BE Statisitiques

Adrien Bonis, Elie Grenier, Victor Mercier, Louis Plumier

20 mai 2018

Sommaire

Table des figures

Liste des tableaux

Ce BE s'intéresse à une étude de la formule de Bréguet, relation fondamentale en aéronautique et plus particulièrement pour le calcul des performances-avion. Cette formule relie la masse de fuel avec quatre constantes (la masse à vide de l'avion, sa masse maximale, l'accélération de la pesanteur, le rayon d'action) et trois variables aléatoires (la vitesse de croisière, la finesse, la consommation spécifique). La finesse modélise la qualité aérodynamique de l'avion et la consommation spécifique la qualité du moteur.

La première partie correspond à une étude descriptive du modèle, à des résultats fondamentaux des probabilités et à une modélisation du bruitage des variables. La deuxième partie permettra grâce à l'étude des indices de Sobol de peser le poids du moteur ou de l'aérodynamisme dans la consommation en carburant. La dernière partie proposera une régression multilinéaire pour une étude sans la connaissance a priori de la formule de Breguet.

1 Simulation du modèle

1.1 Etude descriptive

Commençons par faire quelques remarques sur les variables aléatoires présentées : on peut montrer que si U suit une loi $Beta(\alpha, \beta)$ (sur $[0;1]$) alors $F = (b-a)X + a$ suit une loi $Beta(\alpha, \beta)$ sur $[a;b]$ et que $SFC = 17.23 + U'$ où U' suit une loi $Exp(3.45)$. Comme le langage R peut simuler directement les variables U et U' et pas F et SFC, ces relations seront utiles pour simuler F et SFC. Le fichier `echantillon.R` propose trois fonctions `sample_V`, `sample_F`, `sample_SFC` qui prennent en argument un entier N et renvoient une N-échantillon de la variable considérée.

Résumons dans un tableau les valeurs théoriques (arrondie) des espérances, variance et écart-type des variables aléatoires V, F [1], SFC :

Variable	Densité	Espérance	Variance	Ecart-type
V	$f(x) = \frac{1}{8}\mathbb{1}_{[226,234]}(x)$	230	5.33	2.309
F	$g(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}B(\alpha,\beta)}\mathbb{1}_{[a,b]}(x)$	18.9722	2.11×10^{-3}	0.046
SFC	$h(x) = 3.45e^{-3.45(x-17.23)}\mathbb{1}_{[17.23,+\infty[}(x)$	17.52	0.084	0.2899

Table 1 – Données théoriques sur les variables d'entrée

Le fichier `description.R` permet de décrire les réalisations des N-échantillons des variables V, F, SFC. On obtient alors pour $N = 1000$:

Variable	Moyenne empirique	Variance empirique	Ecart-type empirique	Min	Max
V	229.87	5.12	2.26	226.01	233.99
F	18.98	2.03×10^{-3}	0.045	18.80	19.04
SFC	17.51	0.078	0.28	17.23	20.08

Table 2 – Données empiriques simulées sur les variables d'entrée

En annexe se trouve les fonctions de répartition empiriques et théoriques de ces variables afin de constater l'adéquation en loi d'un grand nombre de réalisations simulées par R avec les lois théoriques.

1.2 Loi forte des grands nombres

Théorème 1 (Loi Forte des Grands Nombres (LFGN)) *Soit $(X_i)_{i\in\mathbb{N}^*}$ des variables aléatoires indépendantes et identiquement distribuées d'espérance finie alors*

$$\bar{X}_n := \frac{\sum_{k=1}^{n} X_n}{n} \xrightarrow[n \to +\infty]{p.s.} \mathbb{E}(X_1)$$

Ce théorème signifie que plus la taille du N-échantillon est grande, plus les réalisations de la moyenne empirique de l'échantillon tendent vers une constante qui est l'espérance (moyenne théorique) de la variable aléatoire dont on tire les échantillons. Pour illustrer ce théorème pour chaque variable, on choisit un N très grand (ici, $N = 1000$), on réalise les réalisations d'un 1-échantillon, d'un 2-échantillon, ... et d'un N-échantillon et on calcule pour chaque réalisation des échantillons la moyenne empirique $\bar{x}_k = \frac{x_1+...+x_k}{k}$. On exprime alors dans un graphe les moyennes empiriques des échantillons en fonction de leur taille. On obtient alors un nuage de points qui converge vers une constante égale à la moyenne théorique.

Rappelons les espérances des trois variables aléatoires de notre problème :

1. Voir annexe pour correction de la densité

Variable aléatoire	Espérance
V	230
F	18.97
SFC	17.52

TABLE 3 – Espérances théoriques des variables d'entrée

Le fichier `FLGN.R` génère les graphes 1, 2, 3 pour les trois variables aléatoires. On obtient bien une convergence des nuages de points vers les valeurs théoriques.

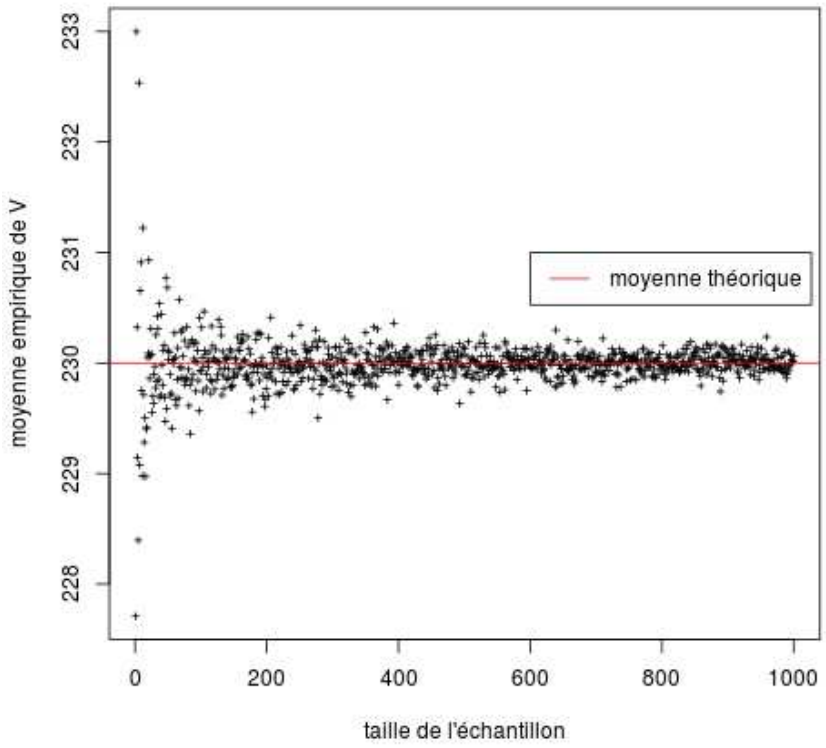


FIGURE 1 – LFGN illustrée avec V

1.3 Théorème central limite

Théorème 2 (Théorème Central Limite (TCL)) *Soit $(X_i)_{i\in\mathbb{N}^*}$ des variables aléatoires indépendantes et identiquement distribuées de variance finie σ^2 alors*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mathbb{E}(X_1)}{\sigma}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Graphiquement, ce théorème signifie que pour un N grand, l'histrogramme décrivant la répartition de la réalisation d'un K-échantillon de la variable $\bar{X}_N$ suit la courbe de la densité d'une $\mathcal{N}(\mathbb{E}(X_1), \sigma/\sqrt{N})$.

Les graphes en figures 4, 5, 6 illustrent le TCL avec les trois variables du problèmes [2]. L'histogramme 4 représente la répartition de K réalisations de $\bar{V}_N$ ($\sigma = 2.309$), l'histogramme 5 la répartition de K réalisations de $\bar{F}_N$ ($\sigma = 0.046$) et l'histogramme 6 la répartition de K réalisations de $S\bar{F}C_N$ ($\sigma = 0.2899$). On constate que la courbe verte représentant la densité d'une loi normale $\mathcal{N}(\mathbb{E}(), \sigma/\sqrt{N})$ épouse bien le contour des histogrammes, conformément au théorème. ($K = 1000$ et $N = 1000$).

2. L'échelle de l'axe des ordonnées (fréquence) est étrange car supérieure à 1. Elle est générée automatiquement par R à partir du moment où on utilise la commande `proba=T` dans `hist`. Cette bizarrerie irrésolue n'empêche pas la compréhension des graphes.

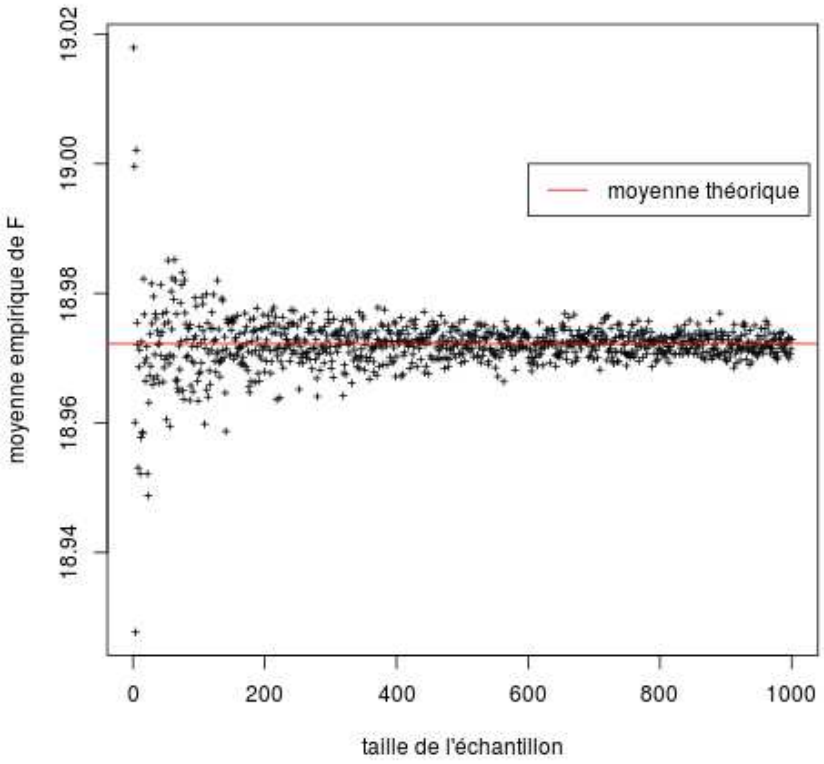


FIGURE 2 – LFGN illustrée avec F

1.4 Bruit gaussien

On suppose que les variables V, F, et SFC sont bruitées :

$$V_i^\sigma = V_i + \sigma\epsilon_i$$

$$F_i^\sigma = F_i + \sigma\mu_i$$

$$V_i^\sigma = V_i + \sigma\theta_i$$

où les variables $(\epsilon_i)_{1...N}$, $(\mu_i)_{1...N}$ et $(\theta_i)_{1...N}$ sont iid de loi $\mathcal{N}(0,1)$

