

MASTER ÉCONOMISTE D'ENTREPRISE



RECHERCHE, RÉALISATION, RESTITUTION

---

## Application des modèles SFA à l'étude des prix

---

*Corentin DUCLOUX et Aybuké BICAT*


29 mars 2024

# | Table des matières

<b>Remerciements</b>	<b>2</b>
<b>Introduction</b>	<b>3</b>
<b>Revue de la littérature</b>	<b>4</b>
Une nouvelle approche de la théorie du consommateur . . . . .	4
Pricing Hédonique . . . . .	5
Aspects théoriques . . . . .	5
Application . . . . .	7
Fonction de production . . . . .	10
Le modèle SFA . . . . .	10
Aspects théoriques . . . . .	10
Utilisation empirique . . . . .	14
SFA & Pricing Hédonique . . . . .	14
<b>Choix et cadrage de la problématique</b>	<b>17</b>
Le marché de la téléphonie mobile, en constante évolution . . . . .	17
Smartphones et Pricing Hédonique . . . . .	20
Effet de réputation . . . . .	21
<b>Acquisition des données</b>	<b>22</b>
Scraping . . . . .	22
Méthodologie . . . . .	23
<b>Statistiques descriptives</b>	<b>24</b>
Analyse des prix . . . . .	24
Mesures de tendance centrale . . . . .	24
Mesures de dispersion . . . . .	25
Prix moyen en fonction d'autres variables . . . . .	25
Etudes des variables catégorielles importantes . . . . .	28
Etude des variables dichotomiques . . . . .	29
Analyse des corrélations . . . . .	30

<b>Modélisation économétrique</b>	<b>31</b>
Sélection de variables . . . . .	31
Modèle Hédonique niveau-niveau . . . . .	33
Interprétations Modèle niveau-niveau . . . . .	36
Modèle Hédonique log-niveau . . . . .	38
Interprétations Modèle log-niveau . . . . .	40
Vérification des hypothèses . . . . .	41
Modèle SFA - Frontière de coût . . . . .	45
Interprétations Modèle frontière de coût . . . . .	47
Analyse comparative des deux modèles . . . . .	49
Quel modèle choisir ? . . . . .	50
Analyse de l'efficacité . . . . .	50
Analyse Factorielle de Données Mixtes . . . . .	54
<b>Comparateur</b>	<b>59</b>
<b>Conclusion</b>	<b>61</b>
<b>Annexe</b>	<b>62</b>
Dérivation de la fonction de vraisemblance . . . . .	62
Dérivation des indices d'efficacité . . . . .	63
<b>Acronymes</b>	<b>65</b>
<b>Glossaire des variables</b>	<b>66</b>
<b>Licence</b>	<b>68</b>
<b>Références</b>	<b>70</b>

# | Remerciements

Nous tenons à remercier chaleureusement Monsieur *Alain BOUSQUET* pour son accompagnement tout au long de ce projet **3R**, qui a toujours été ouvert à l'exploration de nouveaux sujets, à l'expérimentation, et nous a encouragé à creuser diverses pistes de réflexion. Ce sujet a été et sera pour nous l'occasion de mettre en pratique l'ensemble des connaissances acquises dans notre cursus universitaire (microéconomie, économétrie, statistiques, analyse de la concurrence, pricing, développement logiciel sous **R** et **python** ) sur une problématique éminemment appliquée.

\* \* \*

*Note* : Ce **PDF** a été entièrement rédigé en utilisant **Quarto** <sup>(1)</sup>, combinant la puissance et la versatilité de R, Python, et  $\text{\LaTeX}$ . Une présentation interactive **reveal.js** du sujet est aussi disponible. <sup>(2)</sup>

---

<sup>(1)</sup> **Quarto** : Système de publication technique et scientifique *open-source*  $\Rightarrow$  <https://quarto.org/>.

<sup>(2)</sup> Retrouvez la présentation sur [https://corentinducloux.fr/Reveal.js/slides\\_smartphones.html](https://corentinducloux.fr/Reveal.js/slides_smartphones.html).

# | Introduction

En tant que consommateur, nous nous retrouvons souvent face à une question infiniment plus complexe qu'elle n'en a l'air. En des termes simples, elle se traduit par : Pourquoi ce prix ? Pour quelle raison ce stylo, cette nouvelle télévision, ou ce smartphone coûte tant ? Est-ce une simple question de coût de production, de marge ? Ou bien cela prend-il en compte d'autres éléments, tels que la valeur perçue par le consommateur, les caractéristiques spécifiques d'un produit, ou encore le service qu'il rend ?

Dans ce cadre, une approche essentielle dans l'analyse des prix est celle des prix hédoniques. Celle-ci considère que le prix d'un bien ou d'un service est influencé non seulement par ses caractéristiques, mais aussi par la valeur subjective que les individus accordent à ces caractéristiques. Ainsi, la méthode des prix hédoniques examine comment des éléments spécifiques tels que la qualité ou les fonctionnalités d'un produit impactent sa valeur perçue, reflétée dans son prix.

La compréhension des mécanismes sous-jacents à la détermination des prix dans un marché est d'une importance cruciale tant pour les consommateurs que pour les entreprises. Face à ces interrogations, les modèles SFA (*Stochastic Frontier Analysis*) émergent comme un outil puissant, permettant d'évaluer et d'analyser l'efficacité des prix des produits en allant bien au-delà d'une simple évaluation du coût de production. Ces modèles supposent que les prix sont à l'équilibre, et prennent en compte à la fois les valorisations du côté des consommateurs et les coûts de production du côté des producteurs pour comprendre comment se forment réellement les prix.

En ce sens, les modèles SFA, dans le cadre de l'étude des prix, entrent en synergie avec l'approche hédonique en permettant une analyse approfondie des différentes composantes qui influencent la formation des prix.

\* \* \*

Ainsi, cette étude se déroulera en trois étapes clés : une revue de la littérature portant sur les modèles SFA et la notion de prix hédonique. Ensuite, nous nous concentrerons sur le processus de délimitation de notre problématique d'étude (*le marché des smartphones*). Enfin, nous aborderons l'acquisition des données, les statistiques descriptives et la modélisation économétrique pour comprendre plus en détail les mécanismes de fixation des prix dans le contexte complexe et évolutif des smartphones.

# | Revue de la littérature

## Une nouvelle approche de la théorie du consommateur

En microéconomie, dans la théorie du consommateur *classique*, le choix du meilleur ensemble de consommation dépend des préférences d'un individu. Les préférences de cet individu sont classiquement représentées par la fonction d'utilité :

$$U(x) = U(x_1, x_2, \dots, x_n) \quad (1)$$

Avec  $x_1, x_2, \dots, x_n$  un vecteur de  $n$  biens. L'équation 1 exprime donc la relation entre la quantité de biens consommés et le niveau d'utilité que ces biens procurent à un agent. Dès lors, dans ce cadre, la consommation de biens procure **directement** de l'utilité à l'agent. En pratique pourtant, il est difficile de concevoir comment l'achat d'un bien comme une lampe ou un stylo peut nous apporter de l'utilité en tant que consommateur.

Pour répondre à cette difficulté, Lancaster (1966) propose un nouveau cadre conceptuel théorique décrit par les hypothèses suivantes.

### Hypothèses

1. Le bien en lui-même ne procure pas d'utilité au consommateur  $\Rightarrow$  il possède des **caractéristiques** qui procurent de l'utilité.
2. Un bien est un ensemble (*bundle*) de caractéristiques – il possède le plus souvent de nombreuses caractéristiques.
3. Une combinaison de biens peut procurer une utilité qui n'est pas la simple somme des utilités procurées par les biens séparément.

*Illustrons ces points avec quelques exemples :*

- Un ordinateur n'est pas acheté pour le simple plaisir de posséder un ordinateur. Il est acheté car il permet de naviguer sur Internet, écrire des cours, programmer, regarder une série, etc. C'est donc pour les **services qu'il nous rend**, ce qui est modélisé ici par les caractéristiques possédées du bien.
- Les biens possèdent généralement un grand nombre de caractéristiques. Prenons l'exemple d'une gourde : la couleur, la forme, les dimensions et la capacité isothermique sont autant de caractéristiques qui peuvent influencer sur la décision d'achat et la disposition à payer.
- En consommant du lait et du café séparément, les caractéristiques retirées du lait sont de la vitamine D et du calcium, tandis que pour le café les caractéristiques retirées sont de la caféine, une boisson chaude, un "boost" le matin. En revanche, consommer un café latte permettra d'obtenir une boisson plus douce, moins caféiné, un goût différent. En bref, les caractéristiques retirées du mélange sont différentes.

Dans le modèle de Lancaster, on pose une relation **linéaire** entre les prix des biens et leurs caractéristiques. Le prix total  $p$  d'un bien peut donc être considéré comme la somme des prix

individuels associé à chaque caractéristique. Cela découle du fait que les attributs des biens étudiés peuvent être considérés comme des composantes distinctes et séparables.

## Pricing Hédonique

### Aspects théoriques

Rosen (1974) étend ce qui a été apporté par le cadre théorique de Lancaster (1966). La principale différence est qu'il s'intéresse à l'**équilibre de marché de biens différenciés** (là où Lancaster s'intéresse uniquement à la demande) avec :

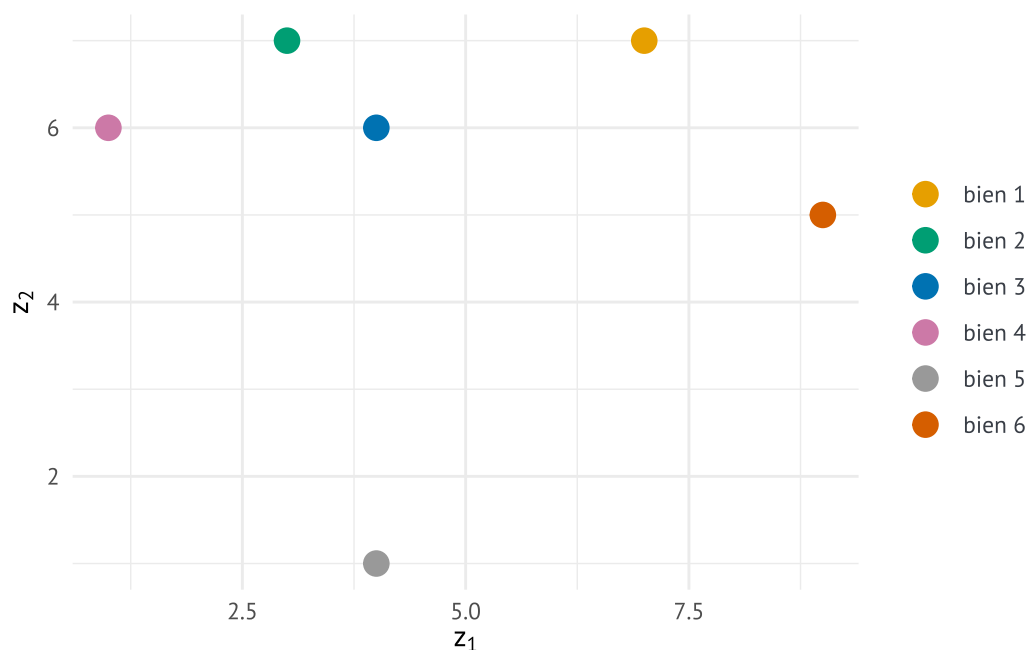
- un continuum de biens du côté de l'offre.
- un continuum de consommateurs hétérogènes du côté de la demande.

Dans ce modèle, la relation entre les prix des biens et leurs attributs peut-être **non-linéaire** et permet aussi de capter des effets d'interaction entre plusieurs variables. Au coût d'une modélisation plus complexe que dans le modèle de Lancaster (1966), les résultats gagnent en robustesse.

L'objet de la contribution de Rosen est d'étudier un bien différencié  $z$  décrit par le vecteur de ses  $n$  caractéristiques mesurables tel que :

$$z = (z_1, z_2, \dots, z_n) \quad (2)$$

Afin de comprendre pourquoi il est important d'étudier des biens différenciés dans ce cadre, regardons en détail le graphique suivant.



**Figure 1** – Plan  $(z_1, z_2)$  de différents biens avec 2 caractéristiques.



En général, nous sommes habitués à représenter les préférences des consommateurs en termes de quantités de biens  $x_1, x_2$ . Ici, on assiste à un changement de paradigme : on va représenter les préférences des consommateurs en termes de caractéristiques de biens, c'est-à-dire dans l'espace  $z_1, z_2$  (on choisit de prendre seulement 2 caractéristiques et 6 biens pour simplifier).

On peut en déduire que les consommateurs achetant le *bien 5* valorisent plus les caractéristiques  $z_1$  que  $z_2$ , et inversement pour le *bien 4*.

*En fait, la différenciation horizontale et verticale des produits implique qu'une vaste gamme de paniers est disponible dans cet espace de consommation !*

- **Différenciation Horizontale**  $\Rightarrow$  A prix donné, il n'y a pas unanimité dans le choix des consommateurs entre 2 biens (jaune et rouge) : ce sont des différences de goûts.
- **Différenciation Verticale**  $\Rightarrow$  A prix donné, il y a unanimité dans le choix des consommateurs entre 2 voitures biens : l'un est meilleur que l'autre.

Il faut aussi noter que dans le modèle de Rosen, le consommateur n'achète qu'**une seule** unité de bien qui est une combinaison d'attributs  $z_1, z_2, \dots, z_n$ . Historiquement, cela s'explique car Rosen s'intéresse principalement aux biens durables (logements, voitures, smartphones...). Il est en effet beaucoup plus simple d'obtenir des caractéristiques observables sur ces biens durables : que ce soit le nombre de pièces pour un logement, la superficie, ou bien la puissance et la longueur d'une voiture.

De toutes ces informations, on peut formuler 2 questions.

- Pour le **producteur**, quelle combinaison de caractéristiques lui permet de maximiser son profit ?
- Pour le **consommateur**, quelle combinaison de caractéristiques lui rapporte le plus d'utilité sous contrainte budgétaire ?

On aboutit à une relation fonctionnelle entre les caractéristiques des biens et leur prix, appelée fonction de prix hédonique  $p(z)$ .

$$p(z) = p(z_1, z_2, \dots, z_n) \quad (3)$$

Un produit est donc défini en chaque point du plan et guide les choix de localisation des consommateurs et des producteurs concernant les ensembles de caractéristiques.

#### Limites

Il n'en reste pas moins qu'il subsiste un problème indéniable : ce qu'on aimerait réellement mesurer c'est le **service rendu par un produit** (*pour lequel le lien précis entre ses fonctionnalités et les services rendus reste inconnu*) et non pas les caractéristiques de ce produit. Mais ce premier est complètement inobservable. Un défi sera donc d'interpréter correctement les résultats des régressions.

## Application

Harrison Jr et Rubinfeld (1978) :

**Objectif** : Examiner comment les données du marché immobilier peuvent être utilisées pour évaluer la *Willingness To Pay* des consommateurs pour une meilleure qualité de l'air.

- Le modèle suppose que les ménages prennent en compte le niveau de pollution de l'air, la quantité et la qualité du logement et d'autres caractéristiques de quartier pour faire leur choix.

- La fonction de la valeur hédonique du logement traduit les attributs du logement en prix, et suppose que les consommateurs perçoivent avec précision ces attributs et que le marché est en équilibre à court terme.

#### Définition des variables

- $W$  = WTP *marginale* pour une meilleure qualité de l'air
- $NOX$  = Concentration des oxydes d'azote<sup>(3)</sup>
- $INC$  = Revenu du ménage en centaine de dollars

Trois niveaux de revenu par an découpés en variable catégorielles :

- **LOW** si  $INC \leq \$ 8500 \Rightarrow Y_0$  (Catégorie de référence)
- **MEDIUM** si  $INC \leq \$ 11500 \Rightarrow Y_1$
- **HIGH** si  $INC \leq \$ 15000 \Rightarrow Y_2$

$$\log(W) = \beta_0 + \beta_1 \log(NOX) + \beta_2 \log(INC) + \beta_3[Y_1 \cdot \log(NOX)] + \beta_4[Y_2 \cdot \log(NOX)] \quad (4)$$

Coefficients estimés pour la régression log – log (significatifs au seuil  $p < 0.01$ ) :

$$\log(W) = \underbrace{2.2}_{\beta_0} + \underbrace{0.97}_{\beta_1} \log(NOX) + \underbrace{0.8}_{\beta_2} \log(INC) - \underbrace{0.03}_{\beta_3}[Y_1 \cdot \log(NOX)] - \underbrace{0.07}_{\beta_4}[Y_2 \cdot \log(NOX)]$$

**Résultats** : La WTP marginale pour une meilleure qualité de l'air augmente avec le niveau de pollution de l'air et avec le niveau de revenu des ménages. Plus précisément, malgré la présence d'effets d'interaction significatifs mais faibles, il est observé que toutes choses égales par ailleurs, lorsque le niveau de  $NOX$  et le revenu du ménage augmentent, le prix a tendance à augmenter également.

---

Pour finir, l'approche hédonique a été utilisée empiriquement dans de très nombreux domaines comme par exemple :

Berndt et Rappaport (2001)  $\Rightarrow$  Secteur informatique.

- L'objectif de cet article est d'examiner l'évolution des prix ajustés en qualité des ordinateurs personnels de bureau et mobiles entre 1976 et 1999.

Chen et Rothschild (2010)  $\Rightarrow$  Secteur de l'hôtellerie.

- Analyse l'impact des caractéristiques des hôtels de Taipei sur leurs tarifs en utilisant les données de 73 hôtels collectées auprès d'un agent de voyage en ligne.

---

<sup>(3)</sup>Variable de pollution,  $NOX$  est un *proxy* pour la qualité de l'air.

Yim, Lee, et Kim (2014)  $\Rightarrow$  Secteur de la restauration.

- Explore l'impact des attributs des restaurants à Séoul sur leurs prix moyens de repas en examinant les données de 185 établissements recueillies via diverses sources.

Dans la littérature, une spécification *semi-log* est généralement préférée en raison de sa capacité à mieux modéliser les relations non linéaires entre les variables. De plus, cette forme permet d'améliorer l'ajustement du modèle aux données observées, et offre un  $R^2$  supérieur à celui obtenu avec d'autres spécifications – voir Bello et Moruf (2010).

# Fonction de production

Avant de passer à l'explication de la seconde partie théorique, c'est-à-dire les modèles SFA, attardons-nous sur la définition d'une fonction de production, fondement important de la SFA.

## Rappel

- Un processus de production représente la transformation d'inputs en outputs.
- Dès lors, une fonction de production  $f(\cdot)$  donne la quantité maximum d'output  $y$  pouvant être produite à partir de combinaison d'inputs.

$$y_i = f(x_i; \beta) \quad (5)$$

Avec  $x_i$  le vecteur d'inputs et  $\beta$  le vecteur de paramètres inconnus à estimer.

$f(x_i; \beta)$  est en fait la frontière de production. Pour l'instant cette frontière ne prend pas en compte l'efficacité technique  $TE_i$  et elle n'est pas *stochastique* car elle n'inclut pas de terme aléatoire.

\* \* \*

Farrell (1957) est le premier auteur à définir cette *Frontière de Production*.

*“When one talks about the efficiency of a firm one usually means its success in producing as large as possible an output from a given set of inputs.”*

Cette définition permet donc d'aboutir à la formulation évoquée à l'équation 5.

## Le modèle SFA

### Aspects théoriques

Aigner, Lovell, et Schmidt (1977) :

**Objectif** : Formulation et estimation de fonctions de frontière de production stochastique.

Avant les travaux de Aigner, Lovell, et Schmidt (1977), les économètres utilisaient principalement dans la littérature des fonctions de production pour étudier le lien entre le niveau de production et la quantité d'inputs utilisés. Cela signifie que la formulation théorique énoncée par Farrell (1957) différait de l'utilisation empirique. En effet, Farrell (1957) a lui introduit la notion d'efficacité au sens du **niveau maximum de production** atteignable étant donné une combinaison spécifique d'inputs.

- On repart de la fonction de production (équation 5), mais en lui ajoutant un terme multiplicatif  $TE_i$ .

$$y_i = f(x_i; \beta) \cdot TE_i$$

$TE_i$  représente l'efficacité technique, définie comme le ratio d'output observé sur l'output maximum réalisable, soit  $TE_i = \frac{y_i}{y_i^*}$ .

- Si  $TE_i = 1$  alors la firme  $i$  produit l'output maximum réalisable, alors que si  $TE_i < 1$ , il existe un écart entre l'output maximum et l'output effectivement observé.

