Université de Poitiers

UFR Sciences Fondamentales et Appliquées
Département de Mathématiques

M1 STDV – Année universitaire 2024-2025

Projet : Modèles Linéaires et Généralisés

Application aux Soras, système Elo aux échecs et les prix d'ordinateurs portables

---

Étudiant :
Soufiane Lmezouari

Encadrant :
Arnaud Poinas

20 avril 2025

Table des matières

1 Sexage de Soras

Etude sur le sexage des Soras

La détermination du sexe chez certaines espèces d'oiseaux peut s'avérer difficile lorsqu'aucun dimorphisme sexuel évident n'est présent. Dans le cas de la Marouette de Caroline (*Porzana carolina*), aussi appelée Sora, les chercheurs ont tenté de mettre en place une méthode fiable de sexage à partir de mesures morphométriques.

L'exercice proposé s'appuie sur des données collectées en 2018 et 2020, issues de l'article scientifique de Dami et al. (2024), dans lequel les auteurs comparent le sexe prédit par observation aux résultats obtenus via des tests ADN.

Marouette de Caroline – Mâle

Marouette de Caroline – Femelle

1.1 Nombre d'individus et de variables

Pour commencer, nous avons chargé le fichier Sora.csv dans R :

le jeu de données composé de **180 individus** et de **9 variables**.

Variable	Description
Capture.Year	Année de capture (2018, 2020)
Our.Guess	Sexe estimé par les chercheurs (Male, Female)
Actual.Sex	Sexe réel selon ADN (Male, Female)
Age	Groupe d'âge : AHY (> 1 an), HY (< 1 an)
Culmen	Longueur du bec supérieur (mm)
Tarsus	Longueur du tarsométatarse (mm)
Toe	Longueur de l'orteil du milieu (mm)
Weight	Masse de l'animal (g)
Fat.Score	Score de graisse (0 = peu, 5 = beaucoup)

TABLE 1.1 – Résumé des variables du jeu de données Sora

D'après le tableau 1.1, et étant donné que nous allons traiter la variable Capture.Year comme une variable qualitative, nous présentons ci-dessous les différentes modalités observées pour les variables qualitatives.

La variable Capture.Year comporte deux modalités : 2018 et 2020. Pour la variable Our.Guess, on observe également deux modalités : Male et Female. De même, la variable Actual.Sex présente les modalités : Male et Female. Enfin, la variable Age contient deux modalités : AHY (plus d'un an) et HY (moins d'un an).

1.2 Étude des erreurs de sexage et facteurs associés

Nous avons d'abord ajouté une nouvelle variable `Mistake` au jeu de données, prenant la valeur 1 lorsque le sexe estimé (`Our.Guess`) diffère du sexe réel (`Actual.Sex`), et 0 sinon.

```
dat$Mistake <- ifelse(dat$Our.Guess != dat$Actual.Sex, 1, 0)
error_rate <- mean(dat$Mistake) * 100
```

Nous avons ensuite calculé le pourcentage d'erreur, qui est **16.11 %**, indiquant que les auteurs se sont trompés dans près de 1 cas sur 6 lors du sexage des individus.

1.3 Ajustement du première modèle logistique

Nous ajustons un modèle logistique ayant pour variable réponse `Mistake`, et comme variables explicatives : `Age`, `Culmen`, `Tarsus`, `Toe`, `Weight` et `Fat.Score`. L'ajustement se fait à l'aide de la fonction `glm` avec une famille binomiale .

```
library(MASS)
mod1 <- glm(Mistake ~ Age + Culmen + Tarsus + Toe + Weight + Fat.Score,
data =dat, family =binomial)
```

Les résultats obtenus sont les suivants :

— **Déviance nulle** : 158,94 sur 179 degrés de liberté.
— **Déviance résiduelle** : 151,39 sur 173 degrés de liberté.

1.4 Significativité du modèle 1 ajusté

Pour ce modèle, on a une déviance résiduelle $D = 151{,}39$ avec 173 degrés de liberté. On rejette l'hypothèse que le modèle est bien ajusté si

$$D > \chi^2_{173,\,0{,}95} \approx 205{,}3.$$

Ce n'est **pas le cas**, donc on **ne rejette pas** l'hypothèse que le modèle est bien ajusté.

On a aussi une déviance nulle $D_0 = 158{,}94$ avec 179 degrés de liberté. On teste si le modèle est significativement meilleur que le modèle nul en comparant :

$$D_0 - D = 7{,}55 \quad \text{et} \quad \chi^2_{6,\,0{,}95} \approx 12{,}59.$$

Ici, $7{,}55 < 12{,}59$, donc on **ne rejette pas l'hypothèse** que le modèle est équivalent au modèle nul : le modèle ajusté n'est **pas significativement meilleur** que le modèle vide au seuil de 5%.

1.5 Création et structuration de la variable Group

Nous créons une variable qualitative `Group` en combinant les modalités des variables `Capture.Year`, `Actual.Sex` et `Age` à l'aide de la fonction `paste()`, en insérant un retour à la ligne pour obtenir le même format que celui utilisé dans l'article original.

```
dat$Group <- paste(dat$Capture.Year, dat$Actual.Sex, dat$Age, sep ="\n")
```

1.6 Reproduction de la Figure 4 du papier

Nous utilisons la fonction factor() pour réordonner manuellement les modalités de la variable Group selon l'année, le sexe et l'âge. comme dans la figure 4 de l'article

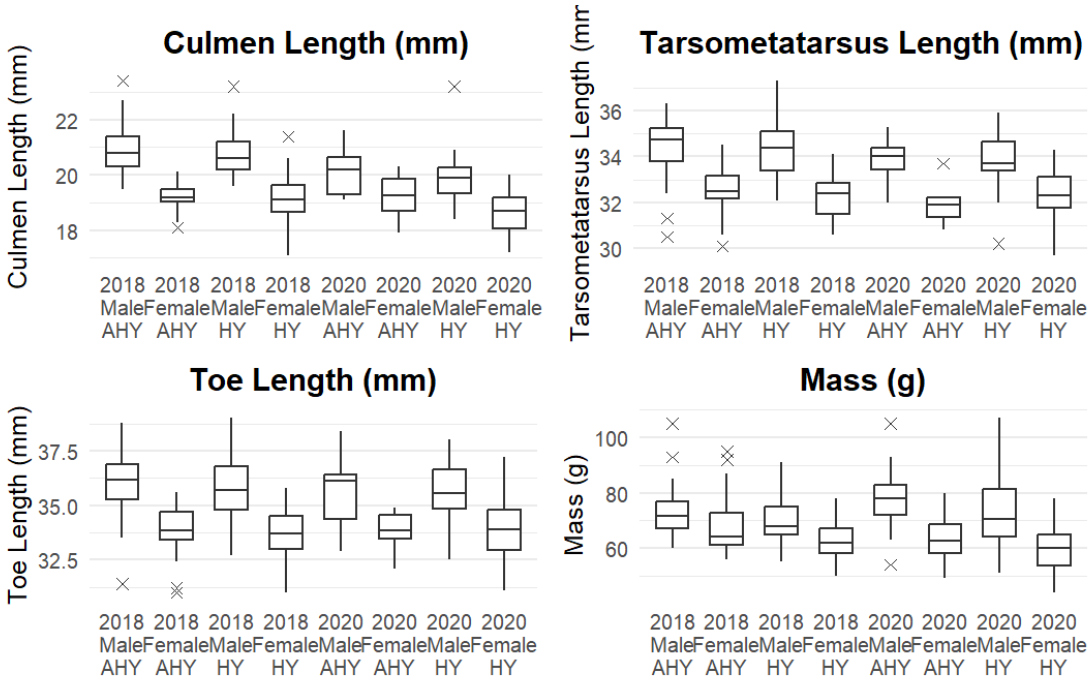


FIGURE 1.1 – Mesures morphométriques des Soras (2018–2020) par sexe et âge.

La figure qu'on a reproduite ressemble beaucoup à celle de l'article, avec des distributions globalement similaires. Il y a juste quelques valeurs extrêmes qui changent un peu entre les deux.

1.7 Modèle logistique ajusté sur les données de 2018

Nous filtrons les données pour ne conserver que celles de l'année 2018, puis nous transformons la variable `Actual.Sex` en une variable binaire `Sex`, où `Male = 1` et `Female = 0`.

```
mod2 <- glm(Sex ~ Culmen + Tarsus + Toe + Weight + Age, data =dat2, family =
    binomial)
```

— Déviance nulle : 149,683 avec 107 degrés de liberté.
— Déviance résiduelle : 57,478 avec 102 degrés de liberté.
— AIC : 69,478.

