Analyse de Régression Multiple : Taux d'Insertion

Salma Lahbati Cyrena Ramdani Yoav Cohen

Table des matières

1 Présentation des données et des variables

Cette étude repose sur un échantillon de diplômés hommes ayant suivi un cursus de Master dans la filière Sciences Technologiques et Sociales (STS) et Droit Économie Gestion (DEG). Les données incluent des observations issues de différentes disciplines, notamment :

— Les Sciences fondamentales

— Ensemble sciences, technologies et santé

— Sciences de la vie et de la terre

— Sciences de l'ingénieur

— Informatique

Dans cette analyse, nous faisons appel à un modèle de régression multiple pour étudier les facteurs influençant le taux d'insertion des diplômés. Ce taux est expliqué par un ensemble de variables explicatives, qui incluent :

— Les caractéristiques spécifiques du domaine d'études des diplômés.

— La situation des diplômés 18 mois après l'obtention de leur diplôme.

— Des indicateurs économiques et sociaux clés tels que le taux de chômage national, les niveaux de salaires et la répartition des emplois en fonction de la stabilité et du type de contrat.

1.1 Première équation du modèle

Le modèle de régression multiple pour expliquer le taux d'insertion est donné par l'équation suivante :

$$Taux\ d'insertion_i = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n + \epsilon_i$$

Où :

— $X_1, X_2, \ldots, X_n$: Représentent les variables explicatives, telles que le taux de chômage, les salaires, etc.

— $\beta_0, \beta_1, \ldots, \beta_n$: Sont les coefficients à estimer.

— ϵ_i : Désigne l'erreur aléatoire ou le résidu pour l'observation i.

1.2 Prétraitement des données

Le prétraitement des données a été réalisé en plusieurs étapes. Tout d'abord, les lignes avec des valeurs manquantes ou non numériques dans la colonne "Taux d'insertion" ont été supprimées. Ensuite, seules les colonnes

pertinentes ont été conservées, et les colonnes contenant plus de 50% de valeurs manquantes ont été éliminées. Pour les autres valeurs manquantes, elles ont été remplacées par la moyenne des valeurs de la colonne.

La colonne "Année" a été convertie en format datetime pour faciliter l'analyse temporelle. Enfin, le fichier nettoyé a été sauvegardé dans un nouveau fichier CSV prêt à être utilisé pour les analyses.

1.3 Variables explicatives sélectionnées

1.3.1 Variables explicatives

Les variables explicatives sélectionnées pour le modèle sont les suivantes :

- Situation
- Part des emplois de niveau cadre
- Part des emplois de niveau cadre ou profession intermédiaire
- Part des emplois à temps plein
- Salaire brut annuel estimé
- Part des diplômés boursiers dans la discipline
- % emplois extérieurs à la région de l'université
- Part des emplois stables
- Code de la discipline
- Code du secteur disciplinaire
- Genre
- Taux de chômage national
- Année (supprimé au 3. car multicolinéarité)

1.3.2 Composantes des variables explicatives

Nous avons également plusieurs composantes au sein de

- "Code de la discipline" :
  - disc12 : Ensemble sciences, technologies et santé
  - disc 13 : Sciences de la vie et de la terre
  - disc 14 : Sciences fondamentales
  - disc 15 : Sciences de l'ingénieur
  - disc 16 : Informatique.
- "Code du secteur disciplinaire" :
  - Disc14_01 : Chimie.

2 Construction du modèle de régression multiple

Les résultats de la régression ordinaire des moindres carrés (OLS) pour le taux d'insertion sont présentés ci-dessous :

OLS Regression Results

Dep. Variable:	Taux d'insertion	R-squared:	0.634
Model:	OLS	Adj. R-squared:	0.632
Method:	Least Squares	F-statistic:	328.2
Date:	Thu, 09 Jan 2025	Prob (F-statistic):	0.00
Time:	22:54:45	Log-Likelihood:	-6580.1
No. Observations:	2481	AIC:	1.319e+04
Df Residuals:	2467	BIC:	1.327e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	660.1697	299.628	2.203	0.028	72.622	1247.718
situation	-0.8207	0.159	-5.148	0.000	-1.133	-0.508
Part des emplois de niveau cadre	0.0620	0.011	5.801	0.000	0.041	0.083
Part des emplois de niveau cadre ou profession intermédiaire	0.1545	0.017	8.975	0.000	0.121	0.188
Part des emplois à temps plein	0.0290	0.015	1.975	0.048	0.000	0.058
Salaire brut annuel estimé	-0.0002	3.63e-05	-4.409	0.000	-0.000	-8.89e-05
Part des diplômés boursiers dans la discipline	-0.0640	0.013	-4.992	0.000	-0.089	-0.039
% emplois extérieurs à la région de l'université	0.0060	0.008	0.739	0.460	-0.010	0.022
Part des emplois stables	0.1962	0.008	24.638	0.000	0.181	0.212
Code de la discipline	-3.5015	0.208	-16.850	0.000	-3.909	-3.094
Code du secteur disciplinaire	-1.8512	0.515	-3.596	0.000	-2.861	-0.842
Genre	-1.9271	0.165	-11.707	0.000	-2.250	-1.604
Taux de chômage national	-1.3573	0.359	-3.776	0.000	-2.062	-0.652

Omnibus:	275.900	Durbin-Watson:	1.656
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1073.926
Skew:	-0.497	Prob(JB):	6.31e-234
Kurtosis:	6.066	Cond. No.	7.80e+05

FIGURE 1 – Résultats OLS

Ce modèle est pas significatif car on a une constante = 660,17 ce qui n'est pas normal comparé aux valeurs du taux d'insertions habituels (entre 75 et 90). Nous concluerons sur la signification des coefficients une fois que nous aurons réduis la multicolinéarité.