Un composant **stochastique**  $\exp\{v_i\}$  est en outre ajouté pour représenter les chocs aléatoires affectant la production. La fonction de production devient alors :

$$y_i = f(x_i; \beta) \cdot TE_i \cdot \exp\{v_i\}$$

On peut ré-écrire l'efficacité technique sous la forme  $TE_i = \exp\{-u_i\}$ . Dès lors :

$$\boxed{y_i = f(x_i; \beta) \cdot \exp\{-u_i\} \cdot \exp\{v_i\}} \quad (6)$$

*Note :* En réarrangeant l'équation 6 avec le logarithme népérien, on obtient :

$$\Leftrightarrow \ln(y_i) = \ln(f(x_i; \beta)) + \underbrace{v_i - u_i}_{\epsilon_i}$$

Le modèle peut alors s'écrire sous la forme suivante :

$$\boxed{\ln(y_i) = \ln(f(x_i; \beta)) + \epsilon_i} \quad (7)$$

L'avantage de cette écriture est qu'elle facilite la manipulation des termes d'erreur, et il est très simple de retrouver le logarithme de l'output maximum. En effet :

$$\Leftrightarrow \ln(y_i) = \underbrace{\ln(f(x_i; \beta))}_{\ln(y_i^*)} + v_i - u_i$$

Et donc le logarithme de l'output observé est simplement  $\ln(y_i) = \ln(y_i^*) - u_i$ .

Les termes d'erreur  $\epsilon_i$  ont ainsi une distribution particulière composée :

- $v_i$  est une **erreur aléatoire**  $\Rightarrow$  variation inexpliquée par les variables indépendantes du modèle, avec  $v_i \sim \mathcal{N}(0, \sigma_v^2)$ .
- $u_i$  est un **composant unilatéral** qui peut être choisi parmi plusieurs distributions<sup>(4)</sup> et  $u_i \geq 0$ , puisqu'il est nécessaire d'avoir  $TE_i \leq 1$ .

---

<sup>(4)</sup>Dans la littérature, deux distributions sont couramment utilisées : la distribution **semi-normale** et **normale tronquée**.

### Conclusion

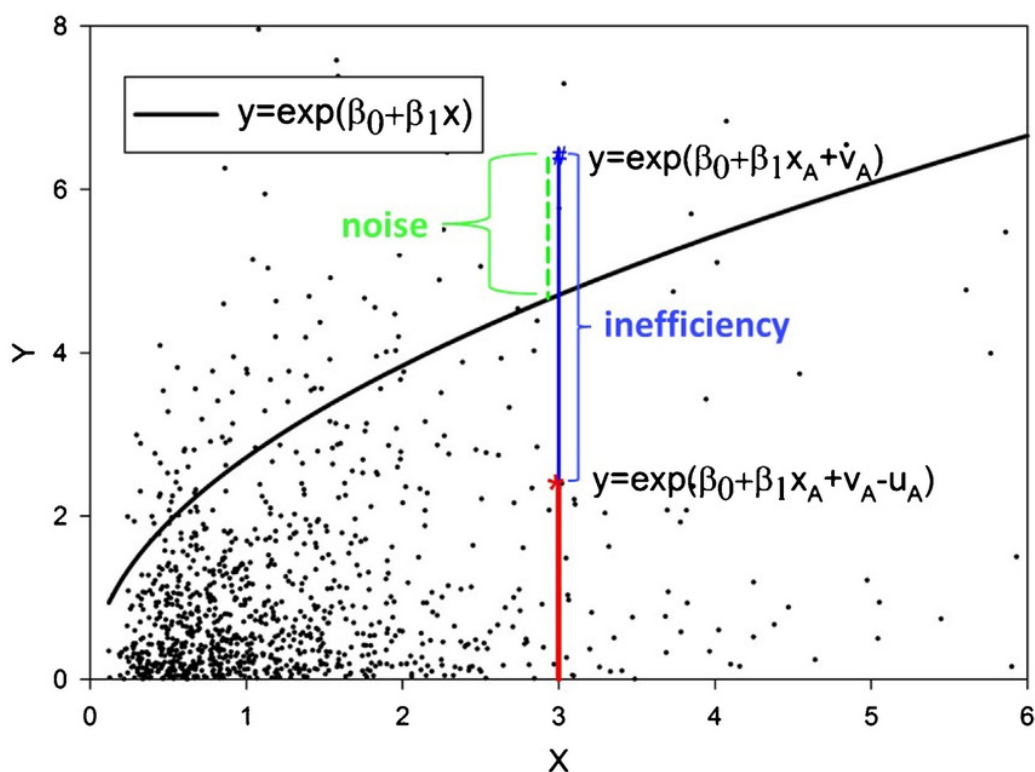
Pour chaque observation dans ce modèle, on récupère  $\epsilon_i$ , qui représente un écart à la frontière. La spécification de cette méthode permet donc d'estimer, à travers l'espérance conditionnelle de  $u_i$  sachant  $\epsilon_i$ , les scores de l'**efficacité technique** de chaque firme.

Enfin, Kumbhakar, Horncastle, et al. (2015) discutent aussi dans la section 3.3 de leur livre des approches dites *distribution-free* sur  $u_i$  dans lesquelles aucune hypothèse ne sont faites sur la distribution que suit les  $u_i$ . Nous ne nous intéresserons pas à ces méthodes puisqu'elles ont le défaut de ne pas pouvoir correctement distinguer les  $v_i$  des  $u_i$ , et donc ne sont pas en mesure d'estimer les scores d'efficacité technique.

On l'a vu ci-dessus, la SFA est une méthode **paramétrique** qui requiert une forme fonctionnelle précise. La SFA n'a cependant pas le monopole dans le domaine de l'estimation des frontières de production.

Un autre modèle (non-paramétrique) a aussi été développé : la Data Envelopment Analysis (DEA). Celui-ci a l'avantage de ne pas exiger d'hypothèse particulière sur des termes d'erreur. La structure du modèle n'est pas spécifiée à priori mais est uniquement déterminée à partir des données.

\* \* \*



**Figure 2** – Représentation graphique d'une SFA.

\* Droits d'auteur : **Lutz Bornmann**

À partir de cette représentation, on peut clairement distinguer les effets de  $v_i$  (**noise**) et ceux de  $u_i$  (**inefficiency**) dans un espace à deux dimensions avec  $X$  la quantité d'inputs et  $Y$  la quantité d'outputs. La frontière optimale de production est ici représentée en noir par  $y = \beta_0 + \beta_1 x$ .

- 2 entreprises utilisant la même quantité d'inputs ( $X = 3$ ) sont mises en évidence dans le graph. La première se situe en dessous de la frontière de production avec  $Y \simeq 2$  et la seconde est au-dessus de celle-ci avec  $Y > 6$ .
- Les 2 firmes utilisent donc la même quantité d'inputs pour une quantité d'output différente, à savoir : la première firme est moins efficace dans l'utilisation optimale de ses inputs, donc son efficacité technique est inférieure à la seconde.



## Utilisation empirique

### Quelques exemples d'application de la SFA dans le cadre de la mesure d'efficacité :

Reinhard, Lovell, et Thijssen (2000)  $\Rightarrow$  Secteur Environnemental.

- L'objectif de cet article est d'estimer l'efficacité environnementale pour les fermes laitières aux Pays-Bas.

Rosko et Mutter (2008)  $\Rightarrow$  Secteur Hospitalier.

- Cet article est quant à lui une méta-analyse de l'ensemble des articles de SFA et de DEA existants sur l'efficacité hospitalière aux Etats-Unis.

Mohamad, Hassan, et Bader (2008)  $\Rightarrow$  Secteur Bancaire.

- Compare l'efficacité des coûts et des profits de 80 banques dans 21 pays comprenant 37 banques conventionnelles et 43 banques islamiques.

---

### En bref, il existe de nombreux domaines d'application !

Un domaine en particulier n'a pourtant pas été évoqué jusqu'ici : pourquoi ne pas utiliser la SFA pour mesurer l'*efficacité* d'un prix (**best-buy frontier**) ?

C'est précisément le cadre du prochain article de notre revue de la littérature.

## SFA & Pricing Hédonique

Arrondo, Garcia, et Gonzalez (2018) :

**Objectif** : déterminer les attributs principaux des prix des sneakers en Espagne et leur efficacité.

Six caractéristiques<sup>(5)</sup> sont étudiées sur  $n = 171$  sneakers.

- **Lightweight** : poids des sneakers.
- **Cushioning** : capacité de la chaussure à absorber les chocs au cours d'une course et tout au long du cycle de vie du produit.
- **Flexibility** : les baskets flexibles s'adaptent mieux à la forme naturelle du pied.
- **Response** : capacité du matériau à retrouver sa forme après les déformations provoquées par l'impact sur le sol.
- **Grip** : l'adhérence donne aux coureurs une certaine assise sur le sol.
- **Stability** : mesure la stabilité du pied à l'intérieur de la chaussure.

En plus de ces 6 caractéristiques techniques, la marque est ajoutée en tant que variable qualitative pour mesurer la *Brand Equity* (la valeur d'une marque pour le consommateur).

---

<sup>(5)</sup>Variables quantitatives discrètes  $\in [1, 10[$ .

Le modèle pour la marque  $k$  s'écrit alors :

$$\ln(p_{ik}) = \alpha_k + \beta X_{ik} + v_{ik} + u_{ik} \quad (8)$$

- $p_{ik}$  est le prix du  $i$ -ème modèle de marque  $k$ .
- $\alpha_k$  est l'effet marque sur le prix de la marque  $k$ .
- $X_{ik}$  est le vecteur des attributs mesurables du  $i$ -ème modèle de marque  $k$ .
- $\beta$  est un vecteur de coefficients pour ces attributs.
- $v_{ik}$  est une erreur aléatoire.
- $u_{ik}$  représente l'inefficacité.

*Note* : On retrouve bien la forme spécifique d'une SFA, caractérisée par la présence des termes  $v_{ik}$  et  $u_{ik}$ . La seule différence est que le terme d'erreur composée est  $\epsilon_{ik} = v_{ik} + u_{ik}$  car nous sommes dans le cadre d'une **frontière de coût** et non de production.

## Résultats :

**Table 1** – Résultats de la régression hédonique

Variables	Coefficient	SE
<i>Lightness</i>	0.007	0.028
<i>Cushioning</i>	0.064 **	0.025
<i>Flexibility</i>	0.058 **	0.026
<i>Response</i>	0.050 *	0.30
<i>Stability</i>	0.070 ***	0.025
<i>Grip</i>	-0.045	0.028
Adidas	2.697 ***	0.401
Asics	2.679 ***	0.389
Saucony	2.779 ***	0.403
Nike	2.714 ***	0.422
Brooks	2.834 ***	0.404
Mizuno	2.524 ***	0.397
New Balance	2.544 ***	0.410
Reebok	2.522 ***	0.403

Les variables *Cushioning*, *Flexibility* et *Stability* sont statistiquement significatives à  $p < 0.05$ . De plus, nous sommes ici dans le cadre d'une régression log-linéaire donc les coefficients peuvent être interprétés comme des **semi-élasticités**, c'est à dire :

⇒ Pour une augmentation d'une unité de *Stability*,  $p_{ik}$  va augmenter de 7%, *cet. par.* <sup>(6)</sup>

⇒ Pour une augmentation d'une unité de *Cushioning*,  $p_{ik}$  va augmenter de 6.4%, *cet. par.*

⇒ Pour une augmentation d'une unité de *Flexibility*,  $p_{ik}$  va augmenter de 5.8%, *cet. par.*

<sup>(6)</sup>Toutes choses égales par ailleurs.

Par conséquent, la caractéristique *Stability* va avoir le plus grand impact sur le prix d'une sneakers, suivi de *Cushioning* et *Flexibility*.

**Table 2** – Indice d'efficacité moyen par marque

Marque	$\hat{\theta}_k$
Adidas ( $n = 28$ )	0.832
Asics ( $n = 35$ )	0.864
Saucony ( $n = 15$ )	0.875
Nike ( $n = 25$ )	0.824
Brooks ( $n = 16$ )	0.860
Mizuno ( $n = 29$ )	0.858
New Balance ( $n = 18$ )	0.848
Reebok ( $n = 5$ )	0.859

$\hat{\theta}_k$  représente l'indice d'efficacité moyen estimé par marque, compris entre 0 et 1.

On remarque tout d'abord que cet indice est compris entre 0.8 et 0.9 pour l'ensemble des marques, c'est à dire qu'il n'y a pas de marque globalement **très inefficace** (si une marque l'était, elle n'arriverait probablement pas à vendre et serait évincée par ses concurrents).

- Nike est la marque qui possède la pire relation prix~attributs de la sélection.
- Saucony est la marque qui possède la meilleure relation prix~attributs de la sélection.

## Résultats

- En estimant l'efficacité des produits, l'article permet de déterminer le montant des réductions à accorder aux sneakers **overprice** afin de les rendre compétitives.
- Il existe une relation inverse entre l'efficacité du produit et la réduction de prix : la réduction de prix est d'autant plus grande que la sneakers est **overprice**.

# | Choix et cadrage de la problématique

L'objectif fixé par notre sujet est de combiner les modèles SFA à une problématique d'étude des prix hédoniques, de manière similaire à ce qui a été entrepris par Arrondo, Garcia, et Gonzalez (2018).

L'ensemble des articles de la littérature exposés ci-dessus ont permis d'affiner notre compréhension théorique des modèles et nous ont aidés à déterminer un marché à étudier. Pour des raisons de disponibilité des caractéristiques et parce que peu d'articles dans la littérature se sont intéressés au pricing hédonique des smartphones, nous avons fait le choix d'analyser le marché de la téléphonie mobile.

Notre problématique est donc la suivante :

**Combinaison d'un modèle SFA et d'une régression hédonique pour évaluer l'écart entre les prix de smartphones et leur valeur (intrinsèque).**

## Le marché de la téléphonie mobile, en constante évolution

Depuis l'apparition des téléphones mobiles au début des années 1990, de nombreuses innovations technologiques ont ajouté des caractéristiques rendant ces téléphones de plus en plus polyvalents. Cette chronologie présente en X les années et les rectangles des différentes catégories correspondent à des **débuts** et des **fin de commercialisation**. L'axe Y permet quant à lui d'améliorer la lisibilité.

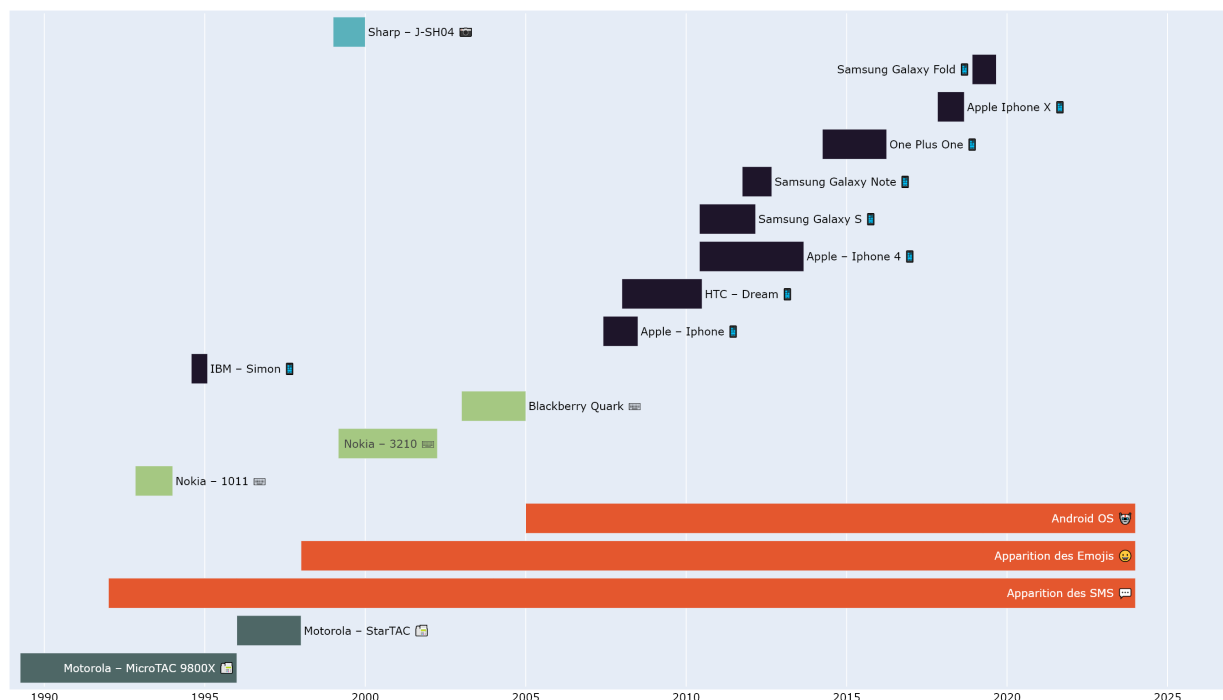




Figure 3 – Smartphone Timeline.

■ Téléphone à clapet ■ Téléphone à clavier ■ Téléphone à appareil photo ■ Smartphone

Examinons quelques modèles de téléphone pour mieux saisir l'impact des innovations majeures sur le marché.

- *Nokia 1011* : Premier écran LCD.
- *IBM Simon* : Premier véritable smartphone avec stylet, commercialisé pendant seulement 6 mois à cause d'un prix élevé de \$899.
- *Nokia 3210* : Premier téléphone à intégrer les SMS et plusieurs jeux. C'est encore aujourd'hui un des téléphones les plus vendus au monde.
- *Sharp J-SH04* : Premier téléphone équipé d'un appareil photo intégré.
- *Blackberry Quark* : Les téléphones  BlackBerry sont les premiers à disposer d'un clavier complet, ce qui, à cette époque, est un avantage majeur. A tel point qu'au début des années 2000 et jusqu'en 2010, BlackBerry devient et reste leader sur le marché de la téléphonie mobile avec 20% de parts de marché à son apogée.
- *Apple Iphone* : En 2007,  Apple annonce l'iPhone. Ce téléphone, qui intègre un écran tactile multitouch, va bouleverser le marché des téléphones mobiles. La vraie révolution, plus que le téléphone en lui-même, est l'*App Store*, qui va permettre d'accélérer le développement de nombreuses applications mobiles.
- *HTC Dream* : Un an après la sortie de l'iPhone, les constructeurs bataillent pour tenter de le concurrencer. HTC est dans ce cadre le premier à intégrer Android OS. Il reste néanmoins un entre-deux (il possède un clavier et un écran tactile).
- *Samsung Galaxy S* : Avec le Galaxy S, Samsung concurrence directement l'Apple iPhone 4 et sort un téléphone meilleur en tout point sur le plan des caractéristiques techniques. L'écran est plus grand, il existe une possibilité d'augmenter le stockage, il possède un meilleur cpu et une meilleure autonomie, tout en étant moins cher.

#### Conclusion

Toutes ces innovations vont avoir un impact dans les caractéristiques les plus valorisées par les consommateurs. Par exemple, il est difficile d'imaginer qu'un consommateur valorisera aujourd'hui un téléphone sans capteur de caméra frontale et arrière ou qui serait incapable d'envoyer des SMS.

Cela permet d'ailleurs d'évoquer une des limites majeures des modèles de pricing hédonique. Comment va-t-on pouvoir modéliser l'arrivée d'une nouvelle caractéristique ? On ne peut pas trouver dans le passé quelle sera la valorisation de cette nouvelle caractéristique.

*Illustrons cette remarque avec l'iPhone.* Un modèle de régression des prix hédoniques réalisé juste avant la sortie de l'iPhone aurait probablement trouvé (sans surprise) que BlackBerry était la marque la plus valorisée par les consommateurs et qu'il faut augmenter la taille du téléphone pour lui permettre d'avoir un plus grand clavier. Il va sans dire que deux mois plus tard, ces résultats sont inutilisables à cause d'une innovation technologique.

Enfin, il existe relativement peu d'articles sur les prix hédoniques des smartphones, ou alors ils sont assez anciens (2004-2005), et on l'a vu, étant donné la vitesse à laquelle

évolue le marché, avoir des données récentes est primordial pour estimer correctement les caractéristiques valorisées par les consommateurs à un instant  $T$ .

# Smartphones et Pricing Hédonique

Il existe néanmoins quelques articles récents traitant du sujet, dont celui de Ahmad, Ahmed, et Ahmad (2019) :

**Objectif :** Pricing des attributs des smartphones au Pakistan

Les données des attributs ont été collectées sur des sites webs et les prix pratiqués relevés dans les magasins de 2 villes du Pakistan ( $n = 348$  smartphones).

Le prix moyen d'un smartphone dans leur étude est de \$136,35. En outre, l'**écart-type** du prix des smartphones est élevé (181), c'est à dire que la dispersion en prix est assez importante, ce qui confirme l'hypothèse que les smartphones sont des biens différenciés.

*Ils proposent alors l'estimation du modèle suivant avec les caractéristiques découpées en variables catégorielles.*

$$\begin{aligned}\ln(PRICE_i) = & \beta_0 + \beta_{1i}BRAND_i + \beta_{2i}WEIGHT_i + \beta_{3i}BATTERY_i \\ & + \beta_{4i}OS_i + \beta_{5i}RAM_i + \beta_{6i}MEMORY_i + \beta_{7i}DISPLAY_i \\ & + \beta_{8i}NETWORK_i + \beta_{9i}BCAM_i + \beta_{10i}FCAM_i + \epsilon_i\end{aligned}$$

**Résultats :**

- La marque, la batterie, le poids, l'OS, la RAM, la mémoire et la taille de l'écran ont un effet positif statistiquement significatif sur les prix des smartphones.