La différence de déviance est de :

$$149{,}683 - 57{,}478 = 92{,}205 \quad \text{avec} \quad 5 \text{ degrés de liberté.}$$

$\frac{92{,}205}{5} = 18{,}441 >> 1$ Le modèle est donc **significativement meilleur** que le modèle vide.

1.8 Sélection de variables selon les auteurs de l'article

Les auteurs de l'article ont utilisé deux critères pour sélectionner les variables explicatives : l'AIC et la validation croisée leave-one-out (CV1).

Le modèle sélectionné incluant uniquement les variables `Culmen` et `Tarsus` avec la plus faible AIC (64,684) ainsi qu'un excellent score CV1 (65,909).

1.9 Sélectionner les variables du modèle avec la méthode ascendante

```
vide_mod =glm(Actual.Sex ~ 1, data =dat2, family =binomial)
modele_selec_AIC=step(model_vide2, scope =list(lower =vide_mod, upper =mod2),
    direction ="forward", trace =1)
```

Le modèle sélectionné par la procédure step() est identique à celui des auteurs, avec les variables Culmen et Tarsus. La légère différence d'AIC (64.68 contre 64.684) est presque négligeable.

1.10 Analyse des facteurs influençant le poids

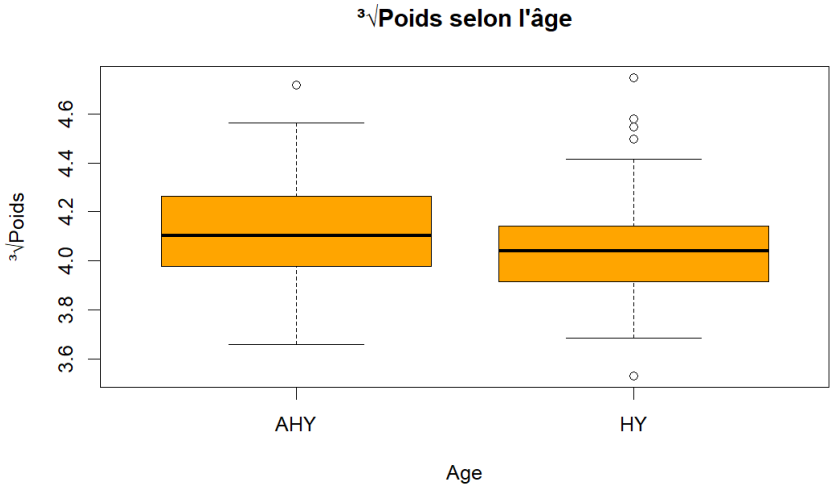


FIGURE 1.2 – ³Poids selon l'âge

```
# Residual standard error: 0.2132 on 178degrees of freedom
# Multiple R-squared: 0.04519, Adjusted R-squared: 0.03982
# F-statistic: 8.424 on 1and 178DF, p-value: 0.004171
```

p-value : 0.004171 <0,05 donc il y'a une difference significative

FIGURE 1.3 – ³Poids selon le sexe

```
# Residual standard error: 0.1957 on 178degrees of freedom
# Multiple R-squared: 0.1947, Adjusted R-squared: 0.1902
# F-statistic: 43.05 on 1and 178DF, p-value: 5.621e-1071
```

p-value : 5.621e-10 <0,05 donc il y'a une difference significative

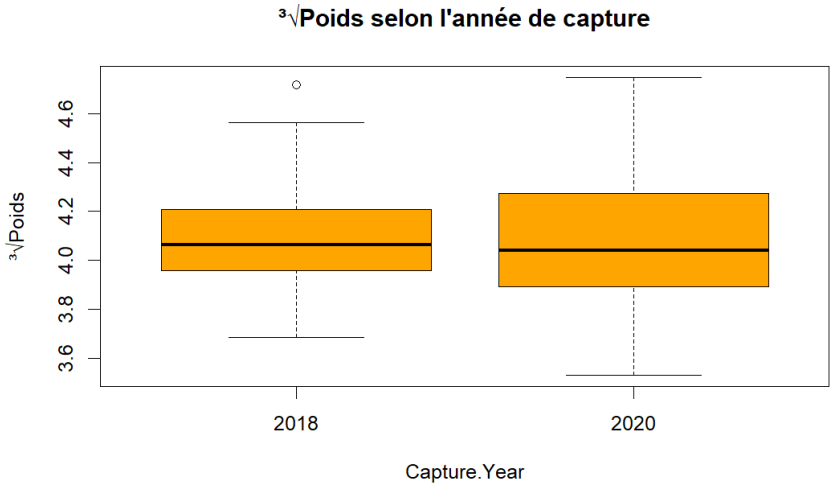


FIGURE 1.4 – ³Poids selon l'année de capture

```
# Residual standard error: 0.2179 on 178degrees of freedom
# Multiple R-squared: 0.001749, Adjusted R-squared: -0.00386
# F-statistic: 0.3118 on 1and 178DF, p-value: 0.5773
```

p-value : 0.5773 >0,05 donc pas difference significative

```
mod_culmen =lm(Poids3 ~ Culmen, data =dat)
summary(mod_culmen)
```

```
# Residual standard error: 0.1999 on 178degrees of freedom
# Multiple R-squared: 0.1602, Adjusted R-squared: 0.1555
# F-statistic: 33.95 on 1and 178DF, p-value: 2.597e-08
```

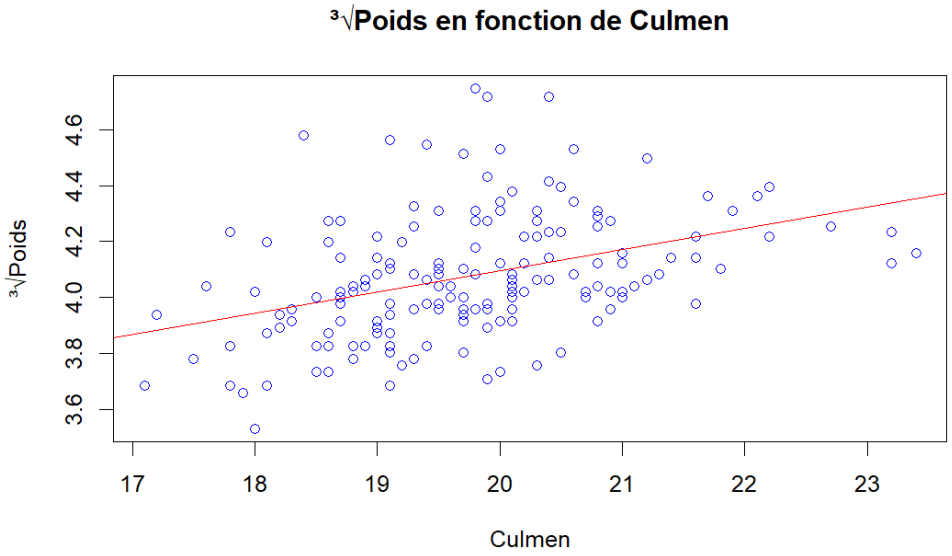


FIGURE 1.5 – Poids en fonction de Culmen

relation linéaire significative

```
mod_toe =lm(Poids3 ~ Toe, data =dat)
summary(mod_toe)
```

```
# Residual standard error: 0.2016 on 178degrees of freedom
# Multiple R-squared: 0.1457, Adjusted R-squared: 0.1409
# F-statistic: 30.35 on 1and 178DF, p-value: 1.25e-07
```

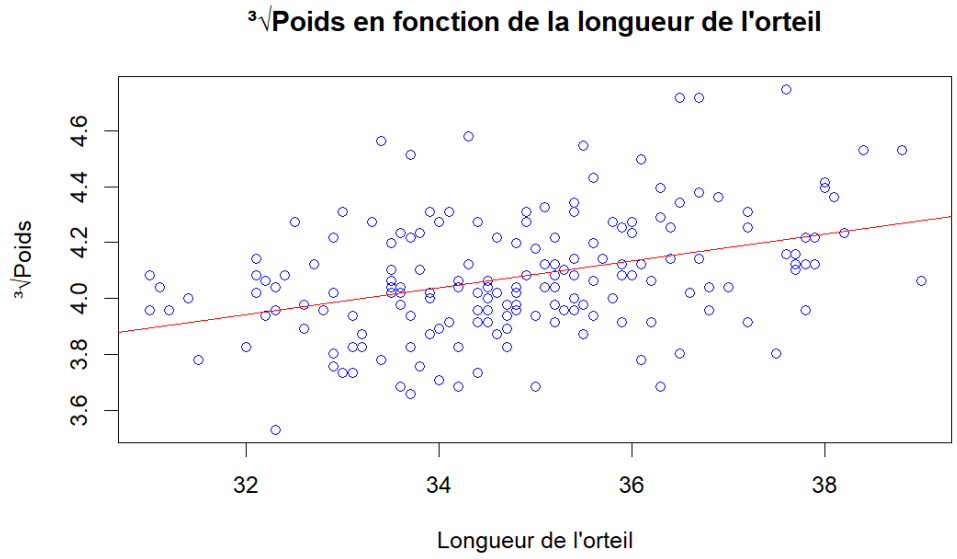


FIGURE 1.6 – ³Poids en fonction de la longueur de l'orteil

relation linéaire significative

```
mod_tarsus =lm(Poids3 ~ Tarsus, data =dat)
summary(mod_tarsus)
```

```
# Residual standard error: 0.2009 on 178degrees of freedom
# Multiple R-squared: 0.152, Adjusted R-squared: 0.1472
# F-statistic: 31.9 on 1and 178DF, p-value: 6.318e-08
```