2.1 Graphique de régression

Voici le graphique représentant la régression réalisée sur le taux d'insertion des diplômés :

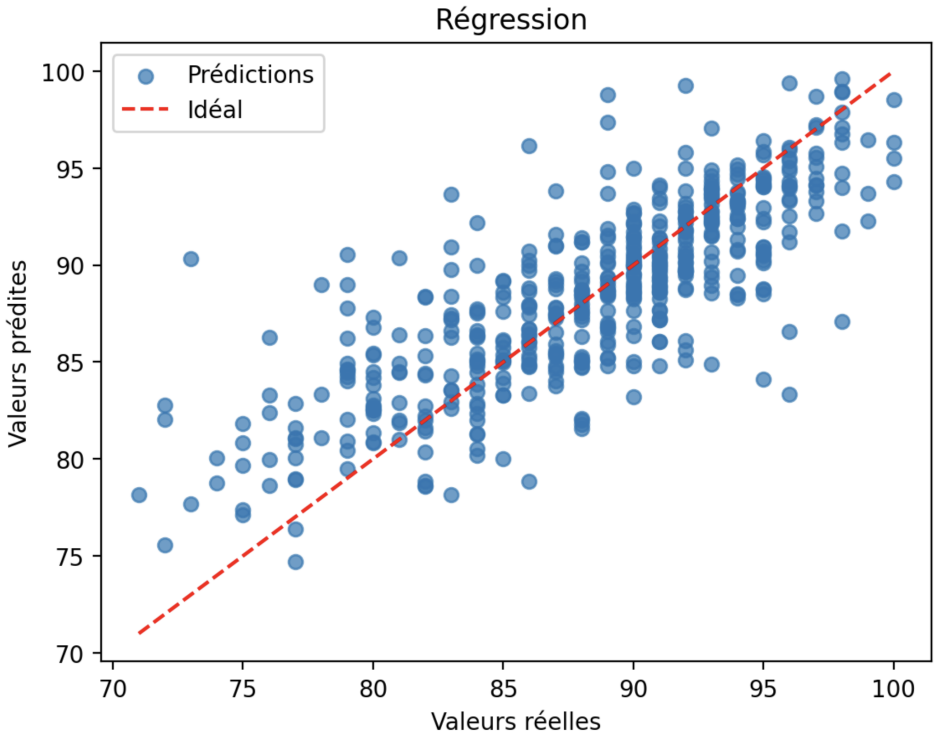


FIGURE 2 – Graphique de la régression du taux d'insertion

3 Validation et diagnostic du modèle

3.1 Analyse de la multicolinéarité (VIF)

3.1.1 Définition du VIF

Le Variance Inflation Factor (VIF) est un indicateur utilisé pour détecter la multicolinéarité dans un modèle de régression. Pour chaque variable explicative X_j, le VIF est défini comme suit :

$$\text{VIF}(X_j) = \frac{1}{1 - R_j^2}$$

où R_j^2 est le coefficient de détermination obtenu en régressant X_j sur toutes les autres variables explicatives du modèle.

3.1.2 Résultats du VIF

Les valeurs du VIF pour chaque variable explicative du modèle sont présentées dans le tableau ci-dessous :

Variable	VIF
Situation	1.3294
Part des emplois de niveau cadre	7.1989
Part des emplois de niveau cadre ou profession intermédiaire	6.6020
Part des emplois à temps plein	2.9160
Salaire brut annuel estimé	4.7502
Part des diplômés boursiers dans la discipline	1.5699
%emplois extérieurs à la région de l'université	1.2924
Part des emplois stables	3.1732
Code de la discipline	1.8881
Code du secteur disciplinaire	1.1170
Genre	1.1112
Taux de chômage national	28.0255
Année	28.7702

3.1.3 Analyse de la Multicolinéarité

Les résultats du Variance Inflation Factor (VIF) montrent que les variables *Année* et *Taux de chômage national* ont des VIF supérieurs à 10, ce qui suggère une forte multicolinéarité. Cela peut entraîner un biais dans les estimations du modèle, car ces variables sont hautement corrélées. De plus,

certaines variables, telles que *Part des emplois de niveau cadre*, présentent une corrélation modérée, mais sans risque significatif de multicolinéarité.

Pour améliorer la robustesse du modèle, nous proposons de supprimer la variable *Année*, car elle est fortement corrélée avec *Taux de chômage national* et explique en grande partie la même chose.