Plus précisément, les résultats indiquent que les fabricants doivent se concentrer sur un téléphone :

- avec une RAM de plus d'1 Go.
- avec une mémoire de plus de 8 Go.
- avec un écran de plus de 5 pouces.
- compatible avec la 4G.
- avec une caméra arrière de plus de 15 mégapixels.

---

Le Pakistan étant un pays en voie de développement, et l'étude datant de 2019, on peut s'attendre à trouver des résultats différents dans nos données.

De plus, sur les 348 smartphones, 127, sont de la marque **QMOBILE**, une société pakistanaise qui vend des smartphones à bas prix, ce qui peut aussi expliquer le prix moyen assez bas.

# Effet de réputation

Dans la section précédente, les résultats de l'étude de Ahmad, Ahmed, et Ahmad (2019) ont permis de discerner que la marque a un effet statistiquement significatif sur le prix des smartphones, c'est pourquoi nous voulions explorer rapidement des questions d'analyse de la concurrence que l'on peut relier à notre sujet.

Boistel (2008) parle spécifiquement de cet effet de réputation.

**Objectif :** Analyser l'impact de la réputation sur les fonctions clés de l'entreprise et son intégration dans le management stratégique actuel.

## Réputation et marketing

- **Considéré comme une priorité majeure :** la réputation est une ressource essentielle, reconnue comme une priorité de recherche par le *Marketing Science Institute*.
- **Influence le comportement des consommateurs:** la réputation impacte l'intention d'achat, la confiance envers les nouveaux produits et est liée à la satisfaction client. Elle agit comme une "garantie".
- **Avantage compétitif :** une solide réputation permet de gagner un avantage compétitif sur le marché, voire un avantage concurrentiel, en attirant les clients et en se différenciant des concurrents.
- **Limitation de la concurrence :** les produits ou services d'une entreprise réputée sont moins facilement remplaçables ou imitables en raison de la perception limitée des consommateurs. Au lieu d'examiner les caractéristiques en détail, ils vont se fier à la marque et à la réputation.
- **Corrélation positive avec le prix :** une meilleure réputation va permettre de fixer des prix plus élevés et d'obtenir un avantage sur les ventes par rapport à la concurrence. *Exemple : Toyota et General Motors forment la joint-venture New United Motor Manufacturing Inc. La société a produit 2 voitures identiques :*
  - la Toyota Corrola.
  - la GM's Geo Prizm.

⇒ La meilleure réputation de Toyota lui a permis de vendre 200000 voitures à 11100 dollars, contre seulement 80000 véhicules à 10700 dollars vendus pour General Motors.



# | Acquisition des données

## Scraping

Il n'existe **évidemment** pas de données directement disponibles regroupant le prix et l'ensemble des caractéristiques des smartphones. Ce qui pourrait le plus s'en rapprocher sont les fiches techniques de téléphones disponibles sur <https://www.01net.com/>. Scraper ce site pourrait être une idée intéressante, mais *01net* n'est pas un revendeur de smartphones.

L'idée est donc de récupérer ces données sur le site d'un revendeur (*Fnac*, *Darty*, *Boulangier*). En effet, l'avantage de la récupération des données sur un site de revente direct est que nous avons une “*photographie*” du marché au moment où le *crawler* récupère et alimente notre base de données. A ce titre, nous avons choisi de récupérer des données sur *Boulangier*.

### Encadrement du web scraping

Le web scraping est encadré en droit français par l'article **L. 342-3<sup>(7)</sup>** du Code de la propriété intellectuelle, qui autorise la pratique suivante :

- L'extraction et la réutilisation d'une partie substantielle, appréciée de façon qualitative ou quantitative, à des fins exclusives d'illustration dans le cadre de l'enseignement et de la recherche et pour un public composé d'élèves, d'étudiants, d'enseignants ou de chercheurs directement concernés. Ainsi, **ce cas de figure étant limité à des fins pédagogiques, il est totalement exclu de faire usage des données extraites à titre commercial.**

Nous précisons donc que nous ne ferons en aucun usage de ces données dans un cadre commercial.

---

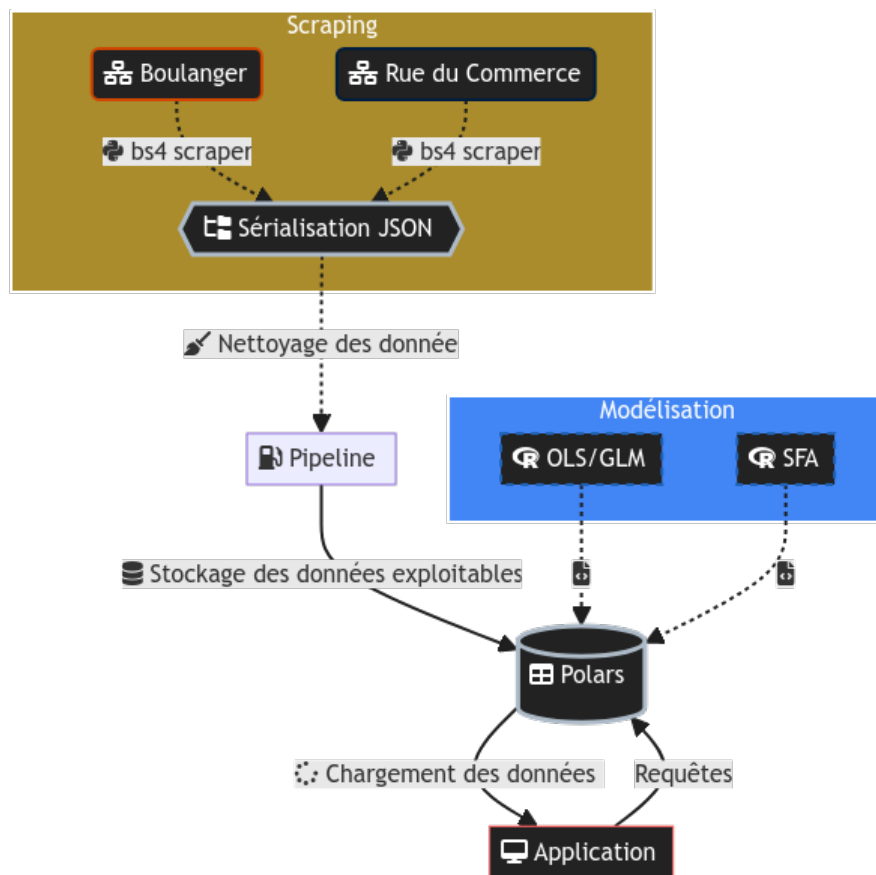
<sup>(7)</sup>Plus de détail sur [legifrance.gouv.fr](https://legifrance.gouv.fr)

## Méthodologie

L'objectif final est de disposer d'une application permettant aux consommateurs ou aux producteurs de comparer l'efficacité des smartphones en fonction de leurs caractéristiques et de leur indiquer quel est le meilleur choix.

### Workflow :

- *Scraping* ⇒ Python
- *Nettoyage des données* ⇒ Python
- *Modélisation* ⇒ R
- *Application* ⇒ Python



**Figure 4** – Diagramme fonctionnel.

Le diagramme fonctionnel permet de comprendre comment interagissent les différents composants logiciels avant leur utilisation dans l'application.

# | Statistiques descriptives

Il y a 487 smartphones valides disponibles dans nos données scrapées sur Boulanger avec 35 variables.

- Variables liées à l'écran : *screen\_type*, *screen\_size*, *screen\_tech*, *diagonal\_pixels*, *ppi*, *resolution\_1*, *resolution\_2*.
- Variables liées à la caméra : *mpx\_backward\_cam*, *cam\_1*, *cam\_2*, *cam\_3*, *sensor*.
- Variables liées aux caractéristiques physiques du téléphone : *color*, *thickness*, *width*, *height*, *net\_weight*.
- Variables liées aux performances : *network*, *cpu*, *ram*, *storage*, *upgrade\_storage*.
- Variables liées à la batterie : *battery*, *fast\_charging*, *induction*, *usb\_type\_c*.
- Variables liées au DAS : *das\_limbs*, *das\_chest*, *das\_head*.
- Autres variables : *repairability\_index*, *model*, *brand*, *made\_in*, *stars*, *reviews*.

Et enfin notre variable à expliquer : **price**.

## Analyse des prix

### Mesures de tendance centrale

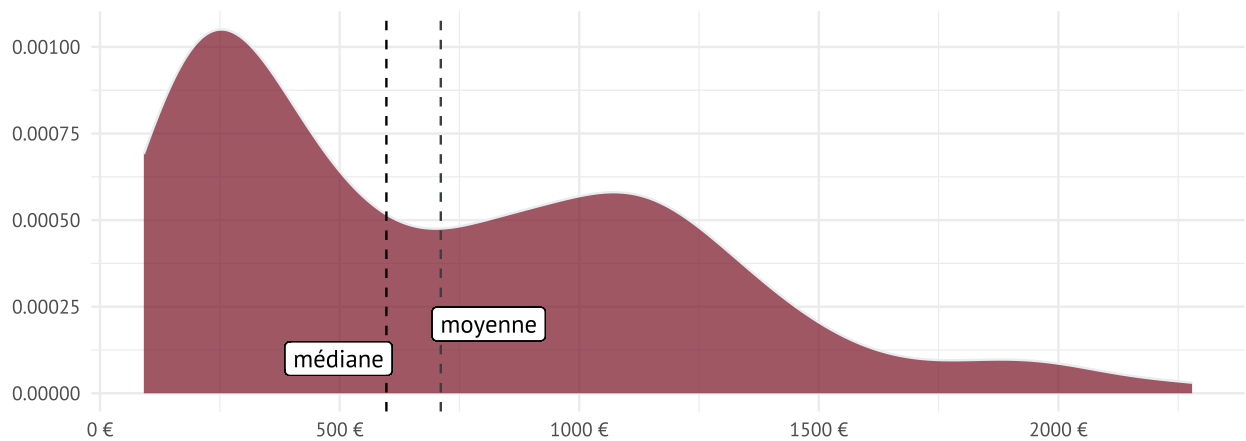
Le prix moyen d'un smartphone de la sélection est de 711.11 €, soit  $\simeq 5$  fois plus élevé que dans l'article de Ahmad, Ahmed, et Ahmad (2019). Cela peut s'expliquer notamment par la différence considérable de **PIB** par habitant.<sup>(8)</sup>

- En 2022 au Pakistan : \$ 1596.7
- En 2022 en France : \$ 40963.8

La médiane est quant à elle de 597.64 €.

---

<sup>(8)</sup>Données issues de la *Banque Mondiale* (PIB par habitant en US dollars courants)



**Figure 5** – Distribution des prix des smartphones.

On peut aussi tester l’asymétrie de la distribution avec le coefficient de *Skewness* de *Pearson* :

$$SK = \frac{3(\bar{x} - \tilde{x})}{\sigma} \simeq 0.67$$

Ce résultat indique que l’asymétrie de la distribution est positive. Il y a beaucoup plus de valeurs concentrées à gauche de la distribution qu’à droite. Si le coefficient était proche de 0, cela signifierait que la distribution est proche d’une loi normale  $\mathcal{N}$ , ce qui n’est pas le cas.

## Mesures de dispersion

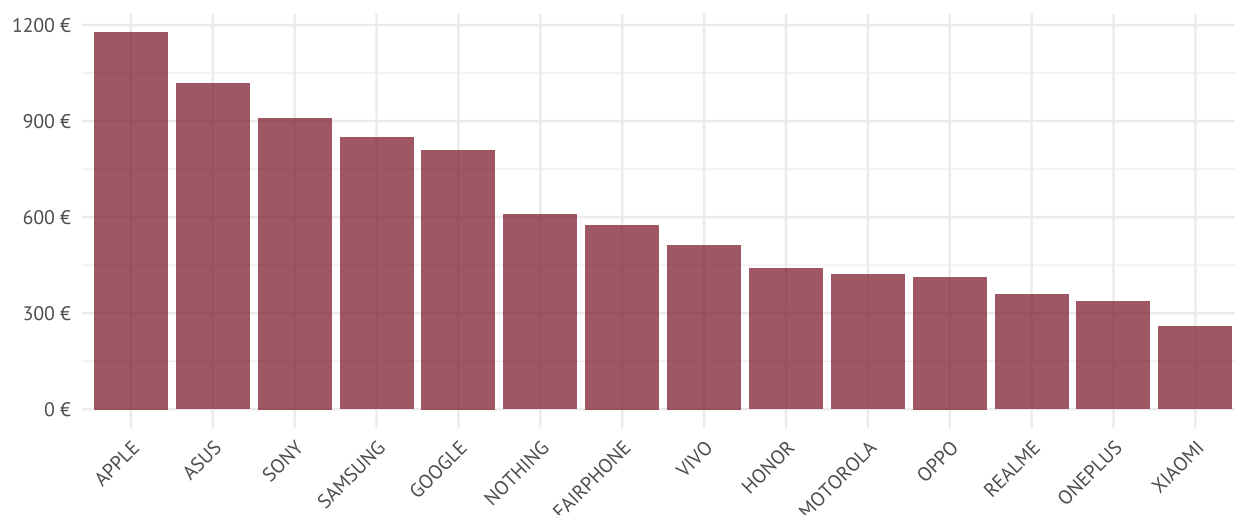
Le prix minimal d’un smartphone dans notre sélection est de 91.4 € pour le modèle **Xiaomi Redmi A1** et le prix maximal est de 2279 € pour le modèle **Samsung Galaxy Z Fold 5**.

Il existe donc une grande étendue de prix, c’est à dire : 2187.6 €. De plus, l’écart-type du prix est très important (506.36).

**Toutes ces mesures nous confirment que la dispersion en prix est très élevée.**

## Prix moyen en fonction d’autres variables

On peut s’intéresser au prix moyen par marque pour regarder s’il existe des différences de prix significatives entre certaines marques pour illustrer l’article de Boistel (2008) cité précédemment.



**Figure 6 – Prix moyen par marque.**

- *Apple* possède en moyenne les téléphones les plus chers dans l'échantillon (1177.75 €).
- Le prix moyen des smartphones commercialisés par *Samsung* est de 848.98 €. C'est certes moins qu'*Apple*, mais cela peut s'expliquer car *Samsung* commercialise à la fois des téléphones très haut de gamme et des téléphones bas de gamme aux prix beaucoup plus attractifs.
- En dernière position, on retrouve *Xiaomi* avec des téléphones à un prix moyen aux alentours de 260 €.

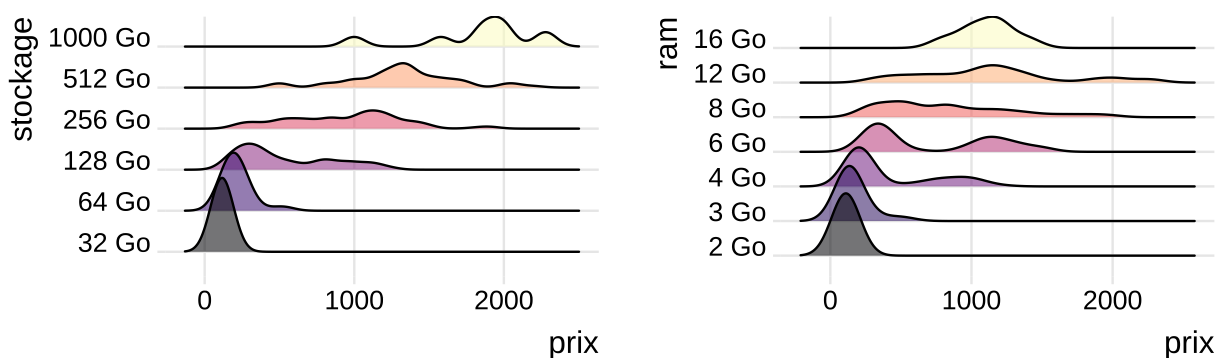
**Table 3** – Prix moyen en fonction de la RAM.

ram	$n$	prix moyen $\bar{p}$
2 Go	21	110.11 €
3 Go	16	165.35 €
4 Go	118	413.4 €
6 Go	87	722.22 €
8 Go	163	835.55 €
12 Go	75	1144.02 €
16 Go	7	1106.14 €

**Table 4** – Prix moyen en fonction du stockage.

stockage	$n$	prix moyen $\bar{p}$
32 Go	27	116.65 €
64 Go	55	214.76 €
128 Go	207	537.44 €
256 Go	127	930.65 €
512 Go	55	1318.86 €
1000 Go	16	1835.65 €

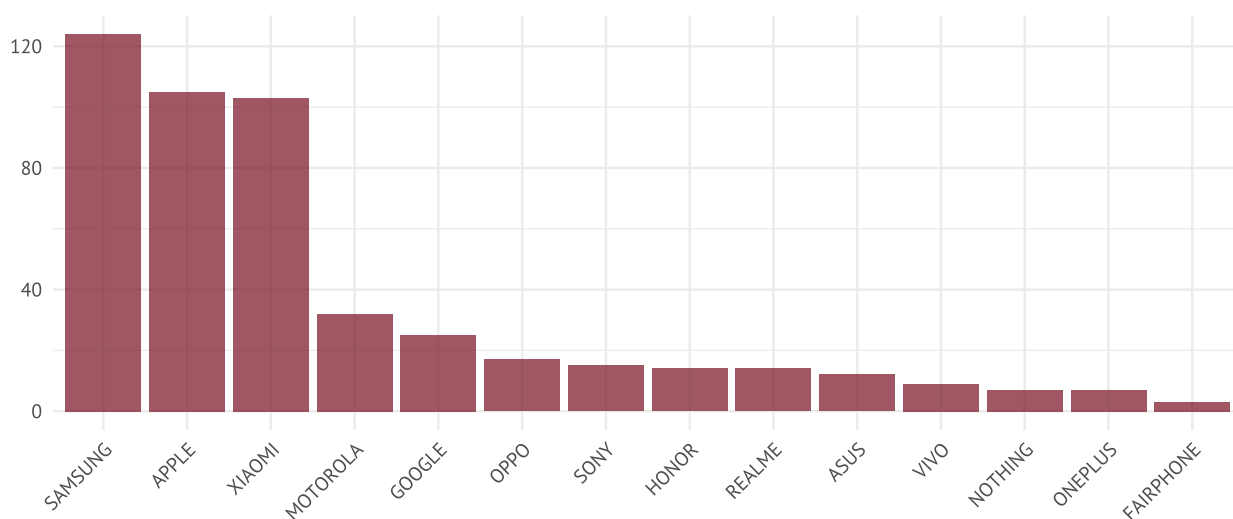
- On peut voir que plus la RAM augmente, plus le prix moyen du téléphone augmente (sauf à partir de 12 Gigaoctets). On a en effet remarqué plus haut que le téléphone le plus cher était le **Samsung Galaxy Z Fold 5**, qui possède 12 Go de RAM, et non 16. Ce constat, en plus du nombre très limité de téléphones disposant de 16 Go de RAM (7), peut expliquer pourquoi le prix moyen des téléphones ayant 16 Go de RAM est inférieur au prix moyen des téléphones disposant de 12 Go de RAM.
- Pour le second tableau, il existe une relation non-linéaire concernant le doublement de la capacité de stockage du téléphone. Par exemple, passer de 64 à 128 Go implique une augmentation du prix moyen de 250% alors que passer de 256 à 512 Go de stockage implique seulement 141% d'augmentation du prix moyen. Il convient aussi de préciser que  $\simeq 70\%$  des téléphones de notre échantillon possèdent entre 128 et 256 Go de capacité de stockage.



**Figure 7** – Ridge plot : Stockage et RAM

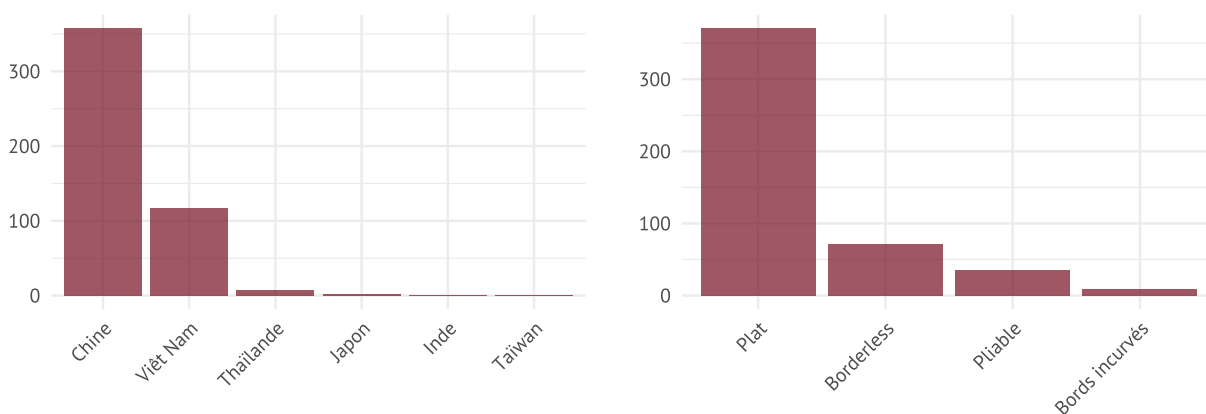
## Etudes des variables catégorielles importantes

Nous allons maintenant étudier les proportions des modalités des variables catégorielles.