Avant d'étudier l'impact du bruit sur M_{fuel}, étudions M_{fuel} sans le bruit. On génére pour cela les réalisations d'un N-échantillon de V, SFC, F et on utilise la formule de BREGUET. On obtient alors :

Variable	Moyenne empirique	Variance empirique	Ecart-type empirique	Min	Max
M_{fuel}	13626.60	85746.17	292.82	13141.29	15336.92

TABLE 4 – Données empiriques simulées sur M_{fuel}

En figure 7 est montré l'histogramme de la répartition de 10000 valeurs de M_{fuel} calculées grâce au script `description.R`. Le trait noir plein est la densité estimée par le logiciel R.

Définissons la fourchette du niveau de bruit qui sera étudiée. On peut faire l'hypothèse que le bruit à le même ordre de grandeur que les variances des trois variables. Un bruit trop grand représenterait des variations excessives et peu réalistes (si par exemple le bruit représente les imprécisions de mesures) ; un bruit trop petit serait transparent dans la variation de la variable aléatoire. Ainsi, on choisit d'étudier un bruit tel que σ^2 soit compris entre 0.01 et 1.

On constate que plus σ^2 augmente, plus la variance empirique de M_{fuel} augmente. De plus, l'écart interquartile augmente. Ces observations rendent compte d'un étalement des réalisations de M_{fuel}. Il est naturel de s'attendre à une dispertion des valeurs si les variables d'entrée sont bruitées. La moyenne, elle, ne

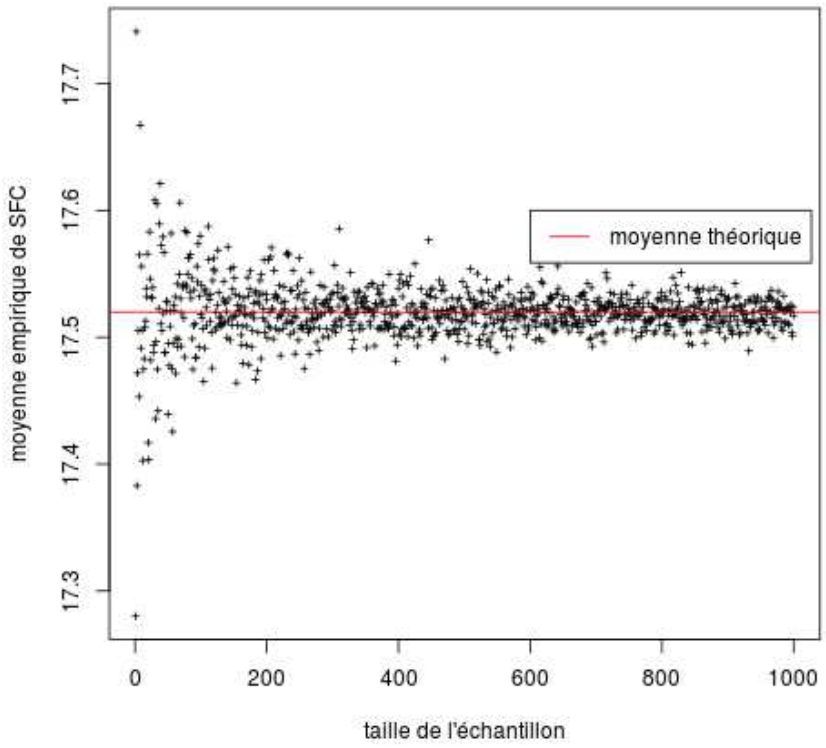


FIGURE 3 – LFGN illustrée avec SFC

varie pas de manière significative. Ceci est certainement dû au fait que les termes d'erreur additionnels soient centrés.

σ^2	Moyenne empirique	Variance empirique	Min	Max	IQR
0.	13625.85	80935.46	13100.87	15865.00	329.37
0.01	13619.93	83323.06	13079.66	15810.94	336.69
0.05	13617.16	81691.08	13009.39	15624.84	336.24
0.1	13621.04	94345.53	12767.53	15647.32	375.83
0.5	13643.27	390720.59	11578.99	16670.52	836.83
1	13664.67	1356399.4	9593.62	19579.52	1560.89

TABLE 5 – Variation des caractéristiques de M_{fuel} en fonction du bruit σ

Les résultats ont été obtenus avec des échantillons de taille 10000 pour obtenir une meilleure précision, qui s'amoindrit sinon à cause du bruit.

La figure 8 montre les différentes densités de M_{fuel} estimées par R pour différents niveaux de bruit. On peut là aussi remarquer un étalement de la densité autour d'une même moyenne qui croit si le bruit augmente ce qui traduit bien la présence de bruit dans les valeurs et illustre les résultats du tableau.

1.5 Conclusion

Cette partie théorique a permi d'illustrer deux théorèmes fondamentaux des probabilités et de confronter des valeurs expérimentales sur les échantillons obtenues à partir d'estimateurs classiques avec les valeurs théoriques.

Pour aller plus loin, on pourrait se demander quelle loi de probabilité correspondrait le mieux pour décrire M_{fuel}, à l'aide par exemple d'un test du χ^2. On peut penser dans un premier temps vu la forme de la densité que M_{fuel} suit une loi normale centrée sur 13625.85 de variance 80935.46.

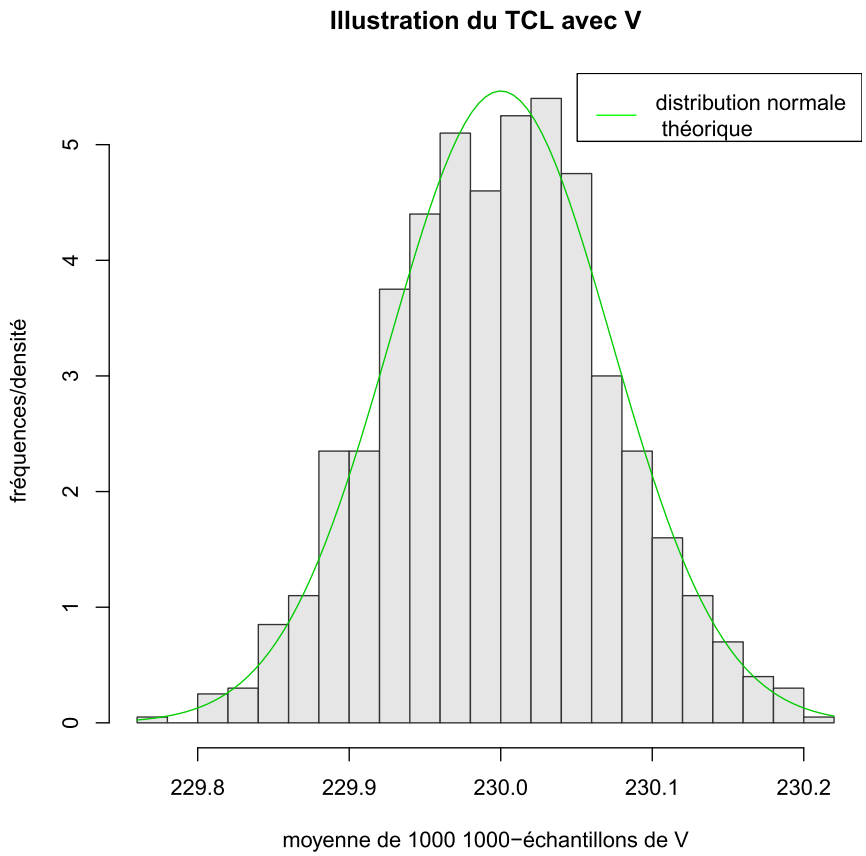


FIGURE 4 – Histogramme de la répartition de $\sqrt{N}\left(\frac{\bar{V}_N - \mathbb{E}(V)}{\sigma}\right)$ avec $N = 1000$, $\mathbb{E}(V) = 230$ et $\sigma = 2.309$

On peut aussi constater que M_{fuel} est peu sensible à un bruit faible ($\sigma^2 \approx 0.01$) mais beaucoup plus pour un bruit de l'ordre de $\sigma^2 \approx 0.5$.

En annexe se trouve des boîtes-à-moustaches de la répartition des valeurs de M_{fuel} afin de rendre plus visuel l'étalement des valeurs à partir d'un bruit élevé.

2 Indices de Sobol

On peut tenter de connaître et quantifier la part d'influence de chacune des variables V, F, SFC sur M_{fuel} dans la relation $M_{fuel} = f(V, S, SFC)$. Pour détecter les variables les plus influentes, on va s'intéresser aux indices de Sobol S. Plus particulièrement, on s'intéressera à l'influence de F et SFC dans notre étude. On cherche alors à estimer simultanément 2 indices de Sobol

$$S := (S^{\{F\}}, S^{\{SFC\}}) = \left(\frac{\text{var}(\mathbb{E}[M_{fuel}|F])}{\text{var}(M_{fuel})}, \frac{\text{var}(\mathbb{E}[M_{fuel}|SFC])}{\text{var}(M_{fuel})}\right)$$

à l'aide la méthode PICK-FREEZE.