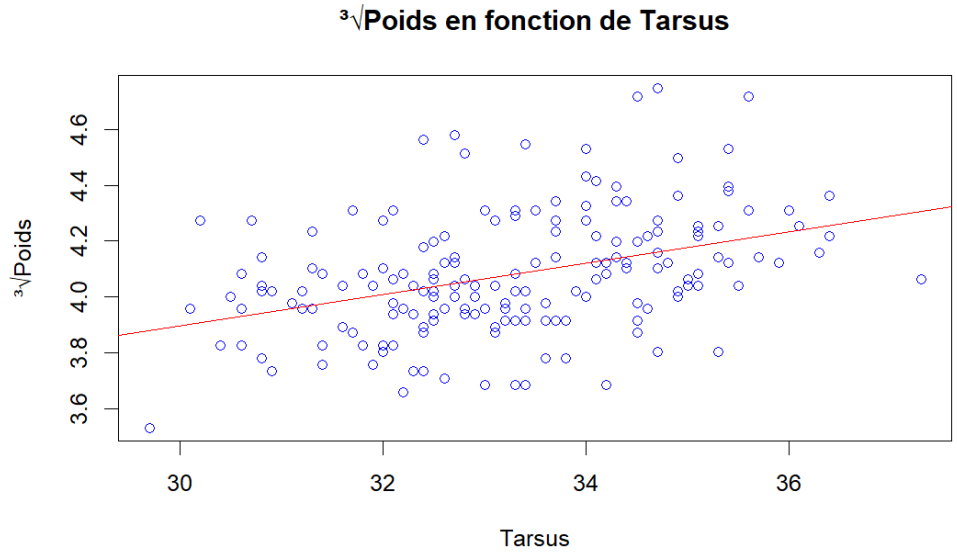


FIGURE 1.7 – ³Poids en fonction de Tarsus

relation linéaire significative

2 Le système Elo aux échecs

Cet exercice explore l'efficacité du système Elo pour prédire les résultats de parties d'échecs. À l'aide de données de tournois français (2020-2023), nous modélisons la probabilité de victoire via des modèles logistiques comparant l'impact des Elos individuels et de leur différence. Nous étudions également les facteurs influençant les égalités, combinant analyse statistique et visualisations pour valider les principes théoriques du système Elo.

2.1 Statistiques descriptives générales

L'ensemble des données analysées comprend un total de 404 166 parties, réparties sur 6 933 tournois différents.

Pour obtenir le nombre de joueurs uniques, nous avons combiné les colonnes contenant les noms des joueurs jouant en blanc et en noir dans une nouvelle variable tous_les_joueurs.

```
tous_les_joueurs <- c(dat$Blanc, dat$Noir)
length(unique(tous_les_joueurs))
```

20 516 joueurs uniques.

2.2 Prétraitement : exclusion des égalités et ajustement des niveaux

Les parties s'étant terminées par une égalité ont été retirées du jeu de données. Par la suite, la fonction droplevels a été utilisée afin de supprimer la modalité '1/2-1/2' de la variable Résultat

```
dat2 <- subset(dat, Résultat != "1/2-1/2")
dat2$Résultat <- droplevels(dat2$Résultat)
```

2.3 Ajustement et évaluation de deux modèles logistiques

Deux modèles ont été ajustés ;
Modèle 1 : Un premier modèle avec le Elo des blancs et celui des noirs comme variables explicatives.

```
mod1 <- glm(Result_Binary ~ Blanc.Elo + Noir.Elo, data =dat2, family =
    binomial(link =logit))
```

Modèle 2 : Un deuxième modèle avec juste la différence d'Elo entre les blancs et les noirs comme variable explicative.

```
dat2$Diff_Elo <- dat2$Blanc.Elo - dat2$Noir.Elo
mod2 <- glm(Result_Binary ~ Diff_Elo, data =dat2, family =binomial)
```

Modèle	AIC	BIC
Modèle 1 (BlancElo + NoirElo)	369 054,2	369 086,5
Modèle 2 (DiffElo)	369 110,6	369 132,2

TABLE 2.1 – Comparaison des critères AIC et BIC pour les deux modèles

d'après les résultats de tableau 2.1, le modèle 1 présente un AIC et un BIC plus faibles que le modèle 2.alors le modèle complet fournit un meilleur ajustement aux données.

Cependant, l'écart entre les valeurs d'AIC et de BIC des deux modèles reste modéré. Cela suggère que le modèle 2 est presque aussi performant

2.4 Test de Wald sur la somme des coefficients Elo

Afin d'évaluer si la somme des effets du Elo des blancs et du Elo des noirs est statistiquement nulle dans le premier modèle logistique, un test de Wald a été réalisé à l'aide de la fonction wald.test de la librairie aod.

$$H_0 : \beta_{\text{BlancElo}} + \beta_{\text{NoirElo}} = 0$$

```
wald.test(b =coefs, Sigma =vcov_mod1, L =matrix(c(0, 1,1), nrow =1))
```

```
#                Chi-squared test:
#       X2 = 58.4, df = 1, P(> X2) = 2.1e-14
```

La p-value est extrêmement petite (bien inférieure à 0,05), ce qui permet de rejeter l'hypothèse nulle. Cela signifie que la somme des coefficients associés aux Elo des deux joueurs est significativement différente de zéro.

Autrement dit, la formulation du modèle logistique ne se réduit pas simplement à une dépendance en fonction de la différence d'Elo : il existe une asymétrie structurelle entre les Elo des blancs et des noirs. Cela peut refléter un avantage intrinsèque aux blancs, souvent observé dans les statistiques d'échecs.

2.5 Analyse de l'intercept du modèle basé sur la différence d'Elo

Terme	Estimate	Std. Error	z-value	p-value
Intercept	0.1172	0.0040	29.06	$< 2 \times 10^{-16}$
Diff_Elo	0.004505	0.00001624	277.35	$< 2 \times 10^{-16}$

TABLE 2.2 – Résumé des coefficients du modèle logistique avec la variable `Diff_Elo`

— **AIC** : 369111
— **Null deviance** : 484835 on 350351 degrees of freedom
— **Déviance résiduelle** : 369107 on 350350 degrees of freedom

Pour ce modèle, on a une déviance résiduelle $D = 369107$ avec 350349 degrés de liberté. On rejette l'hypothèse que le modèle est bien ajusté si

$$D > \chi^2_{350349,\,0,95} \approx 384295{,}1.$$

Ce n'est **pas le cas**, donc on **ne rejette pas** l'hypothèse que le modèle est bien ajusté.

On a aussi une déviance nulle $D_0 = 484835$ avec 350350 degrés de liberté. On teste si le modèle est significativement meilleur que le modèle nul en comparant :

$$D_0 - D = 115728 \quad \text{et} \quad \chi^2_{1,\,0,95} \approx 3{,}84.$$

Ici, $115728 \gg 3{,}84$, donc on **rejette l'hypothèse** que le modèle est équivalent au modèle nul : le modèle avec `Diff_Elo` est **significativement meilleur**.

— L'intercept de 0,1172 indique que lorsque la différence d'Elo est nulle, les *log-odds* de victoire sont de 0,1172.

$$P = \frac{1}{1 + e^{-0{,}1172}} \approx 0{,}529$$

Alors probabilité de victoire des blancs est d'environ 52,9%. ce léger avantage s'explique souvent par le fait que les blancs jouent en premier.
— Le coefficient $\beta = 0{,}004505$ pour `Diff_Elo` signifie qu'une augmentation d'un point d'Elo augmente les *log-odds* de victoire de 0,004505.