**Figure 8** – Proportion des modèles par marque.

- Samsung, Apple & Xiamoi se partagent 70% des téléphones commercialisés sur Boulanger.
- On retrouve la même tendance au niveau des parts de marché mondial des smartphones par rapport à Q3 2022, c'est-à-dire que Samsung, Apple & Xiaomi se partagent respectivement 22, 18 et 14% de parts de marché.



**Figure 9** – Nombre de smartphones par lieu de fabrication et par type d'écran.

- On remarque que la majorité des smartphones sont fabriqués en Chine et au Viêt Nam (97.5%). Les 12 téléphones restants sont fabriqués en Thaïlande, au Japon, en Inde et à Taïwan.
- Concernant les types d'écran, 76.2% des écrans sont *plats*, 14.6% sont *borderless* (bord à bord), il y a 7.4% d'écrans pliables et finalement 1.8% d'écrans à bords incurvés.

## Etude des variables dichotomiques

Nos données contiennent 4 variables dichotomiques aux modalités **TRUE** ou **FALSE**.

Ces variables sont :

- *induction*  $\Rightarrow$  le téléphone dispose-t-il d'une charge à induction ?
- *fast\_charging*  $\Rightarrow$  le téléphone dispose-t-il d'une charge rapide ?
- *upgrade\_storage*  $\Rightarrow$  la capacité de stockage est-elle extensible ?
- *usb\_type\_c*  $\Rightarrow$  le téléphone possède-t-il un port USB type C ?



**Figure 10** – Proportion de modalités des variables dichotomiques.

Si les modalités sont plutôt bien équilibrées pour les variables *induction* et *upgrade\_storage*, ce n'est pas le cas pour la variable *usb\_type\_c* et le déséquilibre est surtout présent pour *fast\_charging*. En effet seulement 19 téléphones n'ont pas de charge rapide (ces 19 téléphones sont beaucoup moins chers et sont principalement des téléphones bas de gamme).

Pour la variable *usb\_type\_c*, il y a 120 téléphones sans chargeur USB type C et le fait de ne pas avoir d'USB type C fait augmenter le prix moyen par rapport aux téléphones qui en ont (les téléphones n'ayant pas d'USB type C sont principalement de marque *Apple*, ce qui explique en partie l'effet).

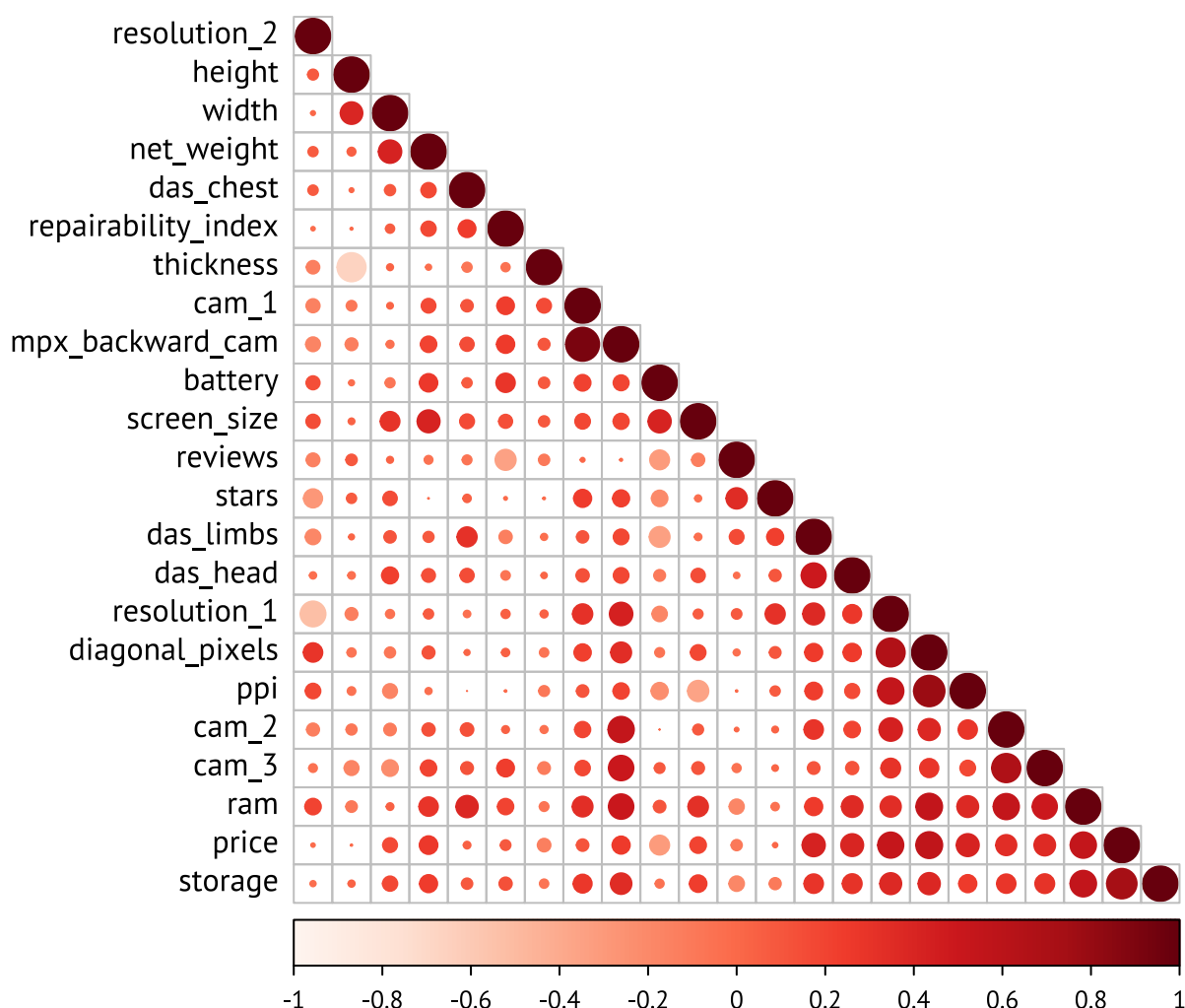
Le même effet d'augmentation du prix moyen s'observe avec la variable *upgrade\_storage*, c'est-à-dire que les téléphones ne possédant pas de système leur permettant d'augmenter leur stockage sont en moyenne plus chers que les téléphones offrant la possibilité de le faire. Ce qui peut paraître à première vue contre-intuitif ne l'est peut-être pas : les téléphones offrant la possibilité d'augmenter le stockage sont ceux qui en ont le moins, d'où la nécessité de laisser au consommateur la possibilité de pouvoir le faire. Inversement, les téléphones qui n'ont pas de système d'augmentation de stockage ont déjà un stockage important.

Les téléphones avec charge à induction sont eux en moyenne beaucoup plus chers que les téléphones ne disposant pas de charge à induction.



## Analyse des corrélations

Une analyse approfondie des corrélations entre l'ensemble des variables numériques disponibles va nous permettre de mesurer la force de la relation linéaire entre paires de variables. Cela nous sera particulièrement utile pour déterminer les variables explicatives fortement corrélées à notre variable à prédire **price**.



**Figure 11** – Matrice des corrélations.

On s'aperçoit que les corrélations les plus importantes avec **price** sont respectivement la capacité de stockage avec un coefficient de corrélation  $r_{price,storage} = 0.73$  et la RAM avec  $r_{price,ram} = 0.55$ . Inversement, la hauteur du téléphone, par exemple, semble ne pas avoir d'incidence sur le prix avec  $r_{price,height} = 0.002$ .

**Compte tenu de la littérature existante et de nos observations, la RAM et la capacité de stockage sont donc des variables incontournables dans la modélisation des prix des smartphones.**

# | Modélisation économétrique

## Sélection de variables

Avec 35 variables explicatives, le choix des variables *essentielles* à retenir est important.

La première question qui se pose, en amont de la modélisation, est celle de la **sélection de variables**. En effet, si nous sélectionnons trop de variables, nous risquons de faire du sur-apprentissage et de modéliser le bruit au lieu des liens statistiques existants. Le but est donc de trouver un ensemble optimal de variables.

Traditionnellement en économétrie appliquée, plusieurs approches heuristiques similaires décrites par Efroymson (1960) ont été utilisées pour le problème de la sélection de variables comme la *Backward Elimination* ou la *Forward Selection*.

Nous allons ici décrire en détail la procédure de *Forward Selection* :

1. On commence par un modèle  $M_0$ , c'est à dire avec constante seulement.
2. On ajoute les variables  $X_i$  une à une dans le modèle.
3. Parmi ces variables, on retient la variable **la plus significative** dans le modèle.
4. On réitère la procédure jusqu'à atteindre un modèle contenant uniquement des variables significatives à un seuil spécifié.

### ⚠ Limites

**Il existe cependant plusieurs problèmes dans ces méthodes.**

- La *Backward Elimination* et la *Forward Selection* ne convergent pas tout le temps vers le même modèle.
- Le modèle final n'est pas forcément optimal.

Pour pallier à ces problèmes, une autre approche possible consiste à faire une recherche exhaustive, c'est-à-dire explorer l'*ensemble* des modèles possibles. Bien que ce soit la meilleure méthode pour obtenir avec certitude le modèle optimal, elle devient rapidement inadaptée dès qu'il y a un nombre de variables trop conséquent, car le nombre de combinaisons possibles de modèles explose.

Néanmoins, une dernière approche existe : les *algorithmes génétiques*. Contrairement à la *forward selection* et la *backward elimination* qui sont des méthodes déterministes, les algorithmes génétiques sont eux stochastiques.

- On peut s'intéresser à plusieurs critères comme : le  $R^2$ , le  $R^2$  ajusté, l'*AIC*, le *BIC*, etc.

Cet algorithme est implémenté dans le package  de `{glmulti}`.

- On l'utilise quand le nombre de variables est très important.

- Consiste à générer au hasard une population de modèles candidats pour ensuite leur permettre d'évoluer. Cette évolution se déroule de génération en génération.
- Permet de trouver le meilleur modèle en explorant seulement un sous-ensemble de modèles (de manière aléatoire) mais avec un biais vers de meilleurs modèles grâce à la sélection.

## Modèle Hédonique niveau-niveau

Pour la facilité des interprétations, dans cette partie, nous commencerons par utiliser un modèle *level – level*, donc notre variable à prédire **price** ne sera pas mise sous forme logarithmique. Dans une seconde partie, nous préférons un modèle *log – level*, qui nous permettra de comparer directement les coefficients de la régression hédonique des prix avec les coefficients de la **SFA**.

Nous allons tout d’abord retirer quelques variables qui ne nous seront pas utiles pour l’analyse et qui risquent de complexifier les régressions : *cpu*, *model*, *sensor*, *screen\_tech*, *stars*, *reviews*, *color*, *cam\_1*, *cam\_2*, *cam\_3*. Par exemple, il y a presque autant de modèles distincts que de nombre d’observations, et *reviews* et *stars* sont si peu corrélées à la variable à expliquer **price** que les ajouter ne semble pas nécessaire.

**Table 5 –** Comparaison des méthodes de sélection (1).

Méthode	Modèle	AIC	AIC <sub>wt</sub>	BIC	BIC <sub>wt</sub>	R <sup>2</sup> <sub>adj</sub>	RMSE
backward	lm	6218.90	0.72	6386.43	0	0.93	132.14
forward	lm	6220.85	0.27	6380.00	0	0.93	132.95
genetic	lm	6229.89	0.00	6363.91	1	0.92	135.85

**Quelques commentaires :** On s’aperçoit que la méthode *backward* et *forward* n’ont pas convergé vers le même modèle : plus précisément, certaines variables comme *diagonal\_pixels* ou *battery* ont été sélectionnées dans la *backward* mais pas dans la *forward* (**On le pressentait, c’est une limite de la méthode**). Plus de détail peut être trouvé dans les résultats des régressions ci-dessous.

On préférera aussi regarder comme critère le *BIC* car il est plus parcimonieux que l’*AIC*.

$$BIC = -2 \ln(\mathcal{L}) + k \cdot \ln(N)$$

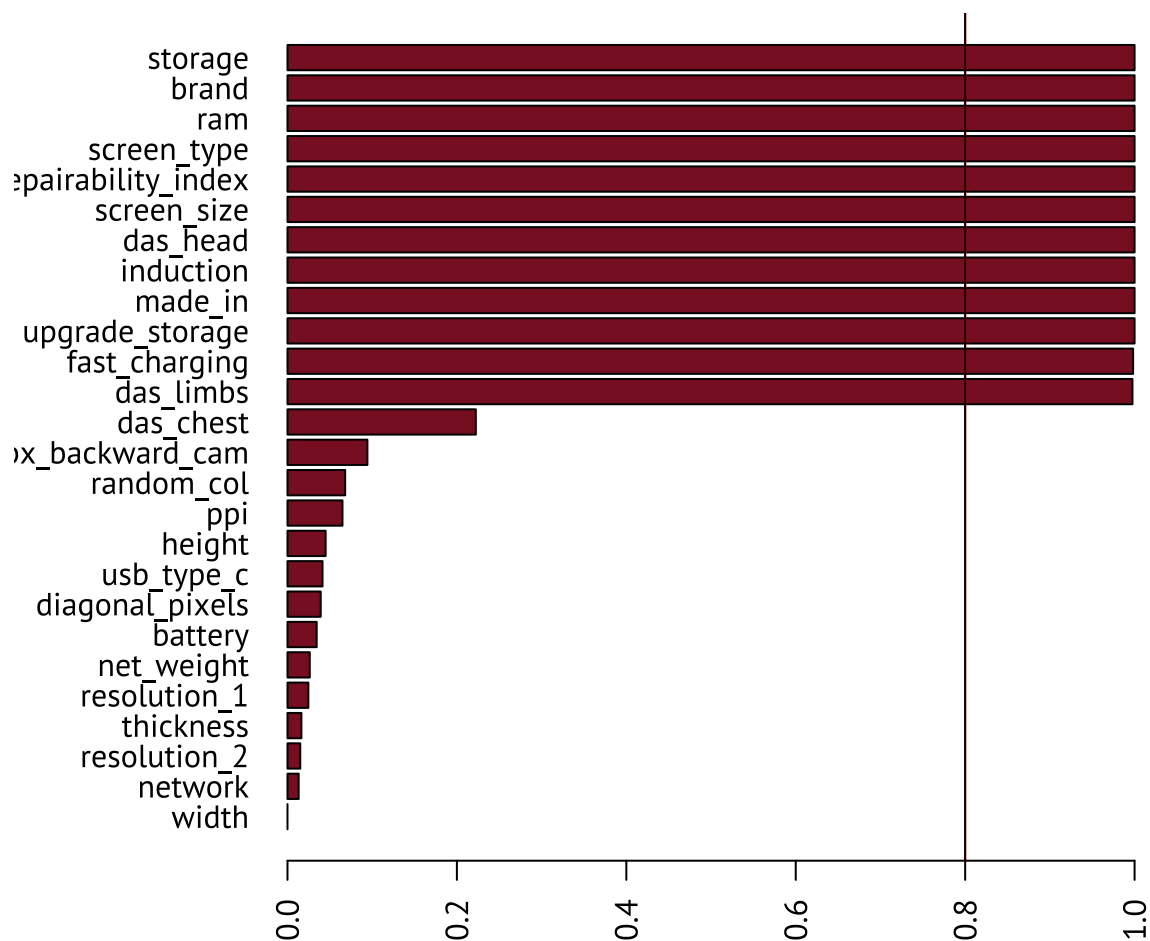
- Avec  $\mathcal{L}$  la vraisemblance du modèle estimée,  $N$  le nombre d’observations dans l’échantillon et  $k$  le nombre de paramètres libres du modèle.

Comparé à l’*AIC*, il pénalise plus le nombre de variables présentes dans le modèle. On voit d’ailleurs dans ce tableau que l’algorithme génétique possède le *BIC* le plus bas, tout en ayant un  $R^2_{adj}$  très légèrement inférieur aux méthodes *backward* et *forward*. On le verra aussi dans le tableau des résultats de régression, mais la régression trouvée par l’algorithme génétique possède moins de variables (12), et celles-ci sont toutes significatives à un seuil  $p < 0.01$ .

Le modèle que nous allons étudier ici sera donc :

$$\begin{aligned}
 price_i = & \beta_0 + \beta_{1i}storage_i + \beta_{2i}brand_i + \beta_{3i}ram_i + \beta_{4i}screen\_type_i \\
 & + \beta_{5i}induction_i + \beta_{6i}screen\_size_i + \beta_{7i}repairability\_index_i \\
 & + \beta_{8i}made\_in_i + \beta_{9i}upgrade\_storage_i + \beta_{10i}fast\_charging_i \\
 & + \beta_{11i}das\_limbs_i + \beta_{12i}das\_chest_i + \epsilon_i
 \end{aligned} \tag{9}$$

On voit que dans le modèle de l’équation 9, il n’y a pour l’instant aucun effet d’interaction ou d’effet non-linéaire. Pour autant, les résultats sont déjà **très satisfaisants** avec un  $R^2_{adj} > 0.9$



**Figure 12** – Importance des variables.

L'importance d'une variable, dans ce contexte, est égale à la somme des probabilités pour les modèles dans lesquels la variable apparaît. Ainsi, une variable présente dans de nombreux modèles avec des poids importants recevra une valeur d'importance élevée. La ligne rouge verticale est tracée à 0.8, correspondant au seuil différenciant les variables importantes des variables moins importantes, mais ce choix (80%) est arbitraire.

Ainsi, on remarque que 11 variables sont présentes dans l'ensemble des modèles, donc ces variables sont très importantes. Ensuite, une seule variable est au-dessus de la ligne de *cutoff* mais inférieure à 1 : *das\_head*. L'ensemble des autres variables sont comprises entre 0 et 0.25. On peut donc considérer que l'importance de ces variables dans notre modèle est négligeable.