2.1 Méthode PICK-FREEZE

La méthode PICK-FREEZE est une méthode, présentée dans [1], qui permet d'estimer S. On présente ici la méthode générale et ses résultats théoriques adaptés à notre problème.

Soit $X = (F, SFC)$ et $X' = (F', SFC')$ une copie indépendante de X. On construit alors les vecteurs PICK-FREEZE de la manière suivante :

$$X^{\{F\}} = (V', F, SFC')$$

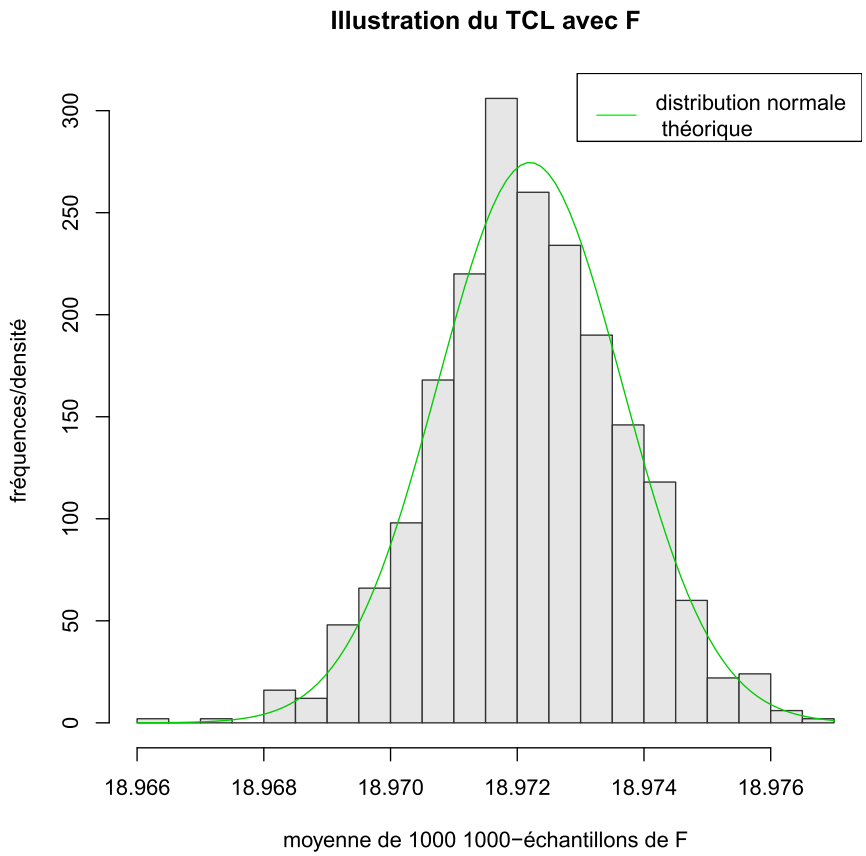


FIGURE 5 – Histogramme de la répartition de $\sqrt{N}\left(\frac{\bar{F}_N - \mathbb{E}(F)}{\sigma}\right)$ avec $N = 1000$, $\mathbb{E}(F) = 18.97$ et $\sigma = 0.046$

Ici, on gèle les valeurs de F et on regénère les valeurs de V et SFC.

$$X^{\{SFC\}} = (V', F', SFC)$$

Ici, on gèle les valeurs de SFC et on regénère les valeurs de V et F.

Puis on calcule

$$M_{fuel} = f(X)$$
$$M_{fuel}^{\{F\}} = f(X^{\{F\}})$$
$$M_{fuel}^{\{SFC\}} = f(X^{\{SFC\}})$$

Pour estimer S, on réalise un N-échantillon de X puis on construit à partir de X les N-échantillons de $X^{\{F\}}$ et de $X^{\{SFC\}}$ puis on calcule les N-échantillons de M_{fuel}, de $M_{fuel}^{\{F\}}$ et de $M_{fuel}^{\{SFC\}}$.

On estime S par les estimateurs suivants :

$$S_N^{\{F\}} = \frac{\frac{1}{N}\sum M_{fuel,i}M_{fuel,i}^{\{F\}} - (\frac{1}{N}\sum M_{fuel,i})(\frac{1}{N}\sum M_{fuel,i}^{\{F\}})}{\frac{1}{N}\sum M_{fuel,i}^2 - (\frac{1}{N}\sum M_{fuel,i})^2}$$

$$S_N^{\{SFC\}} = \frac{\frac{1}{N}\sum M_{fuel,i}M_{fuel,i}^{\{SFC\}} - (\frac{1}{N}\sum M_{fuel,i})(\frac{1}{N}\sum M_{fuel,i}^{\{SFC\}})}{\frac{1}{N}\sum M_{fuel,i}^2 - (\frac{1}{N}\sum M_{fuel,i})^2}$$

$$S_N = \left(S_N^{\{F\}}, S_N^{\{SFC\}}\right)$$

On sait que S_N converge presque sûrement vers S. Par conséquent quand N est grand, les réalisations de S_N tendent vers S. Ceci est montré grâce aux courbes 9 et 10.

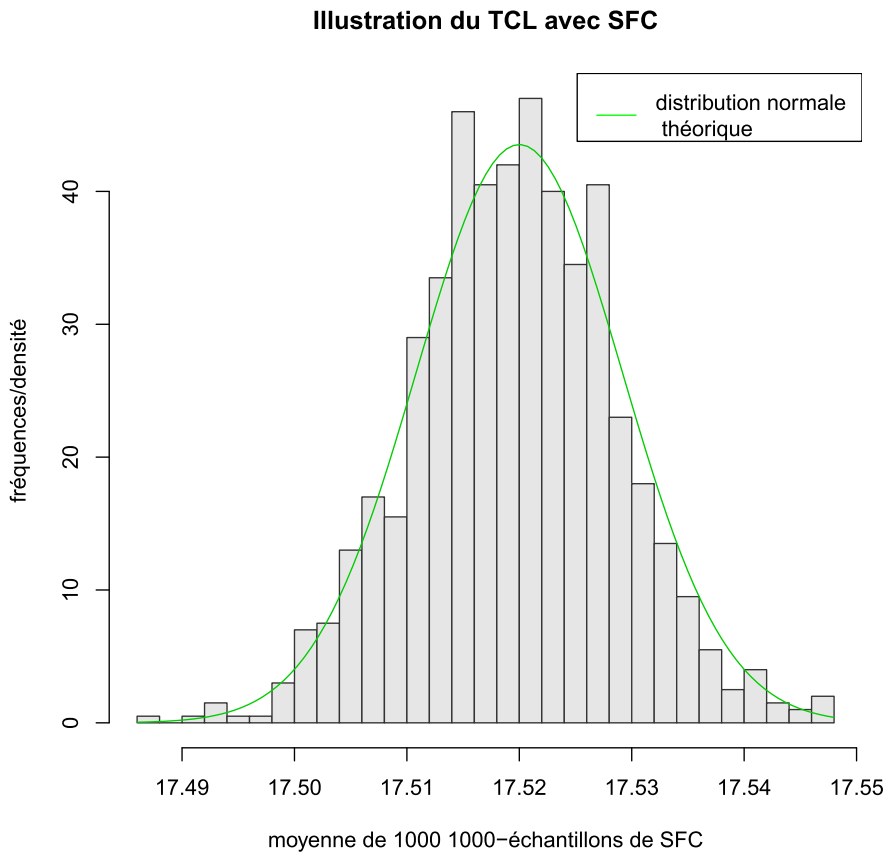


FIGURE 6 – Histogramme de la répartition de $\sqrt{N}\left(\frac{S\bar{F}C_N - \mathbb{E}(SFC)}{\sigma}\right)$ avec $N = 1000$, $\mathbb{E}(SFC) = 17.52$ et $\sigma = 0.2899$

On retiendra alors une réalisation de $S_N^{\{F\}}$ et de $S_N^{\{SFC\}}$ pour $N = 1000$:

$$S^{\{F\}} = 0.0424$$

$$S^{\{SFC\}} = 0.7266$$

En particulier, il semble que

$$S^{\{F\}} < S^{\{SFC\}}$$

ce qui signifie que SFC a plus d'influence que F dans M_{fuel}. Ceci sera discuté dans la partie suivante.

NOTE : il arrive que $S_N^{\{F\}} < 0$ alors qu'un indice de Sobol est toujours positif.

On constate aussi que lorsque le bruit σ augmente, les indices $S^{\{F\}}$ et $S^{\{SFC\}}$ varient. En particulier, $S^{\{F\}}$ augmente et $S^{\{SFC\}}$ diminue comme illustré sur la figure 11. Avec un bruit important, il devient plus difficile de considérer avec certitude que $S^{\{F\}} < S^{\{SFC\}}$. Cela traduit bien le fait que plus les valeurs d'entrée sont bruitées, plus il est difficile de savoir avec certitude quelle variable a le plus d'influence sur la sortie.

2.2 Test

Une question à se poser pour optimiser la consommation d'un avion est s'il vaut mieux améliorer le moteur (SFC) ou l'aérodynamisme (F). On cherche donc à savoir quelle variable à le plus d'influence sur M_{fuel}. On cherche donc à savoir si $S^{\{F\}} < S^{\{SFC\}}$ ou l'inverse.

Avec la méthode PICK-FREEZE, il semble clair que $S^{\{F\}} < S^{\{SFC\}}$. Mais on va tout de même réaliser un test statistique pour s'en assurer (après tout, $S^{\{F\}}$ et $S^{\{SFC\}}$ restent des estimateurs).

