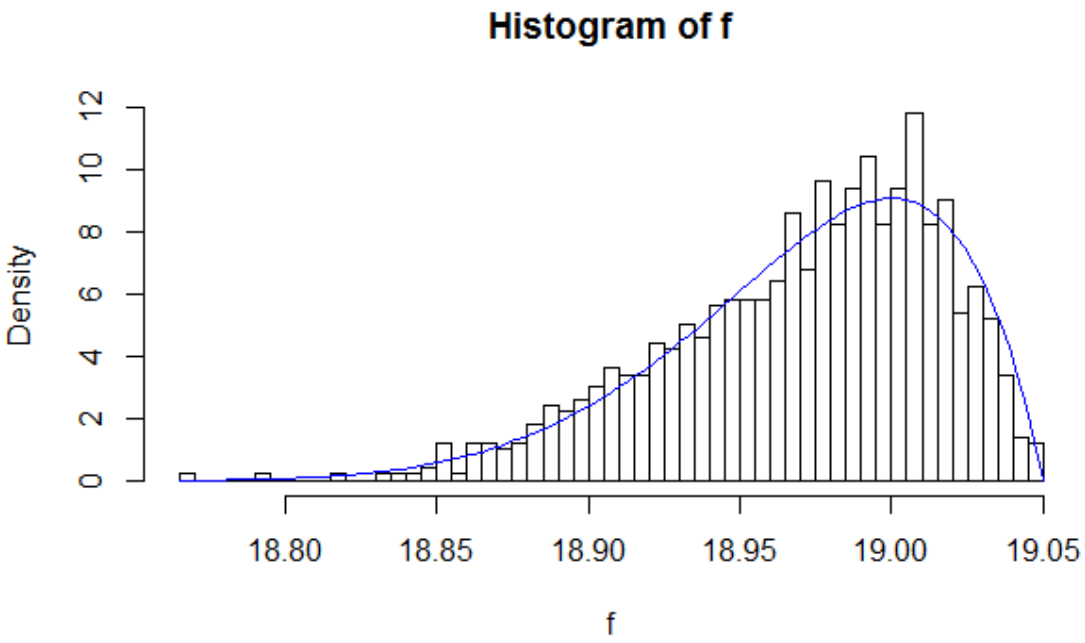
$f_Z(z) = g((b-a)z + a)(b-a)$

$$f_Z(z) = \frac{\big((b-a)z\big)^{\alpha-1}(b-a-(b-a)z)^{\beta-1}}{(b-a)^{\alpha+\beta-1}B(\alpha,\beta)}(b-a)$$

$$f_Z(z) = \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha,\beta)}$$

D'où $Z \sim Béta(\alpha,\beta)$ sur [0,1].
On génère un échantillon de taille 1000 pour Z et on en déduit :
$\mathrm{F} = (\mathrm{b}-\mathrm{a})\mathrm{z} + \mathrm{a}$

On trouve alors, de manière empirique : $\begin{cases} \bar{F} = 18.970 \\ \overline{F^2} - \bar{F}^2 = 0.00223 \end{cases}$.



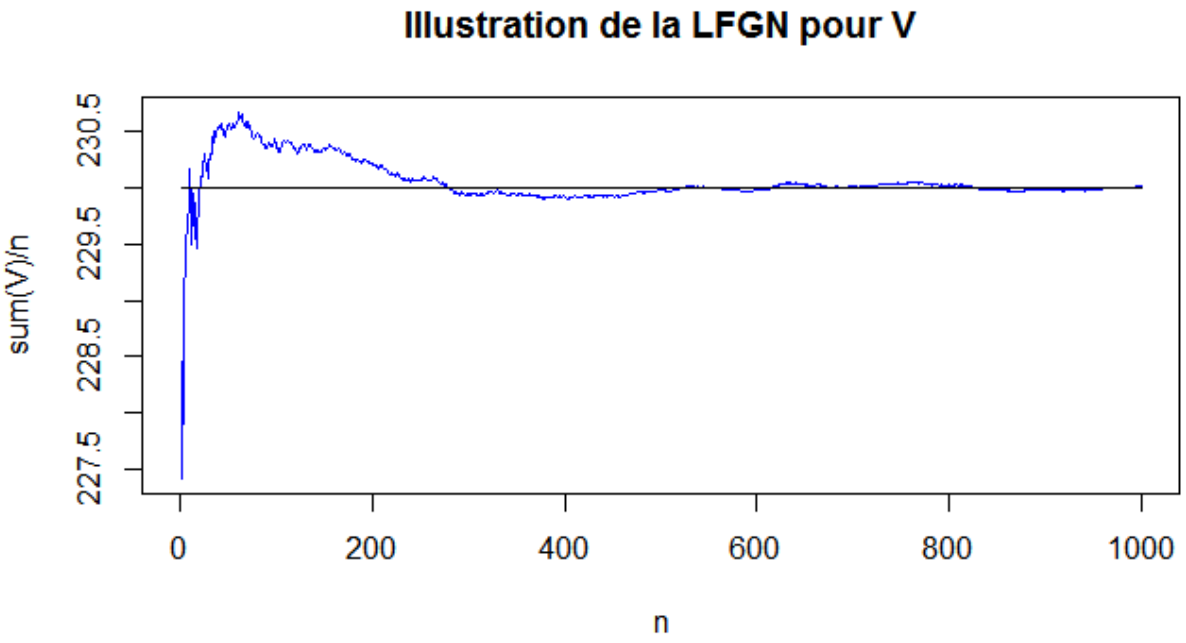
Comme auparavant, la moyenne empirique est proche de l'espérance ($= (b-a)\frac{\alpha}{\alpha+\beta} + a = 18.972$). De même, la variance réelle vaut $(b-a)^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.00212$ soit un écart relatif de 5,2% avec la variance empirique.