**Table 6 – Comparaison des méthodes de sélection (2).**

	<i>Dependent variable:</i>		
	forward	price backward	genetic
	(1)	(2)	(3)
storage	0.749*** (0.047)	0.748*** (0.047)	0.742*** (0.047)
brandASUS	-408.264*** (68.796)	-414.207*** (73.424)	-501.156*** (58.681)
brandFAIRPHONE	-596.063*** (93.349)	-551.526*** (93.980)	-592.163*** (93.040)
brandGOOGLE	-473.395*** (60.739)	-464.841*** (64.695)	-493.572*** (52.844)
brandHONOR	-834.094*** (60.479)	-808.301*** (62.620)	-863.679*** (57.603)
brandMOTOROLA	-632.526*** (50.179)	-606.074*** (51.282)	-678.762*** (39.917)
brandNOTHING	-832.362*** (68.803)	-807.903*** (68.568)	-896.865*** (59.170)
brandONEPLUS	-562.556*** (68.355)	-549.520*** (68.277)	-614.648*** (61.358)
brandOPPO	-593.068*** (50.540)	-576.039*** (50.665)	-628.747*** (45.326)
brandREALME	-480.079*** (52.804)	-471.156*** (53.481)	-469.539*** (49.256)
brandSAMSUNG	-439.536*** (58.339)	-427.905*** (58.337)	-501.056*** (47.367)
brandSONY	-467.249*** (76.229)	-517.148*** (78.426)	-433.391*** (72.424)
brandVIVO	-577.364*** (57.733)	-560.288*** (57.863)	-580.769*** (55.456)
brandXIAOMI	-595.182*** (37.741)	-585.623*** (37.671)	-581.323*** (35.091)
ram	37.750*** (4.407)	37.089*** (4.419)	42.632*** (3.984)
resolution_1		0.435** (0.169)	
resolution_2		0.425** (0.165)	
battery		0.029* (0.017)	
screen_typeBords incurvés	-192.989*** (70.750)	-173.485** (69.119)	-147.108** (67.210)
screen_typePlat	-123.016*** (30.006)	-115.179*** (31.347)	-100.500*** (27.764)
screen_typePliable	304.117*** (44.035)	300.618*** (47.009)	316.851*** (43.490)
induction	168.220*** (23.570)	174.465*** (23.639)	156.811*** (23.219)
screen_size	167.430*** (22.390)	113.845*** (22.974)	154.687*** (18.536)
repairability_index	96.173*** (13.963)	92.514*** (14.347)	108.973*** (13.539)
diagonal_pixels		-0.540** (0.221)	
made_inInde	-114.552 (147.903)	-90.982 (147.403)	-21.635 (144.976)
made_inJapon	562.032*** (122.398)	575.013*** (122.145)	562.520*** (123.872)
made_inTaïwan	364.924** (146.439)	371.614** (147.234)	376.405** (148.161)
made_inThaïlande	520.712*** (84.016)	530.761*** (83.910)	540.906*** (83.742)
made_inViêt Nam	-42.481 (40.358)	-46.251 (40.181)	-27.599 (39.736)
upgrade_storage	-121.057*** (26.861)	-141.890*** (27.060)	-133.052*** (25.669)
fast_charging	-159.191*** (37.214)	-135.971*** (37.830)	-145.993*** (36.887)
das_limbs	87.183*** (19.892)	86.927*** (20.274)	72.879*** (18.845)
das_head	-122.268*** (37.963)	-137.406*** (38.904)	-105.158*** (38.084)
das_chest	-150.742** (59.690)	-142.616** (60.016)	
mpx_backward_cam	0.334 (0.243)		
random_col	-0.018 (0.012)	-0.017 (0.012)	
usb_type_c	39.228* (22.151)	42.748* (21.936)	
height	-0.306* (0.156)	-0.331** (0.158)	
ppi	0.170 (0.114)		
Constant	-924.262*** (195.491)	-697.920*** (174.947)	-1,050.746*** (155.543)
Observations	487	487	487
R <sup>2</sup>	0.931	0.932	0.928
Adjusted R <sup>2</sup>	0.925	0.926	0.923
Residual Std. Error	138.303 (df = 450)	137.768 (df = 448)	140.390 (df = 456)
F Statistic	168.461*** (df = 36; 450)	160.983*** (df = 38; 448)	195.544*** (df = 30; 456)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## Interprétations Modèle niveau-niveau

- Si la capacité de stockage (*storage*) augmente de 1 Go, alors, le prix augmente de 0.74 €, *cet. par.*
- En ayant comme catégorie de référence *Apple* pour la variable marque (*brand*), on peut voir que toutes les marques ont un impact négatif sur le prix, *cet. par.*
  - La marque la plus valorisée derrière *Apple* est *Sony* avec 433.39 € de différence par rapport à *Apple*.
  - La marque la moins valorisée est *Nothing* avec 896.87 € de différence par rapport à *Apple*
- Une augmentation d'un Go de *ram* augmente le prix de 42.63 €, *cet. par.*
- Pour le type d'écran (*screen\_type*), la catégorie de référence est *borderless* (un écran sans bordure).
  - Disposer d'un écran à bord incurvé diminue le prix de 147.11 € par rapport à la catégorie de référence, *cet. par.*
  - Pour l'écran plat, le prix diminue de 100.5 € par rapport à la catégorie de référence, *cet. par.*
  - Avoir un écran pliable fait augmenter le prix de 316.85 € par rapport à la catégorie de référence, *cet. par.*
- Posséder un dispositif de charge à induction (*induction* = **TRUE**) augmente le prix de 156.81 €, *cet. par.*
- Si la taille de l'écran augmente de 1 pouce, alors le prix augmente de 154.69 €, *cet. par.*
- L'indice de réparabilité (*repairability\_index*) est compris entre 1 et 10. Une augmentation de 1 point de cet indice implique une augmentation de 108.97 €, *cet. par.*
- Concernant le lieu de fabrication (*made\_in*), la catégorie de référence est la *Chine*.
  - Les coefficients associés aux catégories *Inde* et *Viêt Nam* ne sont pas significatives, c'est-à-dire qu'il n'y a pas de différence significative de prix avec un smartphone produit en *Chine*.
  - Comparé à un téléphone fabriqué en Chine, un téléphone produit au *Japon* augmente le prix de 562.52 €, suivi de la *Thaïlande* avec une augmentation de 540.91 € et enfin de *Taiwan* avec une augmentation du prix de 376.4 €, *cet. par.*
- Si le téléphone dispose d'un moyen d'augmenter sa capacité de stockage (*upgrade\_storage* = **TRUE**), alors le prix diminue de 133.05 €, *cet. par.*\* Comme nous l'avons remarqué dans la partie de statistiques descriptives.
- Si le téléphone dispose d'une charge rapide (*fast\_charging* = **TRUE**), alors le prix diminue de 145.99 €, *cet. par.*

- Pour les variables liées au DAS, l'effet de *das\_limbs* est positif sur le prix, tandis que l'effet de *das\_head* est quant à lui négatif. L'augmentation d'une unité de Watts par kilogramme du *das\_limbs* augmente le prix de 72.88 € alors que cette même augmentation diminue le prix de 105.16 € pour *das\_head*, *cet.par.* Néanmoins, l'effet peut être globalement vu comme négatif étant donné la valeur des coefficients.

$R_{adj}^2 = 0.923$  donc 92.3% de la variance de la variable expliquée (price) est expliquée par la variance des variables explicatives du modèle.



## Modèle Hédonique log-niveau

Dans cette partie, nous n'effectuerons pas de sélection de variables avec `{glmulti}` car le meilleur modèle de régression hédonique proposé en *log – level* aboutit à des difficultés de convergence pour le modèle SFA. Or, nous voulons comparer les coefficients de cette régression hédonique avec les modèles SFA que nous mettrons en place dans une troisième phase.

### ⚠ Difficultés de convergence d'un modèle SFA

Dans une SFA, les difficultés de convergence peuvent être dues à de vastes zones très plates de la fonction de vraisemblance, pouvant être causées par une forte multicolinéarité entre l'intercept du modèle de frontière et l'intercept du modèle d'inefficacité. Par conséquent, supprimer l'intercept du modèle d'inefficacité améliore parfois la convergence, mais risque aussi d'introduire un biais de variable omise.

Dans le package `{frontier}` que nous utilisons pour effectuer la SFA, ces difficultés sont en général indiquées par le message d'avertissement suivant :

- “le paramètre gamma est proche de la limite de l'espace des paramètres [0,1]”

Cela signifie que l'estimation du paramètre  $\gamma$  est soit proche de la limite inférieure de son espace de paramètres (0), soit proche de la limite supérieure de son espace de paramètres (1). Par exemple, un  $\gamma$  proche de zéro indique qu'il n'y a presque aucune inefficacité, donc qu'il serait possible d'estimer le modèle par MCO, tandis qu'un  $\gamma$  proche de un indique qu'il n'y a presque aucun bruit, donc qu'il serait possible d'utiliser une DEA. Plus globalement, cela indique une mauvaise spécification du modèle ou que d'autres modèles pourraient être plus appropriés dans ce cadre.

Après plusieurs essais, le modèle que nous avons choisi d'étudier est :

$$\begin{aligned} \ln(price_i) = & \beta_0 + \beta_{1i}storage_i + \beta_{2i}brand_i + \beta_{3i}ram_i + \beta_{4i}induction_i \\ & + \beta_{5i}screen\_size_i + \beta_{6i}made\_in_i + \beta_{7i}upgrade\_storage_i \\ & + \beta_{8i}das\_head_i + \beta_{9i}das\_limbs_i + \beta_{10i}das\_chest_i \\ & + \beta_{11i}fast\_charging_i + \beta_{12i}network_i + \beta_{13i}ppi_i + \epsilon_i \end{aligned} \quad (10)$$

Par rapport au modèle en niveau, il n'y a plus les variables *screen\_type* et *repairability\_index*, qui posaient des problèmes de convergence. D'autres variables ont néanmoins été introduites comme *network* ou *ppi*.

**Table 7 –** Modèle de Pricing Hédonique Log-Linéaire

	<i>Dependent variable:</i>
	log(price)
storage	0.001*** (0.0001)
brandASUS	-0.357*** (0.109)
brandFAIRPHONE	-0.179 (0.140)
brandGOOGLE	-0.430*** (0.090)
brandHONOR	-0.987*** (0.080)
brandMOTOROLA	-0.718*** (0.073)
brandNOTHING	-0.877*** (0.107)
brandONEPLUS	-0.652*** (0.107)
brandOPPO	-0.664*** (0.077)
brandREALME	-0.644*** (0.080)
brandSAMSUNG	-0.151* (0.085)
brandSONY	-0.356*** (0.115)
brandVIVO	-0.589*** (0.088)
brandXIAOMI	-0.759*** (0.056)
ram	0.082*** (0.007)
induction	0.272*** (0.037)
screen_size	0.241*** (0.033)
made_inInde	-0.225 (0.233)
made_inJapon	0.414** (0.189)
made_inTaïwan	0.475** (0.234)
made_inThaïlande	0.543*** (0.134)
made_inViêt Nam	-0.149*** (0.053)
upgrade_storage	-0.363*** (0.039)
das_head	-0.209*** (0.061)
das_limbs	0.092*** (0.030)
das_chest	-0.299*** (0.095)
fast_charging	-0.138** (0.058)
network5G	0.306*** (0.037)
ppi	0.001*** (0.0002)
Constant	4.320*** (0.253)
Observations	487
R <sup>2</sup>	0.933
Adjusted R <sup>2</sup>	0.929
Residual Std. Error	0.221 (df = 457)
F Statistic	219.790*** (df = 29; 457)

*Note:* \* p<0.1; \*\* p<0.05; \*\*\* p<0.01

## Interprétations Modèle log-niveau

- Si la capacité de stockage (*storage*) augmente de 1 Go, alors le prix augmente de 0.1%, *cet. par.*
- En ayant comme catégorie de référence *Apple* pour la variable marque (*brand*), on peut voir que toutes les marques ont un impact négatif sur le prix, *cet. par.*
  - La marque la plus valorisée derrière *Apple* est *Samsung* avec une différence de prix de 15.1%.
  - La marque la moins valorisée est *Honor* avec une différence de prix de 98.7% par rapport à *Apple*.
- Si la *ram* augmente de 1 Go, alors le prix augmente de 8.2%, *cet. par.*
- Si le téléphone dispose d'une charge à induction (*induction* = **TRUE**), alors le prix augmente de 27.2%, *cet. par.*
- Si la taille de l'écran augmente de 1 pouce (*screen\_size*), alors le prix augmente de 24.1%, *cet. par.*
- Pour le lieu de fabrication (*made\_in*), la catégorie de référence est la *Chine*.
  - Le coefficient associé à la catégorie *Inde* n'est pas significatif.
  - Un téléphone qui est fabriqué au *Viêt Nam* est 14.9% moins cher qu'un téléphone fabriqué en *Chine*, *cet. par.*
  - Comparé à un téléphone produit en *Chine*, un téléphone produit en *Thaïlande* augmente le prix de 54.3%, suivi de *Taiwan* avec une augmentation de 47.5% et enfin du *Japon* avec une augmentation du prix de 41.4%, *cet. par.*
- Si le téléphone dispose d'un moyen d'augmenter sa capacité de stockage (*upgrade\_storage* = **TRUE**), alors le prix diminue de 36.3%, *cet. par.*
- Pour les variables liées au DAS :
  - l'augmentation d'une unité de Watts par kilogramme du *das\_limbs* augmente le prix de 9.2%, *cet. par.*
  - au contraire, le prix diminue de 20.9% pour le *das\_head* et de 29.9% pour le *das\_chest* lorsque qu'ils augmentent d'une unité de W/kg, *cet. par.*
- Si le téléphone dispose d'une charge rapide (*fast\_charging* = **TRUE**), alors le prix diminue de 13.9%, *cet. par.*
- Si le téléphone est compatible avec la 5G (*network5G* = **TRUE**), alors le prix augmente de 30.6%, *cet. par.*
- Lorsque le *ppi* (*pixels par pouce*) augmente d'une unité, alors le prix augmente de 0.1%, *cet. par.*

$R^2_{adj} = 0.929$  donc 92.9% de la variance de la variable expliquée est expliquée par la variance des variables explicatives du modèle.

## Vérification des hypothèses

Un aspect crucial lors de la construction des modèles de régression linéaire est d'évaluer la qualité de l'ajustement du modèle, mais aussi de vérifier les hypothèses de **Gauss-Markov** :

1. Linéarité dans les paramètres.
2. Les résidus sont d'espérance nulle.
3. Absence de multicollinéarité.
4. Homoscédasticité des résidus.
5. Pas d'autocorrélation des résidus.

**Sous toutes ces hypothèses sont satisfaites, l'estimateur fourni par les MCO est dit BLUE (Best Linear Unbiased Estimator).** Pour les vérifier, on peut utiliser les packages `{performance}` et `{see}` en conjonction<sup>(9)</sup>.

L'hypothèse (1) est vérifiée par la formulation de notre modèle. Pour l'hypothèse (2), on peut effectuer un test de *Student* pour vérifier si  $E(\epsilon_i) = 0$ . Les hypothèses  $H_0$  et  $H_1$  sont les suivantes :

$$\begin{cases} H_0 : E(\epsilon_i) = 0 \\ H_1 : E(\epsilon_i) \neq 0 \end{cases}$$

- La *p-value* issue du test est égale à 1, donc l'hypothèse (2) est vérifiée.

**Table 8** – Vérification de la multicollinéarité.

Variable	VIF	IC {low}	IC {high}	Tolérance
storage	1.99	1.77	2.27	0.50
brand	470.82	398.96	555.65	0.00
ram	4.28	3.71	4.98	0.23
induction	3.23	2.82	3.74	0.31
screen_size	1.95	1.73	2.22	0.51
made_in	20.32	17.30	23.90	0.05
upgrade_storage	3.72	3.23	4.31	0.27
das_head	1.96	1.74	2.23	0.51
das_limbs	2.53	2.23	2.91	0.40
das_chest	4.81	4.16	5.60	0.21
fast_charging	1.27	1.17	1.43	0.79
network	2.30	2.04	2.64	0.43
ppi	2.35	2.08	2.70	0.42

- *Note* : La **Tolérance** correspond à  $\frac{1}{VIF}$ , c'est-à-dire que plus le *VIF* sera élevé, et plus la **Tolérance** sera proche de 0.

<sup>(9)</sup>Plus d'infos sur les packages : `see` et `performance`

### 💡 Mesure de la multicolinéarité

Le Variance Inflation Factor (*VIF*) est une mesure permettant d'analyser l'ampleur de la multicolinéarité des termes du modèle. Plus précisément :

- Un *VIF* inférieur à 5 indique une faible corrélation de ce prédicteur avec d'autres prédicteurs.
- Une valeur comprise entre 5 et 10 indique une corrélation modérée.
- Enfin, des valeurs de *VIF* supérieures à 10 sont le signe d'une corrélation très élevée.

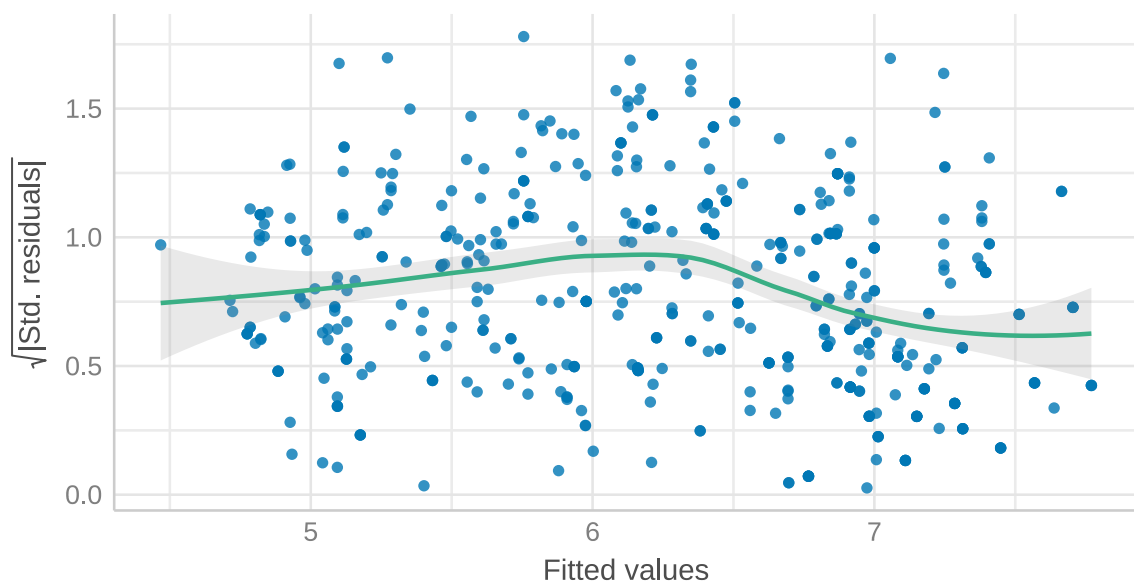
Ce que le *VIF* apporte en plus d'une simple analyse des corrélations est l'analyse d'un cas où une variable est fortement corrélée à une combinaison linéaire de plusieurs variables.

Avec un *VIF* estimé de 470.82, il n'est pas surprenant de constater que la marque est extrêmement corrélée à une combinaison linéaire de plusieurs variables. On peut par exemple penser au choix de localisation de la production, le fait de disposer ou non d'un port USB type C, etc. L'hypothèse (3) n'est donc pas vérifiée, mais cela est cohérent dans le cadre du modèle.

Concernant l'hypothèse (4) - l'Homoscédasticité pour les modèles de régression linéaire signifie que les résidus du modèle ont une variance constante. Si cette hypothèse n'est pas respectée, alors les *erreurs-types* et les *p-values* du modèle ne sont plus fiables. On peut par exemple détecter la présence d'hétéroscédasticité graphiquement :

#### Homogeneity of Variance

Reference line should be flat and horizontal



**Figure 13** – Homoscédasticité des résidus.

- Il est clair que notre modèle présente des résidus hétéroscédastiques. On peut aussi le vérifier avec le test de *Breusch-Pagan*. Les hypothèses  $H_0$  et  $H_1$  sont les suivantes :

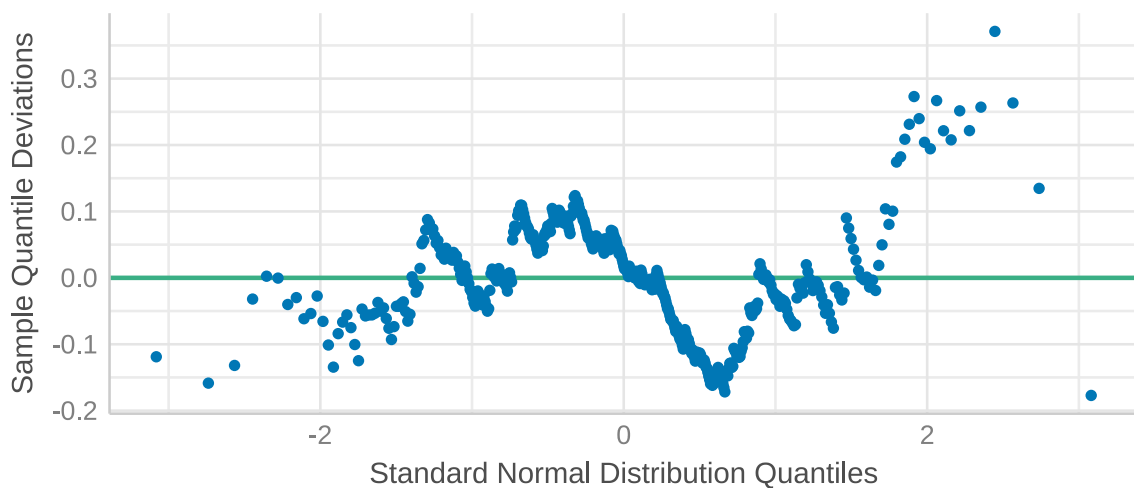
$$\begin{cases} H_0 : V(\epsilon_i) = \sigma^2 \\ H_1 : V(\epsilon_i) = \sigma_i^2 \end{cases}$$

- La  $p$  – *value* issue du test est égale à 0, donc l'hypothèse (4) n'est pas vérifiée.

**Il existe néanmoins plusieurs façons de corriger le problème d'hétéroscédasticité que nous verrons dans la section suivante.**

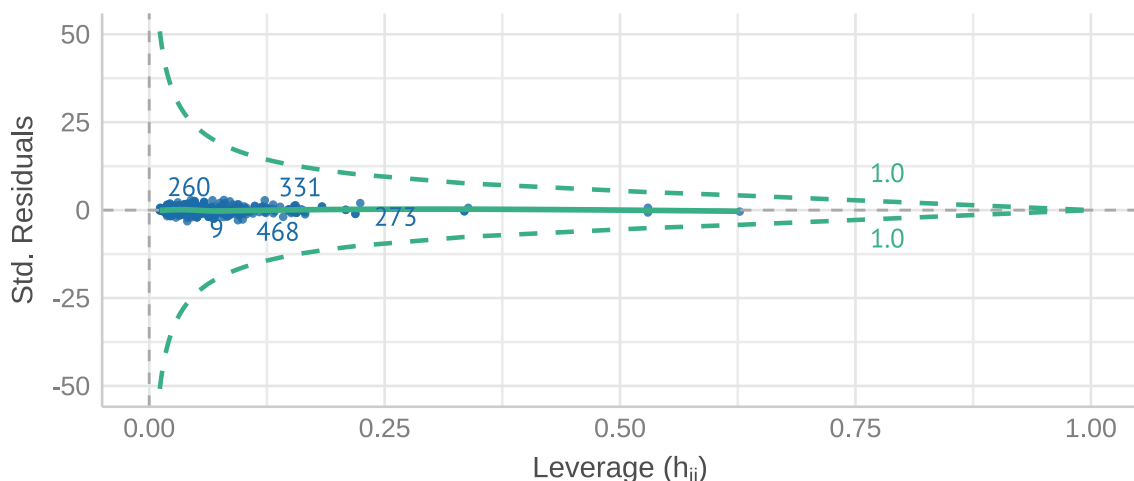
En complément, on peut aussi tester la normalité des résidus, et vérifier si certaines observations ont une influence très importante.