2.6 Diagramme en boîte des différences d'Elo selon le résultat

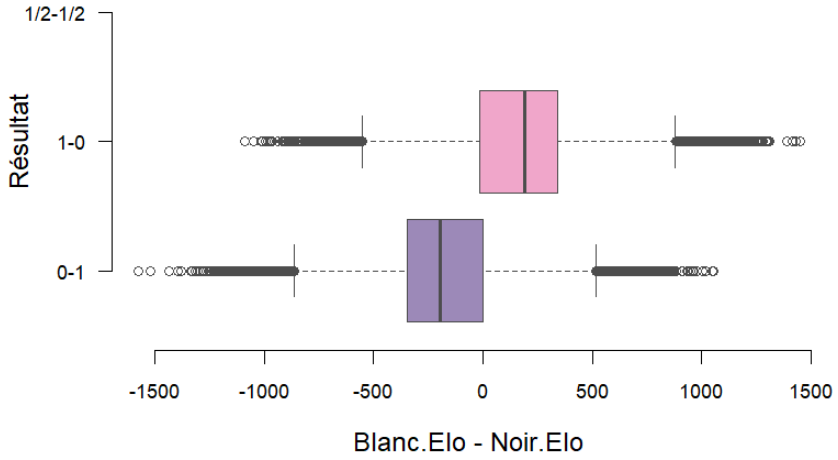


FIGURE 2.1 – Différences d'Elo selon le résultat de la partie

Les blancs ont en moyenne un Elo plus élevé que les noirs lorsqu'ils gagnent (1-0), et plus faible lorsqu'ils perdent (0-1), ce qui reflète une relation cohérente entre la différence d'Elo et les resultas de la partie.

2.7 Différence d'Elo et probabilité de victoire estimée

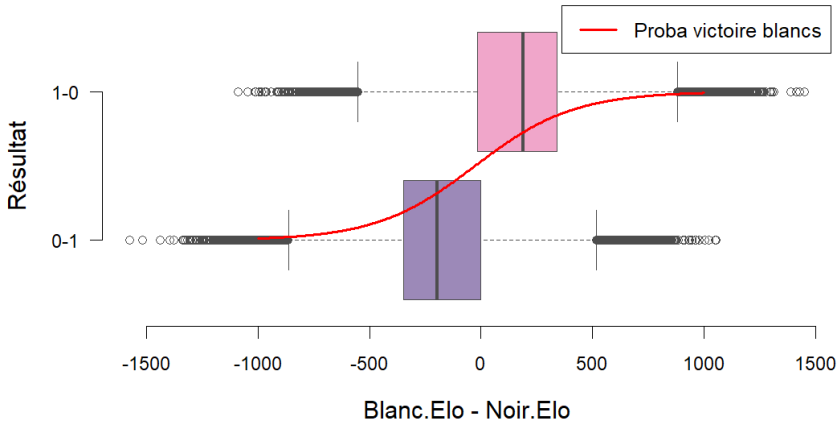


FIGURE 2.2 – Différence d'Elo selon le résultat, avec courbe de probabilité de victoire estimée

La courbe montre que plus les blancs ont un meilleur Elo que les noirs, plus leur chance de gagner augmente.

2.8 Recode du résultat : égalité vs pas égalité

A l'aide de la fonction levels, on'a remplacer les valeurs de la variable Résultat par égalité ou pas égalité selon le résultat de la partie.

```
levels(dat$Résultat) <- c("pas égalité", "pas égalité", "égalité")
table(dat$Résultat)
```

```
#   pas égalité     égalité
>    350352           53814
```

On observe que sur l'ensemble des parties, environ **13 %** se terminent par une égalité.

2.9 Création de la variable `Maitre`

On ajoute une variable `Maitre` au jeu de données. Elle prend la valeur 1 (*Maitre*) lorsque les deux joueurs ont un Elo supérieur à 2200, et 0 (*Non-Maitre*) sinon.

```
dat$Maitre <- ifelse(dat$Blanc.Elo > 2200& dat$Noir.Elo > 2200, 1, 0)
dat$Maitre <- factor(dat$Maitre, levels =c(0,1), labels =c("Non-Maitre","
    Maitre"))
tab <- table(dat$Maitre, dat$Résultat)
```

	pas égalité	**égalité**
Non-Maitre	344383	50513
Maitre	5969	3301

2.10 Rapport des cotes d'égalité entre maîtres et non-maîtres

On estime le rapport des cotes d'obtenir une égalité plutôt qu'un autre résultat pour les parties jouées entre deux joueurs ayant un Elo supérieur à 2200 (catégorie *Maitre*) par rapport aux autres parties (*Non-Maitre*). En utilisant la commande suivante :

```
epitools::oddsratio.wald(tab)
```

Maitre	**pas égalité**	**égalité**	**Total**
Non-Maitre	344383	50513	394896
Maitre	5969	3301	9270
Total	350352	53814	404166

Le rapport des cotes estimé est de **3.77**, avec un intervalle de confiance à 95 % de **[3.61 ; 3.94]**. qui ne contient pas le alors la probabilité d'égalité est environ **3,77 fois plus élevée** dans les parties entre maîtres que dans les autres.

les joueurs ayant une Elo supérieur a 2200 a une gros probalite d'aboutir à une égalité comparent au jouers non-maitre

2.11 Différences d'Elo selon le résultat de la partie : égalité ou non

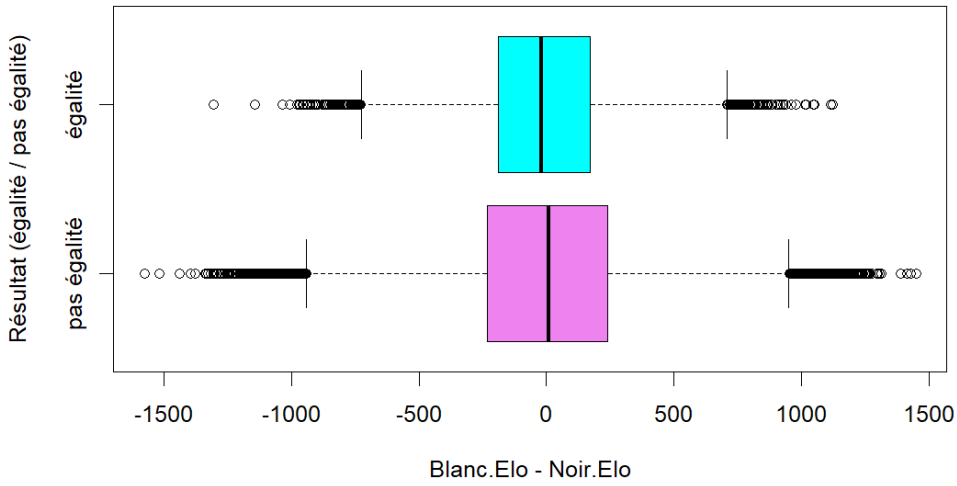


FIGURE 2.3 – Différence d'Elo selon que la partie ait été une égalité ou pas

2.12 modèle logistique quadratique

```
model <- glm(Résultat ~ Diff + I(Diff^2), data =dat2, family =binomial)
```

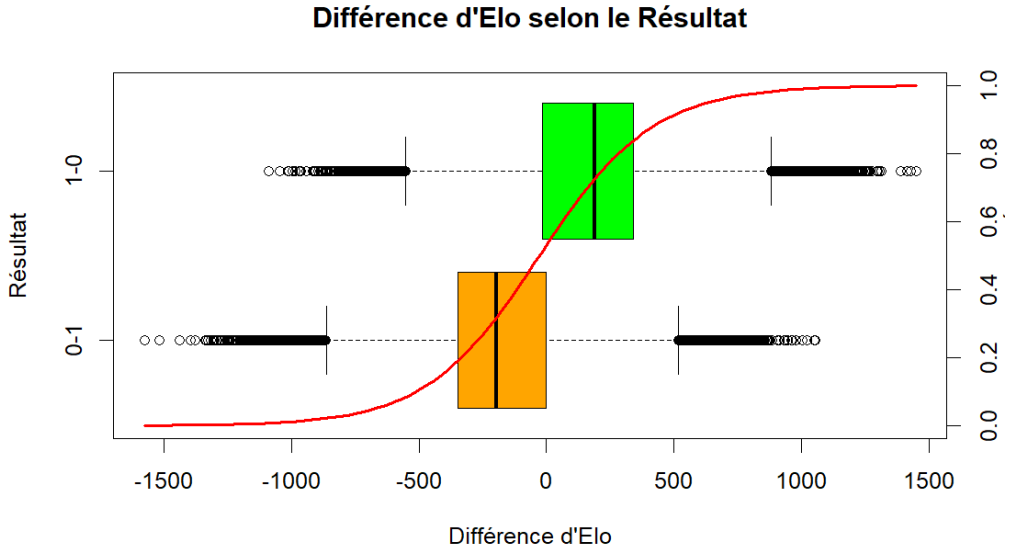


FIGURE 2.4 – Différence d'Elo selon le Résultat

3 Ordinateurs portables.

Notre problématique est la suivante : Quels sont les facteurs qui influencent le prix des laptops ? Nous cherchons à comprendre comment différentes caractéristiques techniques — telles que la taille de l'écran, la mémoire vive, le type de processeur ou encore les options d'affichage — impactent le prix d'un ordinateur portable.