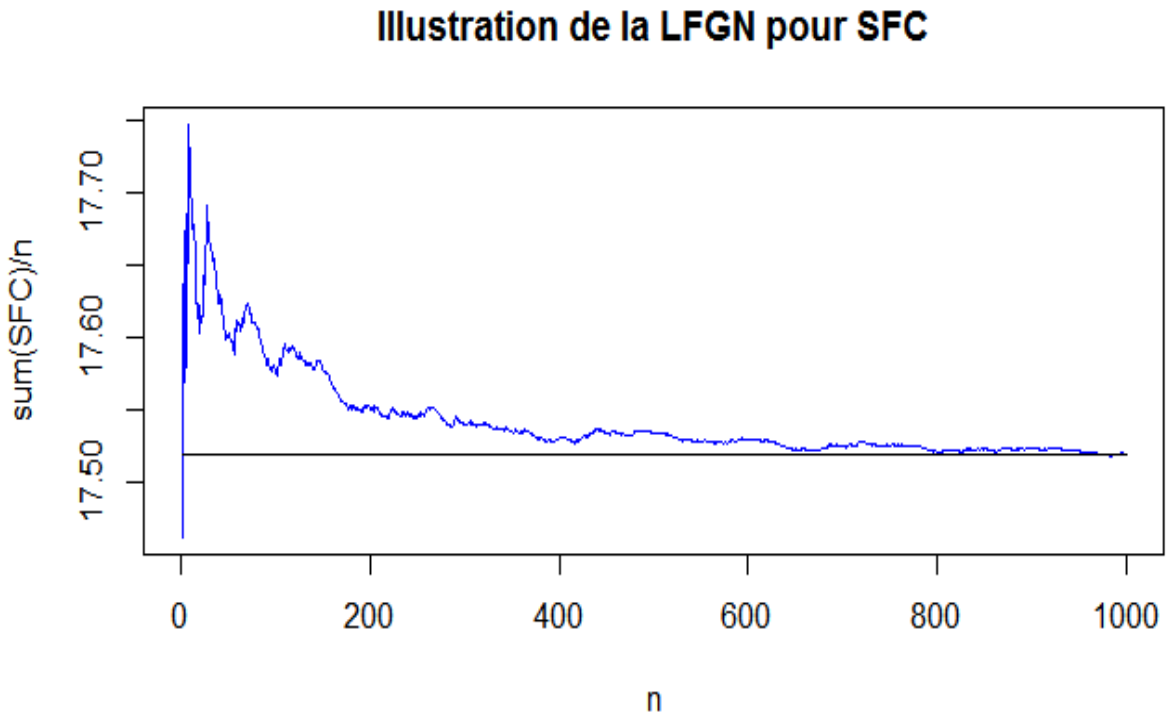
2. Selon la loi forte des grands nombres:
   Soit $(X_i)_{i\in 1..n}$ n variables aléatoires *i.i.d.* $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{p.s}{\to} E(X_1)$.

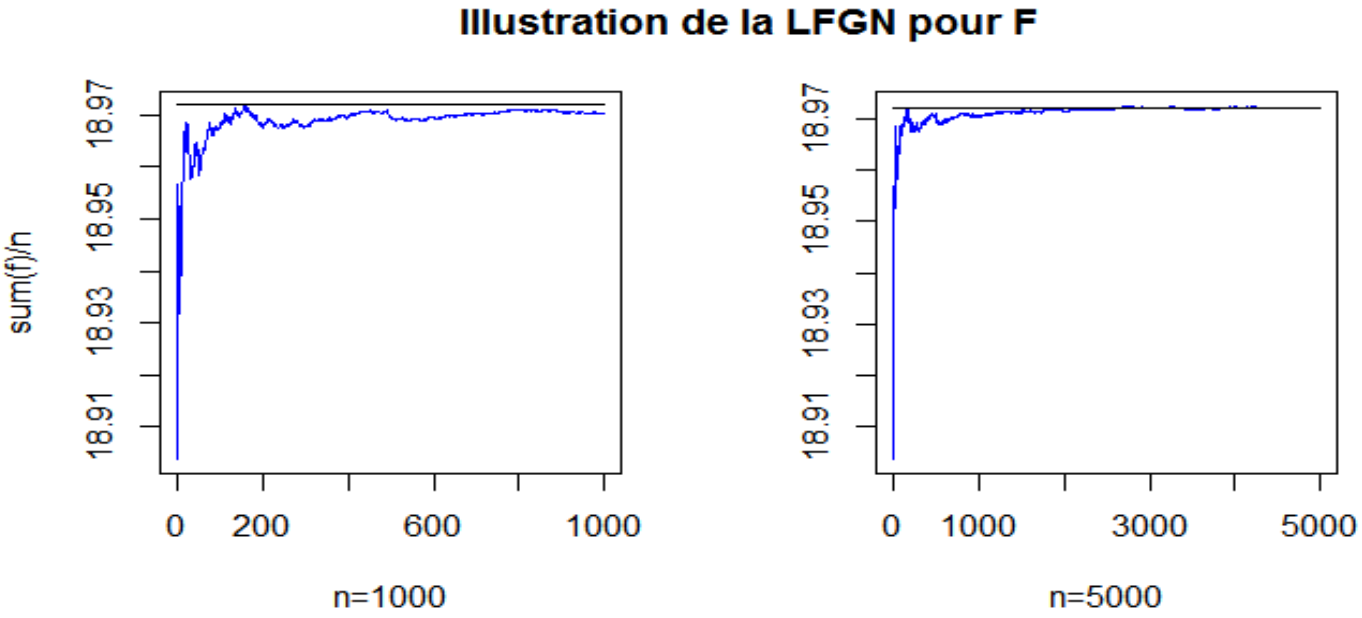
   Pour chaque variable $X \in (V, F, SFC)$, nous créons 1000 échantillons de taille n strictement croissante comprise entre 1 et 1000. Pour chaque

échantillon, nous calculons sa moyenne empirique $\overline{X_n}$. D'après la loi forte des grands nombres, la suite obtenue doit converger vers $E(X_1)$.