3.1.4 Explication : Pourquoi le taux de chômage pourrait expliquer la même chose que l'année ?

Le taux de chômage peut être lié à l'année pour plusieurs raisons :

— **Tendance temporelle commune** : Le taux de chômage suit souvent une évolution régulière au fil des années, reflétant des tendances économiques globales. Par exemple, lors de périodes de croissance économique, le taux de chômage a tendance à diminuer, tandis que durant des récessions, il augmente.

— **Événements économiques ou politiques spécifiques** : Certains événements marquants, comme une récession, une crise financière (ex. : 2008) ou des événements mondiaux (comme la pandémie de 2020-2022), peuvent fortement associer une année donnée à un niveau particulier de chômage.

— **Manque de variables explicatives dans le modèle** : Si d'autres facteurs qui influencent le taux de chômage (comme des politiques économiques spécifiques, des avancées technologiques ou des changements structurels dans l'économie) ne sont pas inclus dans le modèle, la variable *Année* pourrait agir comme une variable de substitution, capturant des effets non pris en compte.

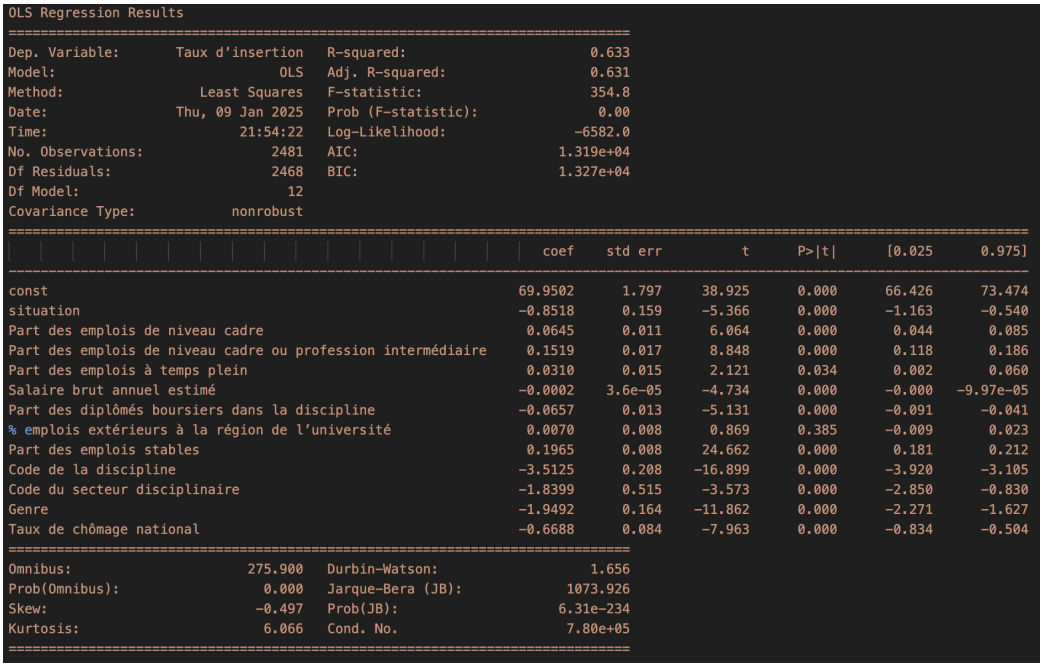
— **Structure des données** : Si le taux de chômage évolue de manière systématique d'une année à l'autre dans les données, la variable *Année* deviendrait inévitablement fortement corrélée avec le taux de chômage, ce qui rend la présence des deux variables dans le modèle redondante.

En conséquence, pour éviter la multicolinéarité, il serait pertinent de retirer la variable *Année* et de se concentrer sur des variables plus distinctes et indépendantes dans l'analyse du taux d'insertion.

3.1.5 Modèle Ajusté : suppréssion de la variable Année

Nous avons donc supprimé la variable "Année" qui introduisait de la multicolinéarité dans le modèle. Voici les résultats du nouveau modèle :

Maintenant, nous pouvons conlclure sur la pertinence des variables (et l'interprétation des coefficients).



OLS Regression Results

Dep. Variable:	Taux d'insertion	R-squared:	0.633
Model:	OLS	Adj. R-squared:	0.631
Method:	Least Squares	F-statistic:	354.8
Date:	Thu, 09 Jan 2025	Prob (F-statistic):	0.00
Time:	21:54:22	Log-Likelihood:	-6582.0
No. Observations:	2481	AIC:	1.319e+04
Df Residuals:	2468	BIC:	1.327e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	69.9502	1.797	38.925	0.000	66.426	73.474
situation	-0.8518	0.159	-5.366	0.000	-1.163	-0.540
Part des emplois de niveau cadre	0.0645	0.011	6.064	0.000	0.044	0.085
Part des emplois de niveau cadre ou profession intermédiaire	0.1519	0.017	8.848	0.000	0.118	0.186
Part des emplois à temps plein	0.0310	0.015	2.121	0.034	0.002	0.060
Salaire brut annuel estimé	-0.0002	3.6e-05	-4.734	0.000	-0.000	-9.97e-05
Part des diplômés boursiers dans la discipline	-0.0657	0.013	-5.131	0.000	-0.091	-0.041
% emplois extérieurs à la région de l'université	0.0070	0.008	0.869	0.385	-0.009	0.023
Part des emplois stables	0.1965	0.008	24.662	0.000	0.181	0.212
Code de la discipline	-3.5125	0.208	-16.899	0.000	-3.920	-3.105
Code du secteur disciplinaire	-1.8399	0.515	-3.573	0.000	-2.850	-0.830
Genre	-1.9492	0.164	-11.862	0.000	-2.271	-1.627
Taux de chômage national	-0.6688	0.084	-7.963	0.000	-0.834	-0.504