On réalise alors le test suivant :

$$H_0 : S^{\{SFC\}} \geq S^{\{F\}} \text{ contre } H_1 : S^{\{SFC\}} < S^{\{F\}}$$

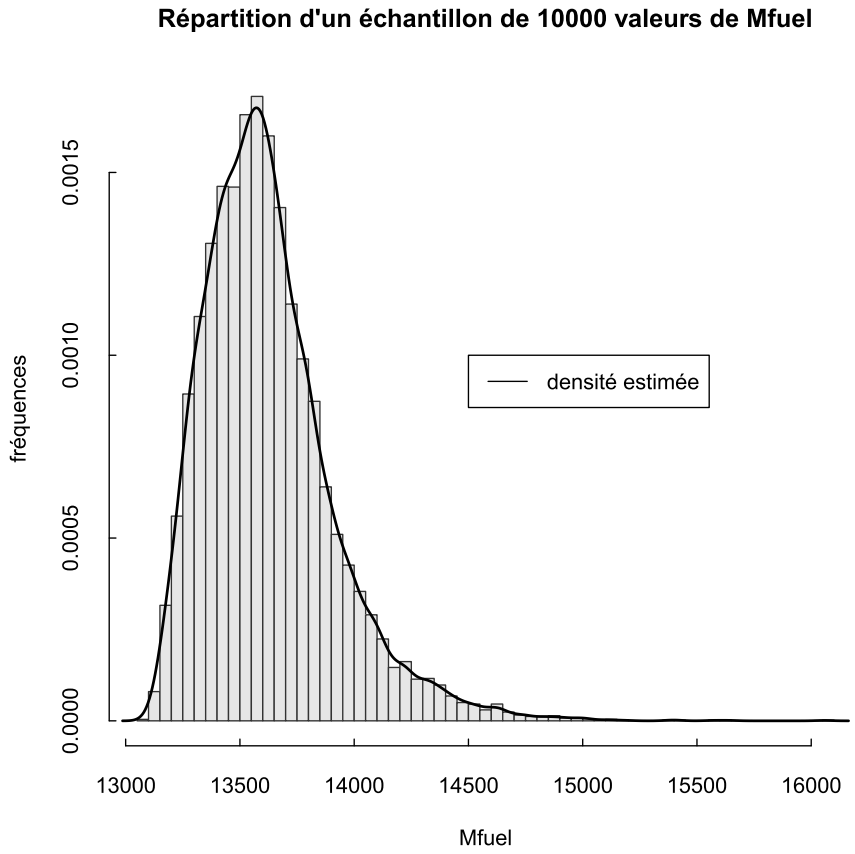


FIGURE 7 – Histogramme de la répartition de 10000 valeurs de M_{fuel}

On rejette H_0 ssi $S^{\{SFC\}} - S^{\{F\}} < -K$ où K est choisit pour avoir un test de taille α. De plus, on se place sous l'hypothèse la moins favorable, c'est-à-dire, $S^{\{SFC\}} - S^{\{F\}} = 0$ (point de H_0 le plus proche de H_1).

On cherche maintenant la loi asymptotique de $S_N^{\{SFC\}} - S_N^{\{F\}}$ sous l'hypothèse que $S^{\{SFC\}} - S^{\{F\}} = 0$.

On sait par [1] que :

$$\sqrt{N}\left(\begin{pmatrix} S_N^{\{F\}} \\ S_N^{\{SFC\}} \end{pmatrix} - \begin{pmatrix} S^{\{F\}} \\ S^{\{SFC\}} \end{pmatrix}\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}_2(0, \Gamma)$$

où Γ est la matrice 2×2 décrite dans l'annexe.

On applique alors la delta-méthode avec la fonction $h(x, y) = y - x$ dont la jacobienne est $A := \begin{pmatrix} -1 & 1 \end{pmatrix}$ pour obtenir (toujours sous l'hypothèse $S^{\{SFC\}} - S^{\{F\}} = 0$) :

$$\sqrt{N}\left(S_N^{\{SFC\}} - S_N^{\{F\}}\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

avec

$$\sigma^2 = A\Gamma A^T$$

On construit alors un estimateur $\widehat{\sigma^2}_N$ convergent de σ^2 en remplaçant dans la matrice Γ les indices de Sobol par leur estimateur respectif et les covariances par leur estimateur classique :

$$\widehat{\mathrm{cov}_N}(X, Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)(Y_i - \bar{Y}_N)$$

Par le théorème de SLUTSKY :

$$\sqrt{N}\left(\frac{S_N^{\{SFC\}} - S_N^{\{F\}}}{\sqrt{\widehat{\sigma^2}_N}}\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

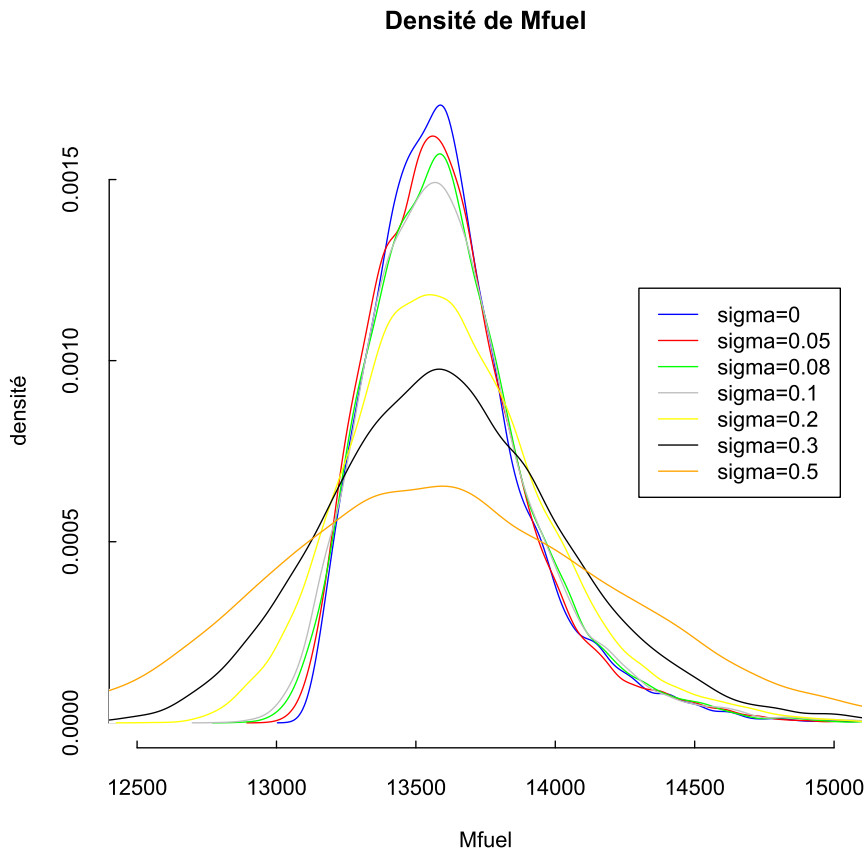


FIGURE 8 – Densité de M_{fuel} pour différents niveaux de bruit. Estimation avec $N = 10000$

On en déduit K :

$$K = -\frac{\Phi^{-1}(\alpha)\sqrt{\widehat{\sigma^2}_N}}{\sqrt{N}}$$

où Φ est la fonction de répartition de la loi normale.

Ainsi, on rejette H_0 ssi $S_N^{\{SFC\}} - S_N^{\{F\}} < \frac{\Phi^{-1}(\alpha)\sqrt{\widehat{\sigma^2}_N}}{\sqrt{N}}$

Le sript `Sobol.R` propose de réaliser le test présenté ci-dessus. En voici une expérience :

bruit : σ^2	$S^{\{F\}}$	$S^{\{SFC\}}$	niveau : α	seuil	rejet
0	0.04762211	0.76175453	0.05	-3.71894734	FALSE

TABLE 6 – Résultat du test

CONCLUSION : on ne peut pas rejeter l'hypothèse H_0.

On peut également répéter cette expérience de niveau $\alpha = 0.05$ 1000 fois et compter le nombre de rejets de H_0 pour vérifier la robustesse de ce test. Le script `Sobol.R` permet de faire cela et on constate qu'on ne peut pas rejeter H_0 1000 fois. Or, le test étant de niveau 0.05, on s'attend à obtenir environ 50 rejets de H_0. Cependant, il ne faut pas oublier que le test a été fait sous l'hypothèse $S^{\{SFC\}} - S^{\{F\}} = 0$. Et on n'a vu que $S^{\{SFC\}} > S^{\{F\}}$. Par conséquent en augmentant le niveau α du test, de plus en plus de tests échoueraient. En effet, avec $\alpha = 0.1$, il y a 0% de rejet, avec $\alpha = 0.5$, il y a 0% de rejet, pour $\alpha = 0.6$, il y a 0.5% de rejet, avec $\alpha = 0.61$, il y a 5.3% de rejet et avec $\alpha = 0.7$, il y a 100% de rejet.

Pour tester la robustesse du test, on peut bruiter les variables d'entrées pour s'assurer d'avoir toujour $S^{\{SFC\}} > S^{\{F\}}$. On sait déjà que $S^{\{SFC\}}$ diminue et que $S^{\{F\}}$ augmente lorsque le bruit σ augmente. Pour σ variant entre 0.01 et 1 et $\alpha = 0.05$, il y a 0% de rejet (tests réalisés avec le même script `Sobol.R`). On en déduit la robustesse du test et qu'on ne peut absolument pas rejeter H_0.