Normality of Residuals  
Dots should fall along the line



**Figure 14** – Normalité des résidus (QQ Plot).

Influential Observations  
Points should be inside the contour lines



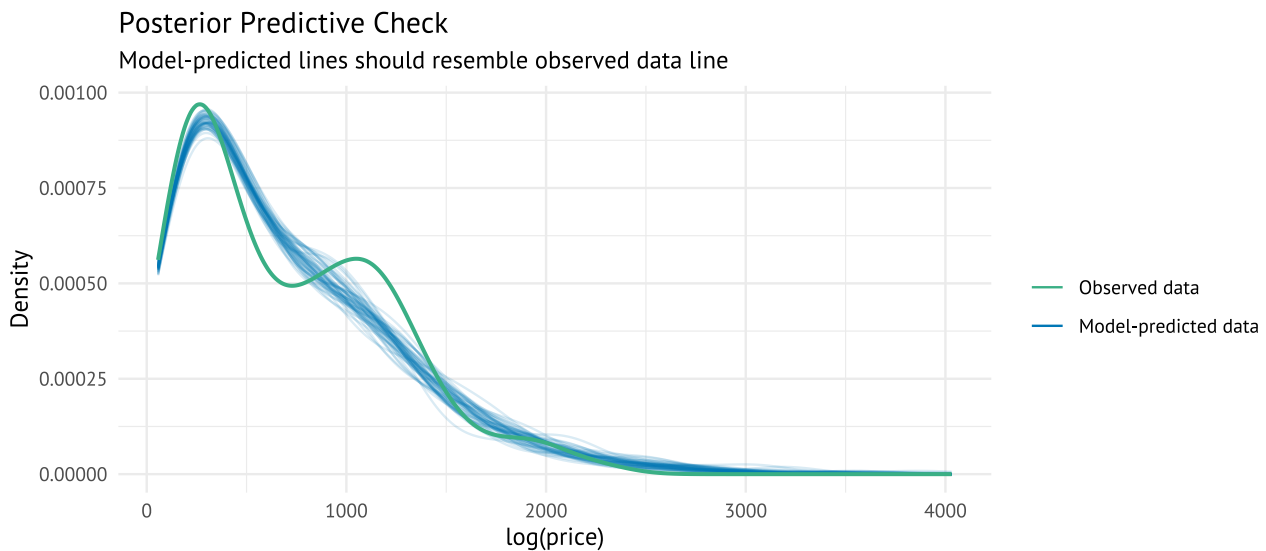
**Figure 15** – Influence des valeurs extrêmes.

- Les résidus ne suivent pas une loi normale et les *outliers* n'influent pas sur les résultats des prédictions.

Enfin, la dernière étape visuelle que l'on peut mettre en œuvre est de vérifier l'ajustement du modèle. Une vérification prédictive a posteriori permet de simuler des données répliquées avec le modèle ajusté puis les comparer aux données observées.

⇒ **L'objectif est de détecter si le modèle est inadéquat pour décrire les données.**

*Note* : Les données utilisées pour la vérification prédictive postérieure sont simulées à partir de la *distribution prédictive postérieure*. Celle-ci est construite après avoir utilisé les données observées  $y$  et les prédicteurs  $X$  pour mettre à jour les croyances sur les paramètres inconnus  $\theta$  dans le modèle. Pour chaque tirage des paramètres  $\theta$  à partir de la distribution a posteriori  $p(\theta|y, X)$ , un vecteur complet de résultats est généré.



- La distribution des prix est bimodale, mais le modèle a du mal à représenter le second pic de la distribution. Cependant, les prédictions sont assez proches de la réalité.

## Modèle SFA - Frontière de coût

Le modèle reste le même que celui de l'équation 10, mais le terme d'erreur  $\epsilon_i$  devient un terme d'erreur composé  $u_i + v_i$ .

Pour modéliser cette frontière de coût, il suffit d'utiliser l'argument `ineffDecrease = FALSE`, autrement dit l'inefficacité augmente la variable endogène.

Il faut maintenant sélectionner la distribution des  $u_i$ . Deux distributions sont disponibles dans le package `{frontier}` :

- La distribution normale tronquée : `truncNorm = TRUE`.
- La distribution semi-normale : `truncNorm = FALSE`.

Nous avons décidé de choisir la distribution *normale tronquée*.

$$\begin{aligned} \ln(\text{price}_i) = & \beta_0 + \beta_{1i}\text{storage}_i + \beta_{2i}\text{brand}_i + \beta_{3i}\text{ram}_i + \beta_{4i}\text{induction}_i \\ & + \beta_{5i}\text{screen\_size}_i + \beta_{6i}\text{made\_in}_i + \beta_{7i}\text{upgrade\_storage}_i \\ & + \beta_{8i}\text{das\_head}_i + \beta_{9i}\text{das\_limbs}_i + \beta_{10i}\text{das\_chest}_i \\ & + \beta_{11i}\text{fast\_charging}_i + \beta_{12i}\text{network}_i + \beta_{13i}\text{ppi}_i + \underbrace{\epsilon_i}_{u_i+v_i} \end{aligned} \quad (11)$$



**Table 9** – Résultats de l'estimation du modèle SFA (Cost Frontier).

	Coefficient	Significativité	Erreur Std
(Intercept)	4.116	***	0.2456
storage	0.0005	***	0.0001
brandASUS	-0.3969	***	0.1093
brandFAIRPHONE	-0.1905		0.1341
brandGOOGLE	-0.5307	***	0.0883
brandHONOR	-1.0333	***	0.0795
brandMOTOROLA	-0.8242	***	0.0754
brandNOTHING	-0.8953	***	0.1029
brandONEPLUS	-0.6939	***	0.104
brandOPPO	-0.7771	***	0.0774
brandREALME	-0.6867	***	0.0777
brandSAMSUNG	-0.2571	**	0.0847
brandSONY	-0.4151	***	0.1127
brandVIVO	-0.6263	***	0.0861
brandXIAOMI	-0.7933	***	0.0559
ram	0.0837	***	0.0065
inductionTRUE	0.289	***	0.0361
screen_size	0.2504	***	0.032
made_inInde	-0.1278		0.2268
made_inJapon	0.4198	*	0.1759
made_inTaïwan	0.5157	*	0.2204
made_inThaïlande	0.5602	***	0.1279
made_inViêt Nam	-0.09		0.0515
upgrade_storageTRUE	-0.3295	***	0.0376
das_head	-0.1986	***	0.059
das_limbs	0.0777	**	0.0296
fast_chargingTRUE	-0.1706	**	0.0558
das_chest	-0.2577	**	0.092
network5G	0.3049	***	0.0357
ppi	0.0009	***	0.0002
sigmaSq	0.1826	*	0.0864
gamma	0.843	***	0.0772
mu	-0.7846		0.6691

- **SigmaSq** ( $\sigma^2$ ) représente la variance de l'erreur composite  $\epsilon_i$  dans le modèle.
- **gamma** ( $\gamma$ ) représente le paramètre d'inefficacité stochastique.  $u_i$  est modélisé en fonction de ce  $\gamma$  estimé pour capturer l'inefficacité inobservable.

## Interprétations Modèle frontière de coût

### 💡 Un estimateur différent

L'estimateur utilisé dans la régression hédonique *log-level* est l'estimateur MCO. Celui-ci vise à minimiser la somme des carrés des différences entre les valeurs observées ( $y_i$ ) et celles prédites par le modèle ( $\hat{y}_i$ ). Dans le modèle SFA, on utilise plutôt l'approche du MV (Maximum de Vraisemblance) pour estimer les paramètres. Au lieu de se concentrer sur la minimisation des résidus, le MV cherche à maximiser la probabilité d'observer les données réellement observées, c'est-à-dire qu'il cherche à maximiser la fonction de vraisemblance.

- Si la capacité de stockage (*storage*) augmente de 1 Go, alors le prix augmente de 0.052%, *cet. par.*
- En ayant comme catégorie de référence *Apple* pour la variable marque (*brand*), on peut voir que toutes les marques ont un impact négatif sur le prix, *cet. par.*
  - La marque la plus valorisée derrière *Apple* est *Samsung* avec une différence de prix de 25.7%.
  - La marque la moins valorisée est *Honor* avec une différence de prix de 103.3% par rapport à *Apple*.
- Si la *ram* augmente de 1 Go, alors le prix augmente de 8.4%, *cet. par.*
- Si le téléphone dispose d'une charge à induction (*induction* = **TRUE**), alors le prix augmente de 28.9%, *cet. par.*
- Si la taille de l'écran augmente de 1 pouce (*screen\_size*), alors le prix augmente de 25%, *cet. par.*
- Pour le lieu de fabrication (*made\_in*), la catégorie de référence est la *Chine*.
  - Le coefficient associé à la catégorie *Inde* et *Viêt Nam* n'est pas significatif.
  - Comparé à un téléphone produit en *Chine*, un téléphone produit en *Thaïlande* augmente le prix de 56%, suivi de *Taiwan* avec une augmentation de 51.6% et enfin du *Japon* avec une augmentation du prix de 42%, *cet. par.*
- Si le téléphone dispose d'un moyen d'augmenter sa capacité de stockage (*upgrade\_storage* = **TRUE**), alors le prix diminue de 32.9%, *cet. par.*
- Pour les variables liées au DAS :
  - L'augmentation d'une unité de Watts par kilogramme du *das\_limbs* augmente le prix de 7.8%, *cet. par.*
  - Au contraire, le prix diminue de 19.9% pour le *das\_head* et de 25.8% pour le *das\_chest* lorsque qu'ils augmentent d'une unité de W/kg, *cet. par.*
- Si le téléphone dispose d'une charge rapide (*fast\_charging* = **TRUE**), alors le prix diminue de 17.1%, *cet. par.*

- Si le téléphone est compatible avec la 5G ( $network5G = \text{TRUE}$ ), alors le prix augmente de 30.5%, *cet. par.*
- Lorsque le  $ppi$  (*pixels par pouce*) augmente d'une unité, alors le prix augmente de 0.09%, *cet. par.*

## Analyse comparative des deux modèles

On peut aussi s'intéresser à l'étude des coefficients obtenus par le modèle de régression log-hédonique et les coefficients obtenus par le modèle SFA.

⇒ On remarque que la significativité et le signe des coefficients restent inchangés. La valeur des coefficients est aussi similaire. Il y a donc une certaine stabilité des résultats.

**Table 10** – Comparaison des coefficients obtenus par les 2 modèles.

	Log-Hedonic			SFA Cost Frontier		
	Coef.	Signif.	Erreur Std	Coef.	Signif.	Erreur Std
(Intercept)	4.3197	***	0.2526	4.116	***	0.2456
storage	0.0005	***	0.0001	0.0005	***	0.0001
brandASUS	-0.357	**	0.1091	-0.3969	***	0.1093
brandFAIRPHONE	-0.179		0.1401	-0.1905		0.1341
brandGOOGLE	-0.4298	***	0.0897	-0.5307	***	0.0883
brandHONOR	-0.9869	***	0.0801	-1.0333	***	0.0795
brandMOTOROLA	-0.7181	***	0.0732	-0.8242	***	0.0754
brandNOTHING	-0.8769	***	0.1065	-0.8953	***	0.1029
brandONEPLUS	-0.6518	***	0.1065	-0.6939	***	0.104
brandOPPO	-0.6639	***	0.0771	-0.7771	***	0.0774
brandREALME	-0.6436	***	0.0795	-0.6867	***	0.0777
brandSAMSUNG	-0.1513		0.0847	-0.2571	**	0.0847
brandSONY	-0.356	**	0.1154	-0.4151	***	0.1127
brandVIVO	-0.5893	***	0.0875	-0.6263	***	0.0861
brandXIAOMI	-0.7591	***	0.0561	-0.7933	***	0.0559
ram	0.0815	***	0.0068	0.0837	***	0.0065
inductionTRUE	0.2724	***	0.0365	0.289	***	0.0361
screen_size	0.241	***	0.0332	0.2504	***	0.032
made_inInde	-0.2249		0.2332	-0.1278		0.2268
made_inJapon	0.4138	*	0.1888	0.4198	*	0.1759
made_inTaïwan	0.4755	*	0.2336	0.5157	*	0.2204
made_inThaïlande	0.5434	***	0.1336	0.5602	***	0.1279
made_inViêt Nam	-0.1489	**	0.0531	-0.09		0.0515
upgrade_storageTRUE	-0.3633	***	0.039	-0.3295	***	0.0376
das_head	-0.209	***	0.0612	-0.1986	***	0.059
das_limbs	0.0921	**	0.0301	0.0777	**	0.0296
das_chest	-0.2993	**	0.0945	-0.1706	**	0.0558
fast_chargingTRUE	-0.1384	*	0.0583	-0.2577	**	0.092
network5G	0.3058	***	0.0365	0.3049	***	0.0357
ppi	0.0008	***	0.0002	0.0009	***	0.0002

## Quel modèle choisir ?

La question qui se pose est la suivante : quel modèle choisir entre la régression hédonique et la SFA ?

Pour répondre à cette question, nous pouvons utiliser le test de rapport de vraisemblance :

$$\begin{cases} H_0 : OLS \text{ log-level} \\ H_1 : SFA \text{ cost frontier} \end{cases}$$

Avec la statistique de test  $\lambda_{RV} = 2 \cdot (\ln \mathcal{L}_1 - \ln \mathcal{L}_2)$

On obtient :

- Log-Vraisemblance pour **OLS log-level** :  $\ln \mathcal{L}_1(\theta; y, X) = 59.03$
- Log-Vraisemblance pour **SFA Cost Frontier** :  $\ln \mathcal{L}_2(\theta; y, X) = 60.62$

Avec  $y$  le vecteur de  $\ln(\text{price})$  et  $X$  la matrice de variables indépendantes.

⇒ On rejette l'hypothèse nulle  $H_0$  car la  $p$  - *value* issue du test = **0.09** < 0.05.

Notre choix se porte donc vers le modèle SFA pour notre analyse puisqu'il offre un meilleur ajustement, suite au rejet de l'hypothèse nulle à l'issue du test de rapport de vraisemblance. En fait, ce n'est pas particulièrement surprenant. Cela souligne simplement le fait que des coefficients significatifs ont été identifiés au niveau des paramètres utilisés pour estimer l'inefficacité dans les résultats de l'analyse de l'estimation SFA Cost Frontier.

## Analyse de l'efficacité

Dans ce modèle, l'efficacité moyenne calculée est de **87.13%**, c'est-à-dire qu'en moyenne, les smartphones de notre échantillon sont *overprice* de **12.87%** !

En utilisant l'espérance conditionnelle  $E(\exp(-u_i) | \epsilon_i)^{(10)}$ , on peut estimer le score d'efficacité pour chaque observation, et donc déterminer quels téléphones sont les plus/moins efficaces de notre sélection. On peut aussi les regrouper par marque.

- *Oppo* est la marque qui possède la pire relation prix~attributs de notre sélection.
- *Nothing* est la marque qui possède la meilleure relation prix~attributs de notre sélection.

*Apple* fait partie du Top 3 des marques les plus efficaces, derrière *Nothing* et *Fairphone* (Il faut néanmoins se souvenir que ces 2 marques ne proposent que très peu de modèles de téléphones, contre 105 pour *Apple*).

**Enfin, on peut aussi s'intéresser aux téléphones les plus proches/les plus éloignés de la frontière de coût.**

---

<sup>(10)</sup>Pour plus de détails de calcul, voir l'**annexe** à la fin de ce document.

**Table 11** – Indice d'efficacité moyen par marque.

Marque	$\hat{\theta}_k$	$n$
OPPO	0.824	17
MOTOROLA	0.834	32
HONOR	0.865	14
SAMSUNG	0.869	124
GOOGLE	0.870	25
ASUS	0.874	12
VIVO	0.875	9
XIAOMI	0.876	103
REALME	0.877	14
ONEPLUS	0.878	7
SONY	0.881	15
APPLE	0.885	105
FAIRPHONE	0.892	3
NOTHING	0.894	7

**Table 12** – Les 5 téléphones les moins efficaces.

model	efficiency	ram	storage	price	logprice	prediction
OPPO Find X5	0.564	8	256	1022.14 €	6.930	6.090
MOTOROLA Razr	0.565	8	256	853.6 €	6.749	5.912
OPPO A77	0.572	4	64	359.16 €	5.884	5.061
Samsung Galaxy S22+	0.606	8	128	1000.33 €	6.908	6.155
Samsung XCOVER 5	0.608	4	64	299 €	5.700	4.950

- Avec une efficacité de 0.564 et un prix de 1022.14 €, le *OPPO Find X5* est très loin de la frontière de coût : son prix est 43.6% *trop cher*.
- Le *Samsung XCOVER 5* est un autre cas *polaire* : même un téléphone moyen de gamme d'une marque comme *Samsung* peut être un mauvais rapport qualité-prix.

**Table 13** – Les 5 téléphones les plus efficaces.

model	efficiency	ram	storage	price	logprice	prediction
iPhone 8 Plus	0.952	3	64	299 €	5.700	6.009
Xiaomi Redmi 12 5G	0.953	8	256	219 €	5.389	5.722
ASUS ROG Phone 6	0.958	16	512	799 €	6.683	7.101
Google Pixel 6 Pro	0.958	12	128	633.67 €	6.452	6.872
XIAOMI Redmi Note 10	0.963	4	64	159 €	5.069	5.591

- L'*ASUS ROG Phone 6* possède un très bon score d'efficacité, supérieur à 0.95. Avec un prix de 799 €, ce téléphone est donc un excellent rapport qualité-prix. Cela peut s'expliquer par sa capacité de stockage de 512 Go et son importante RAM de 16 Go.

- Le *XIAOMI Redmi Note 10* est quant à lui le téléphone le plus efficace de l'ensemble de notre échantillon.

**Table 14** – Les 5 téléphones les moins chers.

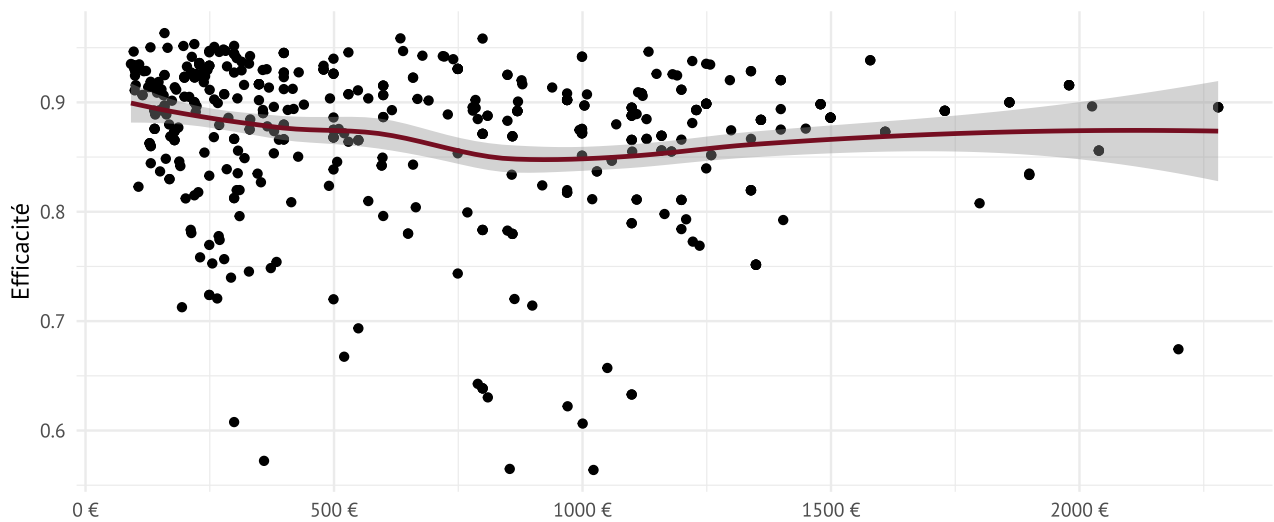
model	efficiency	ram	storage	price	logprice	prediction
Xiaomi Redmi A1	0.935	2	32	91.4 €	4.515	4.640
Xiaomi Redmi 9A	0.934	2	32	95.9 €	4.563	4.679
Xiaomi Redmi 9AT	0.946	2	32	96.91 €	4.574	4.814
Xiaomi Poco C40	0.933	3	32	98.84 €	4.594	4.703
Motorola E13	0.930	2	64	99 €	4.595	4.680

- Parmi ces téléphones les moins chers, on remarque qu'il y a une prépondérance de modèles issus de marque *Xiaomi*. Cela est cohérent avec nos observations dans la partie de **statistiques descriptives**.
- Par ailleurs, les 5 téléphones les moins chers sont en dessous de la barre symbolique des 100 € et pour autant, leur efficacité est supérieure à 0.9 pour l'ensemble des modèles. Les téléphones peu chers ne sont donc pas forcément signe d'un mauvais rapport "qualité-prix".