Omnibus:	275.900	Durbin-Watson:	1.656
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1073.926
Skew:	-0.497	Prob(JB):	6.31e-234
Kurtosis:	6.066	Cond. No.	7.80e+05

FIGURE 3 – Régression ajustée

— **Taux de chômage national** ($\beta = -0.6688$, $p < 0.01$) : Une augmentation de 1 % du taux de chômage national entraîne une diminution moyenne de 0.6688 points du taux d'insertion des diplômés. Cela souligne l'importance du contexte économique sur l'insertion professionnelle.

— **Part des emplois stables** ($\beta = 0.1965$, $p < 0.01$) : Une augmentation de 1 % de la part des emplois stables est associée à une augmentation moyenne de 0.1965 points du taux d'insertion. Cela montre que les diplômés bénéficient directement d'un marché du travail avec plus d'emplois sécurisés.

— **Part des emplois de niveau cadre ou profession intermédiaire** ($\beta = 0.1519$, $p < 0.01$) : Une augmentation de 1 % de la part des emplois de niveau cadre ou intermédiaire entraîne une hausse moyenne de 0.1519 points du taux d'insertion. Les opportunités professionnelles mieux qualifiées jouent un rôle important dans l'employabilité des diplômés.

Ces résultats montrent que l'insertion professionnelle des diplômés dépend fortement du contexte économique et de la qualité des emplois disponibles. Une amélioration des conditions du marché du travail, notamment avec plus d'emplois stables et qualifiés, pourrait significativement augmenter leur taux

d'insertion.

3.2 Analyse de la multicolinéarité

3.3 Test de Breusch-Pagan (Homoscédasticité)

3.3.1 Résultats du test

Les résultats du test de Breusch-Pagan sont les suivants :

— **Statistique de Breusch-Pagan** : 92.98

— **p-valeur** : 1.304×10^{-14}

La p-valeur est inférieure à 0.05, ce qui indique une **hétéroscédasticité** (variance des résidus non constante). Cela suggère que le modèle pourrait ne pas bien représenter la variabilité des erreurs, et qu'un ajustement ou un autre modèle pourrait être nécessaire.

3.3.2 Interprétation visuelle

L'analyse visuelle des résidus peut fournir des informations supplémentaires sur l'homoscédasticité :

— **Cas d'homoscédasticité** : Si les résidus sont dispersés de manière aléatoire autour de la ligne 0, cela soutient l'hypothèse d'homoscédasticité.

— **Cas d'hétéroscédasticité** : Si une forme ou un patron particulier apparaît (par exemple, des "éventails" ou des clusters), cela indiquerait de l'hétéroscédasticité, ce qui signifie que la variance des erreurs n'est pas constante.

3.4 Correction de l'hétéroscédasticité

Ici, il est nécessaire de réduire l'hétéroscédasticité, car la statistique de Breusch-Pagan est de 92.98, indiquant une variance des résidus non constante.

Une approche pour corriger l'hétéroscédasticité consiste à utiliser un modèle comme **XGBoost**, qui réduit l'hétéroscédasticité en ajustant les erreurs de manière adaptative et en capturant des relations complexes entre les variables explicatives. Le modèle minimise une fonction de perte qui inclut à la fois la perte de prédiction et une régularisation pour contrôler la complexité.

La fonction de perte est définie comme suit :