Quantitatives
Price_euros
Inches
ScreenW, ScreenH
PPI
Ram
CPU_freq
PrimaryStorage
SecondaryStorage

Qualitatives
Company
TypeName
Touchscreen
IPSpanel
RetinaDisplay
PrimaryStorageType
SecondaryStorageType
GPU_model
OS

Variables dérivées
log_price
PPI
StorageScore
GPU_perf

TABLE 3.1 – Classification des variables utilisées.

3.1 Analyse exploratoire des données (EDA)

3.1.1 Statistique descriptive

Notre base de données analysée contient **1275 observations** et **23 variables**.

Absence de valeurs manquantes. Aucune variable ne contient de valeurs manquantes.

Les Tableaux 3.2 et 3.3 présentent les principales statistiques descriptives.

Variable	Min	Q1	Médiane	Moyenne	Max
Inches	10.1	14.0	15.6	15.02	18.4
Ram (Go)	2	4	8	8.44	64
CPU_freq (GHz)	0.9	2.0	2.5	2.30	3.6
Weight (kg)	0.69	1.50	2.04	2.04	4.7
Price_euros (€)	174	609	989	1135	6099
ScreenW (px)	1366	1920	1920	1900	3840
ScreenH (px)	768	1080	1080	1074	2160
PrimaryStorage (Go)	8	256	256	444.5	2048
SecondaryStorage (Go)	0	0	0	176.1	2048

TABLE 3.2 – Statistiques descriptives des variables quantitatives (n = 1275).

Les variables numériques montrent une grande variabilité, notamment pour `Ram` (de 2 à 64 Go) et `Price_euros`, qui est très asymétrique.

Variable	Type	Nombre de modalités
Company	Nominale	17
TypeName	Nominale	6
OS	Nominale	7
Touchscreen	Binaire	2
IPSpanel	Binaire	2
RetinaDisplay	Binaire	2
PrimaryStorageType	Nominale	4
SecondaryStorageType	Nominale	4
GPU_company	Nominale	4
GPU_model	Texte libre	109

TABLE 3.3 – Variables qualitatives et nombre de modalités.

La majorité des variables qualitatives ont un nombre de modalités limité, ce qui facilite leur intégration dans un modèle, à l'exception de `GPU_model`, très détaillée, et nécessitant une simplification.

Nous avons également vérifié la présence de doublons dans la base de données , aucune observation dupliquée n'a été détectée.

3.1.2 Visualisation

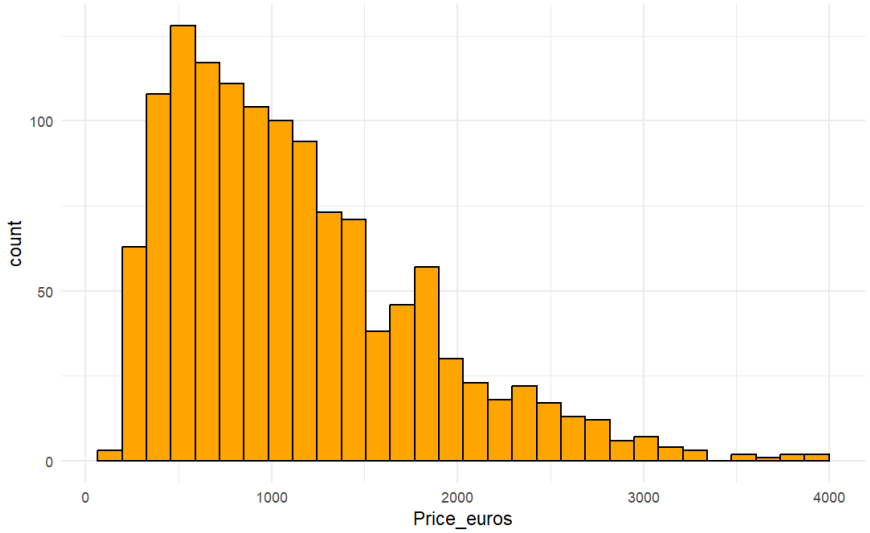


FIGURE 3.1 – Distribution des Prix des Laptops

la Figure 3.1,montre la distribution des prix est fortement asymétrique : la majorité des laptops se situent entre 500€ et 1500€.

Afin de corriger l'asymétrie observée dans la distribution des prix, nous avons appliqué une transformation logarithmique à la variable `Price_euros`. La Figure 3.2 montre la distribution des prix transformés, qui apparaît visiblement plus symétrique.

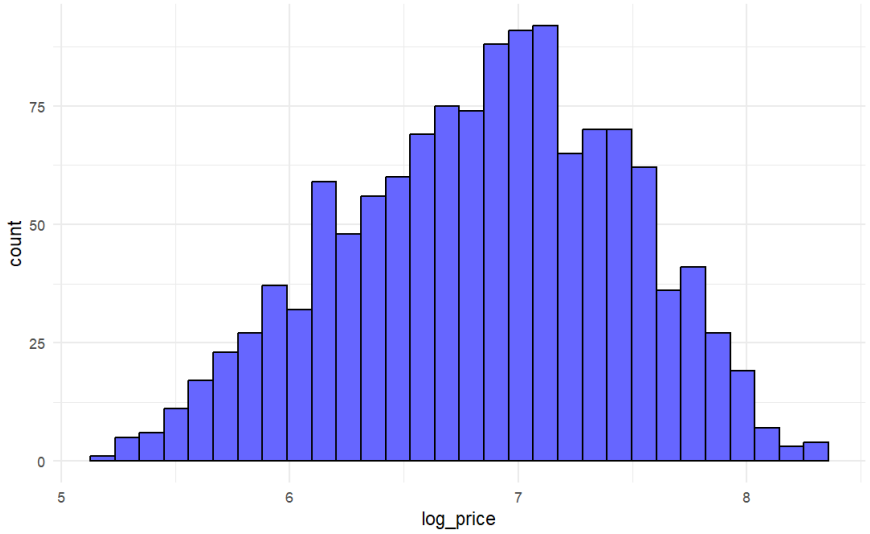


FIGURE 3.2 – Distribution log-transformée des prix des laptops.

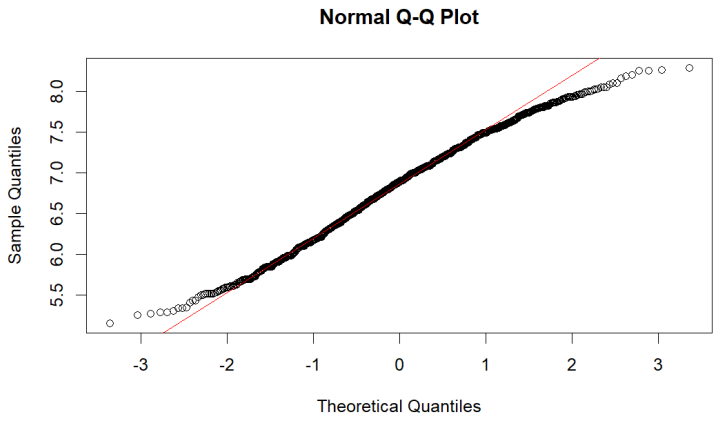


FIGURE 3.3 – Q-Q plot de la variable `log_price`.

Test de normalité (Shapiro-Wilk)

```
Shapiro-Wilk normality test
    data: dat$log_price
        W = 0.98983
    p-value = 9.935e-08
```

Même si le test de Shapiro détecte une légère non-normalité, la transformation logarithmique améliore clairement la distribution. Avec une statistique W proche de 1, on peut considérer que la normalité est raisonnablement (c'est pas 100%).

Le graphique suivant permet de visualiser la distribution des prix des laptops en fonction des marques présentes dans le jeu de données.