**Table 15** – Les 5 téléphones les plus chers.

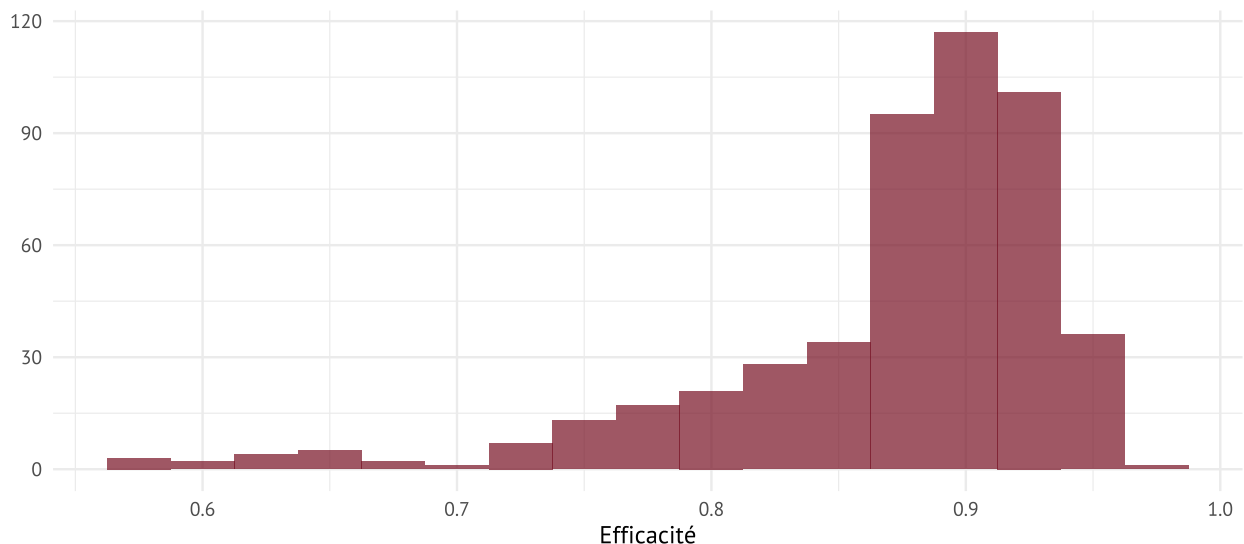
model	efficiency	ram	storage	price	logprice	prediction
iPhone 15 Pro Max	0.916	8	1000	1979 €	7.590	7.578
Samsung Galaxy Z Fold 4	0.896	12	1000	2025.35 €	7.613	7.505
Samsung Galaxy Z Fold 5	0.856	12	512	2039 €	7.620	7.371
Samsung Galaxy Z Fold 3	0.674	12	512	2199.09 €	7.696	7.071
Samsung Galaxy Z Fold 5	0.896	12	1000	2279 €	7.731	7.619

Les tableaux de la section précédente nous donnent un aperçu des valeurs extrêmes, mais on peut synthétiser l'ensemble des informations grâce au nuage de points des efficacités individuelles par téléphone en fonction des prix.



**Figure 16** – Efficacités en fonction des prix.

On remarque une fois de plus que moins les téléphones sont chers, et plus ils sont efficaces. Une des explications que nous proposons est qu'il existe une intensité concurrentielle élevée entre les fabricants dans cette gamme de prix et peu de marge disponible lorsque le prix est bas.



**Figure 17** – Distribution des efficacités.

La distribution est dite *left skewed*, et l'asymétrie de la distribution est négative. Il y a beaucoup plus de valeurs concentrées à droite de la distribution qu'à gauche.



## Analyse Factorielle de Données Mixtes

A posteriori, on cherche à synthétiser l'ensemble des variables de notre modèle de **Stochastic Cost Frontier**. Pour cela, nous allons effectuer une analyse factorielle qui nous permet de prendre en compte simultanément des variables quantitatives et qualitatives en tant que variables actives : une **AFDM** (Analyse Factorielle de Données Mixtes).

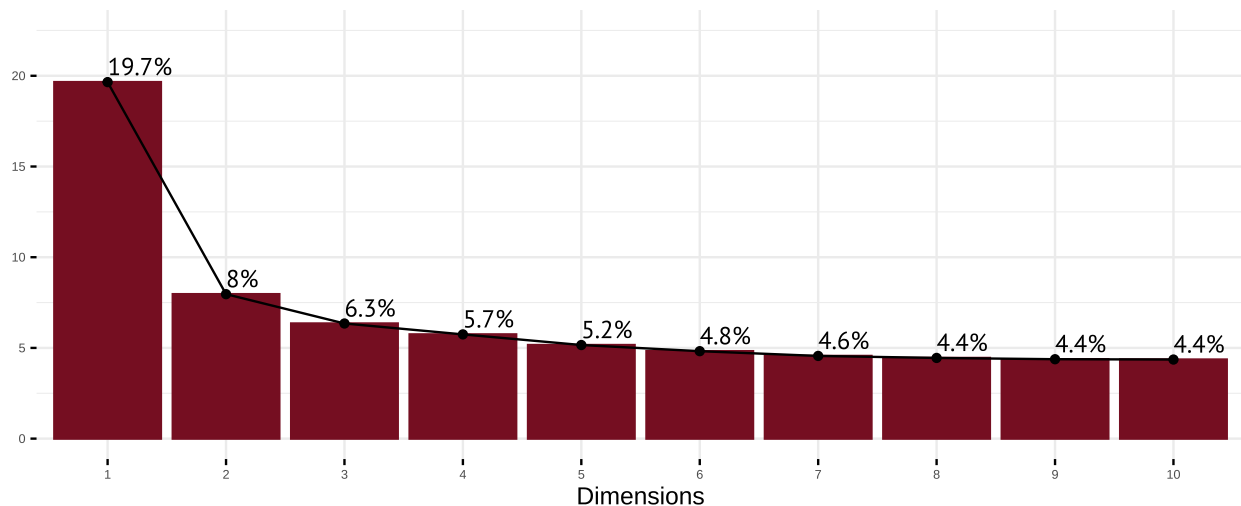


Figure 18 – Inerties.

- Sur les quatre premières dimensions, l'inertie cumulée est de **39.71%**, on peut donc se concentrer sur ces dimensions pour notre analyse.

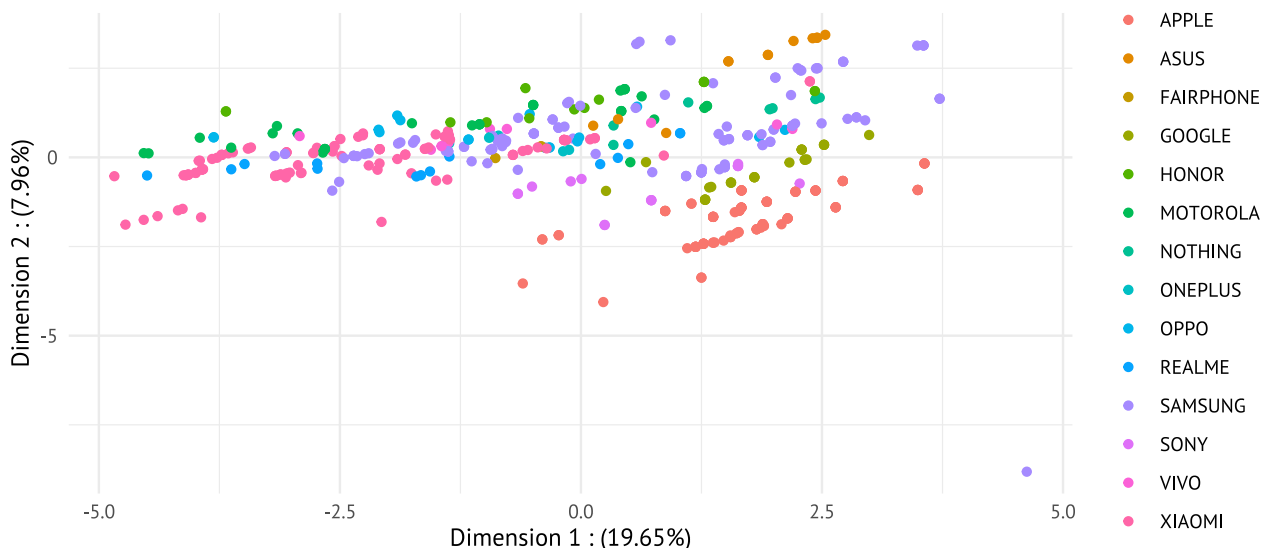
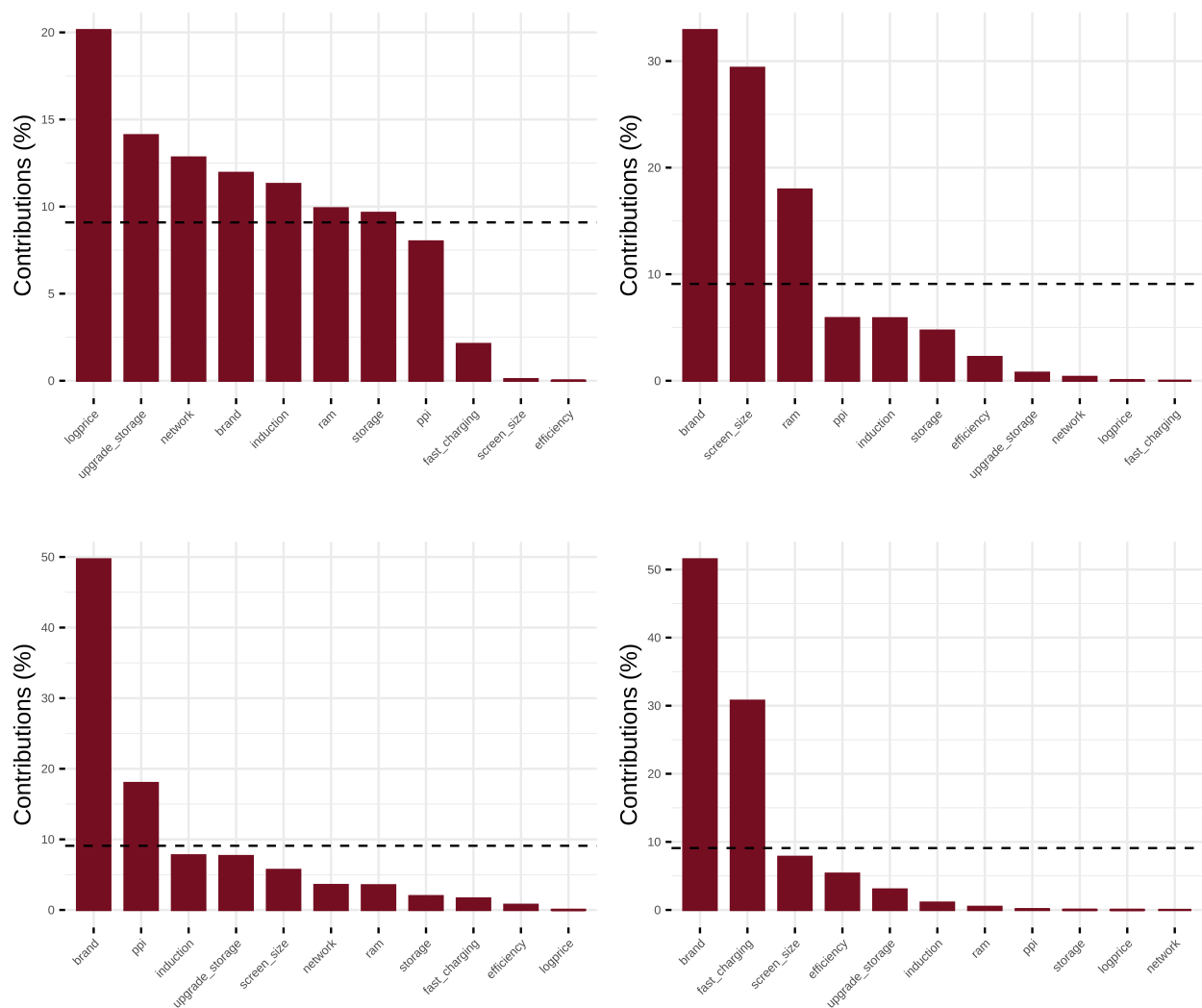


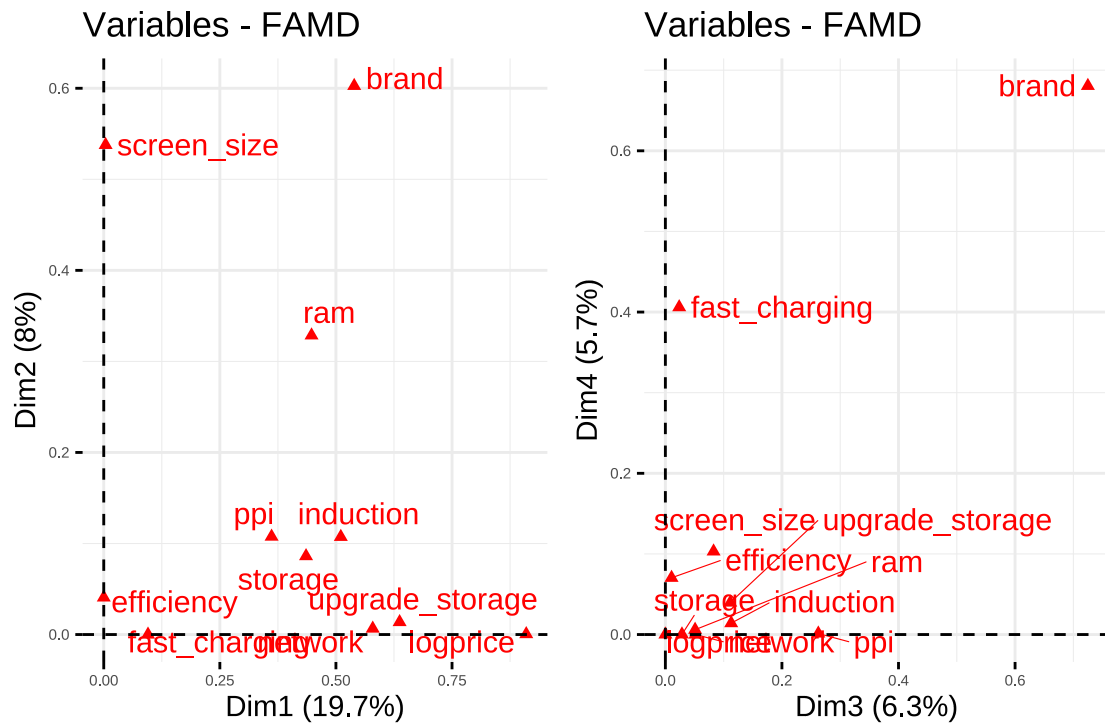
Figure 19 – Nuage des individus (Dimension 1 et 2).

- Le nuage des individus dans la dimension 1 et 2 permet de déterminer que les téléphones sont assez bien regroupés en fonction de leur marque et que les téléphones les moins chers se situent dans la partie négative de l'axe 1 tandis que les téléphones les plus chers se situent dans la partie positive de l'axe 1.



**Figure 20** – Contribution des variables (Dimensions 1 – 4).

- Les variables contribuant le plus à la dimension 1 sont *logprice*, *upgrade\_storage*, *network*, *brand*, *induction*, *ram* et *storage*.
- Dans la dimension 2, les proportions de contribution sont elles partagées entre 3 variables : *brand*, *screen\_size* et *ram*
- Pour la dimension 3 et 4, *brand* est une variable qui contribue encore plus à la construction des axes avec plus de 50% de contribution.

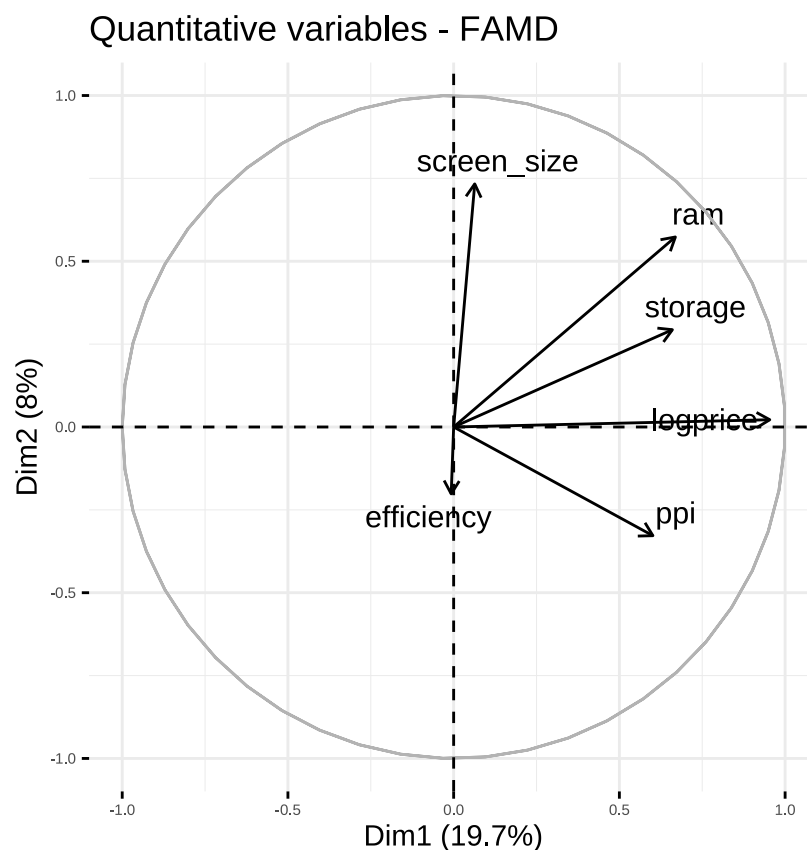


**Figure 21** –  $\eta^2$  des variables (Dimensions 1 – 4).

Les  $\eta^2$  permettent de distinguer les variables les plus structurantes d'un axe

- Dans la dimension 1, la variable *logprice* et *upgrade\_storage* sont les plus structurantes.
- Dans les dimensions 2,3 et 4, on s'aperçoit que c'est la variable *brand* qui est la plus structurante.
- Les variables *ram* et *screen\_size* sont aussi très structurantes dans la dimension 2.

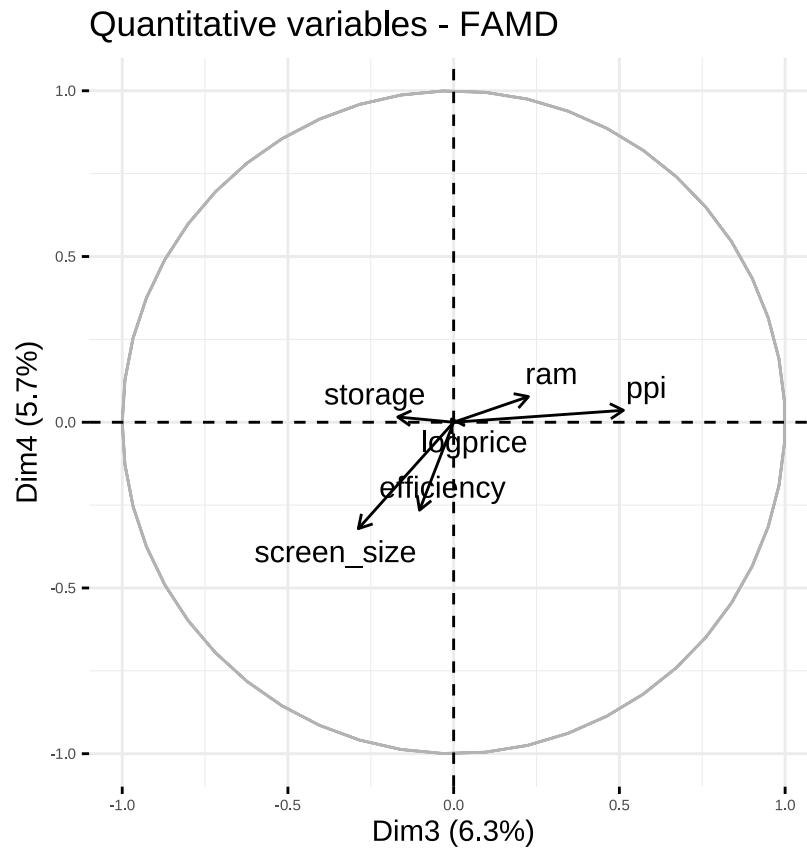
Opposition entre la marque et le prix (dim 1/2)



**Figure 22** – Cercle des corrélations (Dimensions 1 et 2).

D'après le nuage des variables, on peut voir que les variables qui contribuent à l'axe F1 sont le *logprice*, *ppi*, *storage* ainsi que *ram*. On peut donc supposer que les 3 caractéristiques constituent la majeure partie de la variation observée dans la variable *logprice*.

Pour l'axe F2, ce sont principalement les variables *screen\_size* et *efficiency* qui contribuent le plus à cet axe. On peut observer que l'efficacité (*efficiency*) est moins bien représentée que *screen\_size*, comme le suggère la différence dans la longueur des flèches associées à ces variables.