$$L = \sum_{i=1}^{n} \text{loss}(y_i, \hat{y}_i) + \lambda \sum_{t} \|w_t\|^2$$

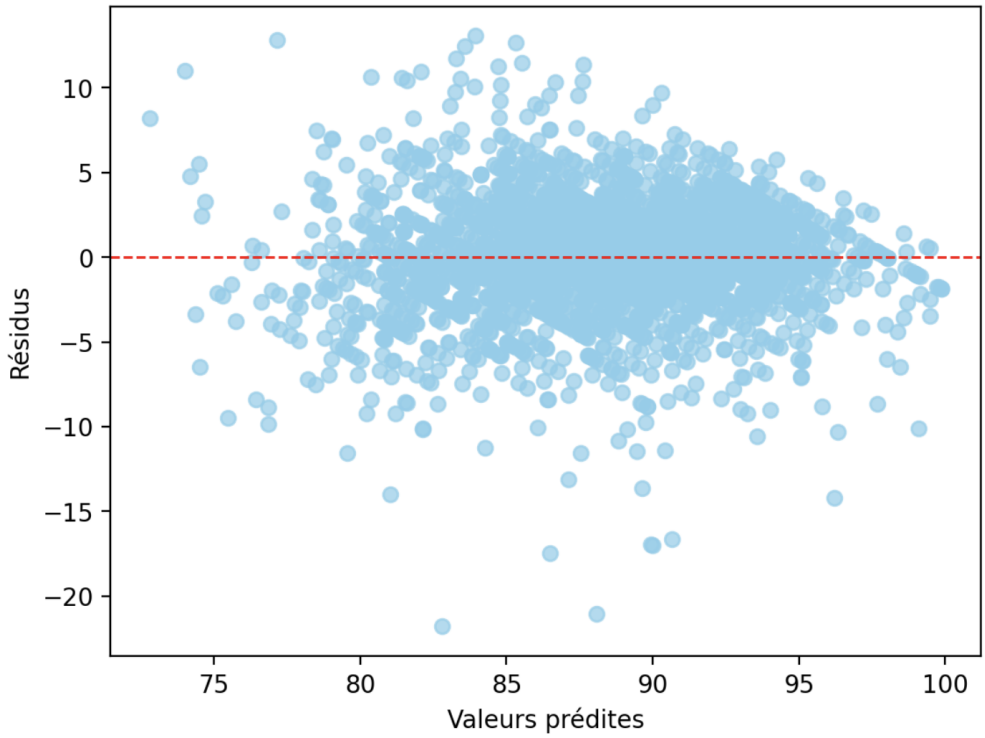


FIGURE 4 – Visualisation de l'hétéroscédasticité

Où :

— $\text{loss}(y_i, \hat{y}_i)$ est la fonction de perte, représentant l'erreur de prédiction entre la valeur réelle y_i et la valeur prédite $\hat{y}_i$.

— λ est un paramètre de régularisation qui contrôle la complexité du modèle.

— w_t est le vecteur de poids des arbres dans le modèle.

Ce processus aide à réduire la variance des résidus, rendant le modèle plus robuste et réduisant ainsi l'hétéroscédasticité.

3.5 Test de Breusch-Pagan pour XGBoost

Après l'application de **XGBoost**, nous avons recalculé la statistique de Breusch-Pagan pour évaluer l'hétéroscédasticité dans le modèle ajusté. Les résultats sont les suivants :

La statistique de Breusch-Pagan (BP) repose sur le coefficient de détermination R^2 de la régression auxiliaire :

$$BP = n \cdot R^2$$

où :

— n est le nombre d'observations,

— R^2 est le coefficient de détermination de la régression auxiliaire.

— **Statistique de Breusch-Pagan :** 13.71

— **p-valeur :** 3.943e-01

La **p-valeur** étant supérieure à 0.05, nous ne rejetons pas l'hypothèse nulle d'homoscédasticité. Cela indique que l'hétéroscédasticité initialement présente a été significativement réduite grâce à l'utilisation de XGBoost, rendant le modèle plus robuste et mieux adapté à nos données.

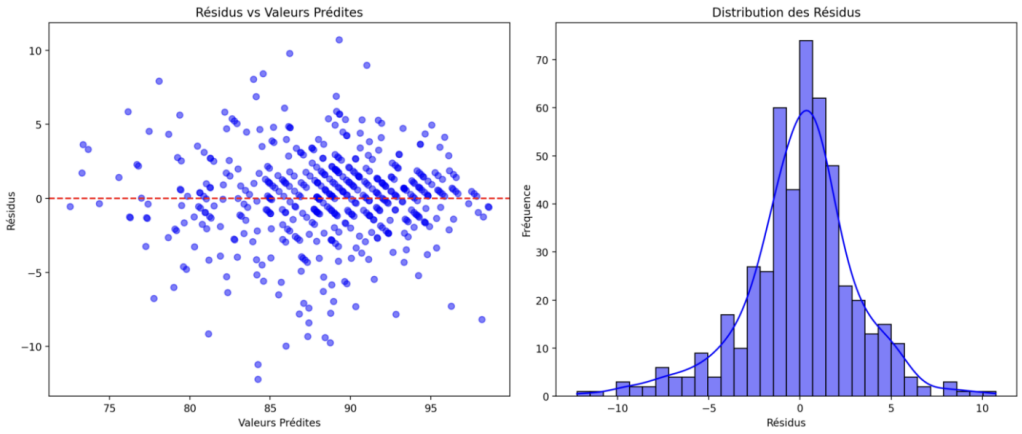


FIGURE 5 – Correction de l'hétéroscédasticité

4 Interprétation des résultats et analyse

4.1 Fondements théoriques de SHAP

SHAP (*Shapley Additive Explanations*) est basé sur la **valeur de Shapley**, une méthode issue de la théorie des jeux coopératifs. Cette méthode permet d'attribuer à chaque caractéristique de manière transparente l'impact qu'elle a sur la prédiction d'un modèle, tout en tenant compte des interactions entre ces caractéristiques.

La valeur de Shapley : La valeur de Shapley pour une caractéristique (i) est calculée comme suit :

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big[v(S \cup \{i\}) - v(S) \Big]$$

Où :

— $v(S)$ est la valeur d'un sous-ensemble S de caractéristiques, indiquant la prédiction ou la performance du modèle en fonction de S.