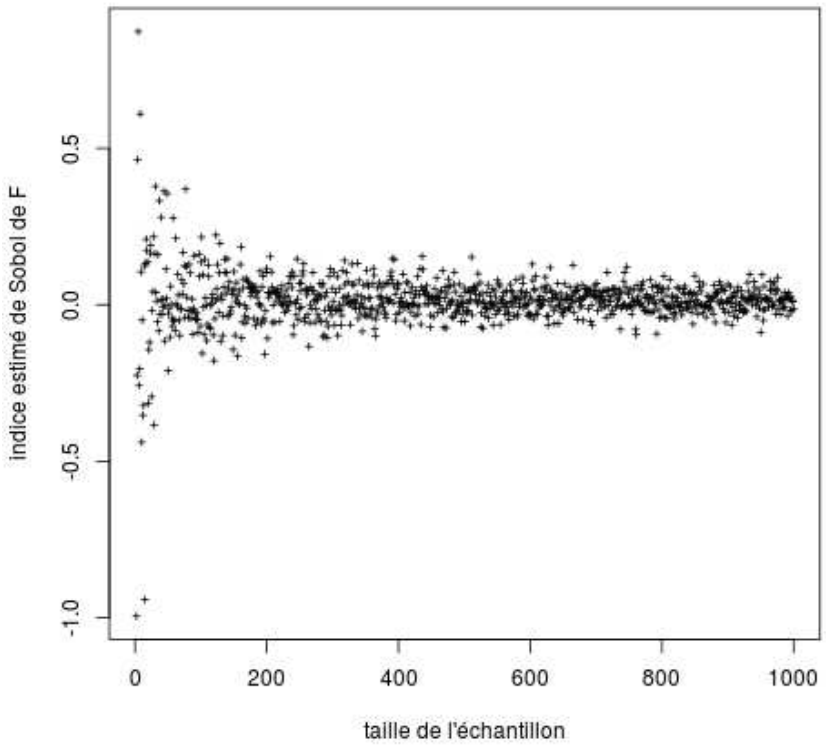


FIGURE 9 – Nuage de points des estimations de $S^{\{F\}}$ en fonction de N, taille de l'échantillon

2.3 Conclusion

Cette partie permet de savoir quelle variable entre S et SFC a le plus d'influence sur M_{fuel} grâce aux indices de Sobol. La méthode PICK-FREEZE permet de fournir un estimateur de $(S^{\{F\}}, S^{\{SFC\}})$. Un test, dont l'interprétation est plutôt délicate, permet de conclure qu'on ne peut absolument pas rejeter l'hypothèse selon laquelle $S^{\{F\}} < S^{\{SFC\}}$, même en présence de bruit sur les variables d'entrée. Un constructeur aéronautique peut conclure de cette étude qu'il vaut mieux dépenser pour améliorer le moteur (SFC) plutôt que l'aérodynamisme de l'avion (F).

3 Modèle linéaire

On suppose inconnue la formule de BREGUET et on ne dispose que des données $(V_i^\sigma, F_i^\sigma, SFC_i^\sigma, M_{fuel,i})_{i=1,...,N}$ simulées ($M_{fuel,i}$ est calculé à partir de la formule de BREGUET). L'objectif est d'exprimer linéairement M_{fuel} à l'aide de V, F et SFC :

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

u représente une erreure gaussienne indépendante centrée.

3.1 Estimateurs

[2] propose une méthode pour estimer les coefficients de la régression linaire. En voici les résulats théoriques adaptés à notre problème :

On range les données dans une matrice $\mathbf{X}$ de N lignes et 4 colonnes où la première colonne ne contient que de 1, la deuxième le vecteur des réalisations $(v_i^\sigma)_{i=1,...,N}$, la troisième le vecteur des réalisations $(f_i^\sigma)_{i=1,...,N}$ et la quatrième le vecteur des réalisations $(sfc_i^\sigma)_{i=1,...,N}$ et dans un vecteur $\mathbf{y}$ composé des réalisations $(m_{fuel,i})_{i=1,...,N}$. On note $\mathbf{u}$ le vecteur $(u_1, ..., u_N)^T$ des erreurs et α le vecteur $(a_0, a_1, a_2, a_3)^T$. Le problème

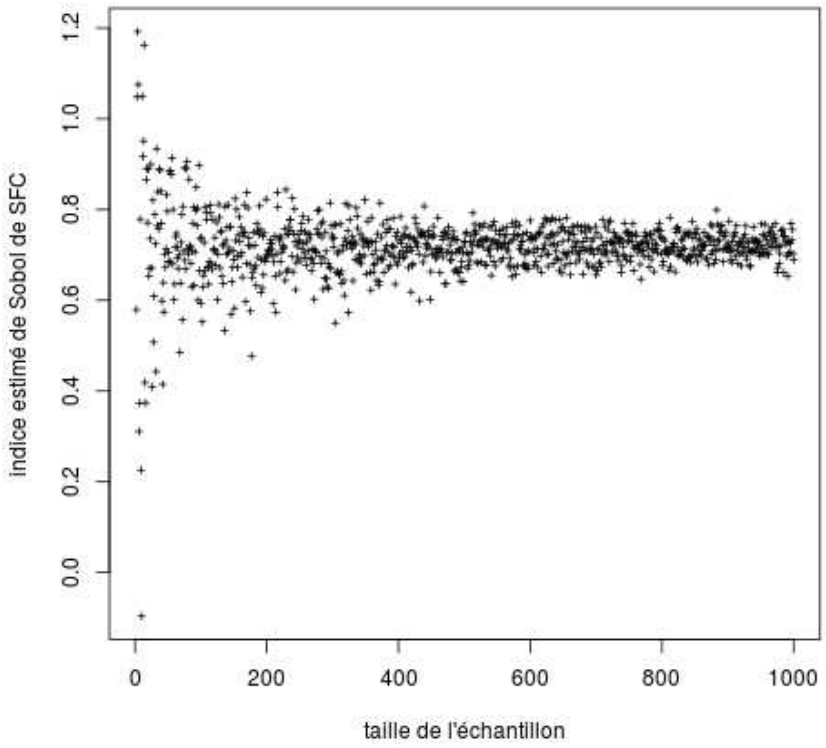


FIGURE 10 – Nuage de points des estimations de $S^{\{SFC\}}$ en fonction de N, taille de l'échantillon

devient alors :

$$\mathbf{y} = \mathbf{X}\alpha + \mathbf{u}$$

Les estimateurs des a_i par la métahode des moindres carrés (MC) est :

$$\mathbf{a} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$$

Il faut pour cela s'assurer que la matrice $X^T X$ soit inversible.

[2] donne également les propriétés des estimateurs. En particulier, $\mathbf{a}$ est sans biais ($\mathbb{E}(\mathbf{a}) = \alpha$) et efficace (i.e. la matrice de covariance atteint la borne inférieure de Cramer-Rao) sous l'hypothèse de normalité des erreurs. Il faudra obtenir K réalisations de $\mathbf{a}$ et les moyenner car rien n'assure la convergence presque sûr de l'estimateur.

3.2 Estimation des coefficients et des erreurs

Le fichier `reg_lin.R` propose un script R pour calculer les valeurs des a_i par la méthode des moindres carrées présentée ci-dessus. A l'aide de la moitié des données $(V_i^\sigma, F_i^\sigma, SFC_i^\sigma, M_{fuel,i})_{i=1,\dots,N}$ fixées au départ, on obtient alors :

$$\begin{cases} a_0 = 28043.18 \\ a_1 = -62.78 \\ a_2 = -761.42 \\ a_3 = 825.64 \end{cases}$$

On peut s'assurer de la vraissemblance de ces résultats. Constatons d'abord que le signe des coefficients a_i correspond bien au sens de variation de M_{fuel} en fonction de la variable considérée. En effet, si SFC augmente, M_{fuel} augmente aussi et le signe de a_3 est positif. De même, M_{fuel} varie dans le sens inverse de F et V et on retrouve bien que les coefficients a_1 et a_2 sont négatifs. De plus, l'espérance théorique de M_{fuel} selon ce modèle est $a_0 + a_1\mathbb{E}(V) + a_2\mathbb{E}(F) + a_3\mathbb{E}(SFC) = 13624.85$ ce qui correspond à l'espérance

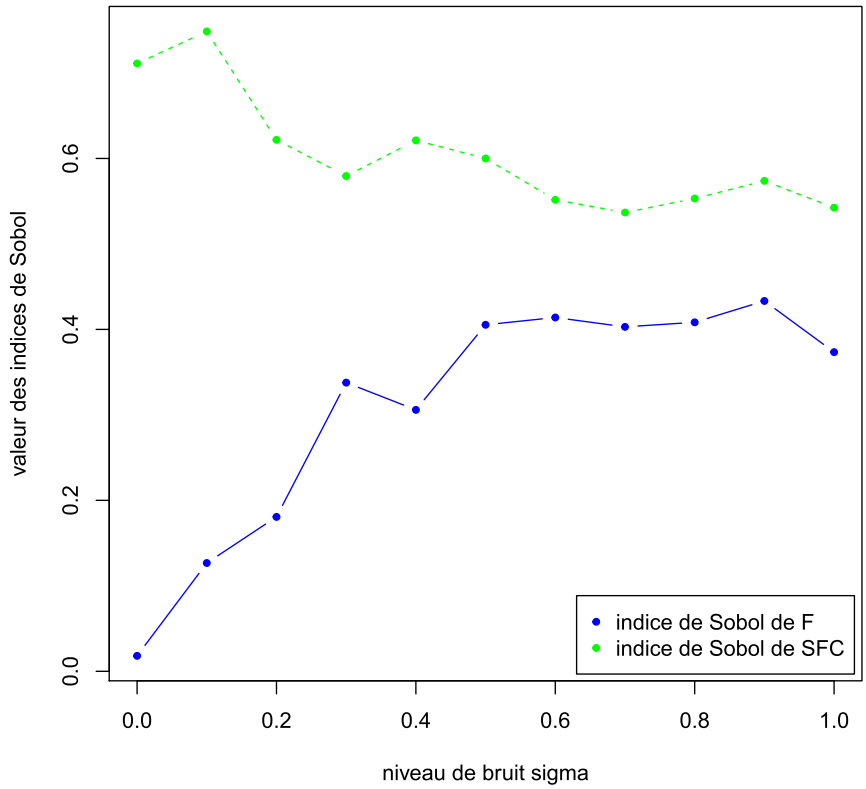


FIGURE 11 – Evolution de $S^{\{SFC\}}$ et de $S^{\{F\}}$ en fonction de σ. $N = 10000$

de M_{fuel} calculée en 1.4. De même pour la variance : la variance théorique de M_{fuel} selon ce modèle est $62.78^2 Var(V) + 761.42^2 Var(F) + 825.64^2 Var(SFC) + \sigma^2 = 79506.42$, ce qui est proche de la variance empirique trouvée en 1.4 avec un bruit de $\sigma^2 = 0.01$, soit 83323.06 (et aussi $79506.42/83323.06 = 0.9541$, proche de R^2).