Rappel : $\begin{cases} E(V) = 230 \\ E(SFC) = 17.52 \\ E(F) = 18.972 \end{cases}$







Pour les deux premières courbes, un échantillon de taille 1000 a suffi pour observer la convergence de la moyenne empirique vers l'espérance. Cependant, pour F, nous avons dû augmenter la taille de l'échantillon jusqu'à 5000. Néanmoins, la LFGN est toujours validée.

3. Selon le théorème centrale limite :

   Soit $(X_i)_{i\in 1..n}$ n variables aléatoires *i.i.d.*, $\overline{X_n} \xrightarrow{\mathcal{L}} \mathcal{N}(E(X_1), \frac{Var(X_1)}{n})$.

   Pour chaque variable $X \in (V, F, SFC)$, nous créons 1000 échantillons de $\overline{X_n}$ à n fixé. D'après le TCL, pour n assez grand (n=1000 par exemple), ces 1000 échantillons doivent être distribués selon une loi normale $\mathcal{N}(E(X), \frac{Var(X)}{n})$.

   Rappel : $\begin{cases} E(V) = 230,\ Var(V) = 5.333 \\ E(SFC) = 17.52,\ Var(SFC) = 0.0840 \\ E(F) = 18.972,\ Var(F) = 0.00212 \end{cases}$

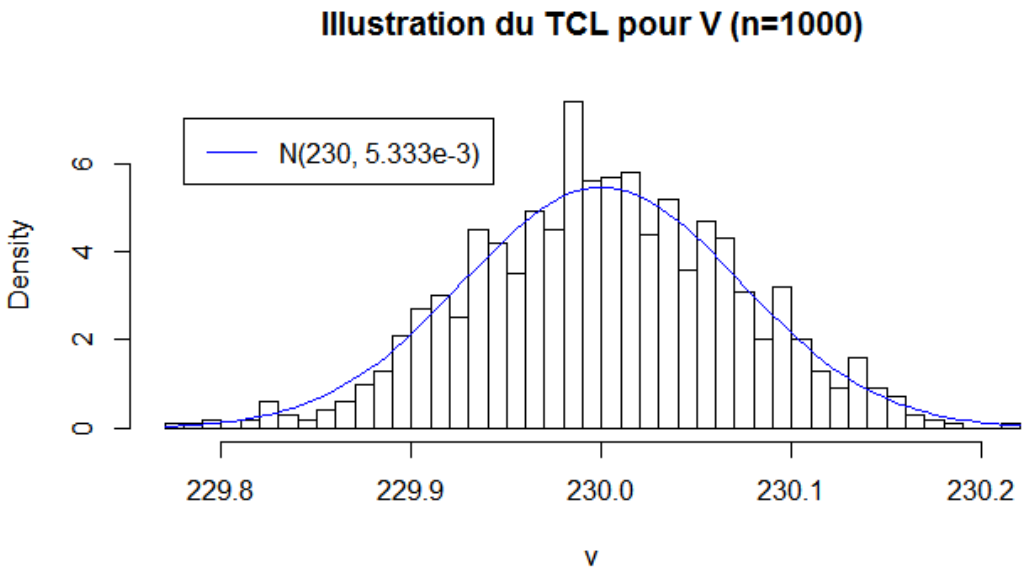


Illustration du TCL pour V (n=1000)
N(230, 5.333e-3)
Density
0
2
4
6
229.8
229.9
230.0
230.1
230.2
v

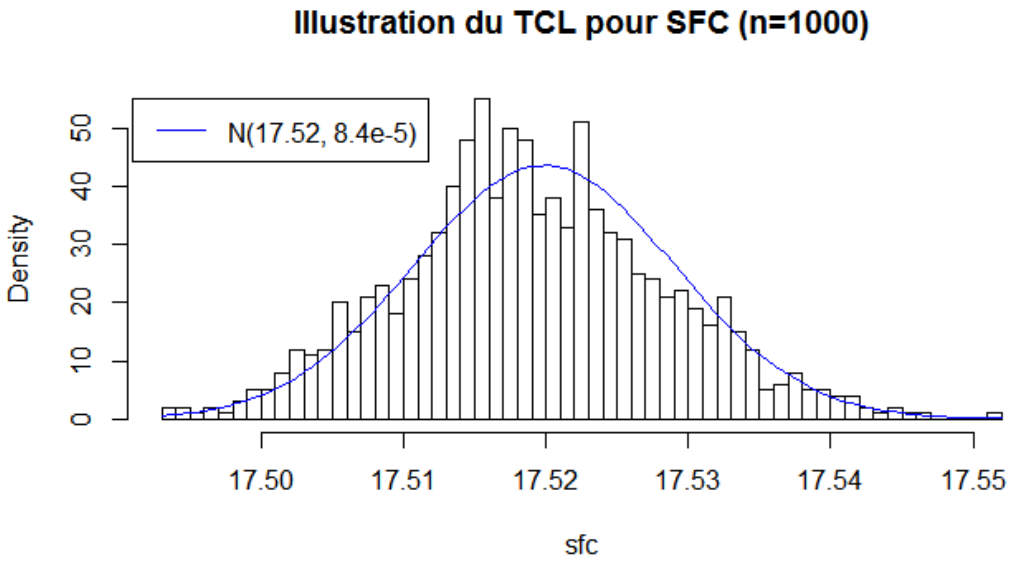


Illustration du TCL pour SFC (n=1000)
N(17.52, 8.4e-5)
Density
0
10
20
30
40
50
17.50
17.51
17.52
17.53
17.54
17.55
sfc