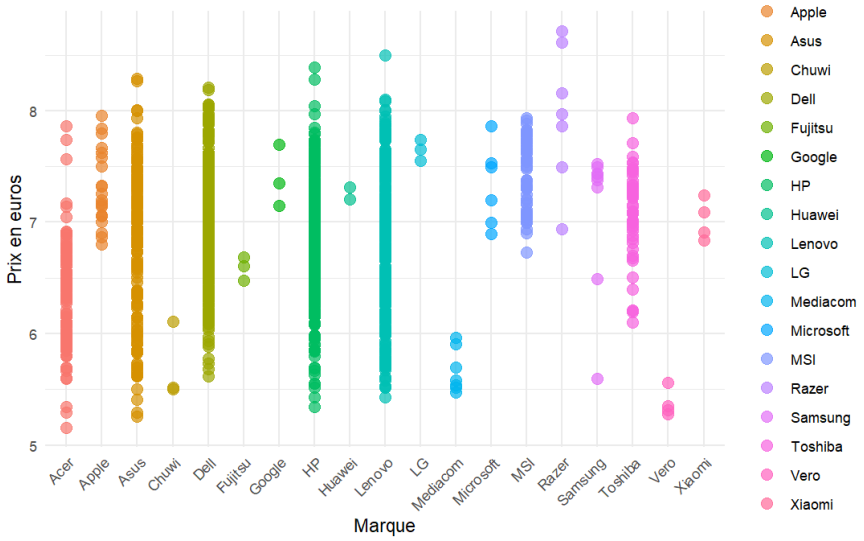


FIGURE 3.4 – Répartition des prix des ordinateurs portables par marque

Avec l'échelle logarithmique, on voit plus clairement que certaines marques comme *Apple*,

Razer ou *MSI* ont des prix généralement plus élevés, tandis que d'autres comme *Chuwi*, *Vero* ou *Mediacom* proposent des ordinateurs à des prix plus bas.

Vérification des valeurs extrêmes Nous avons identifié quelques ordinateurs portables dont le prix est supérieur à 3 900 €, ce qui peut indiquer des valeurs aberrantes. Après vérification manuelle, nous avons trouvé que le prix était raisonnable. Elle est donc justifiée et ne nécessite pas de retrait (modification) de l'analyse.

Marque	**Modèle**	**Prix (€)**
Razer	Blade Pro	6099.0
Lenovo	Thinkpad P51	4899.0
HP	ZBook 17	4389.0
Razer	Blade Pro	5499.0
Asus	ROG G701VO	3975.0
HP	ZBook 17	3949.4

TABLE 3.4 – Laptops avec un prix supérieur à 3900 €.

Nous présentons ci-dessous la distribution des marques dans le jeu de données, afin d'observer leur représentativité au sein de l'échantillon.

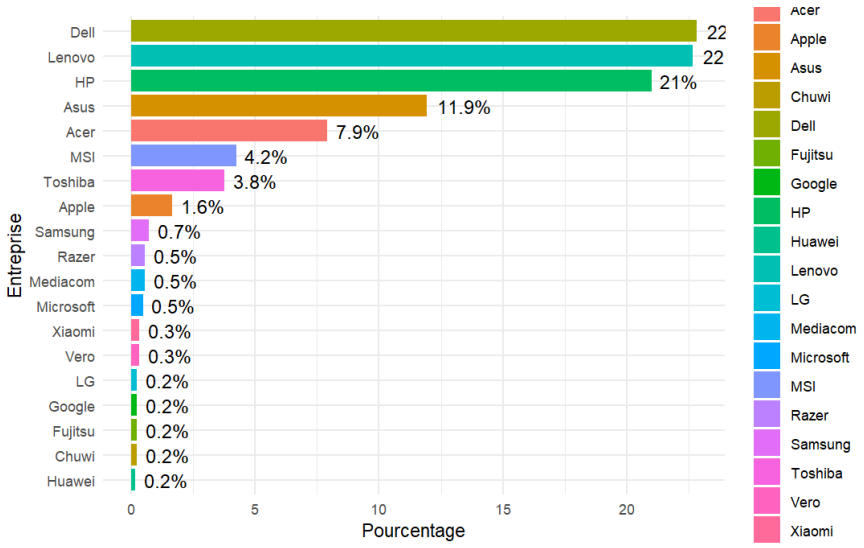


FIGURE 3.5 – Répartition des entreprises

La majorité des laptops provient de quelques grandes marques, notamment *Dell*, *Lenovo* et *HP*, qui représentent à elles seules plus de 60% des observations. À l'inverse, certaines marques comme *Google*, *Huawei* ou *Vero* sont très peu représentées.

Les graphiques ci-dessus montrent la répartition croisée des variables qualitatives (Type, OS, CPU/GPU)

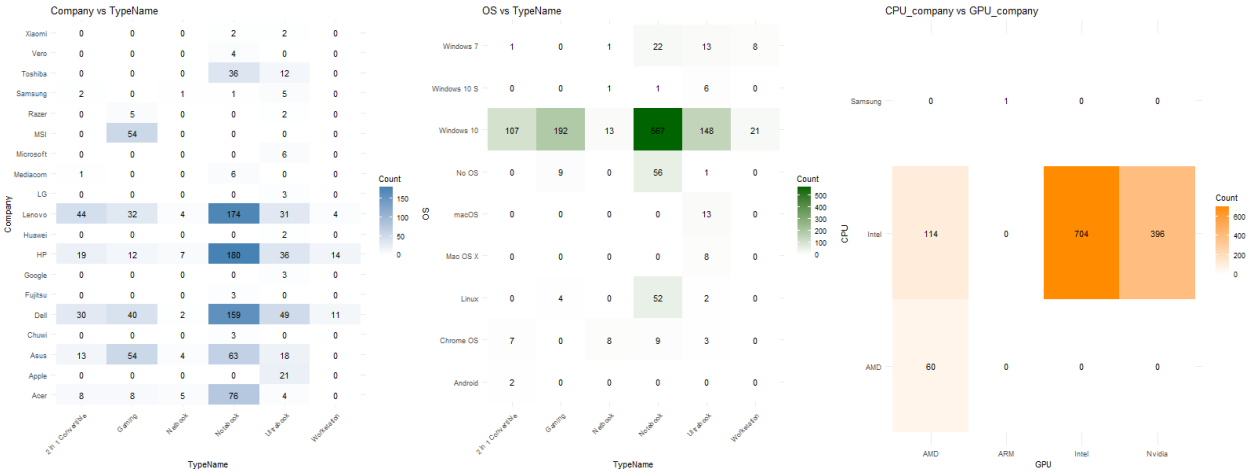


FIGURE 3.6 – Analyse croisée des variables qualitatives (Type, OS, CPU/GPU)

On remarque que certaines marques sont spécialisées dans des types spécifiques de laptops : par exemple, *MSI* produit uniquement des modèles *Gaming*, tandis que *Lenovo* et *HP* proposent une large gamme de types.

La majorité des ordinateurs portables utilisent *Windows 10*, quel que soit le type d'appareil. D'autres systèmes comme *Chrome OS* ou *Linux* restent minoritaires.

Enfin, On voit que les processeurs *Intel* sont souvent associés à des cartes graphiques *Intel* ou *Nvidia*, alors que les processeurs *AMD* apparaissent dans beaucoup moins de configurations.

Etude de la coorélation

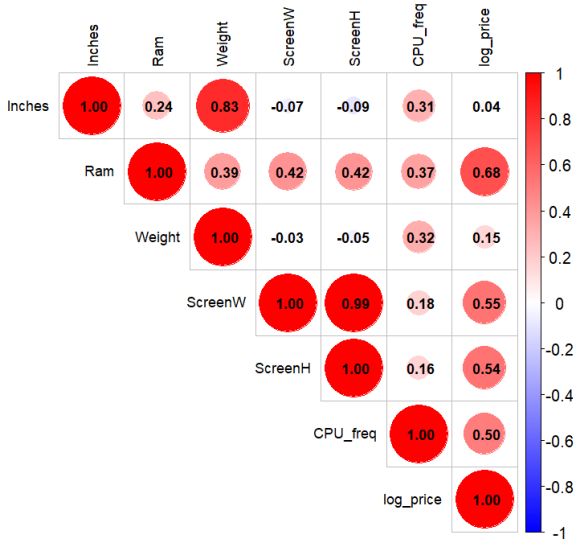


FIGURE 3.7 – Matrice de corrélation

On observe que la variable Ram est fortement corrélée avec le prix (log_price), tout comme la résolution de l'écran (ScreenW et ScreenH) ou encore la fréquence du processeur (CPU_freq).

En revanche, certaines variables comme la taille de l'écran (Inches) ou le poids (Weight) présentent une corrélation beaucoup plus faible avec le prix.

Deux variables, PrimaryStorage et SecondaryStorage, ont été retirées de la matrice de corrélation, car leurs coefficients n'étaient pas significatifs. Cela s'explique par leur relation avec d'autres colonnes, notamment Storage_type, ce qui peut biaiser l'analyse.

3.2 ANOVA

Impact du type et de la capacité de stockage sur le prix

On observe visuellement une différence notable entre les types et capacités de stockage. En particulier, les laptops équipés de *Flash Storage* ou de *SSD* semblent avoir des prix plus élevés.