On fixe alors α aux valeurs ci-dessus, et pour la suite, on utilise l'autre moitié des données obtenues.

[2] donne également une méthode pour estimer la variance σ_u^2 de l'erreur gaussienne $\mathbf{u} \sim \mathcal{N}_N(0, \sigma_u^2 \mathbb{I})$. Pour cela on utilise l'estimateur sans biais de σ_u^2 suivant :

$$\widehat{s^2} = \frac{||\mathbf{y} - \mathbf{X}\alpha||^2}{N - 4}$$

Attention : N est la taille de l'échantillon utilisé pour les calculs. Donc ici, N est égal à la moitié de la taille de l'échantillon de départ.

Le fichier `reg_lin.R` propose un script R pour calculer σ_u^2. En calculant σ_u^2 avec la deuxième moitié de valeurs, on obtient :

$$\sigma_u^2 = 14.61$$

Le document [2] donne également une méthode pour quantifier la qualité de l'approximation. Pour cela, il faut calculer le coefficient de détermination R^2. Ce coefficient permet d'exprimer entre 0 et 1 la part de variation de la sortie (ici M_{fuel}) expliquée par le modèle linéaire. Avec la deuxième moitié des valeurs et à l'aide du fichier, on trouve que

$$R^2 = 0.9995$$

Pour discuter de la pertinence de la régression linéaire, on peut regarder le coefficient R^2. Celui-ci est proche de 1, ce qui signifie qu'une très grande part de la variance de M_{fuel} est expliquée par le modèle : "Plus de 99.9% de la variation de M_{fuel} est expliquée par les variables V, F, et SFC". On peut aussi comparer la

valeur de σ_u^2 avec la variance estimée de M_{fuel} et des variables a_1V, a_2F et a_3SFC (avec l'hypothèse d'un bruit $\sigma^2 = 0.01$

$$\text{var}(M_{fuel}) = 83323.06 \gg \sigma_u^2$$
$$\text{var}(a_1V) = 21046.69 \gg \sigma_u^2$$
$$\text{var}r(a_2F) = 7020.89 \gg \sigma_u^2$$
$$\text{var}(a_3SFC) = 64078.05 \gg \sigma_u^2$$

De même :

$$a_0 = 28043.18 \gg \sqrt{\sigma_u^2} = 3.82$$

On peut en conclure que le modèle obtenu grâce à la régression linéaire ne paraît pas absurde. Dans l'annexe se trouve les valeurs de a_0, a_1, a_2 et a_3 entourées d'un intervalle de confiance pour rendre ces valeurs plus pertinentes.

3.3 Retour sur les indices de Sobol

On peut estimer les indices de Sobol de F et SFC avec le modèle obtenu précédemment avec un niveau de bruit fixé à $\sigma = 0.01$

$$M_{fuel} = a_0 + a_1V + a_2F + a_3SFC$$

pour connaître la part de variation de F et de SFC dans ce modèle et les comparer au modèle original. Le script `Sobol_lin.R` estime les différents indices. On a alors :

$$S^{\{F\}} = 0.0784$$
$$S^{\{SFC\}} = 0.6802$$

Les indices de Sobol obtenus avec la formule de Breguet avec un niveau de bruit de $\sigma = 0.01$ sont :

$$S^{\{F\}} = 0.0391$$
$$S^{\{SFC\}} = 0.7058$$

On constate que les indices calculés avec le modèle linéaire et la formule sont proches. Par conséquent, le modèle linéaire semble traduire la même chose quant à l'influence des variables F et SFC sur M_{fuel} que la formule. Ceci est un argument de plus pour la pertinence du modèle linéaire.

3.4 Conclusion

On a découvert dans cette partie une nouvelle partie des statistiques qui tente d'expliquer une variable par des relations linéaires avec d'autres variables. Un calcul permet de déterminer les coefficients permettant de réaliser une régression linéaire de M_{fuel}, expliquée par V, F et SFC. La modélisation obtenue semble pertinente et s'accorde aux même résulats que la partie 2.

4 Conclusion générale

Ce BE a permis de travailler une bonne partie du programme de statistiques (estimateurs, test et analyse de sensibilité) et de découvrir de nouvelles notions comme la régression linéaire. R a été utilisé comme un langage de programmation et certains graphes ont été difficiles à mettre en forme. Nous avons travaillé tous ensemble sur les trois parties et avons mis en commun avant que Elie Grenier ne rédige le rapport en LaTeX.

On peut pour aller plus loin supposer que l'on ne dispose que des échantillons de V, F, et SFC et inférer les lois des variables (par différents test comme Kolmogorov-Smirnov ou Shapiro-Wilk). Dans le même esprit, il semble pertinent de mettre en place des tests afin de tester la normativité de la loi de M_{fuel} qui a été supposée à la partie 1. On peut aussi chercher d'autres modèles explicatifs de M_{fuel} différents de la régression linéaire, comme des régressions non-linéaires ou non-paramétriques. (problèmes d'optimisation afin de minimiser la fonction d'écart au modèle).

Références

[1] Fabrice Gamboa, Alexandre Janon, Thierry Klein, Agnes Lagnoux-Renaudie, and Clémentine Prieur. Statistical inference for sobol pick freeze monte carlo method. page 26, mar 2013.

[2] Wikistat. Modèle gaussien : regression linéaire multiple — wikistat, 2016. [En ligne ; Page disponible le 2-mai-2018].

Annexe A

La formule de Breguet

La formule de Breguet est une formule mathématique fondamentale en aérodynamique. Elle relie la consommation de fuel aux paramètres de l'avion et opérationnels de la manière suivante :

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

où
- $M_{empty} = 42600kg$ la masse à vide de l'avion
- $M_{pload} = 62500kg$ la masse maximale de l'avion
- $g = 9.8m.s^{-1}$ la constante gravitationnelle
- $R_a = 3000km$ la distance parcourue

sont des constantes et
- V la vitesse de croisière
- F "lift to drag ratio" (finesse)
- SFC "specific fuel consumption"

sont des variables aléatoires.

Plus précisément :
- V suit une loi uniforme sur $[226, 234]$
- F suit une loi Bétâ de paramètres (7,2) sur l'intervalle $[18.7, 19.05]$
- SFC a la densité $h(x) = 3.45e^{-3.45(x-17.23)} \mathbb{1}_{[17.23,+\infty[}(x)$

Le script R `echantillon.R` propose 4 fonctions pour simuler des N-échantillons des différentes variables aléatoires. Un N-échantillon de M_{fuel} est simulé en simulant d'abord un N-échantillon de F, V, et SFC puis en utilisant la formule de Breguet.

Annexe B

Précision sur la loi de F

L'énoncé du BE nous apprend que F suit une loi Bétâ (α, β) sur un intervalle $[a, b]$. Cependant, après avoir vérifié la vraissemblance des formules de densité proposées dans l'énoncé, la densité de F comporte une erreur. En effet, en intégrant la densité sur $\mathbb{R}$ avec la mesure de Lebesgue, on devrait obtenir 1. Or selon l'énoncé, la densité de la loi de F est

$$g(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\beta-1}B(\alpha,\beta)}\mathbb{1}_{[a,b]}(x)$$

En intégrant sur $\mathbb{R}$:

$$\int_{\mathbb{R}} g(x)\,\mathrm{d}x = \int_a^b \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\beta-1}B(\alpha,\beta)}\,\mathrm{d}x$$

changement de variable $x = u(b-a)+a$ avec u sur $[0,1]$ et $\mathrm{d}x = (b-a)\mathrm{d}u$

$$\begin{aligned}
&= \int_0^1 \frac{(u(b-a)+a-a)^{\alpha-1}(b-u(b-a)-a)^{\beta-1}}{(b-a)^{\beta-1}B(\alpha,\beta)}(b-a)\,\mathrm{d}u \\
&= \frac{(b-a)^{\alpha-1}(b-a)^{\beta-1}(b-a)}{(b-a)^{\beta-1}} \underbrace{\int_0^1 \frac{u^{\alpha-1}(1-u)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}u}_{=1} \\
&= (b-a)^{\alpha} \\
&\neq 1
\end{aligned}$$

On en conclut que

$$g(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}B(\alpha,\beta)}\mathbb{1}_{[a,b]}(x)$$

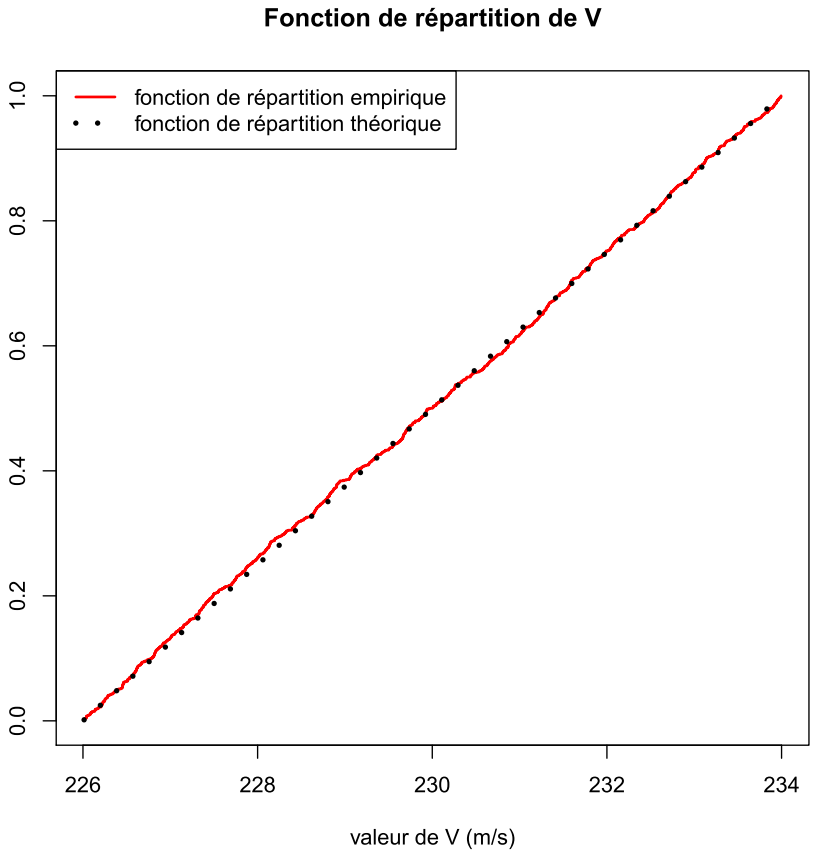
serait une densité convenable pour F.