— N est l'ensemble total des caractéristiques utilisées dans le modèle.

Cette formule mesure l'**impact marginal** de chaque caractéristique (i) en fonction de sa contribution à tous les sous-ensembles possibles de caractéristiques.

Dans le cadre de SHAP : Pour une caractéristique x_i, la contribution à la prédiction est déterminée par la différence entre la prédiction du modèle avec et sans cette caractéristique :

$$\text{Contribution}(x_i) = \hat{y}_{\text{avec } x_i} - \hat{y}_{\text{sans } x_i}$$

Application de SHAP : SHAP est principalement utilisé pour interpréter les modèles complexes tels que :

— Les arbres de décision et forêts aléatoires.

— Les réseaux neuronaux et autres modèles non linéaires.

Pour rendre ces calculs plus efficaces, en particulier dans le cas de modèles à grande échelle, des méthodes d'approximation comme **Tree SHAP** sont employées.

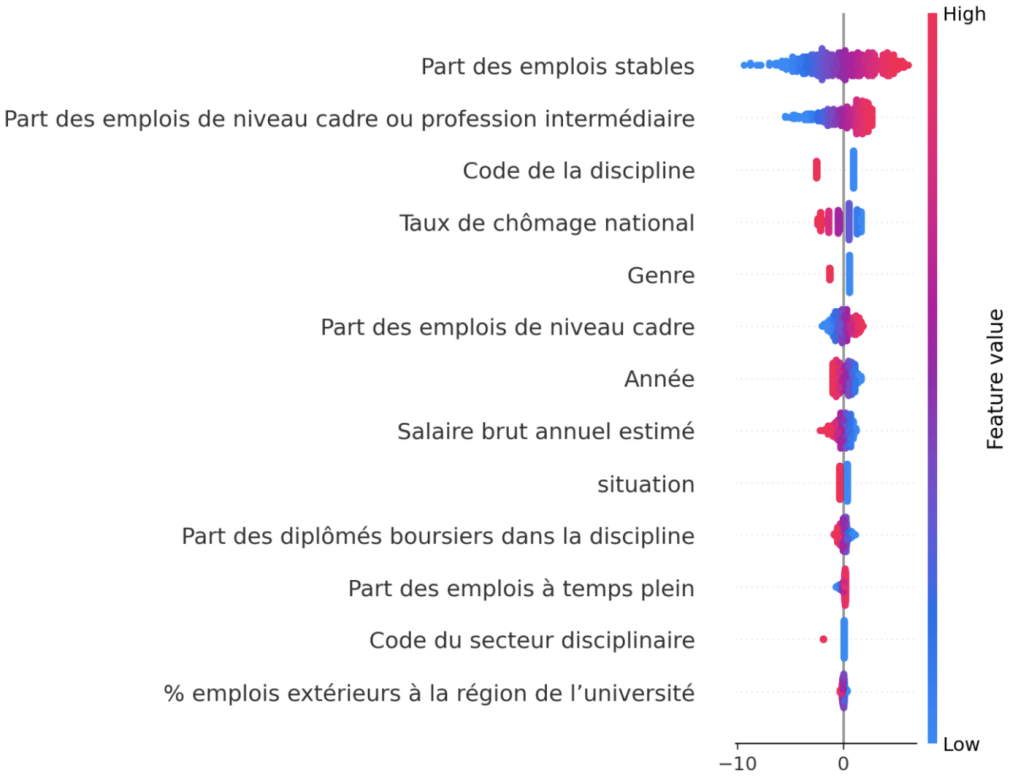


FIGURE 6 – Analyse des contributions des variables avec SHAP

4.2 Test de Durbin-Watson (Autocorrélation)

Le test de Durbin-Watson est utilisé pour détecter l'**autocorrélation** dans les résidus d'un modèle de régression. La statistique Durbin-Watson est calculée selon l'équation suivante :

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Où :

— e_t est l'erreur (résidu) à la période t,

— n est le nombre d'observations,

— e_{t-1} est l'erreur à la période précédente.

Statistique Durbin-Watson : 1.66

Valeurs idéales :

— **Proche de 2 :** Pas d'autocorrélation (les résidus sont indépendants).

— **< 1 ou > 3 :** Présence d'autocorrélation significative (les résidus sont corrélés).

Interprétation des résultats : La statistique Durbin-Watson est proche de 2, ce qui suggère qu'il n'y a pas d'autocorrélation significative. Cela indique que les erreurs du modèle sont **indépendantes**, et que l'hypothèse d'indépendance des erreurs est respectée.

Interprétation visuelle :

— Le graphique de la fonction d'autocorrélation (ACF) des résidus montre les corrélations entre les erreurs à différents décalages.

— Si les barres dépassent les limites de confiance (représentées par des lignes horizontales), cela indique la présence d'autocorrélation à ces décalages.

— Si aucune barre ne dépasse ces limites, cela soutient l'indépendance des résidus et la validité du modèle.