**Figure 23** – Cercle des corrélations (Dimensions 3 et 4).

Les variables ne contribuent pas beaucoup aux dimensions 3 et 4. Pour l'axe F3, la *ppi* contribue pas mal suivi de *storage*, *ram*.

Les variables présentent une contribution relativement faible aux dimensions 3 et 4 de l'analyse. En ce qui concerne l'axe F3, on constate que le **ppi** contribue un peu plus que les autres, suivi par *storage* et *ram*.

Pour l'axe F4, c'est la variable *screen\_size* qui contribue le plus, suivie par *efficiency*, tandis que *logprice* contribue pratiquement pas. Il est également à noter que toutes ces variables sont positionnées du côté négatif de l'axe.

# | Comparateur

En se basant sur notre problématique, nous voulions, en plus d'une estimation économétrique qui présente son lot de complexité, créer une application, et plus précisément un **comparateur d'efficacité**.

En premier lieu, il convient de se poser une question qui semble évidente à première vue, mais qui, quand on s'y penche de plus près, est assez sophistiquée.

- Cette question c'est : *“Qu'est ce qu'un comparateur exactement ?”*

## Définition

- Un comparateur est un site web qui permet de comparer différents produits ou services.
- Principalement axé sur les prix, il peut également prendre en compte les aspects techniques et la qualité des produits.

On retrouve l'utilisation des comparateurs dans de nombreux domaines tels que les comparateurs de vols, d'assurances, de voitures, ou encore de banques.

En bref il existe une variété de cas d'usage et un nombre croissant d'utilisateurs faisant appel à ces comparateurs avant d'acheter un bien ou un service.

D'autre part, il y a des questions juridiques, économiques, statistiques et algorithmiques qui se posent.

En effet, les opérateurs de plateforme en ligne doivent délivrer au consommateur une information **loyale, claire et transparente**, selon l'article **L. 111-7 du code de la consommation**.

**Les comparateurs en ligne sont donc eux aussi soumis à ces obligations de loyauté et de transparence.**

1. Page dédiée accessible depuis toutes les autres pages du comparateur informant le fonctionnement de celui-ci.
2. Faire figurer sur chaque page de résultats les informations concernant les critères de classement.
3. Indiquer via la mention explicite « **annonce** » les résultats liés à des partenariats rémunérés.

Il va sans dire que très peu de comparateurs respectent le premier point, souvent pour une raison simple : ce sont juste des listings de prix de téléphones du moins cher au plus cher, donc il n'y a pas vraiment de fonctionnement à proprement parler. De même, concernant les critères de classement. En revanche le troisième point est presque toujours indiqué quelque part avec la mention *publicité* ou *annonce* quand il y a un partenariat rémunéré, et c'est normal car une grande partie des revenus de ces comparateurs provient de ces partenariats donc ils font assez attention à ce point.

*L'avantage de notre comparateur c'est qu'il respecte les deux premiers points, sans se soucier du troisième puisque nous n'avons pas de partenariats rémunérés. Tout ça permet d'offrir une grande transparence pour le consommateur.*

# | Conclusion

La réalisation d'une étude utilisant les modèles SFA pour évaluer les prix des smartphones a permis de mieux comprendre les dynamiques complexes de ce marché en constante évolution. En explorant la littérature existante sur la SFA et la régression hédonique des prix, nous avons découvert les fondements théoriques de ces modèles, leurs apports ainsi que leurs limites.

Les statistiques descriptives réalisées sur 487 smartphones vendus par un revendeur français (octobre 2023) ont révélé une distribution étalée des prix, confirmant la nature hautement différenciée des smartphones. En utilisant des modèles de régression hédonique, nous avons réussi à prédire avec précision les prix en utilisant différents types de spécifications, notamment *niveau-niveau* en *log-niveau*. Ces résultats ont été extrêmement satisfaisants, démontrant la capacité des modèles à correctement inférer les prix des smartphones en fonction de leurs caractéristiques.

Cependant, pour aller au-delà de la simple prédiction des prix, nous avons également appliqué un modèle SFA, spécifiquement la *Cost Frontier Analysis*, afin d'évaluer l'efficacité des prix des téléphones en tenant compte de leurs caractéristiques. Cela nous a permis de déterminer les téléphones offrant le meilleur rapport qualité-prix, offrant ainsi une perspective supplémentaire dans la compréhension du pricing de ces téléphones.

La prochaine étape de ce projet consistera à créer une application comparative. Celle-ci permettra aux utilisateurs (*tout du moins nous l'espérons*), de spécifier les caractéristiques recherchées dans un téléphone. En utilisant le modèle SFA que nous avons développé, un score sera généré pour chaque téléphone, classant ainsi les appareils du plus efficace au moins efficace en fonction des critères spécifiés. Cette application offrira une solution pratique et personnalisée aux utilisateurs pour prendre des décisions informées lors de l'achat ou de la production de smartphones.

En résumé, cette étude approfondie combinant la régression hédonique, la SFA et l'analyse des prix des smartphones a non seulement permis de prédire précisément les prix en fonction des caractéristiques, mais a également jeté les bases d'une application offrant des recommandations personnalisées basées sur l'efficacité des prix des téléphones. Ce projet ouvre des perspectives intéressantes pour aider les consommateurs et les producteurs à prendre des décisions éclairées dans un marché aussi dynamique et complexe que celui des smartphones.

\* \* \*



# | Annexe

## Dérivation de la fonction de vraisemblance

$$f(v) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(-\frac{v^2}{2\sigma_v^2}\right) \quad (12)$$

$$\begin{aligned} f(u) &= \frac{1}{\sqrt{2\pi}\sigma_u \left[1 - \Phi\left(-\frac{\mu}{\sigma_u}\right)\right]} \exp\left[-\frac{1}{2}\left(\frac{u - \mu}{\sigma_u}\right)^2\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma_u \Phi\left(\frac{\mu}{\sigma_u}\right)} \exp\left[-\frac{1}{2}\left(\frac{u - \mu}{\sigma_u}\right)^2\right] \end{aligned} \quad (13)$$

Etant donné l'hypothèse d'indépendance entre  $u$  et  $v$  :

$$f(v, u) = f(v) \cdot f(u) = \frac{1}{2\pi\sigma_v\sigma_u \Phi\left(\frac{\mu}{\sigma_u}\right)} \exp\left\{-\frac{1}{2}\left[\frac{v^2}{\sigma_v^2} + \left(\frac{u - \mu}{\sigma_u}\right)^2\right]\right\} \quad (14)$$

Donc,

$$f(u - \epsilon, u) = \frac{1}{2\pi\sigma_v\sigma_u \Phi\left(\frac{\mu}{\sigma_u}\right)} \exp\left\{-\frac{1}{2}\left[\left(\frac{u - \epsilon}{\sigma_v}\right)^2 + \left(\frac{u - \mu}{\sigma_u}\right)^2\right]\right\} \quad (15)$$

Cette expression peut être simplifiée. Notez que :

$$\left(\frac{u - \epsilon}{\sigma_v}\right)^2 + \left(\frac{u - \mu}{\sigma_u}\right)^2 = \frac{\sigma_v^2 + \sigma_u^2}{\sigma_v^2\sigma_u^2} \left[u^2 - 2u\left(\frac{\mu\sigma_v^2 + \epsilon\sigma_u^2}{\sigma_v^2 + \sigma_u^2}\right)\right] + \left(\frac{\mu^2}{\sigma_u^2} - \frac{\epsilon^2}{\sigma_v^2}\right) \quad (16)$$

Définissons :

$$\mu_* = \frac{\mu\sigma_v^2 + \epsilon\sigma_u^2}{\sigma_v^2 + \sigma_u^2}$$

$$\sigma_*^2 = \frac{\sigma_v^2\sigma_u^2}{\sigma_v^2 + \sigma_u^2}$$

Dès lors, l'équation 15 est simplifiée pour obtenir :

$$f(u - \epsilon, u) = \frac{1}{2\pi\sigma_v\sigma_u \Phi\left(\frac{\mu}{\sigma_u}\right)} \exp\left\{-\frac{1}{2}\left[\left(\frac{u - \mu_*}{\sigma_*}\right)^2 + \frac{(\mu - \epsilon)^2}{\sigma_u^2 + \sigma_v^2}\right]\right\} \quad (17)$$

Alors, la densité de  $f(\epsilon)$  est :

$$f(\epsilon) = \int_0^\infty f(u - \epsilon, u) du$$

...

$$f(\epsilon) = \frac{\phi\left(\frac{\mu - \epsilon}{\sqrt{\sigma_v^2 + \sigma_u^2}}\right)}{\sqrt{\sigma_v^2 + \sigma_u^2} \left[ \frac{\Phi\left(\frac{\mu}{\sigma_u}\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \right]} \quad (18)$$

En prenant le logarithme de l'équation précédente, on obtient la fonction de log-vraisemblance pour une observation  $i$  dans le cadre d'un modèle avec loi normale tronquée.

$$L_i = -\frac{1}{2} \ln(\sigma_v^2 + \sigma_u^2) + \ln \phi\left(\frac{\mu - \epsilon}{\sqrt{\sigma_v^2 + \sigma_u^2}}\right) + \ln \Phi\left(\frac{\mu_*}{\sigma_*}\right) - \ln \Phi\left(\frac{\mu}{\sigma_u}\right) \quad (19)$$

## Dérivation des indices d'efficacité

$$f(u|\epsilon) = \frac{1}{\sqrt{2\pi}\sigma_*\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \exp\left[-\frac{1}{2}\left(\frac{u - \mu_*}{\sigma_*}\right)^2\right] \quad (20)$$

Avec  $\mu_*$  et  $\sigma_*$  définis plus haut :

$$E(u|\epsilon) = \int_0^\infty \frac{u}{\sqrt{2\pi}\sigma_*\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \exp\left[-\frac{1}{2}\left(\frac{u - \mu_*}{\sigma_*}\right)^2\right] du \quad (21)$$

Définitions :

$$w = \frac{u - \mu_*}{\sigma_*}, w \in \left[-\frac{\mu_*}{\sigma_*}, \infty\right), dw = \frac{du}{\sigma_*}$$

Donc :

$$\begin{aligned} E(u|\epsilon) &= \int_{-\frac{\mu_*}{\sigma_*}}^\infty \frac{\mu_* + w\sigma_*}{\sqrt{2\pi}\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \exp\left(-\frac{1}{2}w^2\right) dw \\ &= \frac{\mu_*}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \int_{-\frac{\mu_*}{\sigma_*}}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) dw + \frac{\sigma_*}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \int_{-\frac{\mu_*}{\sigma_*}}^\infty \frac{w}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) dw \\ &= \mu_* + \frac{\sigma_*}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \cdot \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu_*}{\sigma_*}\right)^2\right] \\ &= \mu_* + \frac{\phi\left(\frac{\mu_*}{\sigma_*}\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)} \sigma_* \end{aligned}$$

Maintenant, pour  $E(\exp(-u)|\epsilon)$ ,

$$\begin{aligned}
E(\exp(-u)|\epsilon) &= \int_0^\infty \exp(-u) f(u|\epsilon) du \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_*\Phi(\frac{\mu_*}{\sigma_*})} \exp\left[-\frac{1}{2}\left(\frac{u-\mu_*}{\sigma_*}\right)^2 - u\right] du \\
&= \frac{1}{\sigma_*\Phi(\frac{\mu_*}{\sigma_*})} \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{u-\mu_*}{\sigma_*}\right)^2 - u\right] du
\end{aligned}$$

Notons que

$$-\frac{1}{2}\left(\frac{u-\mu_*}{\sigma_*}\right)^2 - u = -\frac{[u - (\mu_* - \sigma_*^2)]^2}{2\sigma_*^2} - \frac{1}{2}(2\mu_* - \sigma_*)$$

Dès lors, on peut simplifier la formule :

$$\begin{aligned}
E(\exp(-u)|\epsilon) &= \frac{1}{\sigma_*\Phi(\frac{\mu_*}{\sigma_*})} \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{[u - (\mu_* - \sigma_*^2)]^2}{2\sigma_*^2} - \frac{1}{2}(2\mu_* - \sigma_*)\right\} du \\
&= \frac{\exp\left[-\frac{1}{2}(2\mu_* - \sigma_*)\right]}{\sigma_*\Phi(\frac{\mu_*}{\sigma_*})} \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{u - (\mu_* - \sigma_*^2)}{\sigma_*}\right]^2\right\} du
\end{aligned}$$

Définissons :

$$z = \frac{u - (\mu_* - \sigma_*^2)}{\sigma_*}, z \in \left[-\frac{\mu_*}{\sigma_*} + \sigma_*, \infty\right), dz = \frac{du}{\sigma_*}$$

Enfin,

$$\begin{aligned}
E(\exp(-u)|\epsilon) &= \frac{\exp\left[-\frac{1}{2}(2\mu_* - \sigma_*)\right]}{\sigma_*\Phi(\frac{\mu_*}{\sigma_*})} \int_{-\frac{\mu_*}{\sigma_*} + \sigma_*}^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} dz \\
&= \exp\left(-\mu_* + \frac{1}{2}\sigma_*^2\right) \frac{\Phi\left(\frac{\mu_*}{\sigma_*} - \sigma_*\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}
\end{aligned}$$

# | Acronymes

**WTP** Willingness To Pay

**SFA** Stochastic Frontier Analysis

**DAS** Débit d'Absorption Spécifique

**DEA** Data Envelopment Analysis

**TE** Technical Efficiency

**BLUE** Best Linear Unbiased Estimator

**MCO** Moindres Carrés Ordinaires

**MV** Maximum de Vraisemblance

# | Glossaire des variables

**screen\_type** Type d'écran : Plat, Pliable, etc. ⇒ **str**

**screen\_size** Taille de l'écran, en pouces ⇒ **float**

**screen\_tech** Technologie de l'écran : LCD, OLED, AMOLED. ⇒ **str**

**resolution\_1/2** Résolution verticale/horizontale de l'écran, en nombre de pixels ⇒ **int**

**diagonal\_pixels** Diagonale en nombre de pixels calculée à partir de la résolution ⇒ **float**

**ppi** Pixels Per Inch, calculé à partir de la diagonale et de la taille de l'écran ⇒ **float**

**cam\_1, cam\_2, cam\_3** Résolution de la caméra 1/2/3 en mégapixels ⇒ **int**

**mpx\_backward\_cam** Somme des résolutions des caméras 1, 2 et 3 en mégapixels ⇒ **int**

**sensor** Nombre de caméras arrières équipées sur le téléphone ⇒ **int**

**color** Couleur du téléphone ⇒ **str**

**thickness** Epaisseur du téléphone en millimètres ⇒ **float**

**width** Largeur du téléphone en millimètres ⇒ **float**

**height** Hauteur du téléphone en millimètres ⇒ **float**

**net\_weight** Poids net du téléphone en grammes ⇒ **float**

**network** Prise en charge réseau jusqu'à la 4G ou la 5G ⇒ **str**

**cpu** Type de CPU. *Variable Inexploitable* ⇒ **str**

**ram** Capacité de RAM en Gigaoctets ⇒ **int**

**storage** Capacité de stockage en Gigaoctets ⇒ **int**

**upgrade\_storage** L'appareil dispose t-il d'un moyen d'augmenter son stockage ⇒ **bool**

**battery** Capacité de la batterie en milliampères ⇒ **int**

**fast\_charging** L'appareil dispose t-il d'une charge rapide ⇒ **bool**

**induction** L'appareil dispose t-il d'une charge par induction ⇒ **bool**

**usb\_type\_c** L'appareil dispose t-il d'un port USB type C ⇒ **bool**

**repairability\_index** Indice de réparabilité du téléphone (Note /10) ⇒ **int**

**model** Modèle du téléphone ⇒ **str**

**brand** Marque du téléphone ⇒ **str**

**made\_in** Lieu de fabrication du téléphone ⇒ **str**

*stars* Note sur 5 du téléphone (quand disponible) ⇒ float

*reviews* Nombre de critiques ⇒ int

*das\_head/chest/limbs* Débit d'absorption spécifique tête/corps/membres ⇒ float

*price* Prix du téléphone ⇒ float

# | Licence


**Vous êtes autorisé à :**


**Partager** — copier, distribuer et communiquer le matériel par tous moyens et sous tous formats.


**Adapter** — remixer, transformer et créer à partir du matériel.

L'Offrant ne peut retirer les autorisations concédées par la licence tant que vous appliquez les termes de cette licence.

**Selon les conditions suivantes :**

 **Attribution** — Vous devez créditer l'Œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Œuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Œuvre.

 **Pas d'Utilisation Commerciale** — Vous n'êtes pas autorisé à faire un usage commercial de cette Œuvre, tout ou partie du matériel la composant.

 **Partage dans les Mêmes Conditions** — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'Œuvre originale, vous devez diffuser l'Œuvre modifiée dans les mêmes conditions, c'est à dire avec la même licence avec laquelle l'Œuvre originale a été diffusée.

**Pas de restrictions complémentaires** — Vous n'êtes pas autorisé à appliquer des conditions légales ou des mesures techniques qui restreindraient légalement autrui à utiliser l'Œuvre dans les conditions décrites par la licence.

<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.fr>



# | Références

- Ahmad, Waseem, Tanvir Ahmed, et Bashir Ahmad. 2019. « Pricing of mobile phone attributes at the retail level in a developing country: Hedonic analysis ». *Telecommunications Policy* 43 (4): 299-309. <https://doi.org/10.1016/j.telpol.2018.10.002>.
- Aigner, Dennis, C. A.Knox Lovell, et Peter Schmidt. 1977. « Formulation and estimation of stochastic frontier production function models ». *Journal of Econometrics* 6 (1): 21-37. [https://doi.org/10.1016/0304-4076\(77\)90052-5](https://doi.org/10.1016/0304-4076(77)90052-5).
- Arrondo, Ruben, Nuria Garcia, et Eduardo Gonzalez. 2018. « Estimating product efficiency through a hedonic pricing best practice frontier ». *BRQ Business Research Quarterly* 21 (4): 215-24. <https://doi.org/10.1016/j.brq.2018.08.005>.
- Bello, Ajide K, et Alabi Moruf. 2010. « Does the functional form matter in the estimation of hedonic price model for housing market ». *The Social Sciences* 5 (6): 559-64.
- Berndt, Ernst R, et Neal J Rappaport. 2001. « Price and quality of desktop and mobile personal computers: A quarter-century historical overview ». *American Economic Review* 91 (2): 268-73. <https://www.aeaweb.org/articles?id=10.1257/aer.91.2.268>.
- Boistel, Philippe. 2008. « La réputation d'entreprise: un impact majeur sur les ressources de l'entreprise ». *Revue management et avenir*, n° 3: 9-25.
- Chen, Ching-Fu, et Rochelle Rothschild. 2010. « An application of hedonic pricing analysis to the case of hotel rooms in Taipei ». *Tourism Economics* 16 (3): 685-94. <https://journals.sagepub.com/doi/abs/10.5367/000000010792278310>.
- Efroymson, Michael Alin. 1960. « Multiple regression analysis ». *Mathematical methods for digital computers*, 191-203.
- Farrell, Michael James. 1957. « The measurement of productive efficiency ». *Journal of the Royal Statistical Society Series A: Statistics in Society* 120 (3): 253-81.
- Harrison Jr, David, et Daniel L Rubinfeld. 1978. « Hedonic housing prices and the demand for clean air ». *Journal of environmental economics and management* 5 (1): 81-102. [https://doi.org/10.1016/0095-0696\(78\)90006-2](https://doi.org/10.1016/0095-0696(78)90006-2).
- Kumbhakar, Subal, Alan Horncastle, et al. 2015. *A practitioner's guide to stochastic frontier analysis using Stata*. Cambridge University Press.
- Lancaster, Kelvin J. 1966. « A New Approach to Consumer Theory ». *Journal of Political Economy* 74 (2): 132-57. <https://doi.org/10.1086/259131>.
- Mohamad, Shamsheer, Taufiq Hassan, et Mohamed Khaled I Bader. 2008. « Efficiency of conventional versus Islamic Banks: international evidence using the Stochastic Frontier Approach (SFA) ». *Journal of Islamic economics, banking and finance* 4 (2): 107-30.
- Reinhard, Stijn, CA Knox Lovell, et Geert J Thijssen. 2000. « Environmental efficiency with multiple environmentally detrimental variables; estimated with SFA and DEA ». *European Journal of Operational Research* 121 (2): 287-303.
- Rosen, Sherwin. 1974. « Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition ». *Journal of Political Economy* 82 (1): 34-55. <http://www.jstor.org/stable/1830899>.
- Rosko, Michael D, et Ryan L Mutter. 2008. « Stochastic frontier analysis of hospital inefficiency: a review of empirical issues and an assessment of robustness ». *Medical care research and review* 65 (2): 131-66. <https://doi.org/10.1177/1077558707307580>.
- Yim, Eun Soon, Suna Lee, et Woo Gon Kim. 2014. « Determinants of a restaurant average meal price: An application of the hedonic pricing model ». *International Journal of Hospitality Management* 39: 11-20. <https://www.sciencedirect.com/science/article/abs/>

pii/S027843191400019X.