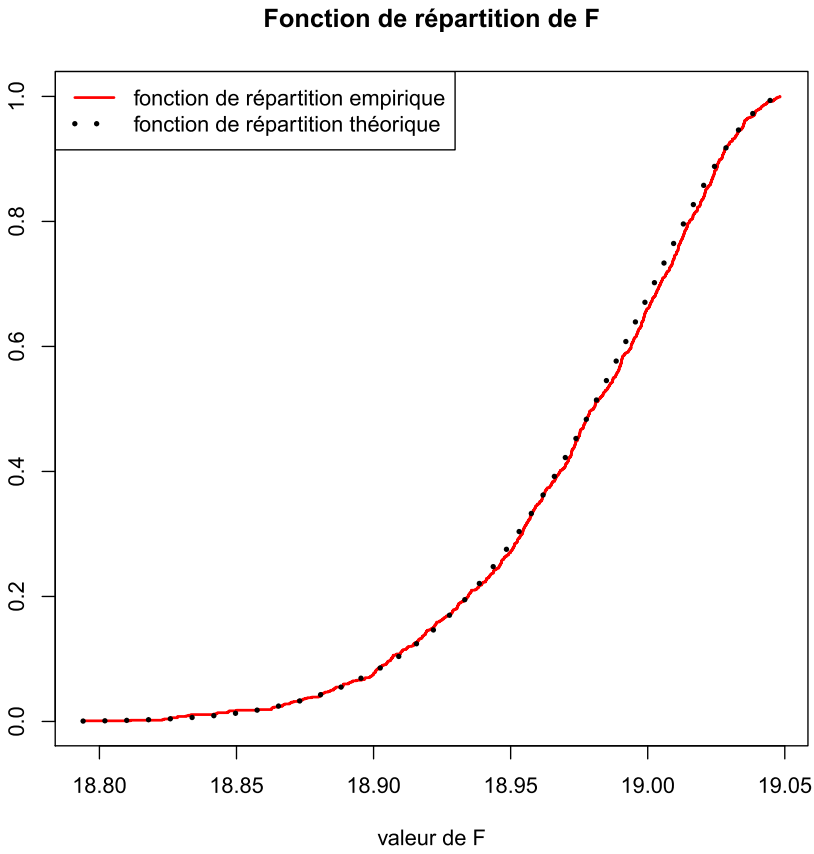
Cette erreur est présente dans l'énoncé du BE mais aussi dans le document [2].

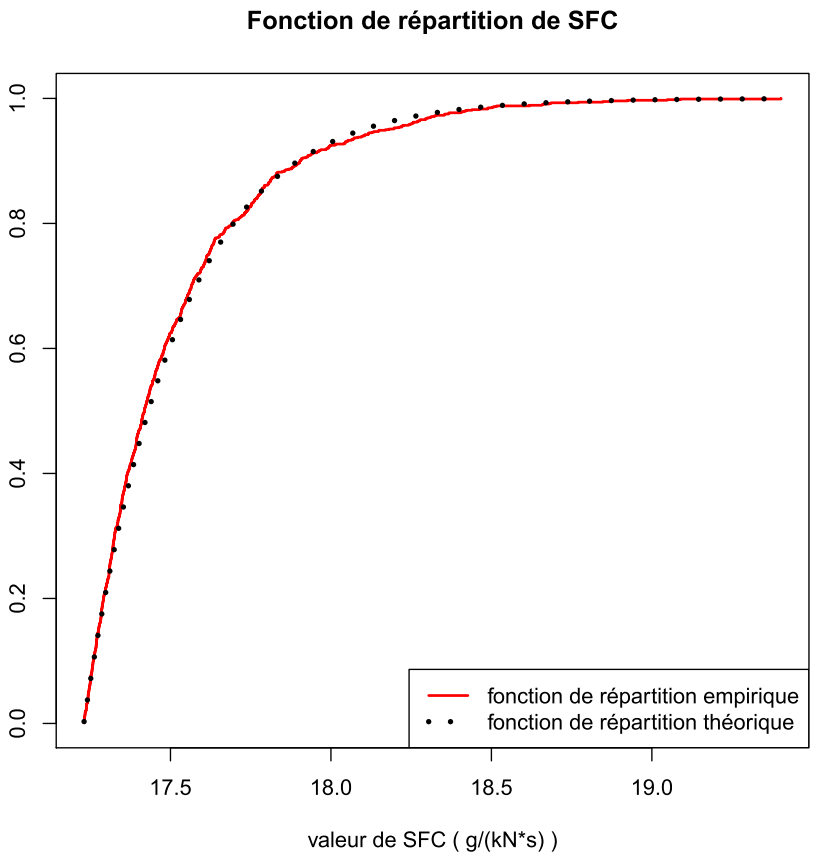
Annexe C

Description des variables

Dans cette annexe, on voulait présenter les histogrammes des échantillons V, F et SFC. Il ne semblait pas pertinent de les inclure dans le corps principal du BE car leurs propriétés sont connues (les lois étant connues) mais ils permetteraient de constater l'adéquation entre les réalisations des variables et les densités théoriques qui ont permies de les simuler. Cependant, un problème du logiciel R irrésolu malgré de longues tentatives répétées reposant sur un défaut persistant de normalisation en fréquence des histogrammes (fonction `hist`) n'a pas permis d'obtenir ces graphes. On propose alors à la place les graphes des fonctions de répartition des variables V, F et SFC. Ceux-ci ne permettent peut-être pas une aussi bonne visualisation de la répartition des échantillons mais sont tout autant pertinents mathématiquement pour montrer l'adéquation (en loi) entre les échantillons empiriques et les lois théoriques.