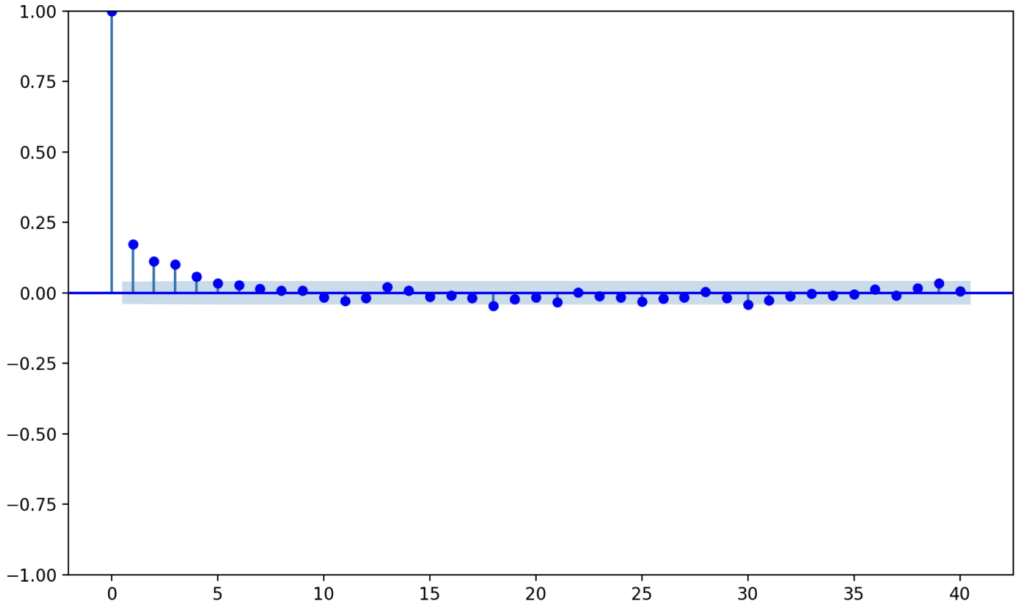


FIGURE 7 – Fonction d'autocorrélation des résidus

5 Prévisions (ARIMA)

5.1 Modèle ARIMA : Explication théorique

Le modèle ARIMA (AutoRegressive Integrated Moving Average) est un modèle statistique utilisé pour prédire les séries temporelles. Il se compose de trois éléments principaux :

— **AR (AutoRegressive)** : Utilisation des valeurs passées pour prédire les valeurs futures. C'est le terme de régression basé sur les observations passées.

— **I (Integrated)** : Processus de différenciation des données pour les rendre stationnaires (éliminer la tendance).

— **MA (Moving Average)** : Modélisation de la relation entre l'observation actuelle et l'erreur de prédiction des périodes passées.

L'équation théorique d'un modèle ARIMA(p, d, q) est la suivante :

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(1 - B)^d Y_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)\epsilon_t$$

Où :

— Y_t est la valeur de la série temporelle au temps t,

— B est l'opérateur de retard ($B^k Y_t = Y_{t-k}$),

— $\phi_1, \phi_2, \ldots, \phi_p$ sont les coefficients de l'auto-régression (AR),

— $\theta_1, \theta_2, \ldots, \theta_q$ sont les coefficients de la moyenne mobile (MA),

— d est le nombre de différenciations pour rendre la série stationnaire,

— ϵ_t est le terme d'erreur à l'instant t.

5.1.1 Détails du modèle ARIMA(3, 1, 3)

Dans ce modèle, les paramètres sont les suivants :

— $p = 4$: Quatres lags (valeurs passées) sont utilisés pour l'auto-régression (AR),

— $d = 2$: La série a été différenciée deux fois pour la rendre stationnaire,

— $q = 4$: Quatres lags de l'erreur sont utilisés pour la moyenne mobile (MA).