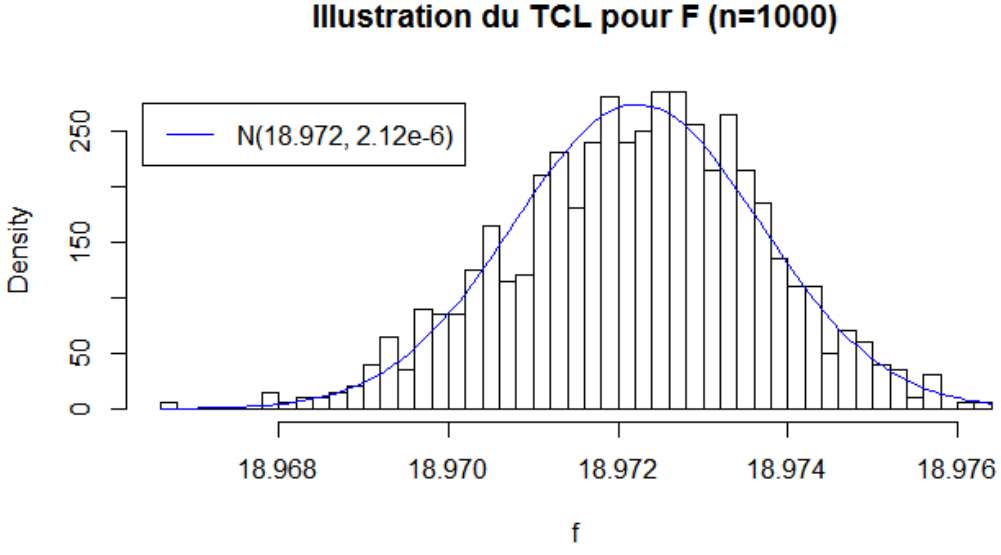
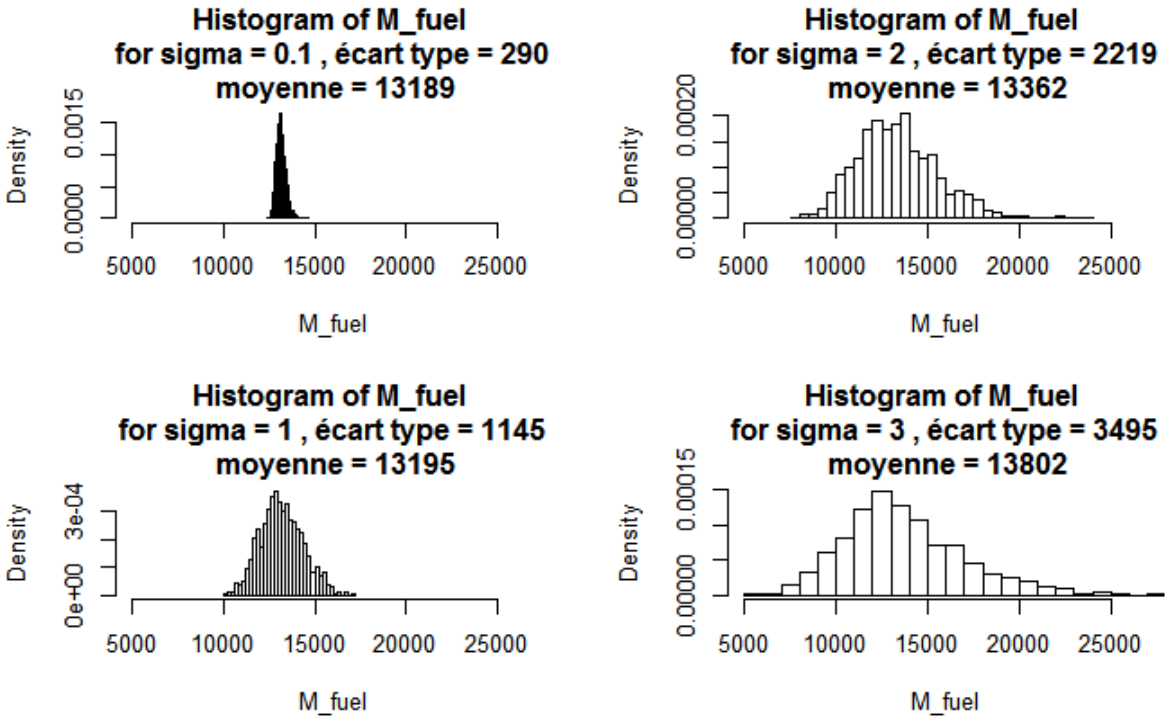


Illustration du TCL pour F (n=1000)
N(18.972, 2.12e-6)
Density
0
50
150
250
18.968
18.970
18.972
18.974
18.976
f

Nous remarquons que la distribution des moyennes empiriques de chaque variable se rapproche d'une loi normale, représenté sur chaque graphique par la courbe bleue.

4. Pour chaque variable on génère un échantillon de taille N = 1000 en tenant compte du bruit (σ). On choisit quatre valeurs possibles de σ : $\sigma \in \{0.1, 1, 2, 3\}$. On en déduit, pour chaque valeur de σ, un échantillon de M_{fuel} en utilisant la formule de Breguet.



Grâce à ces courbes, on peut faire deux observations :

- Un petit bruit entraîne de grandes incertitudes sur M_{fuel}. En effet, pour un bruit σ égal à 0.1 par exemple, l'écart type de M_{fuel} vaut 286. Or pour chaque variable $X \epsilon \{V, SFC, F\}$:

$$\left.\begin{array}{l} \frac{0.1}{\overline{X_n}} < 0{,}57\% \\ \frac{290}{13189} = 2{,}2\% \end{array}\right\} \Rightarrow \frac{écart\ relatif\ de\ M_{fuel}}{écart\ relatif\ de\ X} > 3.9$$

  Le bruit a environ 4 fois plus d'impact sur M_{fuel} que sur chaque variable.
- De plus, grâce aux 4 courbes, on remarque qu'augmenter la valeur du bruit revient à faire croître la quantité moyenne d'essence

consommé. En outre, celle-ci augment de manière « exponentielle » en fonction de σ.
En effet :

Variation de σ	$0.1 \rightarrow 1$	$1 \rightarrow 2$	$2 \rightarrow 3$
Variation de $\overline{M_{fuel}}$	6 kg	167 kg	440 kg

5. De nos jours, les appareils de mesures permettent de calculer avec une grande fidélité la finesse F, la consommation spécifique SFC et la vitesse V. Cependant, chaque appareil possède une incertitude qu'il faut minimiser, et notre étude nous a permis de déterminer la marge d'erreur acceptable. D'après nos résultats, un bruit allant jusqu'à un (σ=1) nous permet tout de même de rester précis dans le calcul de la quantité d'essence nécessaire au vol. Néanmoins, si σ prend des valeurs plus importantes (3 par exemple), on peut commettre une erreur non négligeable, de l'ordre de 400 kg en moyenne(la quantité nécessaire pour transporter un passager). En conclusion, réduire au maximum le bruit σ est primordial dans le calcul de M_{fuel}.