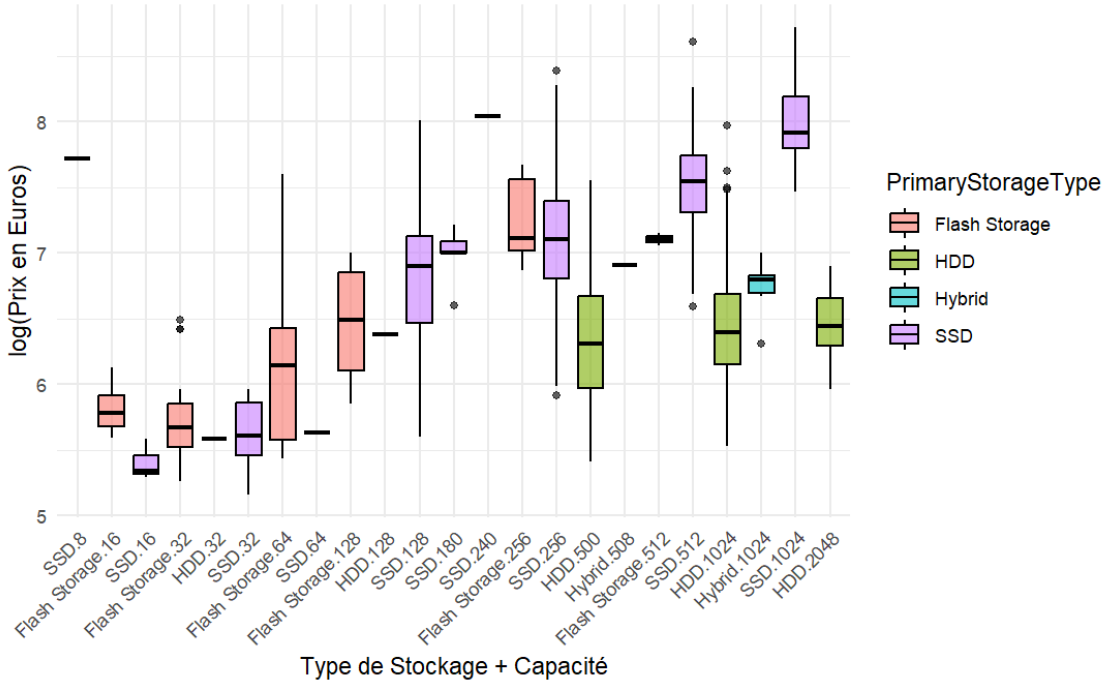
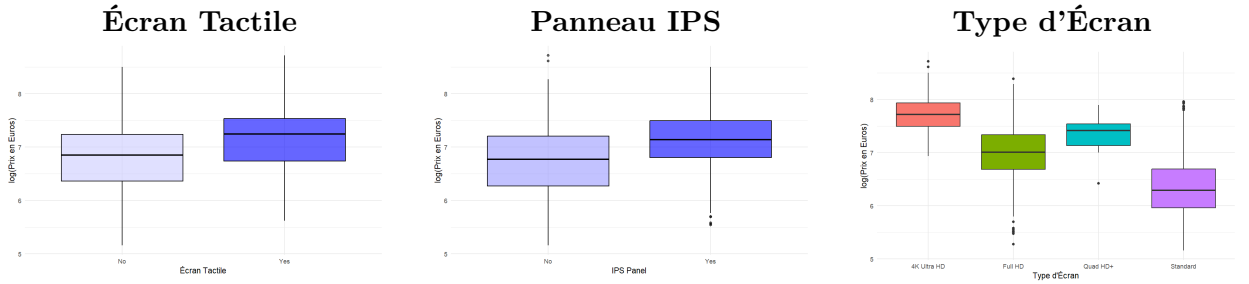


FIGURE 3.8 – Distribution du *log(prix)* selon le type et la capacité de stockage principal.

```
anova_primaty_storage <- aov(log_price ~ PrimaryStorage, data =dat)
summary(anova_primaty_storage)
```

p-value $< 0{,}001$ alors on'a un un effet significatif.

Même résultat pour la variable SecondaryStorage.



ANOVA : p-value $< 0{,}001$ — *effet significatif.*

ANOVA : p-value $< 0{,}001$ — *effet significatif.*

ANOVA : p-value $< 0{,}001$ — *effet significatif.*

FIGURE 3.9 – Effet des caractéristiques de l'écran sur le *log(prix)* des laptops.

3.2.1 AFC

Ce graphique représente une **analyse des correspondances** entre les marques de laptops (en bleu) et les types de modèles (en rouge). On observe que :

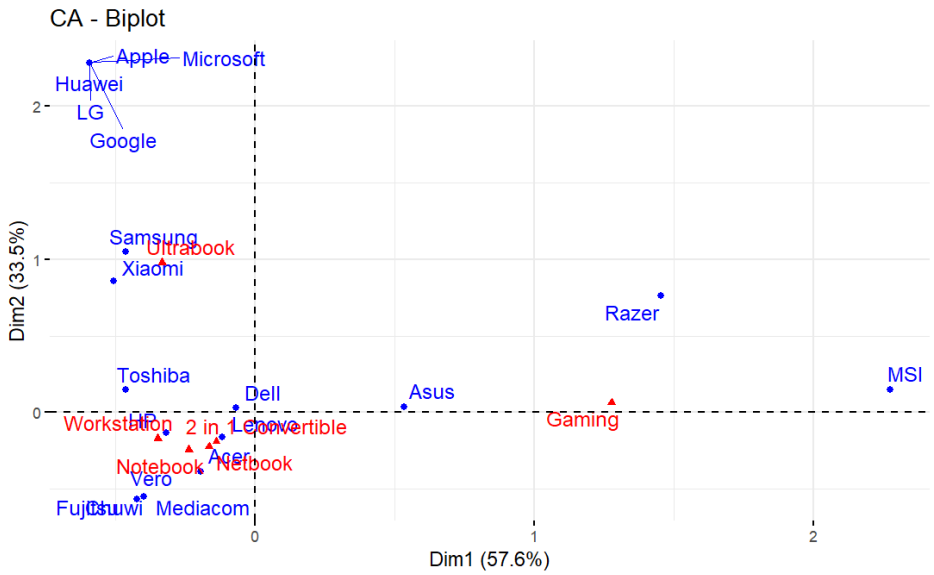


FIGURE 3.10 – Analyse des correspondances entre les marques et les types de laptops.

— **MSI** et **Razer** sont fortement liés aux modèles *Gaming*.
— **Xiomi** et **Samsung** sont proches des catégories *Ultrabook*.

3.3 Modèlisation

Nous avons choisi d'utiliser un première modèle de régression linéaire multiple, dans lequel la variable à expliquer est le logarithme du prix , et les variables explicatives sont des caractéristiques techniques et qualitatives du produit (écran tactile, résolution, stockage, système d'exploitation, marque, etc.).

```
model3 <- lm(log_price ~ Ram + CPU_freq + Inches + ScreenW + ScreenH +
    Touchscreen + IPSpanel + RetinaDisplay + PrimaryStorage + SecondaryStorage
    + Company + TypeName + OS + PrimaryStorageType + SecondaryStorageType, data
     =dat3)
```

```
# Residual standard error: 0.2807 on 1228degrees of freedom
# Multiple R-squared: 0.8023, Adjusted R-squared: 0.7949
# F-statistic: 108.3 on 46and 1228DF, p-value: < 2.2e-16
```

Le modèle obtenu est significatif (p-value $< 2.2 \times 10^{-16}$), avec un R^2 ajusté de 0,7949, ce qui montre que plus de 79% de la variance du prix est expliquée par les variables incluses. Cela indique une excellente capacité explicative du modèle.

Pour évaluer la qualité du modèle, nous avons analysé les résidus.