Erreur quadratique moyenne (MSE) ARIMA (final) : 14.46

5.2 Comparaison des résidus : Régression Linéaire vs ARIMA

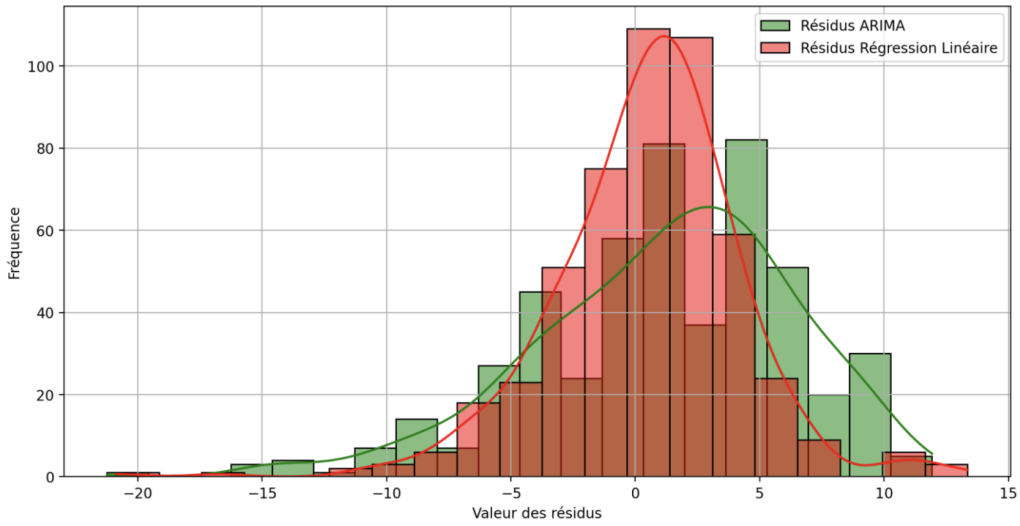


FIGURE 8 – Répartition des résidus : ARIMA vs Régression Linéaire

5.3 Comparaison des prévisions futures : ARIMA vs Régression Linéaire

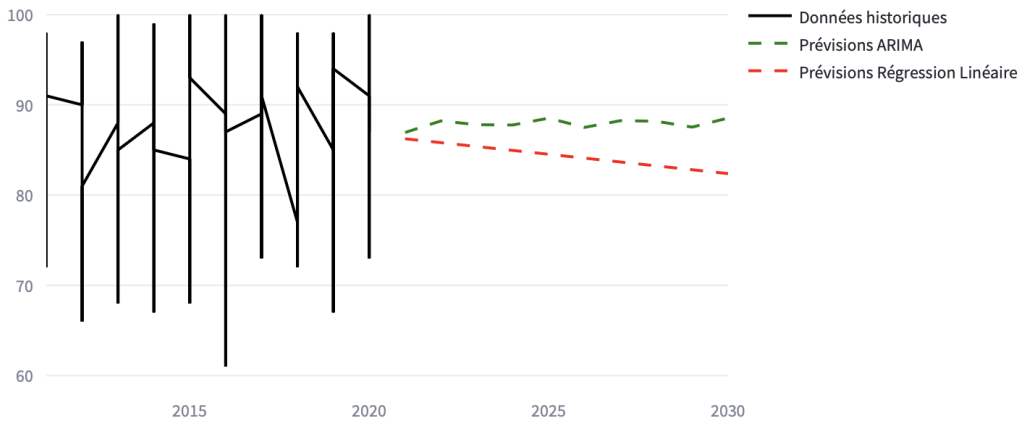


FIGURE 9 – Répartition des résidus : ARIMA vs Régression Linéaire

Année	Prévisions ARIMA	Prévisions Régression Linéaire
2021	86.9378	86.2398
2022	88.2199	85.8110
2023	87.7971	85.3823
2024	87.7652	84.9536
2025	88.5267	84.5249
2026	87.5005	84.0961
2027	88.2470	83.6674
2028	88.1818	83.2387
2029	87.5292	82.8100
2030	88.5374	82.3812

TABLE 1 – Tableau des prévisions futures

5.4 Interprétation et conclusion des prévisions

Les prévisions des modèles ARIMA et Régression Linéaire montrent une légère baisse du Taux d'insertion entre 2021 et 2030, oscillant entre 87 et 88 %. Les deux modèles suivent une tendance similaire, avec des différences minimes dans les valeurs prédites.

— ARIMA prédit des taux légèrement plus élevés que la régression linéaire.

— La baisse progressive du taux d'insertion suggère une stabilisation de la situation, mais les variations restent faibles.

— Le modèle ARIMA fournit des prévisions qui tiennent compte des tendances temporelles et des saisonnalités potentielles dans les données.

— La régression linéaire, bien qu'utile, ne capture pas les effets temporels complexes et pourrait donner des prévisions moins précises sur des périodes futures.

6 Discussion et perspectives

L'analyse menée sur le taux d'insertion des diplômés a permis de mettre en lumière plusieurs facteurs clés influençant leur employabilité. Parmi ces facteurs, le contexte économique, représenté par le taux de chômage national, et les caractéristiques du marché du travail, comme la stabilité et la qualité des emplois disponibles, jouent un rôle déterminant. Une augmentation de la part des emplois stables ou de niveau cadre est directement associée à une amélioration du taux d'insertion, tandis qu'un environnement économique

difficile (taux de chômage élevé) réduit significativement les opportunités pour les diplômés.

L'approche méthodologique, combinant une régression multiple et des outils avancés comme XGBoost et SHAP, a également permis de confirmer la robustesse des résultats. Les tests de diagnostic ont révélé quelques limites, notamment en termes d'hétéroscédasticité, mais des corrections ont été appliquées pour affiner les modèles et améliorer leur interprétabilité.

Cette étude souligne l'importance d'investir dans des politiques favorisant la création d'emplois stables et qualifiés, en particulier dans les secteurs où les diplômés sont les plus représentés. De plus, elle invite à réfléchir aux mesures qui pourraient atténuer l'impact des conditions économiques globales sur l'insertion professionnelle, comme le développement de dispositifs d'accompagnement ou la promotion de la mobilité géographique et sectorielle des diplômés.

Enfin, les perspectives offertes par cette analyse appellent à un approfondissement futur, notamment par l'intégration de données longitudinales pour mieux comprendre les trajectoires professionnelles des diplômés sur le long terme. Cela pourrait permettre de construire des modèles encore plus précis et d'anticiper les évolutions du marché du travail, afin de mieux préparer les générations futures à y évoluer.