II.2. Indice de Sobol

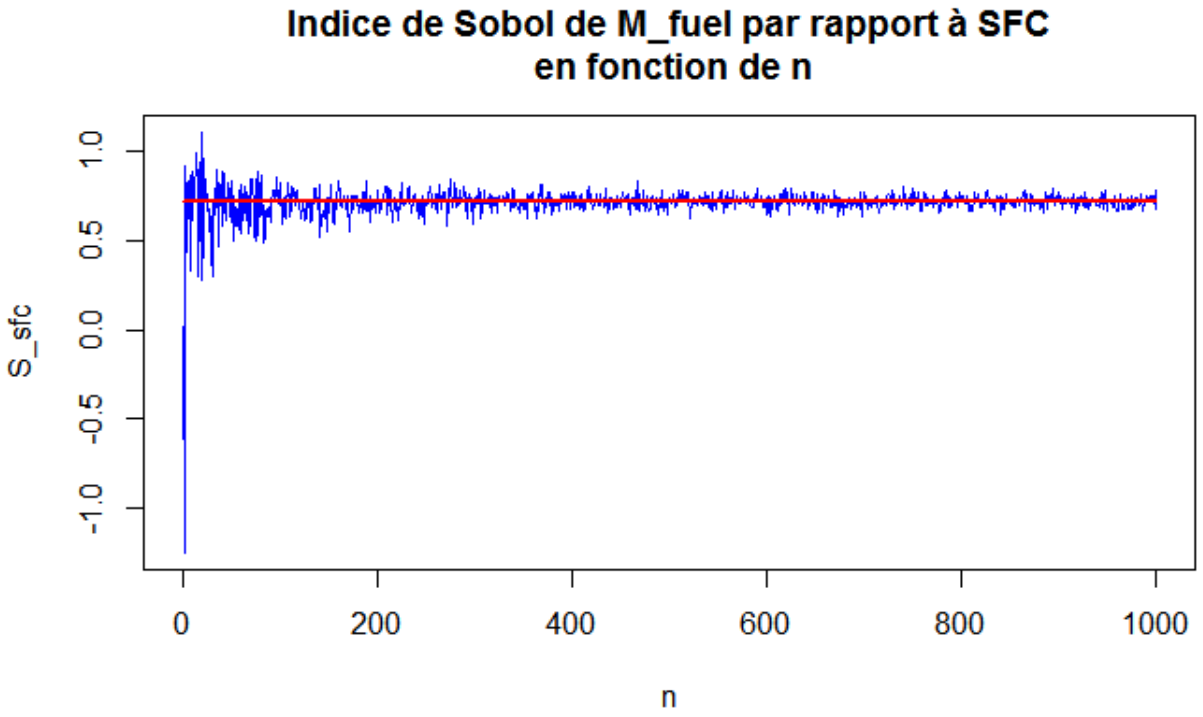
L' objectif de cette partie est de déterminer des deux variables *SFC* et *F* laquelle a le plus d'impact sur M_{fuel}. Pour ce faire, nous allons observer les indices de Sobol S^F et S^{SFC}. Puis effectuer un test statistique pour voir des deux variables laquelle prend les plus grandes valeurs. Et enfin réaliser une régression linéaire multiple.

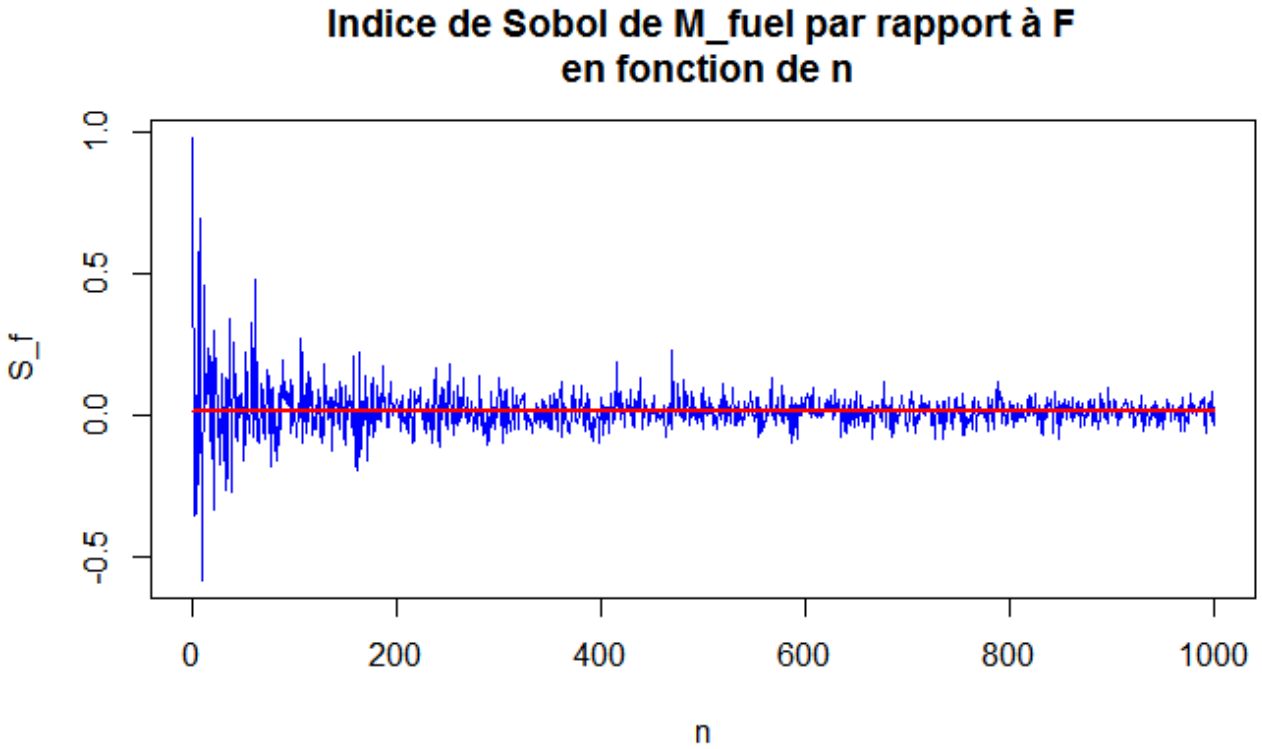
1. On génère deux échantillons *(V, F, SFC) et (V', F', SFC')* de taille n (avec $n \in [\![1, 1000]\!]$). Pour calculer S_n^F, l'indice de Sobol associé à F, on calcule :
$\begin{cases} Y = M_{fuel}(V, F, SFC) \\ Y^F = M_{fuel}(V', F, SFC') \end{cases}$.