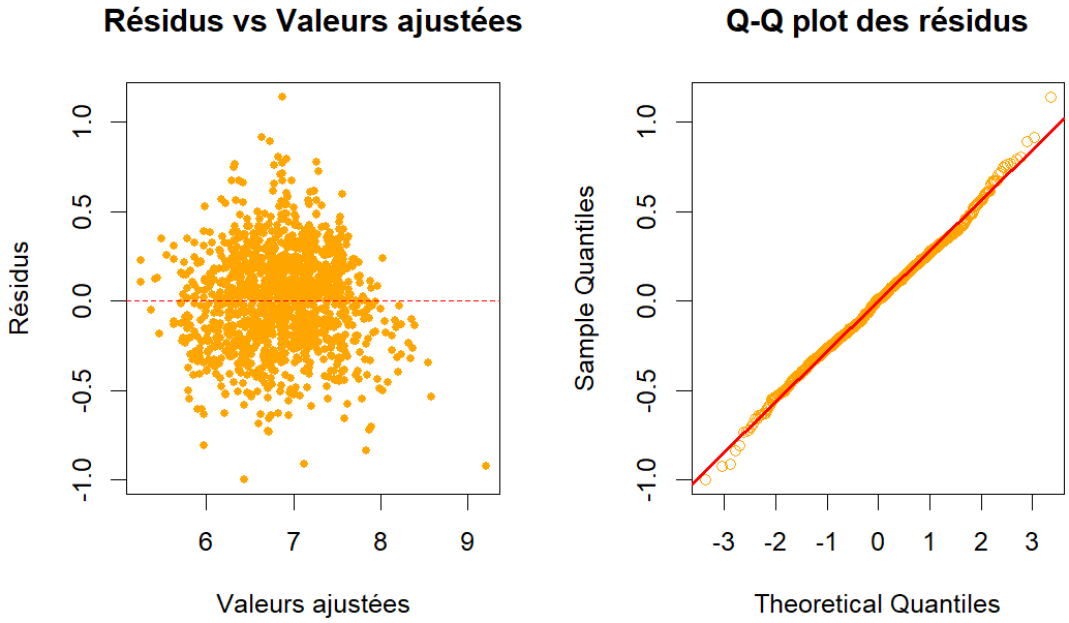


FIGURE 3.11 – analysé les résidus modele 1

Le graphe des résidus vs valeurs ajustées ne montre pas de structure particulière, ce qui indique que l'hypothèse d'homoscédasticité est raisonnablement respectée.

Le Q-Q plot montre que les points suivent bien la droite théorique, ce qui suggère que la normalité des résidus est globalement respectée, à l'exception de légères déviations en queue.

3.4 selection de modele avec la fonction Step

Afin d'identifier le modèle de régression linéaire le plus pertinent, nous avons utilisé la fonction step() en appliquant une sélection automatique des variables basée sur deux critères : AIC et BIC. Les sélections ont été effectuées dans les deux directions : ascendante et descendante .

Les résultats ont montré que les deux approches (ascendante et descendante) ont conduit au même modèle final pour chaque critère (AIC et BIC). Cela confirme la stabilité des modèles sélectionnés.

Le tableau ci-dessous présente les valeurs d'AIC et BIC obtenues, ainsi qu'un résumé des performances des modèles finaux.

TABLE 3.5 – Comparaison des modèles sélectionnés selon les critères AIC et BIC

Critère	**Méthode**	**AIC / BIC**	**RSE**	**R^2 ajusté**
AIC	Ascendante	-2277.704	0.0912	0.9783
AIC	Descendante	-2277.704	0.0912	0.9783
BIC	Ascendante	-1441.493	0.1012	0.9733
BIC	Descendante	-1441.493	0.1012	0.9733

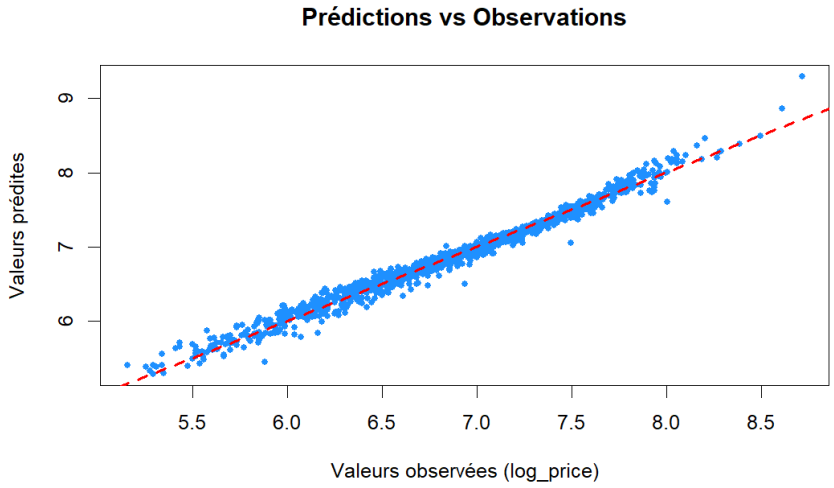


FIGURE 3.12 – Prédictions vs Observations $\text{modele}_s elec_A IC_a sc$

3.5 Division du jeu de données : entraînement et test

Dans le but d'évaluer la performance réelle du modèle et de détecter un éventuel surapprentissage (overfitting), nous avons divisé le jeu de données en deux sous-ensembles distincts train et test.

le model choisi par AIC :

```
Step: AIC=-2471.96
log_price ~ Ram + TypeName + PrimaryStorageType + CPU_freq +
CPU_company + OS + ScreenH + Screen + Company + IPSpanel +
RetinaDisplay
```

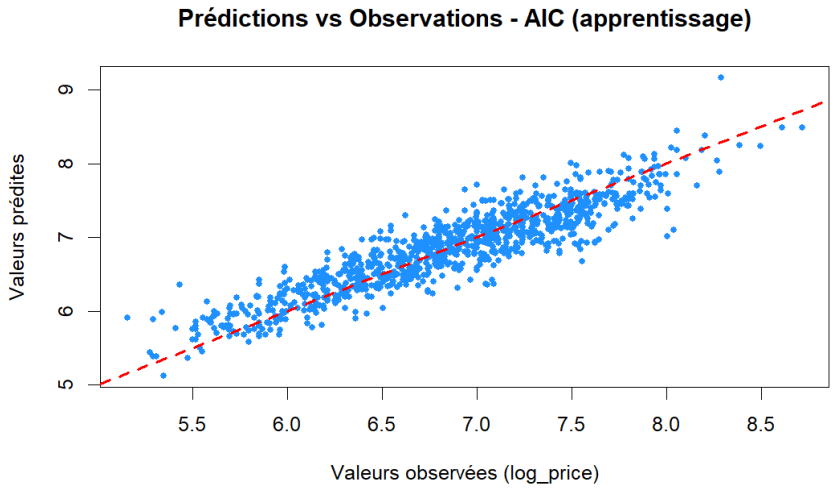


FIGURE 3.13 – Prédictions vs Observations - AIC (apprentissage)

```
summary(modele_AIC_train)

#Residual standard error: 0.2605 on 970degrees of freedom
Multiple R-squared: 0.8319, Adjusted R-squared: 0.8234
F-statistic: 97.98 on 49and 970DF, p-value: < 2.2e-16
```

3.6 Création de la variable PPI et ajustement du modèle

Nous avons créé une nouvelle variable **PPI** (pixels per inch) à partir des dimensions de l'écran, selon la formule :

$$\text{PPI} = \frac{\sqrt{\text{ScreenW}^2 + \text{ScreenH}^2}}{\text{Inches}}$$

Cette mesure permet de refléter la densité de pixels de l'écran, un facteur pertinent pour expliquer le prix.

Un modèle de régression linéaire multiple a ensuite été ajusté.

Le modèle présente un bon ajustement avec un R^2 ajusté de **0.7955**, une erreur standard résiduelle de **0.2802** et une statistique F hautement significative (p-value $< 2.2 \times 10^{-16}$).

3.7 Création de la variable `StorageScore`

Pour mieux représenter la performance globale du stockage, nous avons construit une nouvelle variable continue appelée `StorageScore`, qui combine la capacité de stockage (primaire et secondaire) et le type de technologie utilisée (SSD, HDD, Hybride, ou aucune). Chaque type a été associé à un score basé sur sa rapidité : SSD = 2, HDD = 1, Hybrid = 1.5, None = 0. La formule est la suivante :

```
  score_type <- function(x) {
  ifelse(x =="SSD", 2,
  ifelse(x =="HDD", 1,
  ifelse(x =="Hybrid", 1.5, 0)))
}

ssd_score1 <- score_type(dat3$PrimaryStorageType)
ssd_score2 <- score_type(dat3$SecondaryStorageType)

dat3$StorageScore <- dat3$PrimaryStorage * ssd_score1 +
dat3$SecondaryStorage * ssd_score2
```

Le modèle linéaire ajusté incluant cette nouvelle variable montre un bon pouvoir explicatif avec un R^2 ajusté de **0.7699** et une erreur standard résiduelle de **0.2973**.

Conclusion

Dans ce projet, nous avons utilisé différents modèles statistiques pour étudier trois cas : le sexage des oiseaux Soras, le système de classement Elo aux échecs, et les facteurs influençant le prix des ordinateurs portables. Chaque partie a montré comment les modèles linéaires et logistiques peuvent aider à comprendre les données et faire des prédictions utiles. Grâce à ces analyses, nous avons pu mieux interpréter les résultats et tirer des conclusions claires à partir des données

Liste des tableaux

Table des figures