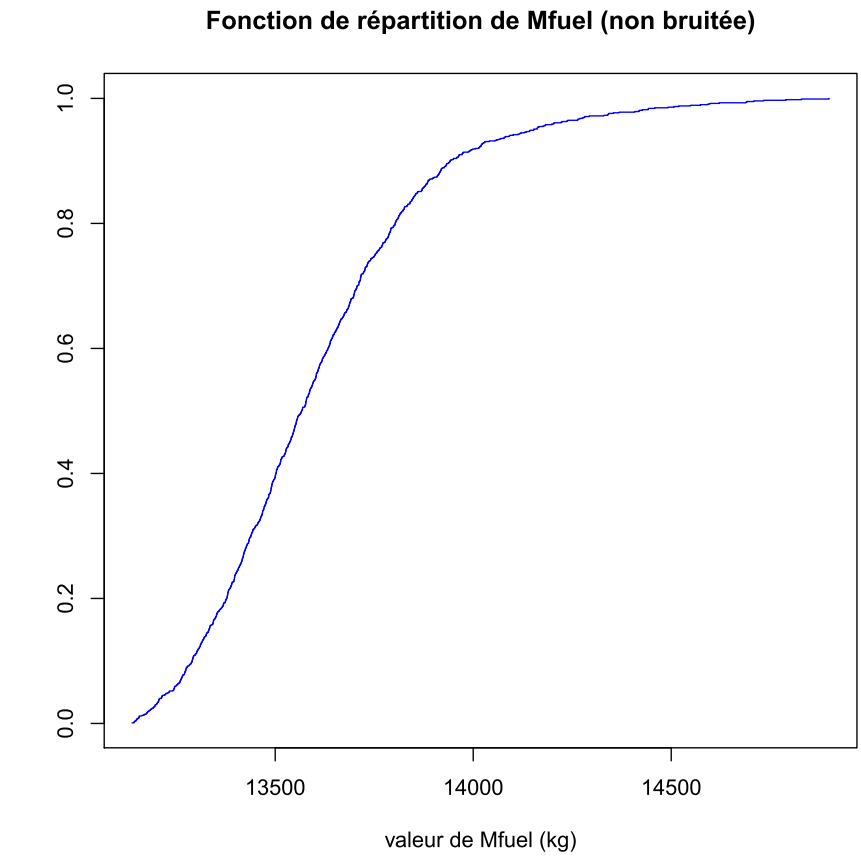


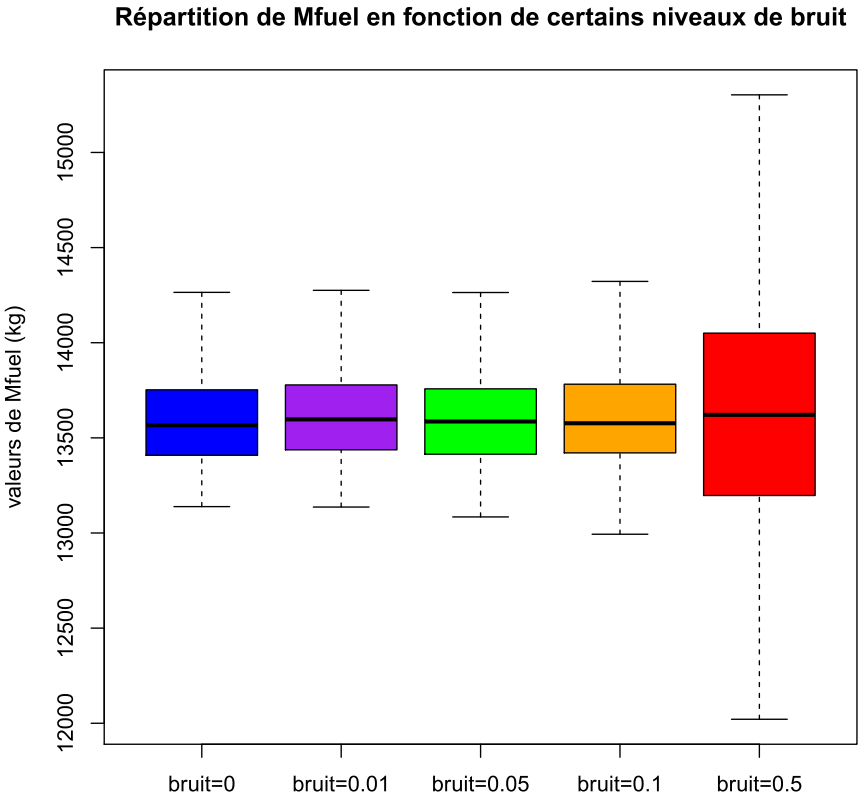

Fonction de répartition de F
fonction de répartition empirique
fonction de répartition théorique
0.0
0.2
0.4
0.6
0.8
1.0
18.80
18.85
18.90
18.95
19.00
19.05
valeur de F

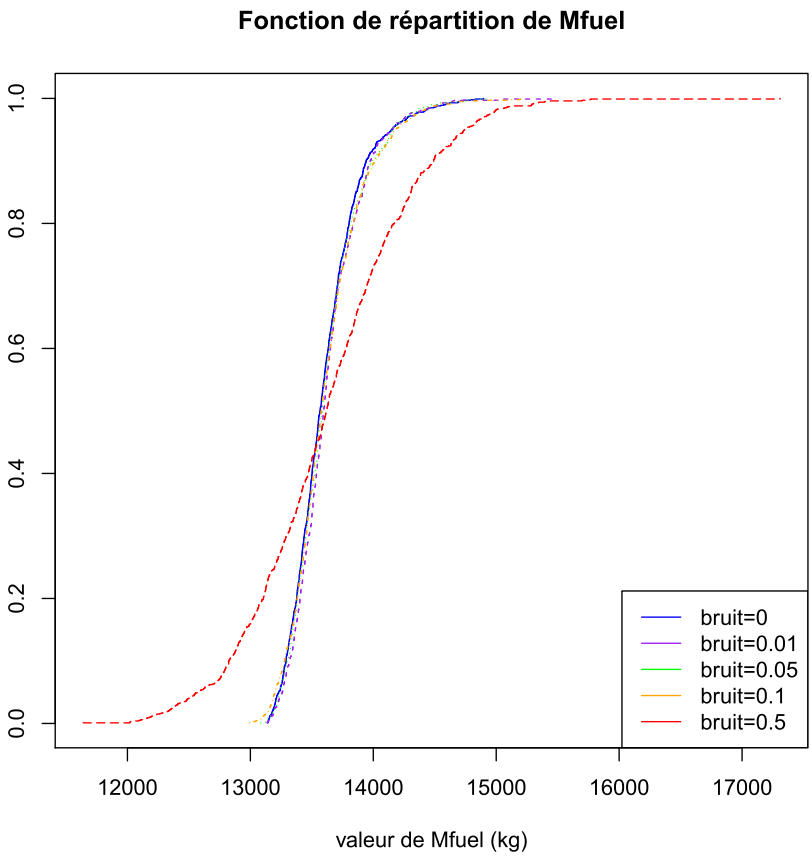


Constatons la superposition entre la fonction de répartition théorique et l'empirique.

On présente également un diagramme boite-à-moustaches (Box-plot) d'un N-échantillon de M_{fuel} et la fonction de répartition estimée de M_{fuel} afin de mieux décrire la seule vaiable pour laquelle il n'est pas donné de modèle théorique. Ces graphes sont répétés avec des niveaux de bruit ("bruit" signifie σ) différents.


Fonction de répartition de Mfuel (non bruitée)
0.0
0.2
0.4
0.6
0.8
1.0
13500
14000
14500
valeur de Mfuel (kg)


Répartition de Mfuel en fonction de certains niveaux de bruit
valeurs de Mfuel (kg)
12000
12500
13000
13500
14000
14500
15000
bruit=0
bruit=0.01
bruit=0.05
bruit=0.1
bruit=0.5



La conclusion est la même que dans la partie 1 : M_{fuel} semble peu sensible au bruit faible ($\sigma \approx 0.01$) mais très pour les bruits dépassant $\sigma = 0.5$. La sensibilité de M_{fuel} au bruitage des variables semble être exponentielle.

Annexe D

Précision sur les indices de Sobol

Dans cette partie est décrite la matrice Γ de la partie 2.2. Le document [2] donne son expression termes à termes qui est ici adaptée à notre problème :

$$\Gamma_{1,1} = \frac{1}{(\mathrm{var}(M_{fuel}))^2}[\mathrm{cov}(M_{fuel}M_{fuel}^{\{F\}}, M_{fuel}M_{fuel}^{\{F\}}) - S^{\{F\}}\mathrm{cov}(M_{fuel}M_{fuel}^{\{F\}}, M_{fuel}^2) \\ -S^{\{F\}}\mathrm{cov}(M_{fuel}M_{fuel}^{\{F\}}, M_{fuel}^2) + S^{\{F\}}S^{\{F\}}\mathrm{var}(M_{fuel})^2]$$

$$\Gamma_{1,2} = \frac{1}{(\mathrm{var}(M_{fuel}))^2}[\mathrm{cov}(M_{fuel}M_{fuel}^{\{F\}}, M_{fuel}M_{fuel}^{\{SFC\}}) - S^{\{F\}}\mathrm{cov}(M_{fuel}M_{fuel}^{\{SFC\}}, M_{fuel}^2) \\ -S^{\{SFC\}}\mathrm{cov}(M_{fuel}M_{fuel}^{\{F\}}, M_{fuel}^2) + S^{\{SFC\}}S^{\{F\}}\mathrm{var}(M_{fuel})^2]$$

$$\Gamma_{2,1} = \frac{1}{(\mathrm{var}(M_{fuel}))^2}[\mathrm{cov}(M_{fuel}M_{fuel}^{\{SFC\}}, M_{fuel}M_{fuel}^{\{F\}}) - S^{\{SFC\}}\mathrm{cov}(M_{fuel}M_{fuel}^{\{F\}}, M_{fuel}^2) \\ -S^{\{F\}}\mathrm{cov}(M_{fuel}M_{fuel}^{\{SFC\}}, M_{fuel}^2) + S^{\{F\}}S^{\{SFC\}}\mathrm{var}(M_{fuel})^2]$$

$$\Gamma_{2,2} = \frac{1}{(\mathrm{var}(M_{fuel}))^2}[\mathrm{cov}(M_{fuel}M_{fuel}^{\{SFC\}}, M_{fuel}M_{fuel}^{\{F\}}) - S^{\{F\}}\mathrm{cov}(M_{fuel}M_{fuel}^{\{F\}}, M_{fuel}^2) \\ -S^{\{SFC\}}\mathrm{cov}(M_{fuel}M_{fuel}^{\{SFC\}}, M_{fuel}^2) + S^{\{SFC\}}S^{\{SFC\}}\mathrm{var}(M_{fuel})^2]$$

Annexe E

Précisions sur la régression linéaire multiple

Voici quelques précisions sur le calcul de R^2. Il faut pour cela calculer quatre nombres :

1. La somme des carrés des résidus (SSE : sum of squared errors)
$$SSE = ||\mathbf{y}\text{-}\mathbf{X}\alpha||^2 (= 5000.29)$$
2. La somme totale des carrés (SST : total sum of squares)
$$SST = ||\mathbf{y} - \bar{y}\mathbf{1}||^2 = \mathbf{y^T y} - N\bar{y} (= 39721761)$$
3. La somme des carrés de la régression (SSR : regression sum of squares)
$$SSR = ||\widehat{\mathbf{y}} - \bar{y}\mathbf{1}||^2 = \alpha^{\mathbf{T}}\mathbf{X^T y} - N\bar{y} (= 39713632)$$
Notons que $SST = SSR + SSE$
4. Le coefficient de détermination R^2
$$R^2 = \frac{SSR}{SST} (= 0.9997)$$

On propose aussi de préciser les estimations des coefficients de la régression linéaire multiple à l'aide du document [1]. Celui-ci donne la méthode pour construire un intervalle de confiance autour de a_i de niveau de sécurité $1 - \alpha$. Cette démarche semble plus pertinente que l'énoncé d'uniques valeurs. On a :

$$IC_{1-\alpha}(a_i) = \mathbf{a}_i \pm t_{1-\frac{\alpha}{2}} \sqrt{\hat{s^2}(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}$$

où $t_{1-\frac{\alpha}{2}}$ est le $1 - \frac{\alpha}{2}$-quantile de la distribution de $Student(N - p)$ (N taille de l'échantillon, p nombre de coefficients, ici $p = 4$).

On trouve pour $\alpha = 0.05$:

$$a_0 = 28099.29 \pm 98.56$$
$$a_1 = -62.77 \pm 0.10$$
$$a_2 = -763.65 \pm 5.00$$
$$a_3 = 824.69 \pm 0.78$$

Bibliographie

[1] C.Chouquet. Modeles lineaires, 2009. Cours de M1 IMAT, Laboratoire de Statistique et Probabilités - Université Paul Sabatier - Toulouse.

[2] Fabrice Gamboa, Alexandre Janon, Thierry Klein, Agnes Lagnoux-Renaudie, and Clémentine Prieur. Statistical inference for sobol pick freeze monte carlo method. page 26, mar 2013.