   On en déduit une estimation de l'indice de Sobol avec la relation :

$$S_n^F = \frac{\frac{1}{n}\sum Y_i * Y_i^F - \frac{1}{n}\sum Y_i * \frac{1}{n}\sum Y_i^F}{\frac{1}{n}\sum Y_i^2 * Y_i^F - (\frac{1}{n}\sum Y_i)^2}$$

On procède de manière analogue pour calculer S_n^{SFC}.





On remarque bien une convergence des indices de Sobol S_n^F et S_n^{SFC} vers leur valeur finale respective $S^F \approx 0.01422$ et $S^{SFC} \approx 0{,}7214$. Pour F, même si l'on a de fortes oscillations autour de la limite, nous pouvons dire qu'à partir de n=400, S^F est environ égale à sa valeur finale. La convergence de S^{SFC} quant à elle est évidente dés n=300.

2. On met en place un test statistique pour tester $H_0 : F \leq SFC$ contre $H_1 : F > SFC$. Pour résoudre ce problème, on utilise la méthode de Wilcoxon. On crée deux échantillons E_1 et E_2 de taille n=1000 suivant respectivement les lois de F et de SFC. En combinant, ces deux échantillons, on obtient une liste *L* de taille 2n que l'on classe selon un ordre croissant.

   On calcule $\sum_{i=1}^{2n} Ri$, avec $\begin{cases} Ri = i \text{ si } L[i] \text{ provient de } E_2 \\ 0 \; sinon \end{cases}$

   Intuitivement, si $F \leq SFC$, on devrait avoir $\sum_{i=1}^{2n} Ri$ très grand.

   D'où la condition de rejet suivante : $\sum_{i=1}^{2n} Ri \leq K, K \in \mathbb{R}_+$.

   Pour construire un test de taille α, on cherche K tel que :

$$P(rejeter\ H_0 \text{ à } tort) = P_{H_0}\left(\sum_{i=1}^{2n} Ri \leq K\right) = \alpha$$

   Or d'après le théorème 4.3. du cours, sous l'hypothèse H_0 :

$$U = \frac{\sum_{i=1}^{2n} Ri - \frac{n(2n+1)}{2}}{\sqrt{\frac{n^2(2n+1)}{12}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

   On fait ici l'approximation (2n=2000): $U \sim \mathcal{N}(0,1)$.

$$\text{Donc} \begin{cases} P_{H_0}\left(\sum_{i=1}^{2n} Ri \leq K\right) = P\left(U \leq \frac{K - \frac{n(2n+1)}{2}}{\sqrt{\frac{n^2(2n+1)}{12}}}\right) = \alpha \\ K = \Phi^{-1}(\alpha)\sqrt{\frac{n^2(2n+1)}{12}} + \frac{n(2n+1)}{2} \end{cases}$$

On obtient :

α	$\Phi^{-1}(\alpha)$	K
0.01	-2.326	970500
0.05	-1.645	979300
0.1	-1.282	983900
0.15	-1.036	987100

Avec R on trouve pour $\sum_{i=1}^{2n} Ri = 496506 \leq K(\alpha = 0.01)$. On rejette donc l'Hypothèse H_0 avec une probabilité $\alpha = 0.01$ d'avoir tort.
D'ailleurs pour vérifier la robustesse de nos résultats nous avons répété l'opération 10 fois, mais on trouvait toujours $\sum_{i=1}^{2n} Ri \leq K(\alpha = 0.01)$. Finalement, on peut affirmer sans trop de risque : $SFC \leq F$.

En introduisant le bruit, on obtient :

σ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\sum_{i=1}^{2n} Ri$	492535	482702	471468	448825	423906	403714	385807	354417	354037	337910

On observe que plus le niveau de bruit est important, plus $\sum_{i=1}^{2n} Ri$ est petit donc plus on est amené à rejeter l'Hypothèse H_0.

3. D'après notre étude sur les indices de Sobol, on a $S^{SFC} > S^F$. On en déduit donc que les variations de SFC ont plus d'impact sur la consommation de carburant que celles de F même si SFC prend des valeurs plus petites que F. En effet, on a pu montrer grâce à un test statistique inspiré du test de Wilcoxon que $SFC \leq F$.

II.3. Modèle linéaire

II.4. Génération des données

Nous avons pu comparer dans les parties précédentes l'influence de *F* et *SFC* sur M_{fuel} de 2 manières différentes : grâce aux calculs des indices de Sobol S_F et S_{SFC} et à la mise en place d'un test visant à comparer les valeurs des deux lois. La dernière étude que nous allons entreprendre se fera à partir d'une linéarisation de $M_{fuel}(V, F, SFC)$.

1. Nous cherchons ici les 4 coefficients a_0, a_1, a_2 et a_3 tels que :

$$M_{fuel} = a_0 + a_1 V + a_2 F + a_3 SFC + \mathrm{u}$$

   Où u est une erreur gaussienne.
   Pour ce faire, nous allons créer un échantillon de taille n=1000 pour chaque variable *F*, *V* et SFC (chacune admettant un bruit σ=0.01). Nous stockerons ces valeurs dans un vecteur *X* de 1000 lignes et de 4 colonnes (la première colonne ne contenant que des 1 afin d'estimer a_0). Grâce à la formule de Breguet, nous pourrons définir un deuxième vecteur *y* tel que pour tout *i* dans $[\![1; 1000]\!]$:

$$y[i] = Mfuel(X(i,2), X(i,3), X(i,4))$$

   Définissons le vecteur *b :*

$$b = (a_0, a_1, a_2, a_3)^T$$

   L'estimation par **minimisation du critère des moindres carrés** permet de trouver ce vecteur grâce à la formule :

$$b = (X^T X)^{-1} X^T y$$

   De plus, cette estimation est sans biais et sa variance est minimale. Un tel estimateur est alors dit "BLUE" : *best linear unbiased estimator.*

2. Nous allons maintenant définir deux jeux de coefficients *b* et *b'*. Le vecteur *b* sera construit à partir des 500 premières valeurs de *X* et de *y*, le vecteur *b'* à partir des 500 dernières. On trouve alors :

$$b = (27123.39, -60.91, -736.42, 801.48)^T$$
$$b' = (27184.83, -60.81, -737.46, 797.79)^T$$

Afin d'évaluer la précision de notre approximation, nous quantifions l'erreur entre les deux jeux de coefficients *b* et *b'* par le calcul des écarts relatifs, puis nous calculons R^2, le *coefficient de détermination*, représentatif de l'efficacité de notre régression linéaire.

Nous trouvons alors les écarts relatifs suivants :

$$\begin{cases} er_{a0} = 1.13e-3 \\ er_{a1} = 8.06e-4 \\ er_{a2} = 7.03e-4 \\ er_{a3} = 2.31e-3 \end{cases}$$

Nous pouvons affirmer devant des écarts relatifs si petits (de l'ordre de 0.1 %) que l'approximation faite sur 500 valeurs des coefficients est très bonne. Cela prouve l'efficacité de la méthode utilisée.
On considère maintenant notre premier vecteur $b = (a_0, a_1, a_2, a_3)^T$. A partir de ce jeu de coefficient, nous déterminerons R^2 pour quantifier l'exactitude de la régression linéaire.

Nous construisons tout d'abord le vecteur $\hat{y}$. Pour tout i dans ⟦1;500⟧ :

$$\hat{y}[i] = a_0 + a_1 X(i,2) + a_2 X(i,3) + a_3 X(i,4)$$

Grâce à ce vecteur, nous pouvons calculer 2 quantités : *SST* et *SSR*.

- *SST* (somme totale des carrés) est définie par la formule :

  $$SST = \|y - \bar{y}\mathbf{1}\|^2 = y'y \ - \mathrm{n}\bar{y}^2$$

  Avec $\bar{y}$, la moyenne du vecteur *y* et n=1000.

- *SSR* (somme des carrés de la régression) est définie par la formule :

  $$SSR = \|\hat{y} - \bar{y}\mathbf{1}\|^2 = \hat{y}'\hat{y} \ - \mathrm{n}\bar{y}^2$$

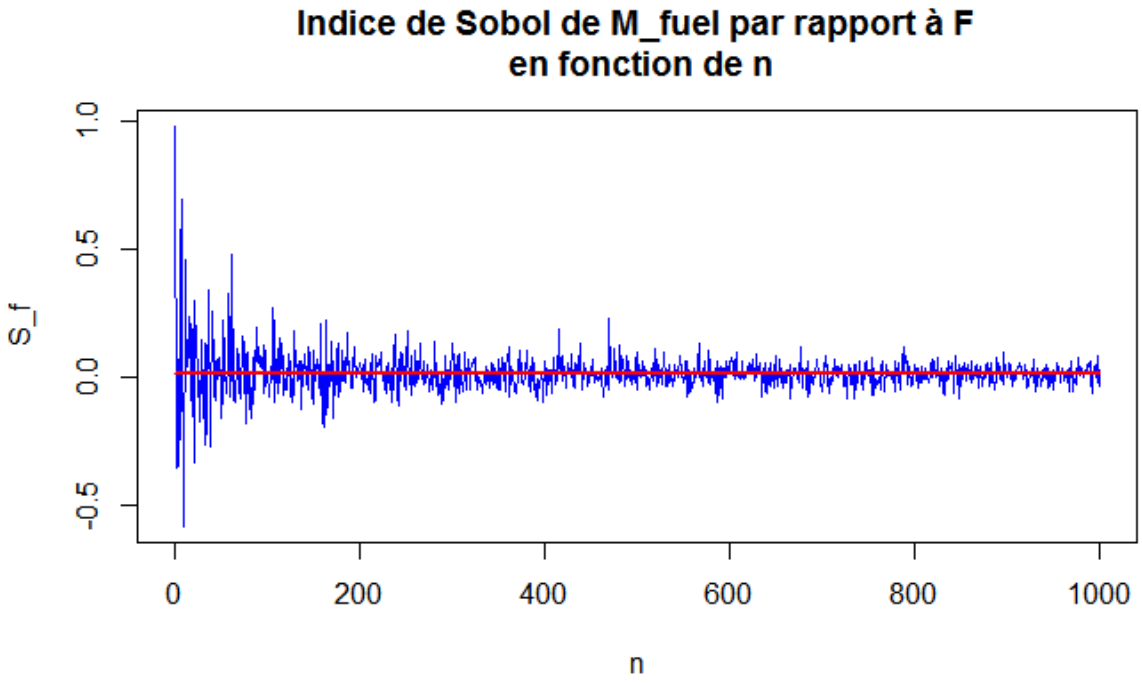
R^2 vaut alors:

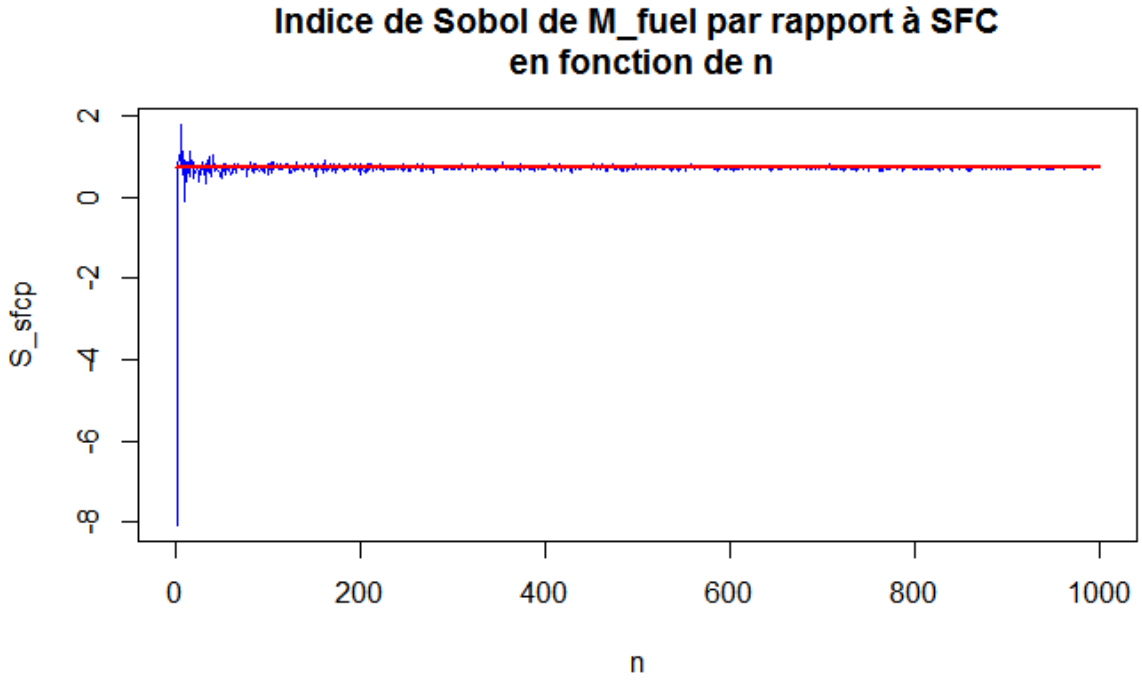
$$R^2 = \frac{SSR}{SST} = \ 0.99988$$

Cette valeur est très proche de 1, la régression linéaire est donc très bonne.
On peut donc retenir ces coefficients et affirmer que :

$$M_{fuel} = 27123.39 - 60.91\,V - 736.42\,F + 801.48\,SFC$$

3. Grâce à cette nouvelle formule, nous pouvons refaire l'étude de la section 2.2 sur le calcul des indices de Sobol, et voir si nous pouvons en tirer de nouvelles conclusions. On obtient les deux graphiques suivants :





On remarque les indices de Sobol S_n^{SFC} convergent vers la même limite S^{SFC} déjà déterminée dans la section 2.2. Les indices de Sobol S_n^F quant à elles convergent vers une limite S^F légèrement supérieure à celle déjà déterminée dans la section 2.2. Nous remarquons également que la convergence se fait plus rapidement (dès la 100ème itération pour S_{SFC} par exemple). En outre, il y a bien moins de variations autour des valeurs finales.

4. En conclusion, la linéarisation de M_{fuel} nous permet d'étudier d'une manière différente l'impact de *F* et de *SFC* sur la quantité d'essence consommée. En effet, même si *F* prend des valeurs environ supérieures à celles de *SFC*, le poids de ces valeurs, et le poids de leurs fluctuations dépendent entièrement des coefficients a_2 et a_3. C'est là tout l'intérêt de la linéarisation : On peut déterminer, en regardant uniquement son expression, l'influence des différentes variables sur M_{fuel}.
   Or, $|a_3|$ est supérieur à $|a_2|$. Par conséquent, selon cette étude, *SFC* a plus d'impact sur la quantité d'essence consommée que *F*

II.5. Conclusion générale

Ce projet s'articule essentiellement autour de la formule de Breguet. Son but principal est de quantifier l'influence de certains paramètres, notamment la finesse Fet la consommation spécifique SFC, sur la consommation totale de carburant M_{fuel}. Nous avons commencé par simuler le modèle avec des échantillons de variables aléatoires grâce au logiciel R, puis nous avons étudié l'impact du bruit sur la loi de M_{fuel}. Cela nous a permis de définir un niveau au-delà duquel l'influence du bruit devient non négligeable ($\sigma > 1$). Physiquement cela illustre l'erreur commise par les appareils de mesure dont la propagation conduit à fausser les mesures. Notre étude s'est ensuite poursuivie par l'estimation des indices de Sobol par rapport aux variables F et SFC dans le but de déterminer laquelle avait le plus d'impact sur la consommation de carburant. Les résultats ont montré que les variations de SFC avaient plus d'influences sur la consommation de carburant que celles de F même si SFC prend des valeurs plus petites que F. Pour finir nous avons utilisé un modèle permettant de linéariser M_{fuel} afin d'étudier l'impact des variables F et SFC et confirmer nos

résultats précédents, en effet, grâce au calcul des coefficients pondérant les variables F et SFC, nous avons pu conclure rapidement de l'impact supérieur qu'avait SFC sur M_{fuel}par rapport à F.

Lors de ce BE l'utilisation du logiciel R était très adapté, étant facile d'utilisation, il offrait tous les outils permettant d'aborder le problème correctement. La difficulté rencontré lors de ce projet à été de rester proche du sujet et de ne pas perdre de vu le but du problème, de surcroit on ne pouvait vraiment se partager le problème avec les membres du groupe car les questions étaient interdépendantes, et seule une compréhension totale de tout le monde fut nécessaire pour mener à bien le projet.